Logistic Regression for Southern Pine Beetle Outbreaks with Spatial and Temporal Autocorrelation

Marcia L. Gumpertz, Chi-tsung Wu, and John M. Pye

ABSTRACT. Regional outbreaks of southern pine beetle (Dendroctonus frontalis Zimm.) show marked spatial and temporal patterns. While these patterns are of interest in themselves, we focus on statistical methods for estimating the effects of underlying environmental factors in the presence of spatial and temporal autocorrelation. The most comprehensive available information on outbreaks consists of binary data, specifically, annual presence or absence of outbreak for individual counties within the southern United States. We demonstrate a method for modeling spatially correlated proportions, such as the proportion of years that a county experiences outbreak, based on annual outbreak presence or absence data for counties in three states (NC, SC, and GA) over 31 yr. In this method, the proportion of years in outbreak is predicted using a marginal logistic regression model with spatial autocorrelation among counties, with adjustment of variance terms to account for temporal autocorrelation. This type of model describes the probability of outbreak as a function of explanatory variables such as host availability, physiography, climate, hurricane incidence, and management type. Explicitly including spatial autocorrelation in the model improves estimates of the probability of outbreak for a particular county and of the importance of the various explanatory variables. **For.** Sci. 46(1):95-107.

Additional Key Words: Generalized estimating equations, spatial prediction, marginal models, correlated proportions, correlated binary data.

1 Introduction

SOUTHERN PINE BEETLE (*Dendroctonus frontalis* Zimm.) outbreaks occur in forests throughout the Southern United States and can cause tremendous economic damage (Holmes 199 1, de Steiguer et al. 1987); hence there is great interest in understanding their causes and in improving tools to predict outbreaks. Visual analysis of maps of outbreaks across the region (Price et al. 1998) reveals striking temporal and spatial patterns, but the autocorrelations associated with these patterns mean the data do not fit the assumptions' required for classical regression. We describe here the development of a statistical model which properly accounts for these autocorrelations and which allows inclusion of additional explanatory covariates in the model.

Previous analyses have examined spatial patterns such as those in Figure I, and often related them to host availability or climate. Temporal patterns such as those in Figure 2 have also been the subject of study, frequently related to weather or to endogenous cycles of the beetle and its predators.

Marcia L. Gumpertz, Associate Professor, Dept. of Statistics, Box 8203, North Carolina State University, Raleigh, NC 27695-8203-Phone: (919) 515-1923; Fax: (919) 515-1169; E-mail: gumpertz@ncsu.edu. Chi-tsung Wu, Assistant Professor, Department of Statistics Feng Chia University, 100 Wenhwa Road, Seatwen, Taichung, Taiwan 407 R.O.C.-Phone: 886-4-451-7250, ext. 405; E-mail: cwu@stat.feu.edu.tw. John M. Pye, Ecologist, USDA Forest Service Southern Research Station, 3041 E. Cornwallis Rd., P. 0. Box 12254, Research Triangle Park, NC 27709-Phone: (919) 549-4013; E-mail: jpye@fs.fed.us

Acknowledgments: Thanks to Thomas P. Holmes of the USDA Forest Service for initiating this research project. We also wish to thank James 0. Paumierfor invaluable help in obtaining and understanding the climatological data, Sharon LeDuc for providing data. and David A. Dickey for helpful suggestions in his review of the manuscript. This work was partially supported by funds from USDA Forest Service Cooperative Agreement No. SRS 33-CA-97-115.

Manuscript received February 5, 1999. Accepted July 21, 1999.

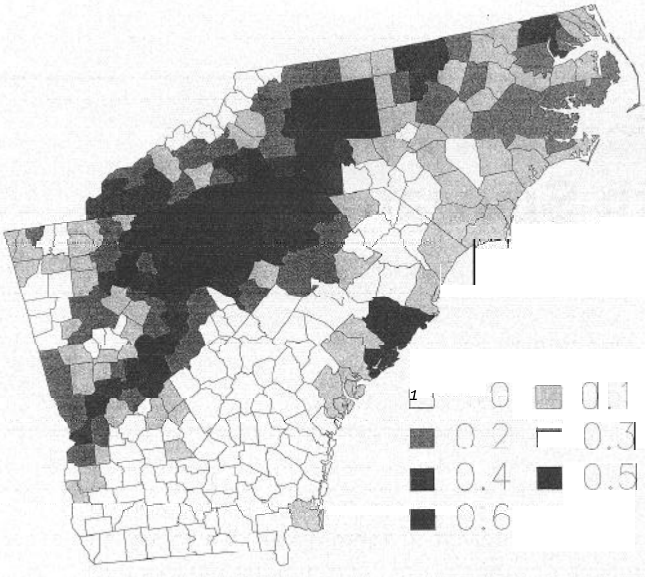

Figure 1. North Carolina, South Carolina, and Georgia proportion of years from 1960 to 1990 with southern pine beetle outbreaks.

Spatial Patterns

Mawby and Gold (1984) found that regional outbreaks of southern pine beetle (SPB) exhibited varying levels of spatial autocorrelation depending on the severity of that year's outbreak. Outbreaks in the United States generally occur in a diagonal band from central Virginia to northeastern Texas, corresponding roughly to the coniferous-broadleaved semi-evergreen forest ecoregion (Bailey 1995). Price and Doggett (1982) visually compared the long-term distribution of outbreaks to the distribution of one of its host species, shortleaf pine (*Pinus echinata* Mill.). Pye (1993) noted a similar correspondence with the timber volumes of shortleaf plus two other important host species, loblolly and Virginia pines (*Pinus taeda* L. and *P. virginiana* Mill.), recognizing that the beetle exploits other southern pine species to varying degrees.

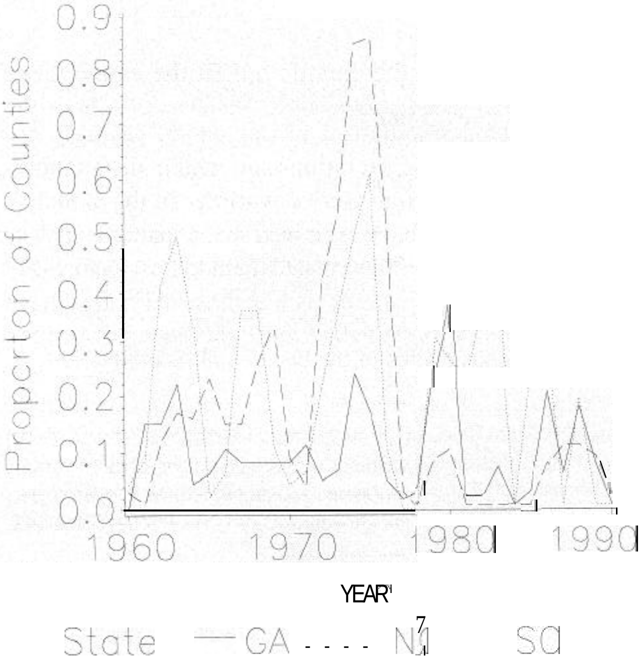

Figure 2. Proportion of counties in outbreak in NC, SC, and GA for each year.

Temporal Patterns

Southern pine beetle populations vary dramatically over time, oscillating between endemic and outbreak conditions, where an outbreak is defined as at least one southern pine beetle spot infestation per 405 ha (1,000 ac) of loblolly/shortleaf or oak/pine type forest (Price et al. 1998). Various researchers have noted a periodicity but have reported these over different spatial scales. Pye (I 993) cited a cycle length of 6-7 yr for recent outbreaks spanning the southern United States, but Mawby and Gold (1984) reported varying periodicities when the region was divided into 24 subregions. Turchin et al. (1991) found temporal autocorrelations at lags of both 1 and 2 yr for populations in East Texas and concluded that delayed density dependence was a more important regulator of populations than density-independent factors such as climate.

Ungerer et al. (1999) have claimed a climatic factor may be important by showing that cold temperature events at the northern limit of outbreaks match experimentally determined lethal tolerances for the beetle. If climate is an important determinant of temporal patterns, it is likely a complex relationship. Climatecan affectpopulation dynamics through direct impact on beetle metabolism, viability, and generation length (Gagne et al. 1980, Hines et al. 1980), or pheromone communication (Fares et al. 1980), or population levels indirectly by modifying the resistance of host species to beetle attack via drought or flooding stress (Kalkstein 1976, Lorio 1986) or disturbances such as lightning (Coulson et al. 1983).

Broad-scale changes in the region could be caused by factors such as (1) regionwide changes in host forest types (Mawby et al. 1989); (2) warmer temperatures from increased atmospheric CO, or other causes (Ungerer et al. 1999); and (3) lengthened rotations on national forests, potentially increasing outbreaks on surrounding private forests (Carter et al. 1991). Evaluation of the likely impacts of such changes requires improved statistical models which simultaneously account for the spatial patterns of beetle range and dispersal, the temporal autocorrelations associated with predator-prey population cycles, and mechanistic measures of host condition and climate. Many studies of southern pine beetle dynamics have been performed in the past, but none have been tailored specifically to data in the form of spatially correlated proportions. Spatial statistical methods for Gaussian (normally distributed) data have begun to be widely used in entomological studies (Liebhold et al. 1993). Methods for non-Gaussian data are also beginning to appear. Recently, Preisler et al. (1997) demonstrated a very flexible generalized additive model to study relationships between twig beetle attacks and explanatory variables, including a function of spatial location as an explanatory variable.

The objective of this study is to demonstrate use of a marginal logistic regression model for spatially correlated proportions, which also incorporates information about temporal autocorrelation. We use the marginal logistic regression model to describe the pattern of southern pine beetle outbreaks and assess the explanatory power of environmental variables for predicting where outbreaks tend to occur in North Carolina, South Carolina, and Georgia. This type of model has two features that distinguish it from an ordinary linear model: (1) the logistic form and (2) incorporation of spatial and temporal autocorrelation.

The advantage of a logistic type of model over a more familiar linear or nonlinear model is that it can be tailored to a binary response, like presence or absence of outbreak, or to a proportion, such as proportion of years in outbreak. With such a model we can make estimates of the odds of an outbreak in a given county or group of counties. A more technical advantage of logistic regression over ordinary regression has to do with the methods of estimation. Logistic regression incorporates information about the variance of binary/proportion data into the estimating equations to provide more efficient estimates than ordinary regression would.

The second distinguishing feature of this type of model, the incorporation of spatial and temporal autocorrelation, also provides advantages both for estimation of the effects of the regressor variables and for estimation of the probability of outbreak in a given county. If spatial and temporal autocorrelation are ignored in fitting the model, the parameter estimates have lower precision (higher variances) than if the correlations are incorporated into the fitting procedure. Furthermore, if ordinary regression software is used the standard errors that are produced are incorrect. Thirdly, and the greatest benefit of all, estimates of the probability of outbreak in a given county are much more precise than if the spatial autocorrelation is ignored. This is because the local patterns of variability are taken into account when making estimates for a site, whereas in ordinary logistic regression they are not.

The bulk of this article demonstrates the fitting and interpretation of a marginal logistic regression model which is described in Section 2. Sections 3 and 4 demonstrate an analysis of the southern pine beetle data starting with ordinary logistic regression and adding complexity as it is needed. We show some typical steps an analyst might go through and some methods for evaluating the adequacy of each model. Examination of the residuals from ordinary logistic regression in Section 3.2 reveals spatial and temporal autocorrelation, indicating that ordinary logistic regression is not the best procedure for these data. In Section 4.1 we show how temporal autocorrelation changes the variance of the proportion of years in outbreak. The ordinary logistic regression model is next modified to account for temporal autocorrelation and fitted using weighted logistic regression, which is readily available in commercial software. Even after accounting for the temporal autocorrelation, spatial correlation remains, so the last step, in Section 4.2, is to incorporate spatial correlation and fit the model using generalized estimating equations. Section 5.1 shows how to use this type of model to interpolate spatially (i.e., to predict the proportion of years in outbreak for a county with missing data). Section 5.2 discusses the use of this model for making predictions into the future, and Section 6 gives a general discussion of our conclusions.

2 Marginal Logistic Regression Model

The general class of models known as marginal models (Liang and Zeger 1986, Diggle et al. 1994) allows for covariates (explanatory variables) and for spatial and temporal correlation but does not require full specification of the joint probability distribution of all sites. Marginal models were initially proposed for longitudinal binary data (Liang and Zeger 1986), but have recently been applied to spatially correlated data as well (Albert and McShane 1995, Gotway and Stroup 1997). The term “marginal” refers to modeling the expected response of a site to the regressor variables, rather than the joint responses of all sites simultaneously. The focus is on the relationships between the explanatory variables and the probability of outbreak. In this type of model, the spatial and temporal correlations are secondary, included to obtain better estimates of the expected response. The marginal logistic model consists of a model for the mean,

$$\text{logit}(p_i) = x_i'\beta \qquad (1)$$

where p_i is the probability with which the ith county experiences outbreaks, and a model for the variances and covariances among the sites. The response variable Y_i is the proportion of years each county experiences outbreaks. The explanatory variables x_i are county-level measurements; that is, we have one measurement of each of the explanatory variables for each county. Hence, the explanatory variables vary over space but not over time, and they can help predict or explain general spatial patterns of southern pine beetle outbreak but cannot shed light on why outbreaks occur in some years but not others. Spatial correlation is explicitly incorporated into the logistic regression model using an exponential covariance function.

In addition, we also know which years each county experienced outbreaks, so we use the outbreak data for individual years to model temporal autocorrelation. The autocorrelation over time within each county is modeled by a first-order Markov process. The temporal autocorrelation then enters into the variance of the proportion of years in outbreak for each county. The methods we present are flexible and can be used with explanatory variables other than those presented here.

3 Preliminary Analysis

Several explanatory variables were considered, all at the county level. These included

- two measures of volume (m3/ha) of host trees: sawtimber volume (22.9 cm dbh minimum for softwoods, 27.9 cm dbh for hardwoods), and poletimber volume (12.7 to 22.8 cm dbh for softwoods and 12.7 to 27.8 cm dbh for hardwoods);
- three physiographic variables: proportion of land area classified as mesic, hydric, and xeric;
- three climate variables computed separately for the fall, winter, spring, and summer seasons: average daily minimum temperature (C), average daily maximum tempera-

ture (C), and average monthly precipitation (cm); number of 6 hour periods with hurricane force winds recorded in the county from 1960 to 1990;

- five management variables: area of land (in 1000 ha) owned by the federal government, forest industry, individuals, private corporations, and states;
- and three location variables: elevation (m), latitude, and longitude.

The host species include loblolly, shortleaf, Virginia, pitch *(Pinus rigida),* sand *(Pinus clausa),* pond *(Pinus serotina),* and Table Mountain *(Pinus pungens)* pines and spruce *(Pinus glabra),* The estimates of host volume, physiographic variables, USDA Forest Service and forest industry land area were obtained from the USDA Forest Service's Forest Inventory and Analysis Data Base Retrieval System (Hansen et al. 1992). Hurricane track data were obtained from NOAA's Atlantic Oceanographic and Meteorological Laboratory's Hurricane Research Division website (Landsea 1995). The climate variables were computed from 30 yr (1960-1991) climatological averages by month for each station, which were obtained from the Southeast Regional Climate Center website for Climatological Normals 1961-1990 (Owenby and Ezell1992). The averages for all stations within a county were averaged together to obtain the county average for each month. Three months were then averaged together to obtain seasonal averages for fall (September through November), winter (December through February), spring (March through May), and summer (June through August) for each county. Climatological averages were not available for many counties; 45 counties were missing precipitation records, and 213 counties were missing temperature records. The value from the available weather station nearest to the county center was substituted for any missing county. Elevation, latitude, and longitude of the stations were obtained from the National Climatic Data Center's cooperative statior master list (National Climatic Data Center 1995).

3.1 Selection of Variables

The general spatial pattern of host volume (Figures 3 and 4) is similar to the pattern of southern pine beetle outbreaks (Figure l), with high incidence of outbreaks in the northwestern part of the region and some high values in counties along the coast, but the correspondence is far from exact. There is also some correspondence between the physiographic variables (Figures 5 and 6) and proportion of years in outbreak. Elevation increases from the coast in the southeast to the mountains in the northwest (Figure 7). Temperature increases from northwest to southeast (Figure 8), and precipitation tends to be highest along the coast and in some parts of the mountains (Figures 9 and 10).

Many transformations and combinations of these explanatory variables are possible. A preliminary analysis using ordinary linear regression, ignoring spatial and temporal correlation, helped to narrow the set of explanatory variables to a manageable size. In the preliminary analysis, a number of models were fit to

$$\sin^{-1}\left(\sqrt{Y}\right)$$

using ordinary linear regression. For a first pass, all variables were included in the model, and then stepwise regression was used to add interaction terms. Next, first-order terms with low F-values were dropped unless they were part of an included interaction term. The set of explanatory variables selected for further consideration included In (elevation), longitude, square roots of saw and poletimber volume per ha, square roots of proportion of 1and area classified as xeric, mesic, and hydric, square roots 0.f land area owned by the federal government and by private forest industry, average number of hurricane wind periods per year, all 12 of the climate variables, and the 5 interactions:

hurricanes per year x $\sqrt{\text{mesic}}$,

maximum summer temperature x $\sqrt{\text{saw volume}}$,

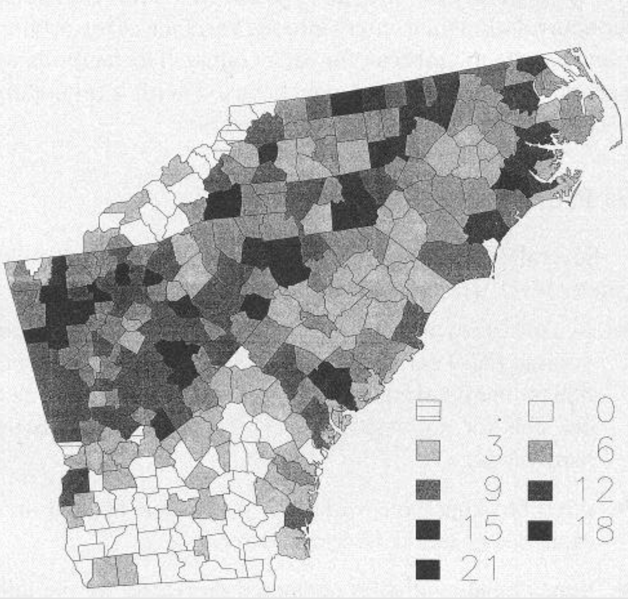

Figure 3. Poletimber volume (m^3/ha) for host species. Four counties were missing data, indicated by a dot in the legend.

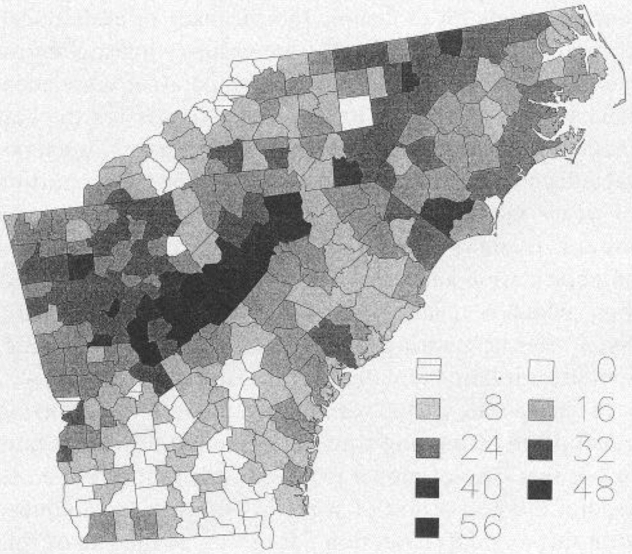

Figure 4. Sawtimber volume (m^3/ha) for host species. A dot in the legend indicates missing data.

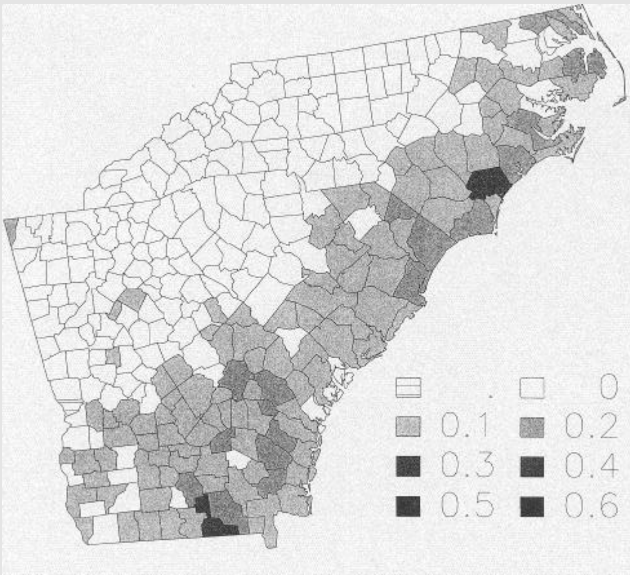

Figure 5. Proportion of land area classified as hydric. A dot in the legend indicates missing data.

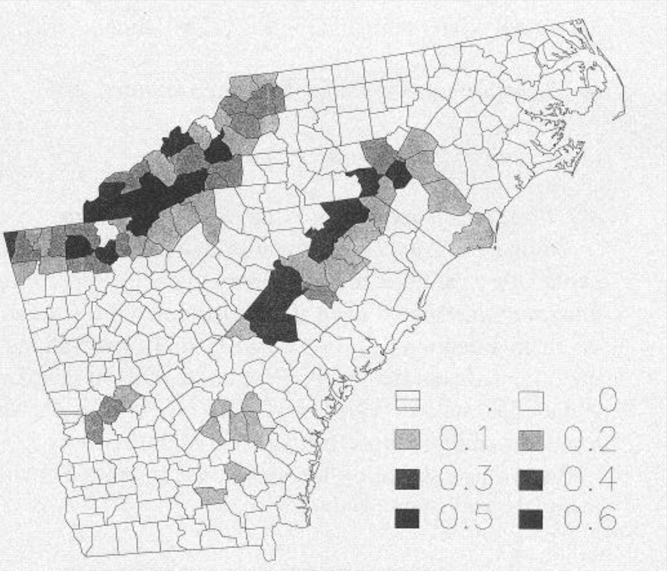

Figure 6. Proportion of land area classified as xeric. A dot in the legend indicates missing data.

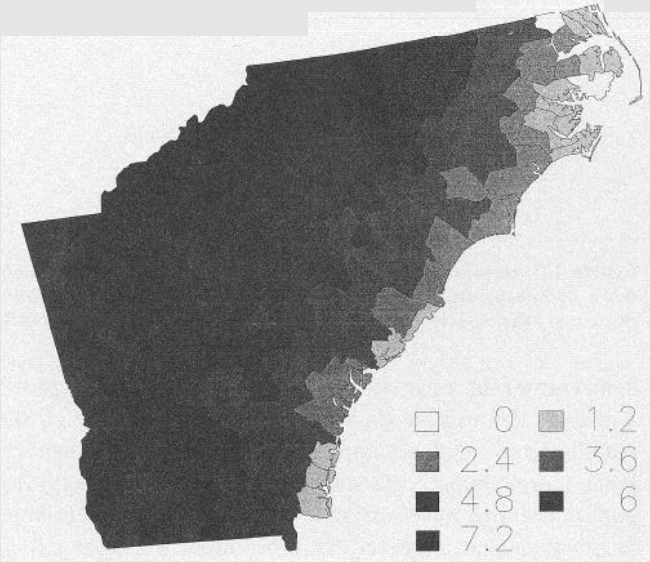

Figure 7. Ln(elevation) in m.

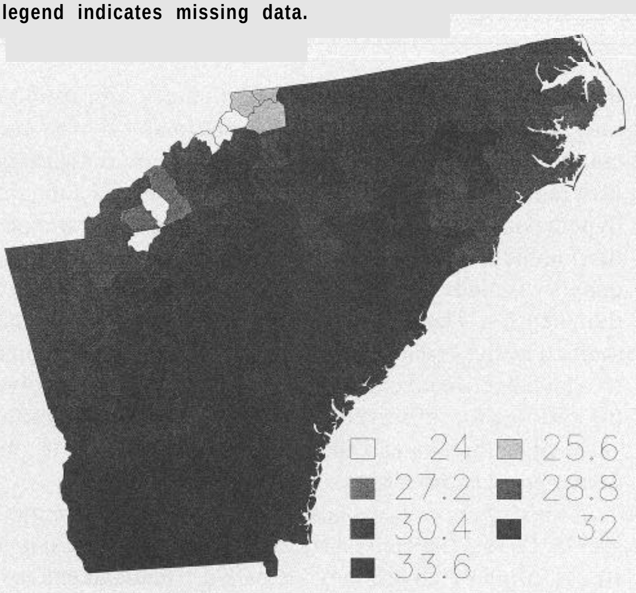

Figure 8. Mean daily maximum temperature (C) for summer months (June, July, August). Climatological average for 1961-1990.

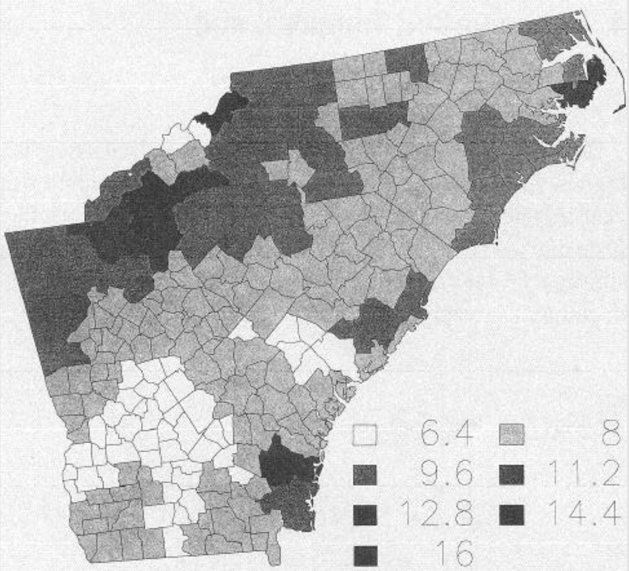

Figure 9. Mean fall(September,October,November) precipitation (cm). Climatological average for 1961-1990.

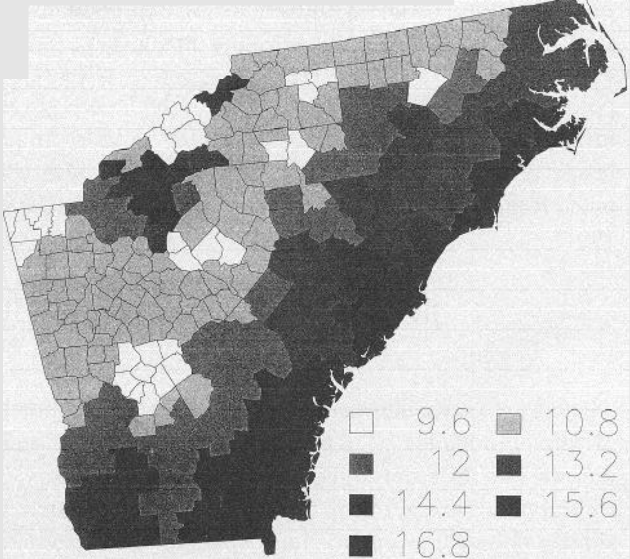

Figure 10. Mean summer precipitation (cm). Climatological average for 1961-1990.

maximum winter temperature × $\sqrt{\text{saw volume}}$,

average spring precipitation x /pole volume, and

minimum fall temperature x $\sqrt{\text{saw volume}}$.

3.2 Residuals from Ordinary Logistic Regression

Ordinary logistic regression was then used to fit this set of explanatory variables to Y_i. Ordinary logistic regression uses maximum likelihood to fit the logistic model for the mean given in Equation (1), but assumes that observations are uncorrelated, and that the proportion of years in outbreak is a binomial random variable with variance $p_i(1-p_i)/n$, where n = 31 yr and p_i = probability of outbreak in county *i*.

To evaluate the fit of this model, we computed deviance residuals. Deviance residuals,

$$d_i = \sqrt{2n_iY_i \log\frac{Y_i}{\hat{p}_i} + 2(n_i - n_iY_i)\log\left(\frac{1-Y_i}{1-\hat{p}_i}\right)},$$

measure the deviations of the fitted values, $\hat{p}_i$, from the observed proportion of years in outbreak, Y_i, for each county (Collett 1991 p. 122). The deviance is twice the difference between the log likelihood of the data under the hypothesized model and the log likelihood under a model that includes a separate parameter for each county. This quantity is made up of a sum of contributions from each of the counties. The deviance residual for a county is then defined as the square root of the county's contribution to the deviance. Notice that the deviance residual involves the ratio of the observed to the predicted proportion of years of outbreaks and the ratio of the observed to predicted proportion of years without outbreaks.

If the deviance residuals are divided by their asymptotic standard errors (theoretical standard errors when the sample size is infinitely large), they are called "standardized deviance residuals" (Collett 199 1). The standardized deviance residuals from the ordinary logistic regression of logit(Y_i) on 27 explanatory variables show some spatial pattern. They tend to be zero or slightly negative in the large area of southern Georgia where no outbreaks were ever observed and positive in the higher elevations; thus this model was not able to account for all of the spatial variability in southern pine beetle outbreaks. An empirical semivariogram was computed from the standardized deviance residuals, d_i (Cressie 1991),

$$\hat{\gamma}(h) = \frac{1}{2N(h)}\sum_{i,i^*}^{N(h)}(d_i - d_{i^*})^2$$

where h = distance between two counties, $N(h)$ is the number of pairs of counties h km apart, and the subscripts *i* and *i** indicate two different counties that are h units apart,

The semivariogram of the standardized deviance residuals shows that counties are spatially correlated to a distance of about 160 km; beyond this distance counties are essentially uncorrelated (Figure 11). The sill of a semivariogram estimates the variance of the response variable. According to Figure 11, the variance of the standardized residuals appears to be about 2.9. However, since these residuals are standardized, we would normally expect them to have variance one. Observing a variance larger than that expected for a binomial model is called "overdispersion."

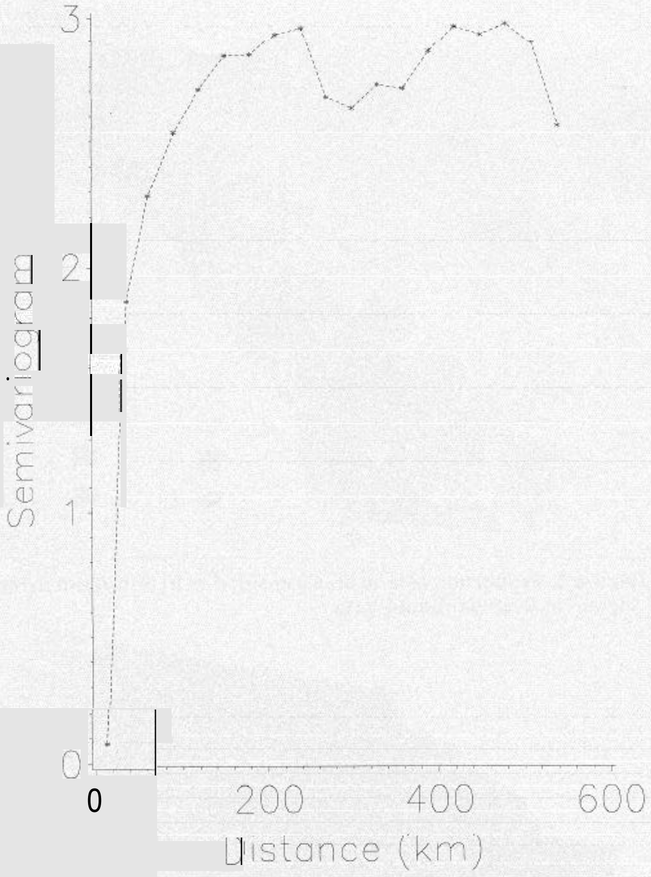

Figure 11. Semivariogram of standardized deviance residuals from ordinary logistic regression on 27 variables. This model does not take spatial or temporal autocorrelation into account.

4 Incorporating Temporal and Spatial Autocorrelation

4.1 Temporal Autocorrelation

A possible reason that the sill (the observed variance) is so high is that outbreaks in a given county are correlated from year to year. Since the outbreaks in a county are correlated from one year to the next, the variance of the proportion is not simply $p_i(1-p_i)/n$. Letting U_{ij} be the outbreak status (0 or 1) in county i in year j, the variance of the proportion is

$$\begin{aligned}\text{Var}(Y_i) &= \text{Var}\left(\frac{1}{n}\sum_j U_{ij}\right)\\ &= \frac{1}{n^2}\sum_j \text{Var}(U_{ij}) + \frac{1}{n^2}\sum_j\sum_{k\neq j}\text{Cov}(U_{ij},U_{ik})\\ &= \frac{p_i(1-p_i)}{n} + \frac{p_i(1-p_i)}{n^2}\sum_j\sum_{k\neq j}\text{Corr}(U_{ij},U_{ik})\end{aligned}$$

The additional term is the total of all of the correlations among different years. If years are independent, this term is zero, and we get the usual binomial variance, $\mathrm{Var}(Y_i) = p_i(1 - p_i)/n_i$. If the correlation from one time to another within a county is positive, the variance of Y_i is larger than $p_i(1 - p_i)/n_i$.

It is possible to account for the temporal autocorrelation by fitting a model to it. We allow each site to have a separate time correlation pattern, but within a site the correlation model is the same for all years. For simplicity, the probability of outbreak for the ith site, p_i, is assumed to be constant over time for each site; that is, there are no overall increasing or decreasing trends in the number of outbreaks over time.

A simple model for the time series for each site is a first-order Markov process, where the probability of outbreak in one year depends only on whether there was an outbreak in the previous year. Under a first-order Markov process, the correlation between observations at time j and time k at site i is

$$\rho_i^{|j-k|}.$$

The correlation parameter ρ_i gives the correlation between a particular year and the previous year, so if the correlation between two consecutive years is 0.6, then the correlation between observations two years apart is $0.6^2 = 0.36$ under the first-order Markov model. The correlation parameter is estimated for each site by computing the first-order autocorrelation coefficient

$$\hat{\rho}_i = \frac{\dfrac{1}{n_i - 1}\displaystyle\sum_{j=1}^{n_i - 1}(U_{ij} - Y_i)(U_{i,j+1} - Y_i)}{Y_i(1 - Y_i)}$$

Kedem (1980, p. 70).

The total correlation among pairs of years within a site is

$$\begin{aligned} TOTCORR_i &= \sum_j \sum_{k \neq j} Corr(U_{ij}, U_{ik}) \\ &= \sum_j \sum_{k \neq j} \rho_i^{|j-k|} \\ &= 2\frac{\rho_i}{1 - \rho_i}\left((n_i - 1) - \frac{\rho_i}{1 - \rho_i}(1 - \rho_i^{n_i - 1})\right) \end{aligned}$$

Estimated temporal autocorrelation coefficients range from −0.21 to 0.78 for counties that ever had any southern pine beetle outbreaks, with 50% of the cour having $\hat{\rho}_i > 0.32$. Thus there is a substantial amount of correlation over time within a county and me amount or autocorrelation varies widely among counties. The next question is whether the first-order Markov process adequately describes the observed temporal correlation pattern. In a second-order Markov process, the probability of outbreak depends on the outbreak status of the previous 2 yr, not just the previous year. We compared the first-order model to a second-order Markov process using a Chi-square test for goodness of fit (Guttorp 1995 p. 71) with significance level 0.05 for each county. The first-order model appears to be appropriate for this process. In only 7 (4%) of the 182 counties that ever experienced any southern pine beetle outbreak was the first-order model rejected in favor of the second-order model.

The standard method of estimation for logistic regression implemented in software packages such as SAS® PROC LOGISTIC (SAS Institute 1997) is maximum likelihood under the assumption that the response variable has a binomial distribution. Our response variable is not binomial since the observations are correlated over time and the joint distribution is not known, so maximum likelihood estimation is not possible. However, a method of estimation called "quasi-likelihood estimation" makes use of what is known about the variance of Y. The quasi-likelihood method involves iteratively solving the system of equations

$$D'V^{-1}(y - p) = 0$$

where the ith element of y is Y_i, the ith element of p is p_i,

$$D = \frac{\partial p}{\partial \beta'}$$

and V is a diagonal matrix with ith diagonal element $\mathrm{Var}(Y_i)$ (Diggle et al. 1994, Appendix A.6). In this particular problem, where we want to account for correlation over time but we still assume that observations are spatially uncorrelated, this method of estimation amounts to weighted logistic regression, replacing the binomial variance expression

$$\frac{p_i(1 - p_i)}{n_i},$$

with the appropriate variance of

$$Y_i, \frac{p_i(1 - p_i)}{n_i}(1 + TOTCORR_i / n_i).$$

It is easy to implement using software that allows weighted logistic regression. For example, in SAS® Proc Logistic (SAS Institute 1997), one would regress Y_i on the explanatory variables and specify weights $w_i = 1/(1 + TOTCORR_i/n_i)$ where w_i is a multiplier for the inverse of the variance. Several counties experienced no outbreaks between 1960 and 1990. For these counties, we set the correlation between any 2 yr to be 0.99, which gives *TOTCORR = (30)(3 1)(0.99).*

The standardized deviance residuals resulting from fitting this model still show spatial autocorrelation, but now the sill is close to one (Figure 12). The exponential semivariogram fitted to the empirical semivariogram by weighted nonlinear least squares (Cressie 1991) is

$$\hat{\gamma}(h) = 0.26 + 0.81(1 - e^{-3h/76.8}).$$

This function has a sill of 1.07 and a range of 76.8 km. Note that in the exponential correlation model, the range is the distance at which the semivariogram is 95% of the sill. Incorporating the temporal autocorrelation into the variance of Y_i seems to account for all of the "overdispersion" seen in the data.

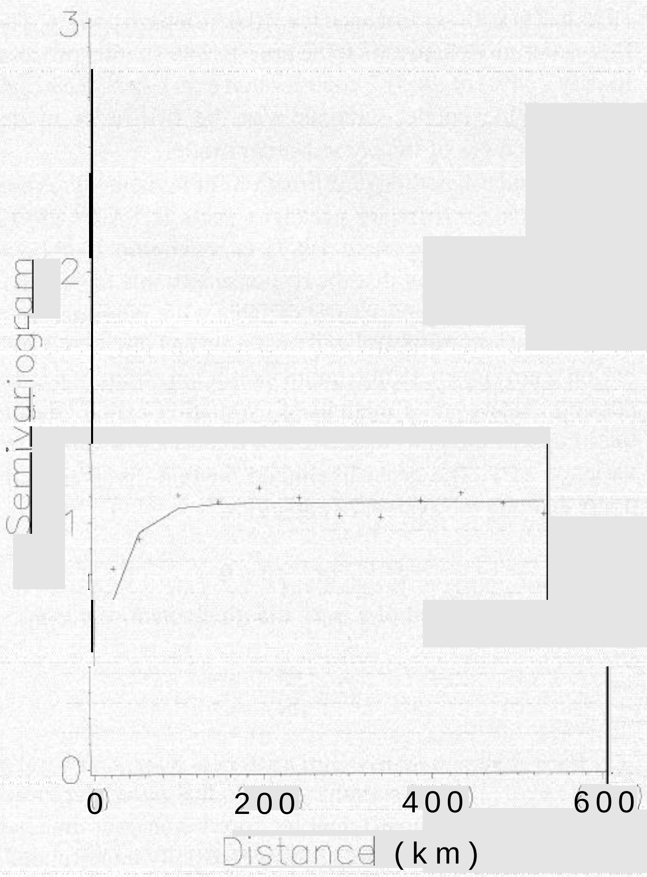

Figure 12. Empirical and fitted semivariograms of standardized deviance residuals from weighted logistic regression on 27 variables. Temporal autocorrelation is incorporated into the variance of Y_i.

4.2 Incorporating Spatial Autocorreh

The last refinement to the model is to incorporate the spatial autocorrelation, resulting in the model

$$\mathrm{logit}(p_i) = x_i'\beta \text{ with}$$
$$V = \mathrm{Var}(y) = A^{1/2}RA^{1/2} \qquad (2)$$

where

$$A = diag\left(\frac{p_i(1-p_i)}{n_i}(1+TOTCORR_i/n_i)\right),$$

$$R_{i,i*} = CORR(Y_i, Y_{i*}) = \frac{C_1}{C_0+C_1}e^{-3h/a},$$

h is the distance in km between counties i and $i*$, and a is the range of spatial correlation.

We use the fitted exponential semivariogram function to estimate the pairwise covariances among the counties. The method of generalized estimating equations can be used to estimate the parameters of a marginal logistic regression model with spatial correlation. The procedure is to iteratively solve the equation:

$$D'V^{-1}(y-p) = 0$$

(Liang and Zeger 1986; Gotway and Stroup 1997).

Next we sequentially dropped nonsignificant terms from the spatial logistic model. At each step we dropped the least significant term, with the exception that we retained any variables that were involved in a significant interaction. Then we refit the semivariogram model and the spatial logistic model before proceeding to drop another term. The Wald statistic was used for all tests. For testing one variable, the Wald statistic has the familiar form of a normal score; it is just

$$\frac{\hat{\beta}_k}{s(\hat{\beta}_k)},$$

where β_k is the parameter being tested. This statistic is compared to the standard normal distribution. Any linear combination of parameters, $L\beta$, where L is a matrix of coefficients selecting elements of β, may be tested using the Wald statistic (Gotway and Stroup 1997). The general form is

$$W = (L\hat{\beta})'(L\mathrm{Var}(\hat{\beta})L')^{-1}L\hat{\beta},$$

where $\mathrm{Var}(\hat{\beta}) \approx (D'V^{-1}D)^{-1}$, and the hypothesis being tested is H_0: $L\beta = 0$. For large samples, the Wald statistic approaches a chi-square distribution with degrees of freedom equal to the rank of L.

Thirteen variables were retained in the sequential procedure described above using an $\alpha = 0.10$ significance level (Table 1). As a final check, the 14 variables that had been dropped from the original 27-variable model were tested simultaneously using a Wald test with 14 degrees of freedom. The P-value for dropping all 14 variables from the model was 0.55, indicating that together they do not contain significant explanatory power beyond that contained in the final model of Table 1.

In the fitted model, the probability of southern pine beetle outbreaks increases with the amount of fall precipitation; this is the single strongest predictor of outbreak probability and also visually corresponds well with the pattern of outbreaks (Figure 9). The estimated probability of outbreak tends to be higher for areas with dry summers and lower for areas with high summer precipitation. For a given volume of sawtimber, the estimated probability of outbreak increases as summer or winter daily maximum temperature increases. Looked at the other way, probability of outbreak increases with volume of sawtimber per ha, but the volume of timber needed before outbreaks begin depends on the mean daily maximum temperature (Figure 13). Note that the summer and winter mean daily maxima are very highly correlated with each other, making it difficult to determine which of these variables might be responsible for the observed pattern of outbreaks.

The final model reproduces the general spatial pattern of southern pine beetle outbreaks fairly well (compare Figures 14 and 1), but smooths the proportions somewhat, The result is that the estimated probabilities in the

Table 1. Final logistic regression model with spatial autocorrelation [Equation (2).] Parameter estimates. standard errors, and P-values from Wald tests using the exponential covariance model. Temporal autocorrelation within counties is assumed to be a first-order Markov process.

Parameter	Estimate	SE	P-value
Intercent	-8.7	6.9	0.20
Ln[elevation (m)]	0.21	0.10	0.04
Longitude	0.21	0.053	0.00008
$\sqrt{\text{saw volume}(m^3/ha)}$	3.61	1.09	0.0009
$\sqrt{\text{hydric}}$ proportion	-1.49	0.54	0.006
$\sqrt{\text{xeric}}$ proportion	-0.92	0.38	0.02
$\sqrt{}$ national forest (thousand ha)	0.095	0.034	0.005
Mean daily maximum fall temp (C)	-0.75	0.34	0.03
Mean fall precipitation (cm)	0.35	0.062	1 E-8
Mean daily maximum winter temp (C)	-0.19	0.25	0.45
Mean daily maximum summer temp (C)	1.34	0.27	1 E-6
Mean summer precipitation (cm)	-0.12	0.053	0.03
Max summer temp x $\sqrt{\text{saw}}$ volume per ha	-0.18	51	0.0003
Max winter temp x $\sqrt{\text{saw volume per ha}}$	0.16	0.046	0.0005
c_0, nugget parameter	0.24		
c_1, sill parameter	0.85		
a, range parameter	97		

high-outbreak counties are lower than the observed proportions of years in outbreak, and the reverse is true in the low-outbreak counties.

5. Prediction and Evaluation

5.1 Spatial Prediction for Individual Counties

The estimates given in the previous section represent the average effects of the explanatory variables. The spatial correlation has been incorporated into the parameter estimates, β, and the estimated probability for an individual county is

$$\hat{p}_i = \frac{e^{x_i\hat{\beta}}}{1+e^{x_i\hat{\beta}}}.$$

This quantity estimates the mean probability of outbreak for a county with given values of the explanatory variables. We can construct a predictor that also takes into account the responses of surrounding counties. In linear regression models, the best linear unbiased predictor (BLUP), also known as the *kriging predictor* or *kriging with external drift* (Christensen 1991, Goovaerts 1997), does this. In the BLUP, the estimate for a particular county, say $\hat{p}_0$, is adjusted according to its location and the correlations among the counties. Letting

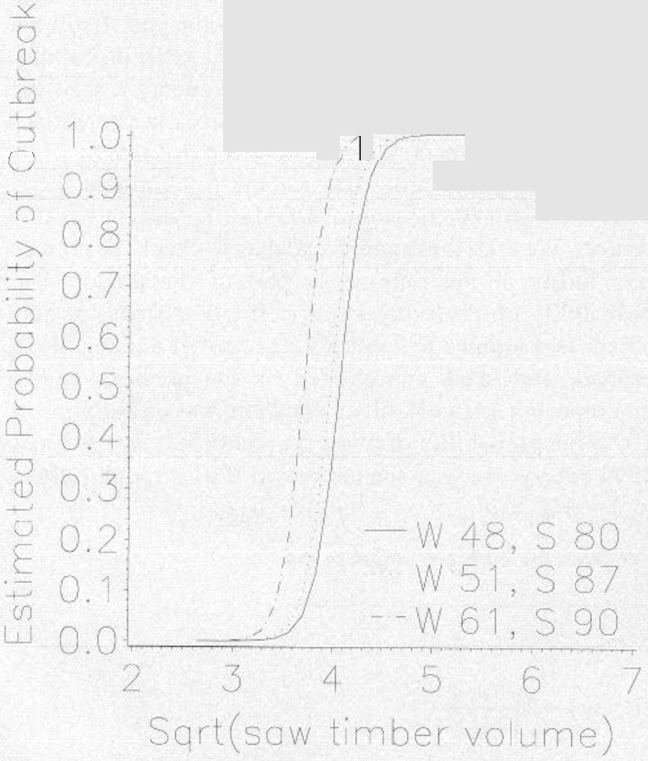

Figure 13. Estimated probability, $\hat{p}$, of outbreak from model (2) vs. sawtimber volume (m3/ha), with separate lines for different average daily maximum summer and winter temperatures (C). The temperature values are climatological normals for 1961-1990.

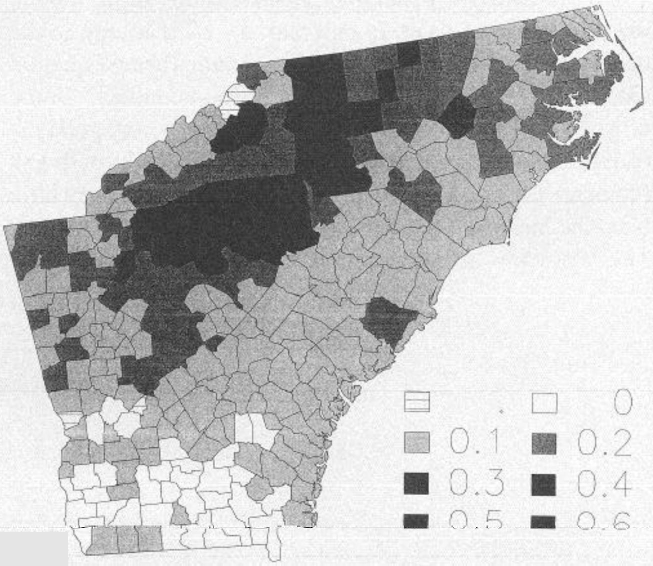

Figure 14. Estimated probability, $\hat{p}$, of outbreak from spatial logistic regression, model (2).

x_0' = design matrix for county to be predicted

Y_0 = response for county to be predicted

X_1 = design matrix for other counties

y_1 = responses for other counties,

the best linear unbiased predictor is

$$x_0'\hat{\beta} + \mathrm{Cov}(Y_0, y_1)\mathrm{Var}(y_1)^{-1}(y_1 - X_1\hat{\beta}).$$

The logistic regression model is linear for the logits; that is, $\mathrm{logit}(p) = X\beta$. We can obtain an approximate best linear unbiased predictor of the logits and transform this back to the original scale to get the spatial prediction of the probability of outbreak [also see Gotway and Stroup (1997) for a slightly different predictor]. Denoting the logit of the probability of outbreak in the county to be predicted as ν_0 and in the other counties as ν_1, the best linear unbiased predictor of ν_0 is

$$\hat{\nu}_0 = x_0'\hat{\beta} + \Sigma_{\nu 01}\Sigma_{\nu 11}^{-1}(\nu_1 - X_1\hat{\beta}),$$

where the covariance matrix of ν, Σ_ν, is approximated by

$$\Sigma_\nu = \mathrm{Var}(\nu) \approx \frac{d\nu(p)}{dp}\, Var(y) \left(\frac{d\nu(p)}{dp'}\right)'$$

and it is partitioned into

$$\Sigma_\nu = \begin{bmatrix} \Sigma_{\nu 00} & \Sigma_{\nu 01} \\ \Sigma_{\nu 10} & \Sigma_{\nu 11} \end{bmatrix}.$$

This approximation for the covariance matrix of ν comes from a first-order Taylor series expansion of ν as a function of p. To obtain the predicted logit, substitute $\hat{\beta}$ into Σ_ν.

The prediction of the urobabilitv of outbreak is obtained by transforming the predicted logit to the original scale,

To evaluate the spatial interpolation ability of the model, the proportion of years in outbreak for each county in the dataset was predicted from the other counties using Equation (3). This is called "leave-one-out cross validation." In the cross-validation, the spatial predictor $\ddot{p}$ does a very good job of reproducing the map of proportion of years in outbreak (compare Figures 1 and 15). The predictions show very little bias; the mean of the prediction errors, $Y_i - \ddot{p}_i$, is 0.00018. They also show little variability, with mean square prediction error 0.08 1, and 75% of the prediction errors lie between –0.037 and 0.035.

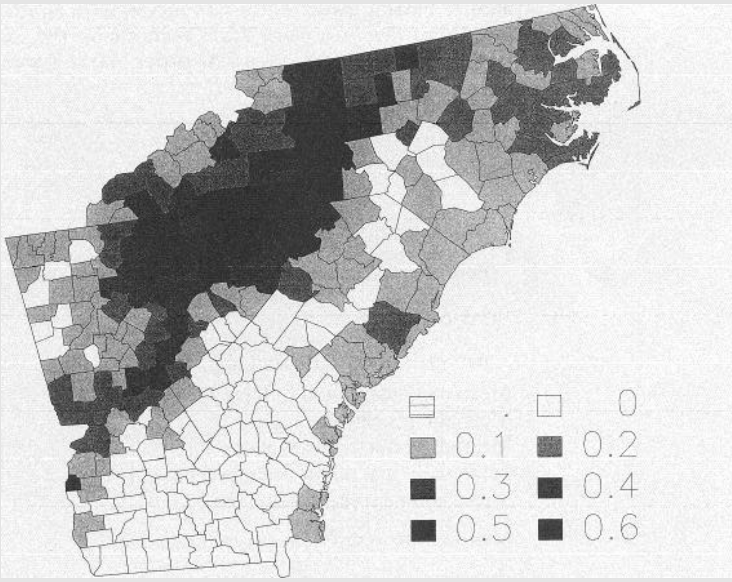

Figure 15. Predicted probability, $\ddot{p}$, [Equation (3)] of outbreak from spatial logistic regression, model (2). In $\ddot{p}$ the estimate $\hat{p}$ is adjusted for the responses of surrounding counties.

5.2 Predicting into the Future

We might also be interested in predicting the proportion of years that a particular county will experience outbreaks or the probability that a county will experience at least one outbreak in the next several years. The temporal part of this model allows us to make these types of forecasts. We have assumed that each county follows a first-order Markov process, which means that the probability of outbreak in year t depends only on whether there was an outbreak the previous year. If p_i is the marginal probability of outbreak, and p_{i11} is the probability of outbreaks in 2 consecutive years in county i, then the probabilities of an outbreak or of no outbreak in any particular year given the previous year are given in Table 2.

Southern pine beetle data for the 6 additional years 1991 through 1996 became available after the start of this project. We used these additional data to check the predictive utility of the time-series part of the model. The probability of observing any specific 6 yr string of outcomes is computed by multiplying together a string of six probabilities, each conditioned on the previous year's outcome. For example, given that there was no outbreak in 1990, the probability of observing outbreaks in 1993 and 1994 but no other years in the period 199 1 through 1996 is $p_{i0|0} \times p_{i0|0} \times p_{i1|0} \times p_{i1|1} \times p_{i0|1} \times p_{i0|0}$.

Table 2. Conditional probability of outbreak given the outbreak status of the previous year, first-order Markov process.

$$p_{i1|1} = \Pr\{\text{outbreak} \mid \text{outbreak in previous year}\} = \frac{p_{i11}}{p_i}$$

$$p_{i1|0} = \Pr\{\text{outbreak} \mid \text{no outbreak in previous year}\} = \frac{p_i - p_{i11}}{1 - p_i}$$

$$p_{i0|1} = \Pr\{\text{no outbreak} \mid \text{outbreak in previous year}\} = \frac{p_i - p_{i11}}{p_i}$$

$$p_{i0|0} = \Pr\{\text{no outbreak} \mid \text{no outbreak in previous year}\} = \frac{1 - 2p_i + p_{i11}}{1 - p_i}$$

The probability of at least one outbreak in the ith county in the 6 year period is 1 – Pr{ no outbreaks I $U_{i,1990}$}, where U_{ij} is the outbreak status of county i in year j. For each county we computed the conditional probability, labeled $P1_{i|1990}$, of at least one outbreak in 1991-1996. The 31 yr of data for each county prior to the period being forecast provide baseline information from which we can obtain a historical estimate of the probability of at least one outbreak in any 6 yr stretch given the preceding year's outbreak status. For each year that experienced an outbreak in the period 1960-1984 we tallied whether there was at least one outbreak in the 6 yr period immediately following that year. The proportion of such 6 yr periods that included at least one outbreak is labeled π_i. For counties that had an outbreak in 1990, we predict that there will be at least one outbreak in the 6 yr 1991-1996 if

$$\frac{P1_{i|1990}}{1-P1_{i|1990}} \geq \frac{1-\pi_i}{\pi_i}.$$

A similar computation was done for counties that had no outbreak in 1990. This is called a "Bayes discriminant rule." The effect of using

$$\frac{1-\pi_i}{\pi_i}$$

rather than a cutoff of 1 is to make it harder to predict an outbreak for a county if historically there have been few 6 yr periods with outbreaks and easier to predict an outbreak if there have been many 6 yr periods with at least one outbreak. Table 3 summarizes the predictions based on the first-order Markov chain model compared to the actual numbers of counties that experienced at least one outbreak in 1991-1996.

The probability that a county will experience m years of outbreaks is obtained by adding together the probabilities of all strings that contain exactly m outbreaks. The expected proportion of years in outbreak, given the 1990 data, is then { Pr(1 outbreak I $U_{i,1990}$) + 2 x Pr(2 outbreaks I $U_{i,1990}$) + . . . + 6 x Pr(6 outbreaks | $U_{i,1990}$)}/6. Over all counties the average proportion of years with outbreaks in 1991-1996 was 0.097. Using the first-order Markov model, the average predicted proportion of years in outbreak was 0.131. The estimate not using any model would be the average proportion of years in outbreak from 1960-1990, 0.143. The Markov chain model provided a modest improvement, from a 47% overprediction to a 35% overprediction. The Markov chain estimate was just as variable, however, as the naive estimate Y_i; both had root mean square prediction error close to 0.145.

Table 3 Prediction of whether a county will experience at least one southern pine beetle outbreak in the years 1991-1996 crosstabulated with observed outcome. The predictions are based on the first-order Markov process conditional on the 1990 outbreak status for each county.

Observed at	Predict at least one outbreak		No. of
least one outbreak	No	Yes	counties
No	0.82	0.18	192
Yes	0.36	0.64	109

6 Discussion

The marginal logistic regression model has several features that make it a good tool for describing the spatial pattern of southern pine beetle outbreaks. Traditional logistic regression models include the assumption that the observations are independent of each other. Data on patterns of outbreaks of pests and diseases tend to be correlated spatially and temporally, rather than being independent. Marginal models allow specification of a correlation structure in addition to a model for the mean response function. Software for fitting these types of models is rapidly becoming available (Wolfinger and O'Connell 1993, Littell et al. 1996).

We did find both spatial and temporal autocorrelation in the southern pine beetle data. The spatial and temporal correlations were incorporated into the model by assuming that the process is a first-order Markov process over time and that spatial correlation among sites has an exponential form. These are simple assumptions about the correlation structure; however, they appear to fit the data well. The modeling of spatial correlation structures is currently an active area of statistical research. One of our assumptions that is probably too simple to be realistic is that the spatial autocorrelation is stationary over the entire three-state region, and it may be possible to improve the model by relaxing that assumption.

We found that we could adequately estimate the mean probability of southern pine beetle outbreak for a county with given characteristics, and do an excellent job of spatial interpolation (spatial prediction) using the marginal logistic regression model with spatial and temporal autocorrelation. The set of variables including elevation, longitude, sawtimber volume per ha, area of national forestland, some of the physiographic variables, precipitation in fall and summer, and average daily maximum temperature in fall, winter, and summer together provided the best fit to the observed data. This should be interpreted to mean that this set of explanatory variables does a good job of describing the spatial pattern of outbreak probabilities. There are other sets of variables that would do an equally good job of predicting the outbreak probabilities. Many of the explanatory relationships revealed as significant in this model agree with observations elsewhere: the volume of sawtimber-size pines is a better predictor of SPB attacks than the volume of the smaller pulpwood stems, and the more sawtimber-size pines the better for beetle populations. National forests are managed on longer rotations than those of the forest industry, and older pine are generally more vulnerable than younger trees. Lastly, the moderate moisture status found on mesic sites may generate less resistance to attack than the more chronically stressful xeric or hydric sites. Other relationships are harder to interpret without either examining outbreaks over a broader spatial scale or including in the model the temporal dynamics of outbreaks and time-varying explanatory variables.

In this study the climate variables are highly correlated with each other, with elevation, and with the physiographic variables. It is not possible to determine which are causal effects. What we can do, however, is to conclude that counties with certain characteristics are more or less likely to see outbreaks. When doing regression on correlated variables, it is usually informative to examine more than one set of explanatory variables that have about the same amount of predictive power. Examining different models can give a better understanding of the clusters of variables that tend to appear together and tend to be highly correlated with the same phenomena. There is also room for examining other sets of variables and other functional forms of variables. Our interest here was to demonstrate the potential utility of this statistical approach rather than to definitively model the causal factors of outbreaks. Models of similar form might also be useful in other forestry applications where binary dependent variables occur, such as models of harvesting, regeneration, or land conversion decisions, or disturbance processes such as landslides or windthrow.

Finally, we mention two areas for improvement. In order to deal with the different spatial scales of the climate variables we took the simplest possible route and used averages of stations within a county to represent the county. For counties with no climatological stations, we used the value of the station nearest to the county center, thus the climate variables incorporate a high degree of error. The usual effect of large errors inregressor variables is that regression coefficients tend to appear smaller and less significant than they would if the explanatory variables were measured without error. Location of stations may also be biased toward lower elevations. More complete climate data and/or more sophisticated spatial interpolation methods for the climate variables would certainly improve parameter estimation. Even greater improvements could be obtained by developing methods of fitting and evaluating models for binary data from individual years, as opposed to data in the form of proportions. We used the proportion form of the response variable rather than the binary response for individual years for two reasons: (1) variograms and correlation functions of binary variables depend in a complicated way on the probability of outbreak and are difficult to work with; and (2) 30 yr averages of the climate variables were more readily available than averages for individual years. Judging by the cross-validation results, we were able to obtain excellent spatial interpolations, but predictions of future outbreak probabilities were not as good. To do a betterjob of predicting into the future, measurements of climate and other explanatory variables each year are needed. Because temperatures and precipitation change each year, but physiographic characteristics and elevation do not, using time-varying climate data should give better ability to distinguish between the temperature and precipitation effects and the effects of physiographic variables and elevation. It should also allow us to test whether temperatures and precipitation during the preceding year affect the probability of SPB outbreak.

Literature Cited

ALBERT, P., AND L. MCSHANE 1995. A generalized estimating equations approach for spatially correlated binary data: Applications to the analysis of neuroimaging data. Biometrics 51:627-638.

BAILEY, R. 1995. Description of the ecoregions of the United States. Misc. Publ. 1391, USDA For. Serv., Washington, DC. 108 p.

CARTER, D., J. O'LAUGHLIN, AND C. MCKINNEY. 1991. Southern pine beetle impacts and control policy in the 1982-1986 Texas epidemic. South. J. Appl. For. 15:145-153.

CHRISTENSEN, R. 1991. Linear models for multivariate, time series, and spatial data. Springer-Verlag, New York. 317 p.

COLLETT, D. 199 1. Modelling binary data. Chapman and Hall, London. 369 p.

COULSON, R., ET AL. 1983. The role of lightning in the epidemiology of the southern pine beetle. Z. Ang. Ent. 96:182-193.

CRESSIE, N. 1991. Statistics for spatial data. Wiley, New York. 900 p.

DE STEIGUER, J., R. HEDDEN, AND J. PYE. 1987. Optimal level of expenditure to control the southern pine beetle. USDA For. Serv. Res. Pap. SE-263. 30 p.

DIGGLE, P., K. LIANG, AND S.L. ZEGER. 1994. Analysis of longitudinal data. Clarendon Press, Oxford. 253 p.

FARES, Y. P. SHARPE, and C. MAGNUSON. 1980. Pheromone dispersion in a forested ecosystem. P. 75-93 In Modeling Southern Pine Beetle populations, Stephen, F.M., J.L. Searcy, and G.D. Hertel (eds.). USDA For. Serv.Tech. B ull. 1630. Washington, DC.

GAGNE, J, R. COULSON, J. FOLTZ, T. WAGNER, AND L.J. EDSON. 1980. Attackand survival of *Dendroctonus frontalis* in relation to weather during three years in east Texas. Environ. Entomol. 9:222-229.

GOTWAY, C., AND W. STROUP. 1997. A generalized linear model approach to spatial data analysis and prediction. J. Agric. Biol. Environ. Statist. 2:157-178.

GOOVAERTS, P. 1997. Geostatistics for natural resources evaluation. Oxford University Press, New York. 483 p.

GUTTORP, P. 1995 Stochastic modeling of scientific data. Chapman and Hall, London. 372 p.

HANSEN, M.H., T. FRIESWYK, J.F. CLOVER, AND J.F. KELLY. 1992. The Eastwide forest inventory database: Users manual. USDA Forest Service Gen. Tech. Rep. NC-151 48 p. Database is available online at http://www.srsfia.u sfs.msstate.edu/scripts/ew.htm Accessed 8/2/96 (host volume), 4/9/98 (physiographic variables), 3/13/98 (forestland ownership).

HINES, G., H. TAHA, AND F. STEPHEN. 1980. Model for predicting southern pine beetle population growth and tree mortality. In Modeling Southern Pine Beetle populations, Stephen, F.M., J.L. Searcy, and G.D. Hertel (eds.). USDA For. Serv.Tech. Bull. 1630. Washington, DC.

HOLMES, T. 1991. Price and welfareeffectsofcatastrophicforestdamagefrom southern pine beetle epidemics. For. Sci. 37:500-5 16.

KALKSTEIN, L. 1976. Effects of climatic stress upon outbreaks of the southern pine beetle. Environ. Entomol. 5:653-658.

KEDEM, B. 1980. Binary time series. Marcel Dekker, New York. 140 p.

LANDSEA, C. 1995. Atlantic Basin best track documentation. ftp:hrd-type42.nhc.noaa.gov/data/ linked to http://www.aoml.noaa.gov. Accessed on 1/19/98.

LIANG, K., AND S. ZEGER. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73: 13-22.

LIEBHOLD, A., R. ROSSI, AND W. KEMP. 1993. Geostatistics and geographic information systems in applied insect ecology. Annu. Rev. Entomol. 38:303-327.

LITTELL, R., G. MILLIKEN, W. STROUP, AND R. WOLFINGER. 1996. SAS System for mixed models. SAS Institute, Cary, NC. 633 p.

LORIO, JR., P. 1986. Growth-differentiation balance: a basis for understanding southern pine beetle - tree interactions. For. Ecol. Manage. 14:259-273.

MAWBY, W., AND H. GOLD. 1984. A reference curve and space-time series analysis of the regional population dynamics of the southern pine beetle (*Dendroctonus frontalis* Zimmermann). Res. Popul. Ecol. 26:261-274.

MAWBY, W.D., F.P. HAIN, AND C.A. DOGGETT. 1989. Endemic and epidemic populations of southern pine beetle: implications of the two-phase model of forest man agers. For. Sci. 35:1075-1087.

National Climatic Data Center. 1995. Cooperative Station Data CD-ROM. Also available online at NCDC Online Inventory COOP.TXT. http://www.ncdc.noaa.gov/pub/data/inventories/COOP-ACT.TXT. Accessed 5/21/98.

OWENRY, J.R., AND D.S. EZELL. 1992. Monthly station normals of temperature, precipitation and heating and cooling degree days 1961-90. Climatography of the United States, no. 81. National Ocean and Atmospheric Administration, National Climatic Data Center, Ashville, NC. Available online at http://www.dnr.state.sc.us/water/climate/sercc/climate_atlasdata.html. Accessed 5/l 5/98.

PREISLER, H., N. RAPPAPORT, AND D. WOOD. 1997. Regression methods for spatially correlated data: An example using beetle attacks in a seed orchard. For. Sci. 43:71-77.

PRICE, T., AND C. DOGGETT. 1982. A history of southern pine beetle outbreaks in the southeastern United States. Tech. rep., South. For. Insect Work. Group, GA For. Comm., Macon, GA. 35 p.

PRICE, T., C. DOGGETT, J. PYE, AND B. SMITH. 1998. A history of southern pine beetle outbreaks in the southeastern United States. Tech. rep., South. For. Insect Work. Group, GA For. Comm., Macon, GA. 71 p.

PYE, J.M. 1993. Regional dvnamics of southern pine beetle *populations* P. I1 l-l 24 in Proc. Spatial analysis and forest pest management, Liebhold, A.M., and H.R. Barrett (eds.). USDA For. Serv. Gen. Tech. Rep. NE-17.5.

SAS INSTITUTE INC. 1997. SAS/STAT® Software: Changes and enhancements through Release 6.12. SAS Institute Inc., Cary, NC

TURCHIN, P., P. LORIO, A. TAYLOR, AND R. BILLINGS. 1991. Why do populations of southern pine beetles (Coleoptera: Scolytidae) fluctuate? Environ. Entomol. 20:4 01-409.

UNGERER, M.J., M.P. AYRES, AND M.J. LOMBARDERO. 1999. Climate and the northern distribution limits of *Dendroctonus frontalis* Zimmermann (Coleoptera: Scolytidae). J. Biogeography. In press.

WOLFINGER, R., AND M. O'CONNELL. 1993. Generalized linear mixed models: a pseudolikelihood approach. J. Stat. Comput. Simulat. 48.233-243.