

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

by

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

Institute of Statistics Mimeo Series Number 2513

January, 1999

NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina

The Library of the Department of Statistics
North Carolina State University

Mimeo # 2513

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

BY:

Marcia L. Gumpertz, Chi-tsung Wu,
and John M. Pye
Institute of Statistics Mimeo
Series No. 2513, Jan. 1999
North Carolina State University

Name

Date

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

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

Author Citation

- Marcia L. Gumpertz, Associate Professor, Dept. of Statistics, Box 8203, North Carolina State University, Raleigh, NC 27695-8203, E-mail: gumpertz@ncsu.edu, phone: (919)859-1753, fax: (919)859-1169
- Chi-tsung Wu, Postdoctoral Associate, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan 11529 R.O.C., Email: cwu@stat.sinica.edu.tw
- John M. Pye, Ecologist, USDA Forest Service Southern Research Station, 3041 Cornwallis Rd., P. O. Box 12254, Research Triangle Park, NC 27709, Email: jpye/srs_rtp@fs.fed.us, phone:(919)549-4013

Acknowledgments

Thanks to Thomas P. Holmes of the U.S. Forest Service for initiating this research project. We also wish to thank James O. Paumier for invaluable help in obtaining and understanding the climatological data, to Sharon LeDuc for providing data, and to 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.

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 consist of binary data, 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 years. In the proposed 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 yields improved estimates of the probability of outbreak for a particular county and of the importance of the various explanatory variables than would otherwise be obtained.

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 1991; 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 1, 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.

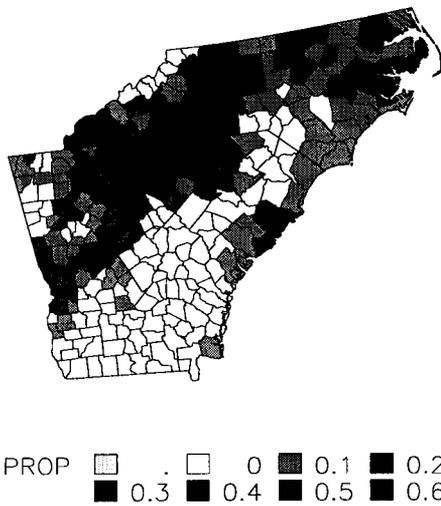


Figure 1: North Carolina, South Carolina, and Georgia proportion of years in outbreak.

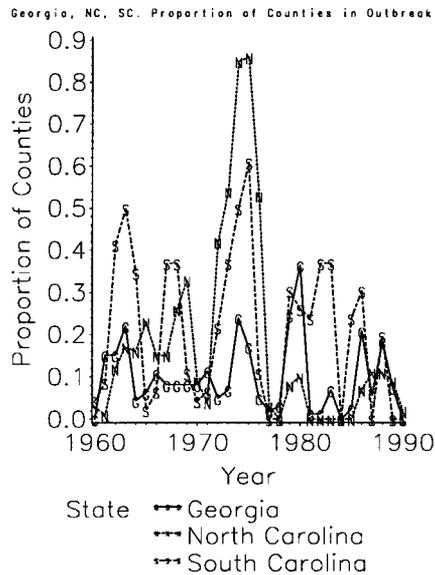


Figure 2: Proportion of counties in outbreak in NC, SC, and GA

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), and this distribution has been generally attributed to host factors. 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.

Temporal patterns: Southern pine beetle populations vary dramatically over time, oscillating between endemic and outbreak years, where outbreak is defined as at least one southern pine beetle spot infestation per thousand acres 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 (1993) cited a cycle length of 6-7 years for recent outbreaks spanning the Southern U.S. 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 one and two years 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, Ayres and Lombardero (in review) 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. Climate can affect population 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 affect population level 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).

Evaluation of the likely impacts of broad scale changes in the region, changes such as those due to: 1) regionwide changes in host forest types (Mawby et al. 1989), 2) warmer temperatures from increased atmospheric CO₂ or other causes (Ungerer, Ayres, and Lombardero in review), and 3) lengthened rotations on National Forests, potentially increasing outbreaks on surrounding private forests (Carter et al. 1991) requires improved statistical models which

simultaneously account for the spatial processes 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. 1997). 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 the present study is to demonstrate use of logistic regression for spatially correlated proportions, while also incorporating information about temporal autocorrelation. We use a 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. The response variable is the proportion of years each county experiences outbreaks. The explanatory variables include two measures of volume of host timber, three physiographic variables, three climate variables averaged over fall, winter, spring, and summer months, average number of hurricanes, amount of national forest acreage, elevation and latitude and longitude. As used here, the explanatory variables vary over space but not over time. That is, we have one measurement of each of the explanatory variables for each county. Thus, the explanatory variables 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 other explanatory variables than those presented here.

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. The parameters of the logistic regression model quantify the change in log odds of outbreak for changes in the regressor variables. A more technical advantage of logistic regression over ordinary regression has to do with the methods of estimation. For uncorrelated normally distributed data, ordinary least squares is the most efficient method of estimation, but it is not efficient for binary and proportion data. Logistic regression incorporates information about the variance of binary/proportion data into the estimating equations to provide more efficient estimates than ordinary regression would.

Using a model that incorporates spatial autocorrelation provides advantages both for estimation of the effects of the regressor variables and for estimation of the probability of outbreak in a given county. Incorporating the spatial autocorrelation has the potential to provide more efficient (i.e. more precise) estimates of the effects of the regressor variables. Just as important is the fact that if spatial and temporal autocorrelation are present but not acknowledged, the standard errors of regression coefficients from ordinary regression could seriously misrepresent the actual variability of the regression coefficients. Fitting a model that incorporates spatial and temporal autocorrelation into the covariance matrix guarantees that the autocorrelation will be accounted for in parameter estimates and their standard errors. The greatest benefit of incorporating spatial autocorrelation into the model is that 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 logistic regression model with correlated observations that we use belongs to the

general class of models known as marginal models (Liang and Zeger 1986; Diggle et al. 1994). The model allows for covariates (explanatory variables) and for spatial and temporal correlation. 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 mean response 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.

In this paper we demonstrate the fitting and interpretation of a marginal logistic regression model for describing the conditions and variables associated with southern pine beetle outbreaks. Sections 2 and 3 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 2.2 reveals spatial and temporal autocorrelation, indicating that ordinary logistic regression is not the best procedure for these data. In section 3.1 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 3.2, is to incorporate spatial correlation and fit the model using generalized estimating equations. Section 3.3 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 3.4 discusses the use of this model for making predictions into the future, and section 4 gives a general discussion of our conclusions.

2 Preliminary Analysis

Several explanatory variables were considered, all at the county level. These included two measures of volume of host trees: sawtimber volume, suitable to make boards and other lumber (million cu ft/acre) and poletimber volume, smaller trees of commercial value (million cu. ft/acre); three physiographic variables: percent 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, average daily maximum temperature, and average daily precipitation; number of 6 hour periods with hurricane force winds recorded in the county from 1960 to 1990; two management variables: national forest acreage and forest industry acreage; and three location variables: elevation, latitude and longitude. The estimates of host volume, physiographic variables, U.S. Forest Service and forest industry acreage per county were obtained from the U.S. Forest Service's Forest Inventory and Analysis Data Base Retrieval System (<http://www.srsfia.usfs.msstate.edu/scripts/ew.htm>, 8/2/96 (host volume), 4/9/98 (physiographic variables), 3/13/98 (forestland ownership)). The 30 year monthly climatological averages for each station were obtained from the Southeast Regional Climate Center website for Climatological Normals 1961-1990 (http://water.dnr.state.sc.us/climate/sercc/norm_station.html 5/14/98). All stations within a county were averaged together to obtain the monthly county average. Sets of 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, thus the climate variables incorporate a high degree of error. The usual effect of large errors in regressor variables is that regression coefficients tend to appear smaller and less significant than they would if the explanatory variables were measured without error. Elevation, latitude and longitude of the stations were obtained from the National Climatic Data Center's coop station master list

(<http://www.ncdc.noaa.gov/pub/data/inventories/COOP-ACT.TXT> 5/21/98).

2.1 Selection of Variables

The general impression of the spatial pattern of host volume (Figs. 3, 4) is similar to the pattern of southern pine beetle outbreaks (Fig. 1), 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 (Figs. 5,6) and proportion of years (out of 31) in outbreak. Elevation increases from the coast in the southeast to the mountains in the northwest (Fig. 7). Precipitation roughly follows the elevation pattern (Fig. 9,10), and temperature also increases from northwest to southeast (e.g. Fig. 8).

Many transformations and combinations of these explanatory variables are possible. A preliminary analysis using ordinary logistic 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 the logit of \tilde{p}_i using ordinary logistic regression, where \tilde{p}_i equals the proportion of years the i^{th} county was in outbreak. Variables that appeared in several of the models with high coefficients of determination (R^2) and low C_p values were selected. The set of explanatory variables selected for further consideration included $\ln(\text{elevation})$, longitude, square roots of saw and pole timber volume per acre, square roots of percent land area classified as xeric, mesic, and hydric, square roots of number of acres owned by the Federal government and number of acres owned by private forest industry, average number of hurricane wind periods per year, all twelve of the climate variables, and the five interactions that had the smallest “p-values”: hurricanes per year $\times \sqrt{\text{mesic}}$, maximum summer temperature $\times \sqrt{\text{saw volume}}$, maximum winter temperature $\times \sqrt{\text{saw volume}}$, average spring precipitation $\times \sqrt{\text{pole volume}}$, and minimum fall temperature $\times \sqrt{\text{saw volume}}$.

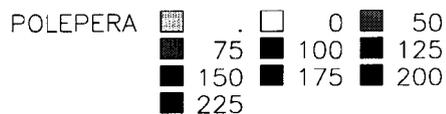
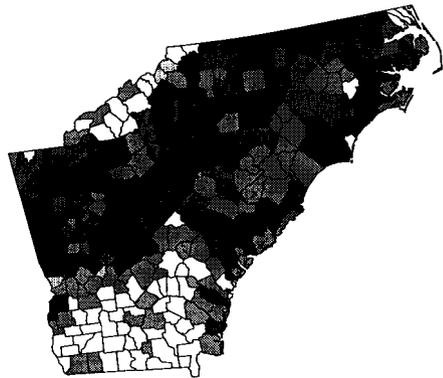


Figure 3: Pole volume per acre
(million cu. ft. per acre)

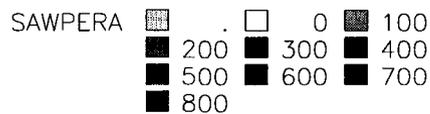
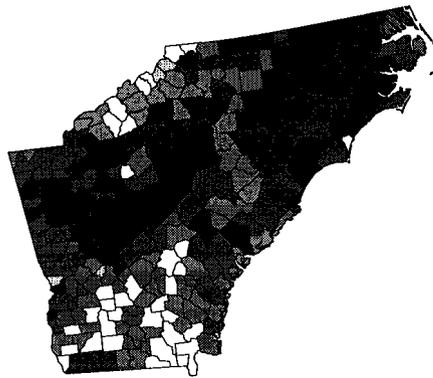


Figure 4: Saw volume per acre (mil-
lion cu. ft. per acre)

2.2 Residuals from Ordinary Logistic Regression

Deviance residuals measure the deviations of the fitted values, \hat{p}_i , from the observed proportion of years in outbreak, \tilde{p}_i , for each county (see Crockett 1991),

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

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. The form of the deviance residuals is not entirely intuitive, but we can see that it 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 1991). The standardized deviance residuals from the ordinary logistic regression of

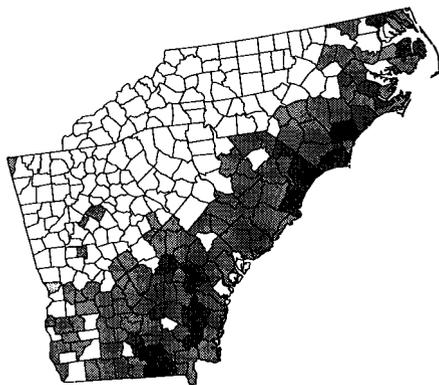


Figure 5: Proportion of land area classified as hydric.

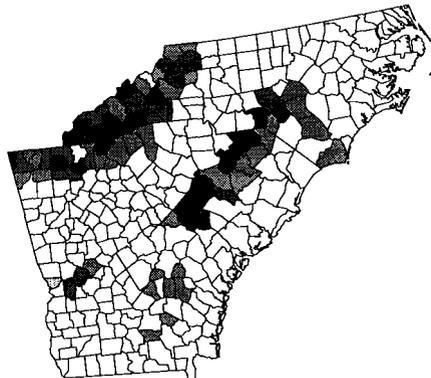


Figure 6: Proportion of land area classified as xeric.

$\text{logit}(\tilde{p}_i)$ 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),

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

where h =distance between two counties and $N(h)$ is the number of pairs of counties h miles apart. Note that the distance should be computed in some linear units such as miles or meters, not in degrees of latitude and longitude. This may be accomplished either by (1) computing great circle distances between points with latitude and longitude coordinates, or (2) first projecting latitude and longitude onto a Euclidean coordinate system such as Universal Transverse Mercator (UTM) and then computing Euclidean distances. The great circle distances were used here.

The semivariogram of the standardized deviance residuals shows that counties are spatially correlated to a distance of about 100 miles; beyond this distance counties are essentially

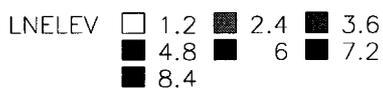


Figure 7: Ln(elevation) in ft.



Figure 8: Mean daily maximum temperature (F), Summer.

uncorrelated (Fig. 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”.

3 Incorporating Temporal and Spatial Autocorrelation

3.1 Temporal Autocorrelation

The reason that the sill (the observed variance) is so high is that outbreaks in a given county are correlated from year to year. Ordinary logistic regression assumes that observations are uncorrelated and assumes that the proportion of years in outbreak is a binomial random variable with variance $p_i(1 - p_i)/n$, where $n=31$ years and p_i =probability of outbreak in county i . Since the outbreaks in a county are correlated from one year to the next, the variance of the proportion is not simply $p_i(1 - p_i)/n$, but rather it is

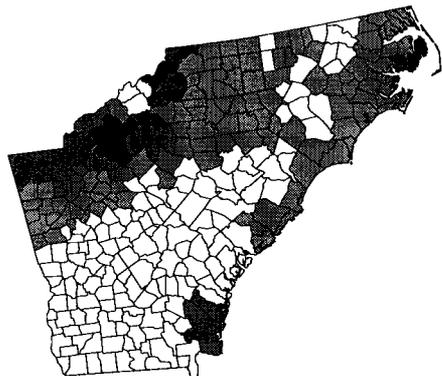


Figure 9: Mean Fall precipitation (in).

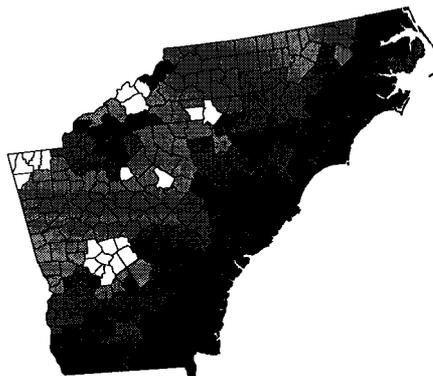


Figure 10: Mean Summer precipitation (in).

$$\text{Var}(\tilde{p}_i) = \frac{p_i(1-p_i)}{n} + \frac{p_i(1-p_i)}{n^2} \sum_j \sum_{k \neq j} \text{Corr}(U_{i,j}, U_{i,k}),$$

where $U_{i,j}$ is the outbreak status (0 or 1) in county i in year j . 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, $\text{Var}(\tilde{p}_i) = p_i(1-p_i)/n$. If the correlation from one time to another within a county is positive, the variance of \tilde{p}_i is larger than $p_i(1-p_i)/n$.

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. The probability of outbreak for the i^{th} 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

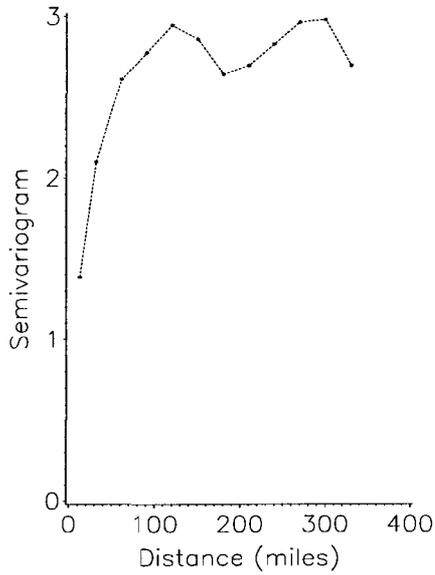


Figure 11: Semivariogram of standardized deviance residuals from ordinary logistic regression on 27 variables. This model does not take spatial or temporal autocorrelation into account.

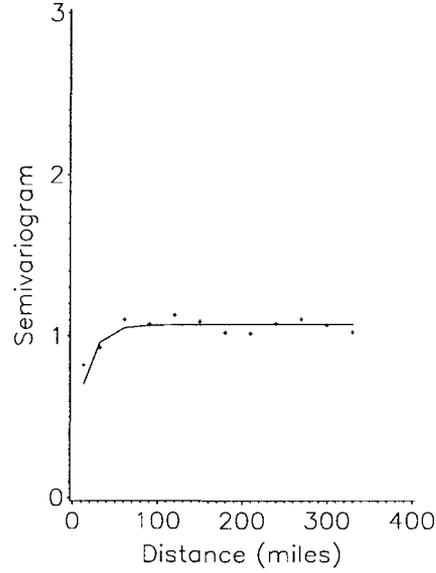


Figure 12: Empirical and fitted semivariograms of standardized deviance residuals from weighted logistic regression on 27 variables. Temporal autocorrelation is incorporated into the variance of \tilde{p}_i .

between a particular year and the previous year, so if the correlation between two consecutive years is .6, then the correlation between observations two years apart is $.6^2 = .36$ under the first order Markov model. The correlation parameter is estimated for each site by computing the first order autocorrelation coefficient

$$\hat{\rho}_i = \frac{\frac{1}{n_i-1} \sum_{j=1}^{n_i-1} (U_{ij} - \tilde{p}_i)(U_{i,j+1} - \tilde{p}_i)}{\tilde{p}_i (1 - \tilde{p}_i)}.$$

(Kedem 1980, p. 70).

The total correlation among pairs of years within a site is

$$\begin{aligned} \text{TOTCORR}_i &= \sum_j \sum_{k \neq j} \text{Corr}(U_{ij}, U_{ik}) \\ &= \sum_j \sum_{k \neq j} \rho_i^{\|j-k\|} \end{aligned}$$

$$= 2 \frac{\rho_i}{1 - \rho_i} ((n_i - 1) - \frac{\rho_i}{1 - \rho_i} (1 - \rho_i^{n_i - 1})).$$

Estimated temporal autocorrelation coefficients range from -.21 to .78 for counties that ever had any southern pine beetle outbreaks, with 50% of the counties having $\hat{\rho}_i > .32$. Thus there is a substantial amount of correlation over time within a county and the amount of 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 two years, 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) (see p. 71) with significance level .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 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 \tilde{p} . The quasi-likelihood method involves iteratively solving the system of equations

$$D'V^{-1}(y - \mu) = 0,$$

where the i^{th} element of y is \tilde{p}_i , $D = \frac{\partial \mu}{\partial \beta}$ and V is a diagonal matrix with i^{th} diagonal element $\text{Var}(\tilde{p}_i)$. 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 doing weighted logistic regression, replacing the binomial variance expression with the appropriate variance of \tilde{p}_i . It is easy to implement using software that allows weighted logistic regression. For example in SAS Proc Logistic, one would regress \tilde{p}_i on the

explanatory variables and specify weights $w_i = 1/(1 + TOTCORR_i/n_i)$.

The standardized deviance residuals resulting from fitting this model still show spatial autocorrelation, but now the sill is close to one (Fig. 12). The exponential semivariogram fitted to the empirical semivariogram by weighted nonlinear least squares (Cressie 1991) is $\hat{\gamma}(h) = .24 + .83(1 - e^{-3h/46.8})$. This function has a sill of 1.07 and a range of 47 miles. 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 \tilde{p}_i seems to account for all of the “overdispersion” seen in the data.

3.2 Incorporating Spatial Autocorrelation

The last refinement to the model is to incorporate the spatial autocorrelation. We use the fitted exponential semivariogram function to estimate the pairwise covariances among the counties. The method of generalized estimating equations (Liang and Zeger 1986; Gotway and Stroup 1997) can be used to estimate the parameters of a marginal logistic regression model with spatial correlation. The procedure is to iteratively solve the equations

$$D'V^{-1}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}$$

where

$$\begin{aligned} V &= A^{1/2}RA^{1/2} \\ A &= \text{diag}\left(\frac{p_i(1-p_i)}{n_i}(1 + TOTCORR_i/n_i)\right) \\ R_{i,i^*} &= \text{Corr}(Y_i, Y_{i^*}) = \frac{C_1}{C_0 + C_1}e^{-3h/a}, \end{aligned}$$

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

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 may be tested using the Wald statistic. The general form is $W = (L\hat{\beta})'\text{Var}(\hat{\beta})^{-1}L\hat{\beta}$, where $\text{Var}(\hat{\beta}) = (D'V^{-1}D)^{-1}$, and the hypothesis being tested is $H_0: L\beta = \mathbf{0}$. For large samples it has 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 $\alpha = .10$ 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 .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 visually corresponds most closely with the pattern of outbreaks (Fig. 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 saw timber the estimated probability of outbreak increases as summer daily maximum temperature increases. Looked at the other way, probability of outbreak increases with volume of saw timber per acre but the rate of increase depends on the daily maximum temperature (Fig. 13). Note that the summer and winter mean daily maxima are very highly correlated with each other, making it next to impossible 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 Figs. 14 and 1), but smoothes the proportions somewhat. The result is that the estimated probabilities in the high-outbreak counties are lower than the observed proportions of years in outbreak, and the reverse is true in the low-outbreak counties.

Parameter	Estimate	Std. Error	P-value
Intercept	-16.16	7.58	0.033
ln(elevation)	0.21	0.10	0.039
longitude	0.21	0.052	0.00007
$\sqrt{\text{saw volume}}$	1058	363	0.0035
$\sqrt{\text{hydric}}$	-1.48	0.53	0.0055
$\sqrt{\text{xeric}}$	-0.92	0.38	0.016
$\sqrt{\text{national forest}}$	0.061	0.022	0.0049
ave. maximum fall temperature	-0.42	0.19	0.025
average fall precipitation	0.89	0.16	1.1E-8
ave. maximum winter temperature	-0.10	0.14	0.45
ave. maximum summer temperature	0.74	0.15	9.1E-7
average summer precipitation	-0.30	0.13	0.026
max summer temp $\times \sqrt{\text{saw volume}}$	-26.76	7.41	0.00031
max winter temp $\times \sqrt{\text{saw volume}}$	23.52	6.71	0.00045
c_0	.24		
c_1	.84		
a	60		

Table 1: Final logistic regression model with spatial autocorrelation. Parameter estimates, standard errors, and p-values from Wald tests using the exponential covariance model.

3.3 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, $\hat{\beta}$, and the estimated probability for an individual county is $\hat{p} = \frac{e^{X\beta}}{1+e^{X\beta}}$. 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

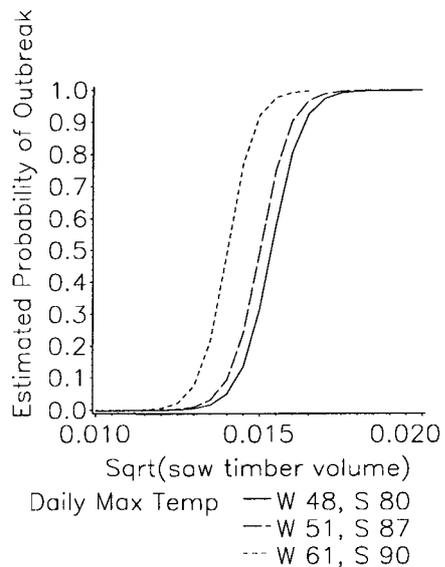


Figure 13: Estimated probability of outbreak vs. saw timber volume per acre, with separate lines for different maximum summer and winter temperatures.

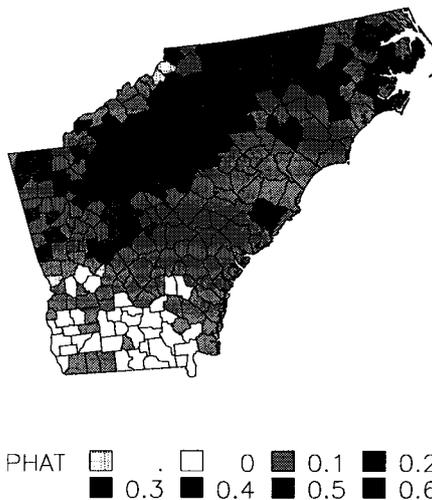


Figure 14: Estimated probability, \hat{p} , of outbreak from spatial logistic regression.

kriging predictor, 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

- \mathbf{x}'_0 = design matrix for county to be predicted
- y_0 = response for county to be predicted
- \mathbf{X}_1 = design matrix for other counties
- \mathbf{y}_1 = responses for other counties,

the best linear unbiased predictor is

$$\mathbf{x}'_0 \hat{\boldsymbol{\beta}} + \text{Cov}(y_0, \mathbf{y}_1) \text{Var}(\mathbf{y}_1)^{-1} (\mathbf{y}_1 - \mathbf{X}_1 \hat{\boldsymbol{\beta}}). \quad (1)$$

The logistic regression model is linear for the logits; that is, $\text{logit}(p) = X\boldsymbol{\beta}$. 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 $\boldsymbol{\nu}_1$, the best linear unbiased predictor of ν_0 is

$$\hat{\nu}_0 = X_0 \hat{\boldsymbol{\beta}} + \Sigma_{\nu 01} \Sigma_{\nu 11}^{-1} (\boldsymbol{\nu}_1 - X_1 \hat{\boldsymbol{\beta}}), \quad (2)$$

where the covariance matrix of $\boldsymbol{\nu}$, Σ_{ν} , is approximated by

$$\Sigma_{\nu} = \text{Var}(\boldsymbol{\nu}) \approx \frac{d\boldsymbol{\nu}(\mathbf{p})}{d\mathbf{p}'} \text{Var}(\mathbf{y}) \left(\frac{d\boldsymbol{\nu}(\mathbf{p})}{d\mathbf{p}'} \right)' \quad (3)$$

and it is partitioned into

$$\Sigma_{\nu} = \begin{bmatrix} \Sigma_{\nu 00} & \Sigma_{\nu 01} \\ \Sigma_{\nu 10} & \Sigma_{\nu 11} \end{bmatrix}. \quad (4)$$

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

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

$$\hat{p}_0 = \frac{e^{\hat{\nu}_0}}{1 + e^{\hat{\nu}_0}}.$$

In a leave-one-out crossvalidation, the spatial predictor, \hat{p} does a very good job of reproducing the map of proportion of years in outbreak (compare Figs. 1 and 3.3). The predictions show very little bias; the mean of the prediction errors, $\tilde{p}_i - \hat{p}_i$, is .00018. They also show little variability with mean square prediction error .081, and 75% of the prediction errors lie between -.037 and .035.

3.4 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

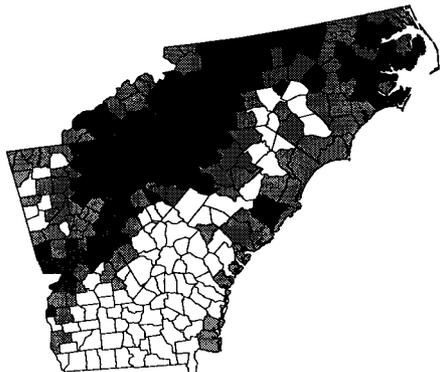


Figure 15: Predicted probability, \hat{p} , of outbreak from spatial logistic regression. In \hat{p} the estimate (\hat{p}) is adjusted for the responses of surrounding counties.



$p_{i1 1}$	=	$\Pr \{ \text{outbreak} \mid \text{outbreak in previous year} \}$	=	$\frac{p_{i11}}{p_i}$
$p_{i1 0}$	=	$\Pr \{ \text{outbreak} \mid \text{no outbreak previous year} \}$	=	$\frac{p_i - p_{i11}}{1 - p_i}$
$p_{i0 1}$	=	$\Pr \{ \text{no outbreak} \mid \text{outbreak previous year} \}$	=	$\frac{p_i - p_{i11}}{p_i}$
$p_{i0 0}$	=	$\Pr \{ \text{no outbreak} \mid \text{no outbreak previous year} \}$	=	$\frac{1 - 2p_i + p_{i11}}{1 - p_i}$

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

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 two 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 six additional years 1991 through 1996 became available after the start of this project. We use these additional data to check the predictive utility of the time-series part of the model. The probability of observing any specific 6-year string of

outcomes is computed by multiplying together a string of 6 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 1991 through 1996 is $p_{i0|0} \times p_{i0|0} \times p_{i1|0} \times p_{i1|1} \times p_{i0|1} \times p_{i0|0}$.

The probability of at least one outbreak in the i^{th} county in the six year period is $1 - \Pr\{\text{no outbreaks} \mid U_{i,1990}\}$, 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 years 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-year 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-year period immediately following that year. The proportion of such 6-year periods that included at least one outbreak is labeled π_i . We predict that there will be at least one outbreak in the 6 years 1991-1996 if $\frac{P1_{i|1990}}{1-P1_{i|1990}} \geq \frac{1-\pi_i}{\pi_i}$. 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-year periods with outbreaks and easier to predict an outbreak if there have been many 6-year 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.

Observed at Least One Outbreak	Predict at Least One Outbreak		Number of Counties
	No	Yes	
No	.82	.18	192
Yes	.36	.64	109

Table 3: Predicted vs observed proportion of counties that experienced at least one southern pine beetle outbreak in the years 1991-1996. The predictions are based on the first order Markov process conditional on the 1990 outbreak status for each county.

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 $\{ \Pr(1 \text{ outbreak} | U_{i,1990}) + 2 \times \Pr(2 \text{ outbreaks} | U_{i,1990}) + \dots + 6 \times \Pr(6 \text{ outbreaks} | U_{i,1990}) \} / 6$. Over all counties the average proportion of years with outbreaks in 1991-1996 was .097. Using the first order Markov model the average predicted proportion of years in outbreak was .131. The estimate not using any model would be the average proportion of years in outbreak from 1960-1990, .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 \tilde{p}_i ; both had root mean square prediction error close to .145.

4 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, saw timber volume per acre, number of acres of national forest land, 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: sawtimber-size pines more commonly are attacked by SPB than the smaller pulpwood stems, and the more of them the better for beetle populations. National Forests manage their forests on longer rotations than 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 say 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. Looking at 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 definitely 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. 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. 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 better job 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.

References

- 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, D.C. 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.

- Collett, D. (1991). *Modelling Binary Data*. London: Chapman and Hall.
- Coulson, R., P. Hennier, R. Flamm, E. Rykiel, L. Hu, and T. Payne (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*. New York: John Wiley.
- de Steiguer, J., R. Hedden, and J. Pye (1987). Optimal level of expenditure to control the southern pine beetle. Research Paper SE-263, USDA For. Serv., Southeastern Forest Experiment Station, Asheville, NC. 30 p.
- Diggle, P., K. Liang, and S. L. Zeger (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.
- Fares, Y., P. Sharpe, and C. Magnuson (1980). Pheromone dispersion in a forested ecosystem. Tech. Bull. 1630, USDA For. Serv., Washington, D.C. P. 75-93 in Stephen, F.M., Searcy, J.L. and Hertel, G.D. (eds.) *Modeling Southern Pine Beetle Populations*.
- Gagne, J., R. Coulson, J. Foltz, T. Wagner, and L. J. Edson (1980). Attack and 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. *JABES* 2, 157–178.
- Guttorp, P. (1995). *Stochastic Modeling of Scientific Data*. London: Chapman and Hall.
- Hines, G., H. Taha, and F. Stephen (1980). Model for predicting southern pine beetle population growth and tree mortality. Tech. Bull. 1630, USDA For. Serv., Washington, D.C. P. 4-12 in Stephen, F.M., Searcy, J.L. and Hertel, G.D. (eds.) *Modeling Southern Pine Beetle Populations: Symposium Proceedings*.
- Holmes, T. (1991). Price and welfare effects of catastrophic forest damage from southern pine beetle epidemics. *For. Sci.* 37, 500–516.
- Kalkstein, L. (1976). Effects of climatic stress upon outbreaks of southern pine beetle. *Environ. Entomol.* 5, 653–658.

- Liang, K. and S. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.
- Liebhold, A., R. Rossi, and W. Kemp (1997). 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*. Cary, NC: SAS Institute.
- Lorio, Jr., P. (1986). Growth-differentiation balance: a basis for understanding southern pine beetle - tree interaction. *For. Ecol. Man.* 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 zimmermanni*). *Res. Popul. Ecol.* 26, 261–274.
- Mawby, W. D., R. P. Hain, and C. A. Doggett (1989). Endemic and epidemic populations of southern pine beetle: implications of the two=phase model for forest managers. *For. Sci.* 35, 1075–1087.
- 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. Technical report, GA For. Comm., Macon, GA. By the Southern Forest Insect Working Group, 35 p.
- Price, T., C. Doggett, J. Pye, and B. Smith (1998). A history of southern pine beetle outbreaks in the southeastern united states. Technical report, GA For. Comm., Macon, GA. By the Southern Forest Insect Working Group, 71 p.
- Turchin, P., P. Lorio, A. Taylor, and R. Billings (1991). Why do populations of southern pine beetles (coleoptera: scolytidae) fluctuate? *Environ. Entomol.* 20, 401–409.
- Wolfinger, R. and M. O’Connell (1993). Generalized linear mixed models—a pseudolikelihood approach. *J. Stat. Comput. Simulat.* 48, 233–243.