Regression Methods for Spatially Correlated Data: An Example Using Beetle Attacks in a Seed Orchard

Haiganoush K. Preisler, Nancy G. Rappaport, and David L. Wood

ABSTRACT. We present a statistical procedure for studying the simultaneous effects of observed covariates and unmeasured spatial variables on responses of interest. The procedure uses regression type analyses that can be used with existing statistical software packages. An example using the rate of twig beetle attacks on Douglas-fir trees in a seed orchard illustrates the problem and the recommended statistical procedures. Use of this procedure revealed spatial patterns in beetle attacks that were not otherwise evident. We also found that the covariates, tree vigor, cone crop, and host genotype (i.e., clone) apparently had significant effects on number of beetle attacks within a model that also included location effects. This procedure should prove useful in other situations where spatial trends are of interest in themselves and are not simply viewed as nuisance parameters to be dealt with by the proper experimental design. For. Sci. 43(1):71–77.

Additional Key Words: Douglas-fir, nonparametric regression, Poisson counts, spatial statistics, twig beetles, *Pityophthorus orarius*.

Entomological data are often observed across geographic space (e.g., numbers and characteristics of insects on trees in a stand, or in plots within an experimental field or in regions within a country). Most insect data observed over a geographic area show a spatial structure (clustering) because of the similarity of observations at neighboring points. Correlations among values in contiguous areas can sometimes be explained in terms of observed covariates that are also spatially correlated. Similarities between contiguous areas can also result from some unsuspected or unmeasured variables. In traditional experimental design, potential spatial effects are usually dealt with by randomization or blocking of treatments. However, when values from neighboring points are not independent of one another, such methods are not optimal because they fail to estimate or characterize the role of spatial location which may be of intrinsic interest to the scientist. In addition, such methods may confound the interpretation of other measured effects, which may be simply artifacts resulting from undetected spatial correlations.

Analyses of insect populations commonly ignore the spatial structure of the data; multiple regression type methods are used to study the effects of observed covariates on the response variable. Such methods could be adequate if correlation among the observations were fully explained by the known covariates. Some recent studies have used geostatistical methods to quantify spatial patterns of insect populations. With this approach, correlations among neighboring values are modeled as a function of the geographic distances between the points (variagram). Values of responses at unsampled locations are estimated by a weighted average of nearby samples (kriging). Geographic information systems (GIS) are then used to produce maps of insect populations that are overlaid with maps of other covariates (e.g., weather or soil characteristics) for studies of the effects of known covariates and other unmeasured spatial variables on the insect populations [see Liebhold et al. (1993) for a list of references on the application of GIS and geostatistics to entomological data]. Another approach is to model the conditional probabilities of outcomes at a particular location given the observed values at neighboring locations. Examples of such models are the auto-logistic or auto-Poisson models (Besag 1972, Cressie 1993, Preisler 1993, Preisler and Mitchell 1993). Other related methods include the marked point process procedures described in Penttinen et al. (1992).

Here, we present a regression method for the analysis of spatially correlated data. This approach uses spatial location as a predictor variable in a regression model with the other

Haiganoush K. Preisler and Nancy G. Rappaport, USDA Forest Service, Pacific Southwest Research Station, P.O. Box 245, Berkeley, CA 94701, and David L. Wood, Department of Environmental Science, Policy, and Management, University of California Berkeley, CA 94720.

Manuscript received June 9, 1995. Accepted December 20, 1995.

Reprinted from *Forest Science*, Vol. 43, No. 1, February 1997

usual set of observed covariates. Cressie (1993) suggests analyzing data collected on a grid by fitting a model with additive row and column effects. Another model with explanatory variables can then be used to investigate reasons for the spatial clusterings that might be detected from the purely spatial model. Cressie (1993) also recommends using variance stabilizing transformations before fitting the various models. In this paper we used a regression method that fitted spatial trends and explanatory variables simultaneously in one model. We used a nonparametric local regression routine to estimate the spatial trend as a smooth surface. Consequently, we were able to fit spatial trends with more general functional forms than those with additive row and column effects. We also used the techniques of generalized additive models (see Section 2.1) which enabled us to model directly the distribution of the observed counts without having to resort to variance stabilizing transformations. Other articles that describe the fitting of spatial and other covariates in a regression model include Cook and Pocock (1983), Cleveland et al. (1992), and Brillinger (1994).

We demonstrate the regression approach for analysis of spatial data with an application involving the number of twig beetle attacks per tree in a Douglas-fir seed orchard. The Douglas-fir twig beetle, *Pityophthorus orarius* Bright, mines and kills the tips of developing twigs that bear two generations of cones, i.e., the primordia of the next season's cone crop and the currently developing conelets. The beetles have only one generation per year. Females emerge in the spring, mate, then attack developing shoots. Only one egg is laid in each shoot (Hedlin and Ruth 1970).

The absolute number of cones killed is probably too low to be of economic concern, even in a production seed orchard (Rappaport and Wood 1994). This insect-host system is of considerable biological interest, however, because (1) beetle distributions can be easily and accurately measured, (2) seed orchards are usually clonal (Cress and Daniels 1990), allowing analysis of the effect of host genotype, (3) seed orchards are planted in grids, providing a regular framework for the study of spatial distributions, and (4) repeated measurements can easily be made at annual increments, facilitating analysis of temporal changes in insect distribution and abundance. This system, then, provides a model for the study of spatial and temporal aspects of insect-host interactions. Spatial effects are potentially extremely important in the assessment of treatment effects in seed orchards, but they are rarely accounted for in such analyses.

Besides the practical issues described above, the study of twig beetles is important because of the recent introduction of pitch canker, *Fusarium subglutinans* f. sp. *pini*, to coastal California (Storer et al. 1995). Twig beetles in the genus *Pityophthorus* Eichhoff and cone beetles in the closely related genus *Conophthorus* Hopkins have been implicated as vectors of pitch canker to California pines (Hoover et al. 1995, Dallara et al. 1995). No treatment is known for this disease, which infects nearly all pine species that have been tested; it was recently found in Douglas-fir as well (Storer et al. 1995), raising the danger of transmission to commercial Douglas-fir forests in the Sierra Nevada and Coast Ranges of California. These issues underscore the importance of developing a complete understanding of twig-beetle-host relationships, especially with regard to those factors that influence the development of beetle infestations.

1 Field Data

Field data, originally reported in Rappaport and Wood (1994), were collected at the Little River Seed Orchard (Louisiana-Pacific Corporation), which is located about 12 km inland from the Pacific Ocean in Humboldt County, California, at an elevation of 91 m. The orchard consists of ca. 2000 Douglas-fir ramets (a ramet is a cutting that is taken from a superior clone and grafted onto rootstock) that were grafted in 1968. In 1987, we surveyed each tree in 12 contiguous plots. There were a total of 636 trees in these 12 plots, with 1 to 29 ramets per clone and 64 clones altogether. We counted the number of twig beetle attacks per tree, documented the clone number of each tree, and counted the numbers of cones per tree up to 100. Fewer than 2% of trees sampled had more than 100 cones per tree. We also rated tree vigor according to the following visual index: 0 = tree apparently healthy; 1 = foliage slightly chlorotic, with internodes shorter than normal; 2 = noticeably chlorotic foliage, internodes stunted; 3 = extremely chlorotic foliage, stunted internodes, and noticeable needle-drop (Rappaport and Wood 1994)

The spatial distribution of beetle attacks per tree (Figure 1), unadjusted for the effects of covariates, was particularly interesting because some areas of the orchard appeared to

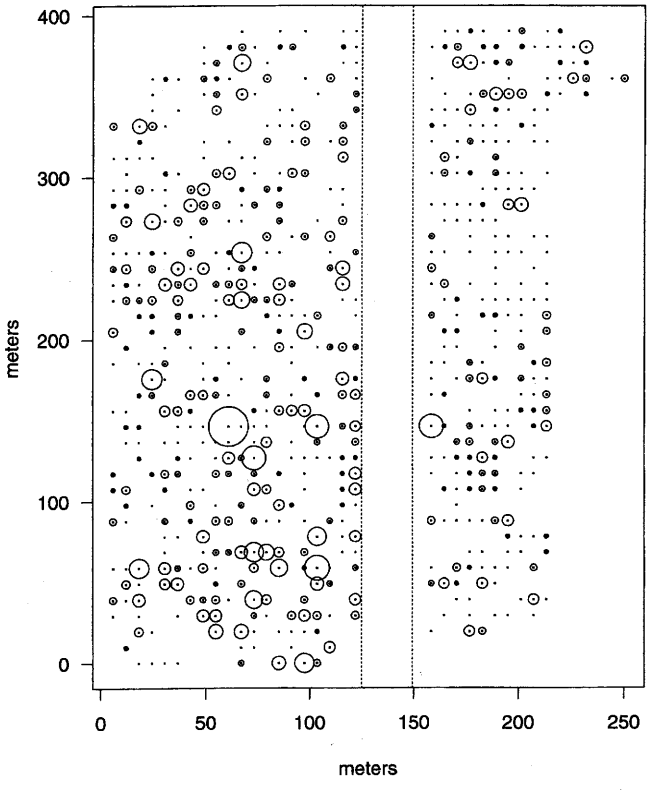

Figure 1. Twig beetle attacks in the seed orchard. Radii of circles are proportional to the square root of the number of beetle attacks on the tree located at the center of the circle. The blank strip in the middle represents the location of a waterline.

have higher concentrations of beetle attacks than others, and attack rates in neighboring trees were apparently correlated. Most seed orchards are designed such that ramets are balanced over blocks of clones from a limited geographic region, so that self-pollination by ramets from the same clone will be avoided and yet pollination contamination from off-site trees will be minimized. To accomplish this end, ramets were planted in rows so that clones 1 through 64 are planted in order, then the same order was repeated until the end of the orchard.

A plot of number of beetle attacks per tree versus number of cones per tree (Figure 2) suggested no relationship between the number of cones per tree and the number of attacking beetles for trees with fewer than 20 cones, but the relationship for trees with more than 20 cones appeared to be linear (albeit with a very small slope). The smooth curve through the points was generated using the nonparametric local regression routine, **loess** (Cleveland and Devlin 1988). We were interested in whether this relationship holds when other variables, such as spatial location, clone number, and vigor, are introduced in the model. Finally, a plot of the logarithm of beetle attacks per tree by vigor class (Figure 3) suggested that stressed trees might be more susceptible to beetle attack than unstressed trees.

2 Statistical Methods

2.1 A Generalized Additive Model

In this section we model the response variable, number of beetles per tree, by using generalized additive regression techniques (GAM) (Hastie 1992). The formal statement for a generalized additive regression model is

$$g(\mu) = \alpha + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p) \qquad (1)$$

where μ is the expected value of the response variable; g is a link function such as a logit or logarithmic transform; $\{\mathbf{x} = x1, x_2, \cdots x_p\}$ is a vector of covariates; and $f_1(\cdot), f_2(\cdot), \cdots, f_p(\cdot)$ are either parametric or nonparametric smooth functions. Examples of parametric functions include simple transformations such as polynomials and logarithms, as well as step functions introduced by categorical covariates. Nonparametric functions are smooth curves produced by nonparametric local regression routines such as **loess** (Cleveland and Devlin 1988, Cleveland et al. 1992).

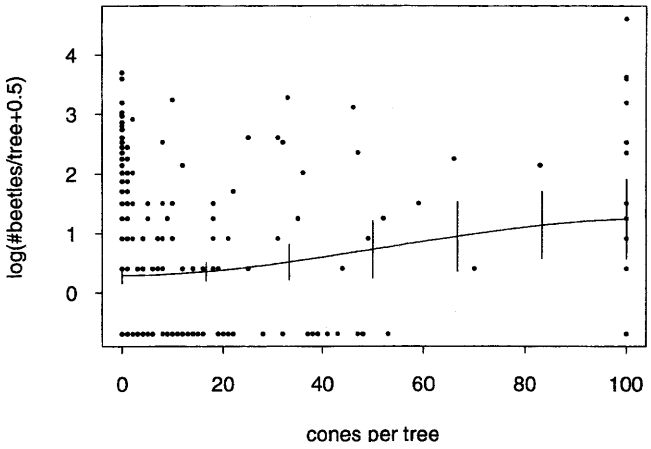

Figure 2. Relationship between number of beetles per tree and number of cones per tree. The curve through the points was developed using a nonparametric local regression smoother. The vertical bars indicate approximate 95% pointwise confidence intervals of the curve.

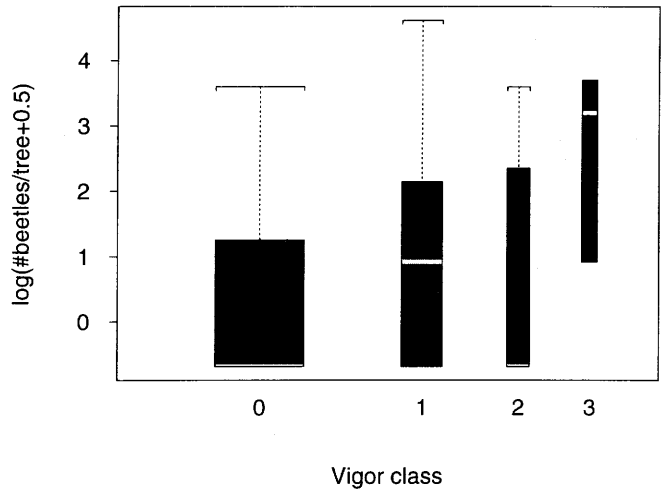

Figure 3. Boxplots of the logarithms of number of beetles per tree for each of the vigor categories indicate the range of the beetle counts for each vigor class. Boxes include the middle 50% of the data. The median is indicated as a white bar across the box. The width of the boxes are proportional to the square root of the number of trees in each category.

Classical linear regression is a special case of Equation (1), that is obtained when the link function g is the identity function, the functions $f_1, f_2, \cdots, f_p$ are linear transformations, and the response variable is approximately normally distributed. Generalized additive models extend linear regression models in two ways. First, the distribution of the response variable may come from an exponential family (e.g., normal, binomial, and Poisson distributions). Second, the linear relationship between responses and the covariates may be replaced by the more flexible link function described in (1).

For the counts of beetles, let n_i be the observed number of beetles in tree i, with $i = 1, \cdots, 637$. Let (x_i, y_i) be the spatial coordinates of tree i. Let $z_i = \{vigor_i, cone_i, clone_i\}$ be a vector of the observed covariates tree vigor, number of cones per tree, and clone type. We assume that n_i $(i = 1, \cdots, 637)$ is a realization of a random variable N such that

$$E(N) = \mu = e^{\lambda} \quad \text{and} \quad \text{var}(N) = k \cdot \mu \qquad (2)$$

where λ is a function of the covariates and the spatial location of trees and k is a scale parameter (i.e., N is assumed to be an overdispersed Poisson variable). We also assume that λ is an additive predictor given by

$$\lambda = \alpha + f_1(cone) + f_2(x, y) + vigor + clone \qquad (3)$$

where α is the intercept parameter; $f_1(cone)$ is a (nonparametric) smooth function describing the relationship between the number of cones per tree and the number of beetle attacks; (x, y) are the spatial coordinates of the tree;

$f_2(x, y)$ is a nonparametric smooth surface that describes the effect of location due to unobserved factors; *vigor* is a categorical variable with four levels; and *clone* is a categorical variable with 63 levels.

2.2 Estimation

We used the generalized additive routine, **gam,** within the S-PLUS statistical package (Statistical Sciences 1993), to estimate the parameters and the smooth functions in Equation (3). The generalized additive routine uses locally weighted polynomial regression and iteratively reweighted least square algorithms to arrive at values of the parameters that minimize the appropriate deviance function (Hastie 1992). The deviance is a linear function of the log likelihood. Specifically, the deviance is twice the log-likelihood ratio statistic for testing the adequacy of the fitted model against the saturated model. A saturated model is a model where the predicted value of each observation is the observed response. For the normal (i.e., Gaussian) distribution the deviance reduces to the residual sum of squares (McCullagh and Nelder 1989). For Poisson counts data, the deviance is given by

$$D = 2\sum_i \left[n_i \ln \frac{n_i}{\hat{\mu}_i} - (n_i - \hat{\mu}_i) \right] \quad (4)$$

The parameter values that minimize the deviance function in (4) are the same as those maximizing the likelihood function. Consequently, the routine **gam** produces approximate maximum likelihood estimates of the parameters.

We used analysis of deviance tables (McCullagh and Nelder 1989) to check the significance of each of the variables in Equation (3). Analysis of deviance is a generalization of analysis of variance to a wider class of models than those assuming normal errors. While analysis of variance uses the difference in residual sums of squares between models to test the significance of particular terms, analysis of deviance uses the difference in the deviances between two models. Let M be the model defined by the equations in (2) and (3) and M_1 be model M less the effect of one of the covariates. If in addition the scale parameter k is known, then we can use the test statistic

$$\frac{D_{M_1} - D_M}{kq} \quad (5)$$

based on the usual likelihood ratio test (Venables and Ripley 1995), where D_{M_1} and D_M are deviances of models M_1 and M respectively. The above test statistic is compared with the corresponding value from χ^2_q table to test the significance of the removed covariate, where q is the difference in the degrees of freedom between the two models. When k is unknown, as is the case here, we can use the test statistic in (5) with k replaced by an estimate such as

$$\hat{k} = \frac{D_M}{p},$$

where p is the degrees of freedom of model M. The test statistic will now be compared with the corresponding value from F_{qp} table. The resulting significant levels are approximate.

2.3 Goodness-of-Fit

The adequacy of the Poisson model was checked by producing normal probability plots of the residuals

$$residual = \sqrt{n + 3/8} - \sqrt{\hat{\mu} + 3/8} \quad (6)$$

where n is the observed number of beetles and $\hat{\mu}$ is the fitted number of beetles. If the Poisson model described in Section 2.1 is correct and μ is not small, then the residuals in (6) are expected to behave like a sample from the normal distribution (Rao 1973). Consequently, the normal probability plot of the residuals will appear to be roughly linear.

3 Results

All the variables in Equation (3) were found to be significant (Table 1). Figure 4a gives a plot of the observed and fitted values graphed against values of the additive predictor [Equation (3)]. Figure 4b is the normal probability plot of the residuals from the Poisson model using Equation (6). Except for a few possible outliers, the fit appears to be adequate. The outlier with the largest negative residual (residual = –3.8) is a tree that had no attacks although it had the maximum number of observed cones (100 cones).

A plot of the estimated nonparametric smooth function $\hat{f}(cone)$ against the number of cones per tree (Figure 5) indicated that the relationship observed in Figure 2 between the number of attacks and the number of cones per tree was persistent even within a model that includes the spatial locations of trees and the additive effects of the other observed variables.

A significant spatial trend in beetle attacks was apparent (P-value < 10^{-12}). This trend can be seen in the plot of the estimated smooth surface $\hat{f}_2(x, y)$ (Figure 6), which describes the relationship between number of attacks and spatial location of trees in a model that also includes the additive effects of the other observed covariates. Beetle attacks tended to be highest in the lower left (southwestern) portion of the stand and lowest in the upper right portion. Prevailing winds were from the southwestern quadrant of the seed orchard, so this phenomenon may be the result of a source of beetle infestation to the windward of that quadrant, or from anemotaxis of ovipositing adults from within the orchard.

Except for four clones, the effects of the different clones on the susceptibility of a tree to twig beetle attacks was apparently random, as shown in a normal probability plot of the estimated clone effects (Figure 7). All ramets of the four outlying clones had no beetle attacks. Three of the clones included too few cases (a total of five trees) to draw any meaningful conclusions. The fourth clone in this group, however, had 11 ramets distributed throughout the stand. Further studies of the resistance of this latter clone to beetle attacks might be justified.

Table 1. Analysis of deviance table for testing the significance of the spatial and other explanatory variables in the generalized additive model.

Model	Model		Test			
	Deviance	d.f.	Deviance[a]	d.f.[b]	F-value[c]	P-value
M includes						
All covariates induding spatial	1,972.7	558.4				
M_1 is model M less						
Spatial effect	2,355.0	566.0	382.3	7.63	14.8	$< 10^{-12}$
Cone effect	2,187.6	562.4	214.9	3.98	15.29	7.96×10^{-12}
Vigor effect	2,108.8	561.4	136.1	3.00	12.85	3.98×10^{-8}
Clone effect	2,986.6	621.4	1,013.9	62.98	4.56	$< 10^{-12}$

[a] Test Deviance = $D_{M_1} - D_M$

[b] Test d.f. = d.f. of M_1-d.f. of M

[c] F-value = $\dfrac{test\,deviance / d.f.}{D_M / d.f.}$

4 Discussion

Our understanding of ecological systems can be improved if a spatial component is included among predictive variables, and if nonlinear methods of modeling and estimation are used when analyzing spatially compiled data. Until recently, however, researchers have tended to ignore the spatial structures in their studies because the few statistical methods that are appropriate for analysis of spatially correlated data are not included in major statistical packages and are often hard to use (Roberts et al. 1993). Here we presented a method for analyzing spatially correlated data by using an existing statistical package. The method can be implemented and interpreted directly, primarily because it is simply an extension of linear regression techniques to nonlinear and spatial relationships. All analyses and graphical outputs were done using the S-PLUS package; we did not have to write new programs. We used a Poisson model to analyze the particular data in this paper; however, the methods for incorporating a spatial surface into a regression model can also be used for other data types, e.g., binary response data or continuous (e.g., Gaussian) response data.

In our analysis, we identified a single clone that appears to be resistant to beetle attack, and three other potential candidates. Neither location, vigor, or cone crop seems to account for the absence of beetle attacks on trees of the "resistant" clone. Differences in the number of attacks among the rest of the clones, however, appear to be simply due to random fluctuations.

Researchers have been concerned about the potential impact of spatial correlations, and have consequently attempted to compensate for this potential problem by randomization or blocking of treatments. However, even these precautions cannot guarantee that some observed effects do not result from spatially correlated covariates.

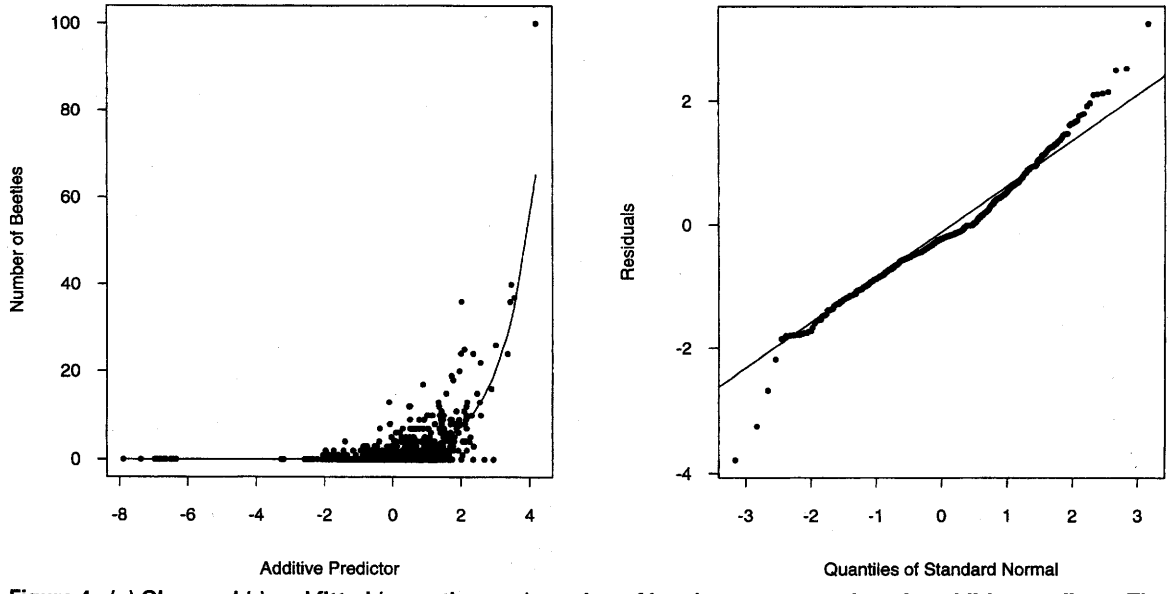

Figure 4. (a) Observed (•) and fitted (smooth curve) number of beetles per tree against the additive predictor. The fitted curve was developed using the Poisson model described in the text. (b) Normal probability plot of the residuals from the Poisson fit.

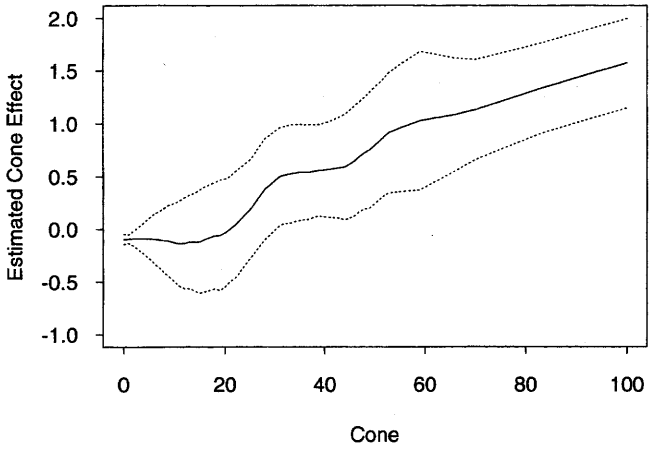

Figure 5. Estimated smooth curve describing the relationship between the number of cones per tree and the number of beetle attacks. The dotted lines indicate an approximate 95% pointwise confidence band of the estimated curve. Note that a straight line could be drawn within the bands.

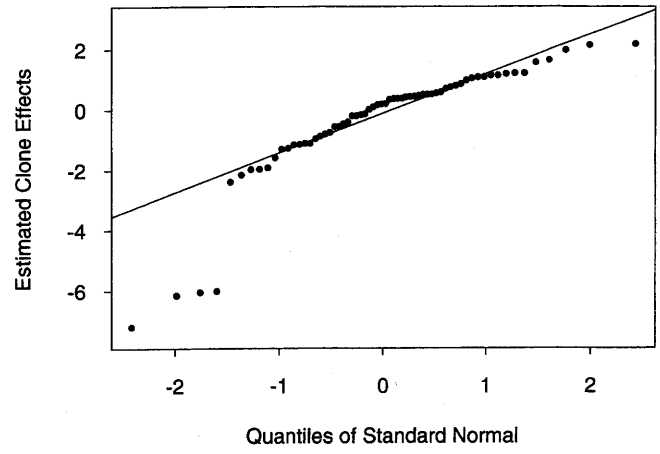

Figure 7. Normal probability plot of the estimated clone effects. Except for four clones, the effects appear to be consistent with a random normal sample.

With this method, we accomplished two results. First, we where able to estimate spatial (direction, location) effects, which were of interest in themselves, instead of just being nuisance parameters.

Second, we estimated the effects of other measured covariates within a model that also included possible effects of unmeasured covariates (such as prevailing winds or soil types) that may vary spatially. Without the availability of such a method, scientists would have to resort to informal (and often laborious) ways of verifying that the effects seen are not an artifact of some spatial parameter (in this instance orchard design).

The results of this study have important implications for analyses involving organisms with spatially aggregated distributions. Given the presence of covariates (e.g., clone) that may have an additional spatial component, it is especially important to consider the effect of location to make sound inferences about those covariates. Similar arguments can be made about temporal effects, which, like spatial location, might be expected to play an important role in insect distribution.

For instance, in seed orchards, beetle abundance and distribution the previous year might affect current year distributions, and cone population dynamics might also influence cone and seed insect population dynamics in subsequent years. Silvicultural practices such as insecticide treatments, fertilization, and mowing might also be expected to affect insect distributions. The methods presented here will facilitate the analysis of such covariates.

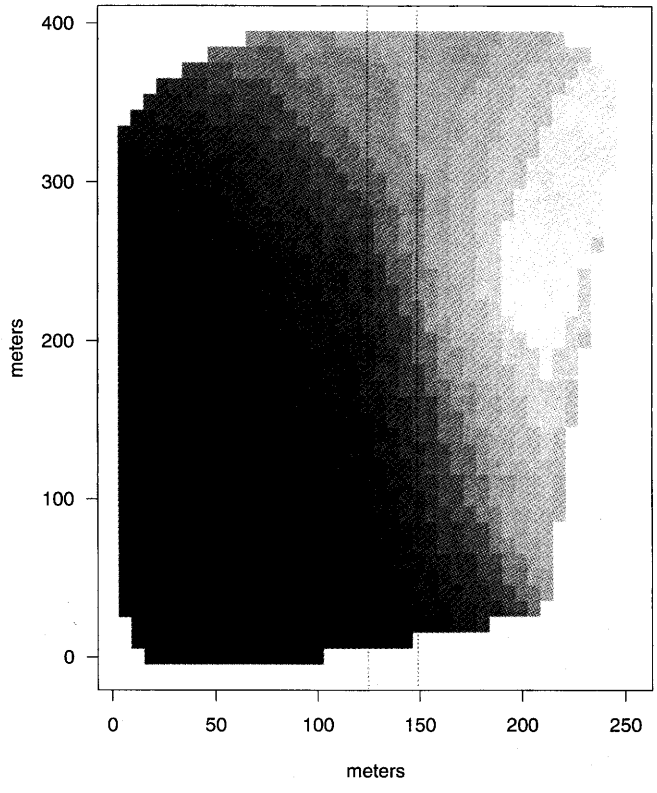

Figure 6. Estimated smooth surface describing the relationship between spatial location and number of beetle attacks. The darker shades indicate larger estimated attacks.

Literature Cited

BESAG, J.E. 1972. Nearest-neighbour systems and the auto-logistic model for binary data. J. Roy. Stat. Soc. B. 34:75–83.

BRILLINGER, D.R. 1994. Examples of scientific problems and data analyses in demography, neurophysiology, and seismology. J. Comp. Graph. Stat. 3:1–22.

CLEVELAND, W.S., AND S.J. DEVLIN. 1988. Locally weighted regression: An approach to regression analysis by local fitting. J. Am. Stat. Assoc. 83:596–610.

CLEVELAND, W.S., E. GROSS, AND W.M. SHYU. 1992. Local regression models, *in* Statistical analysis in *S*, Chambers, J.M., and T.J. Hastie (eds.). Wadsworth, Pacific Grove, CA.

COOK, D.G., AND S.J. POCOCK. 1983. Multiple regression in geographical mortality studies, with allowance for spatially correlated errors. Biometrics 39: 361–371.

CRESS, D.W., AND J.D. DANIELS. 1990. Forest seed orchards of Western Oregon, Western Washington, and Northern California. Daniels and Associates, Centralia, WA. 188 p.

CRESSIE, N. 1993. Statistics for spatial data. Wiley, New York. 900 p.

DALLARA, P.L., S.J. SEYBOLD, W. FRANCKE, AND D.L. WOOD. 1995. The chemical ecology of *Pityophthorus* Eichhoff (Coleoptera: Scolytidae) in central coastal California. Proc. of 43rd annual meet. of the California Forest Pest Council. Append. p. 68–69.

HASTIE, T.J. 1992. Generalized additive models. P. 249–307 *in* Statistical models in *S*, Chambers, J.M., and T.J. Hastie (eds.). Wadsworth, Pacific Grove, CA.

HEDLIN, A.F., AND D.S. RUTH. 1970. A Douglas-fir twig mining beetle, *Pityophthorus orarius* (Coleoptera: Scolytidae). Can. Entomol. 102:105–108.

HOOVER, K., D.L. WOOD, J.W. FOX, AND W.E. BROS. 1995. Quantitative and seasonal association of the pitch canker fungus, *Fusarium subglutinans* f. sp. *pini* with *Conophthorus radiatae* (Coleoptera: Scolytidae) and *Ernobius punctulatus* (Coleoptera: Anobiidae) which infest *Pinus radiata*. Can. Entomol. 1: 79–91.

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

MCCULLAGH, P., AND J.A. NELDER. 1989. Generalized linear models. Ed. 2. Chapman and Hall, London. 511 p.

PENTTINEN, A., D. STOYAN, AND H. M. HENTTONEN. 1992. Marked point processes in forest statistics. For. Sci. 38(4):806–824.

PREISLER, H.K. 1993. Modelling spatial patterns of trees attacked by bark-beetles. Appl. Statist. 42(3):501–514.

PREISLER, H.K., AND R.G. MITCHELL. 1993. Colonization patterns of the mountain pine beetle in thinned and unthinned lodgepole pine stands. For. Sci. 39(3):528–545.

RAO, C.R. 1973. Linear statistical inferences and its applications. Ed. 2. Wiley, New York. P. 426–427

RAPPAPORT, N.G., AND D.L. WOOD. 1994. *Pityophthorus orarius* Bright (Coleoptera: Scolytidae) in a northern California Douglas-fir seed orchard: Effect of clone, tree vigor, and cone crop on rate of attack. Can. Entomol. 126:1111–1118.

ROBERTS, E.A., F.W. RAVIN, AND S.J. FLEISCHER. 1993. Spatial data representation for integrated pest management programs. Am. Entomol. 39(2):92–107.

STATISTICAL SCIENCES. 1993. S-PLUS User's Manual, Version 3.2. Stat Sci, a Division of Math Soft, Inc., Seattle.

STORER, A.J., T.R. GORDON, P.L. DALLARA, AND D.L. WOOD. 1994. Pitch canker kills pines, spreads to new species and regions. Calif. Agric. 48(6):9–13.

VENABLES, W.N., AND B.D. RIPLEY. 1995. Modern applied statistics with S-Plus. Springer-Verlag, New York. 462 p.