Conference on Applied Statistics in Agriculture

1994 - 6th Annual Conference Proceedings

AUTOLOGISTIC MODEL OF SPATIAL PATTERN OF PHYTOPHTHORA EPIDEMIC IN BELL PEPPER: EFFECTS OF SOIL VARIABLES ON DISEASE PRESENCE

M. L. Gumpertz

J. Graham

J. B. Ristaino

AUTOLOGISTIC MODEL OF SPATIAL PATTERN OF PHYTOPHTHORA EPIDEMIC IN BELL PEPPER: EFFECTS OF SOIL VARIABLES ON DISEASE PRESENCE

M. L. Gumpertz, J. Graham, and J. B. Ristaino
North Carolina State University

ABSTRACT

The pathogen *Phytophthora capsici* causes lesions on the crown, stem, and leaves of bell pepper, and rapidly causes the plant to die. The spatial patterns of disease in an agricultural field contain information about pathogen dispersal mechanisms and can be useful for developing methods of control of disease. Soil water content, soil pathogen population density, and disease incidence data were collected on a 20 × 20 grid in two naturally infested commercial bell pepper fields. In one field the initial pattern of disease closely matched the soil water content pattern and disease developed in areas where the pathogen population levels were high. In the other field no such correspondence was obvious from maps of disease and soil water content.

The autologistic model is a flexible model for predicting presence or absence of disease based on soil water content and soil pathogen population, while taking spatial correlation into account. In the autologistic model the log odds of disease in a particular quadrat are modeled as a linear combination of disease in neighboring quadrats and the soil variables. Neighboring quadrats can be defined as adjacent quadrats within a row, quadrats in adjacent rows, quadrats two rows away, and so forth. The regression coefficients give estimates of the increase in odds of disease if neighbors within a row or in adjacent rows show disease symptoms; thus we obtain information about the degree of spread in different directions. The coefficients for the soil variables give estimates of the increase in odds of disease as soil water content or pathogen population density increase. In this problem, soil water content is also highly correlated over quadrats. This introduces a kind of collinearity between water content and the disease in neighboring quadrats, making estimation and interpretation of the parameters of the autologistic model more difficult. We discuss fitting and evaluating the autologistic model when the covariates are themselves spatially correlated.

Additional keywords: spatial correlation, disease incidence, Markov random field, multidimensional binary data, pseudolikelihood estimation.

1. INTRODUCTION

Statistical models of the spatial patterns of disease in an agricultural field can be useful for understanding dispersal mechanisms and for developing methods of control of disease. This paper describes and demonstrates the use of the autologistic model for spatial pattern of *Phytophthora* epidemics in bell pepper. There are two features of the autologistic model that make it well suited to study of spatial pattern of disease: (1) it applies specifically to binary response variables such as disease presence or absence; and (2) explanatory variables can be incorporated into the model.

Statistical methods for continuous response variables have been more widely used in plant disease research than models developed specifically for binary data. For example, spatio-temporal autocorrelation analysis, described in Reynolds and Madden (1988), and kriging (see Lecoustre et al. 1989) have been proposed for describing spatial correlations in disease epidemics. These methods involve modeling the spatial or spatial-temporal covariance structure of the disease data with autoregressive-moving average type models and variogram models, respectively. They are used to study the effects of distance (and time) on the spread of the epidemic and to map the disease. Although designed for data measured on a continuous scale, they have also been applied to categorical data.

The model upon which this paper focuses, the autologistic, was originally developed by physicists to model ferromagnetism, which involves binary data on a lattice (Cressie 1991). Besag (1972, 1974) developed much of the statistical theory of autologistic models and gave some examples involving plant disease. Some autologistic models incorporating time have also been proposed (Besag 1977, Chadoeuf et al. 1992). For example, Besag (1977) demonstrated the use of the autologistic model to describe incidence of footrot in endive as a function of disease in neighboring plots at both the current and the previous times. The autologistic model has also been extended to ordered categorical data, such as disease rankings on a scale of 1 to 4 (Strauss 1992). Smyth et al. (1992) applied a similar model to anthracnose progress in tropical pasture legumes.

The studies cited above modeled spatial pattern of disease as a function solely of proximity of diseased plants or of spatial and temporal relationships, but did not incorporate other environmental information. Measures of pathogen population density in the soil and environmental covariates, such as soil moisture, microclimate variables, elevation gradient of the field, soil nutrient concentrations, and soil compaction, should provide additional information for predicting the presence or absence of disease and elucidating the conditions under which the epidemic spreads. Concomitant information can theoretically be incorporated into any of the spatio-temporal autocorrelation, kriging and logistic regression models.

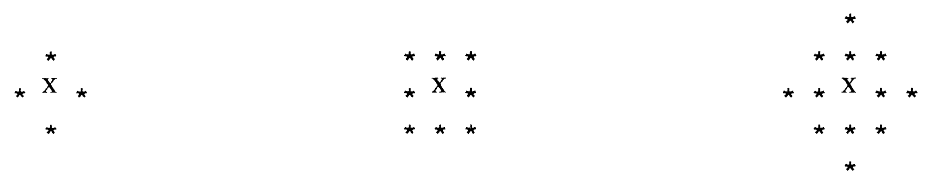
In the autologistic model presented here, the log odds of disease presence in a particular quadrat, also called the logit,

$$\text{logit (p)} = \ln\left(\frac{\text{Pr(disease present)}}{\text{Pr(disease absent)}}\right),$$

is modeled as a linear combination of soil water content and pathogen population density in the quadrat, and disease presence in neighboring quadrats. In the next section we describe the autologistic model with covariates in more detail and briefly discuss available methods of estimation of the parameters. Section 3 demonstrates the fitting and interpretation of this model for a *Phytophthora* epidemic in two naturally infested fields of bell pepper. Section 4 offers some practical methods for determining how well this model captures the spatial relationships and highlights some important issues in practical application of this model to plant disease epidemics. Finally, we list some problems in actual application of these methods that call for future research.

2. THE AUTOLOGISTIC MODEL

In the autologistic model the conditional probability of occurrence of disease depends on disease in neighboring quadrats. The definition of neighboring quadrats is very flexible and can be tailored to the particular situation under study. For rectangular lattices there are some standard systems of neighbors. A first order system includes only the four adjacent quadrats in the set of neighbors, two within the row and two in adjacent rows. A second order model includes the four diagonal neighbors in addition to the quadrats of the first order model. A third order model includes quadrats two rows or columns away (Besag 1974).



first order second order third order

However, it is not necessary to use the standard neighbor systems and it is not necessary that the lattice have any regular shape. What is required is that a set of neighbors be defined for each quadrat in the lattice, and if quadrat *i* is a neighbor of quadrat *j*, the converse is also true.

The autologistic model is a simple generalization to spatial data of the standard logistic model for independent binary data. In the standard logistic model for binary data the log odds of disease are modeled as a linear function of some regressor variables, X_1, ..., X_r.

$$\mathrm{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_r X_r = \sum_{k=o}^{r} \beta_k X_k, \qquad (1)$$

where *p* is the probability of a "success"; *e.g.*, the probability of disease being present.

In the context of designed experiments with plants grown in pots, if inoculum were introduced into each pot individually and all pots were physically separated from each other, then the observations would be independent of each other, the standard logistic model (1) would be appropriate, and the parameters of model (1) could be estimated by maximum likelihood using standard software such as SAS® PROC LOGISTIC.

In the context of a disease epidemic in an agricultural field, quadrats are spatially correlated. One general way of incorporating spatial correlation into models is to predict the response at a particular site conditioned on the responses at neighboring sites. For binary data, if the response at site *i* depends on *r* covariates and on the responses in neighboring sites in a pairwise fashion, then the conditional probability of a particular response, $y_i = 1$ (disease present) or $y_i = 0$ (disease absent), given the observed amount of disease in the neighbors (the set of neighbors of the i^{th} site is denoted N_i) is:

$$\Pr(Y_i = y_i \mid y_j, \text{ responses at neighboring sites } j \neq i)$$

$$= \frac{\exp\left\{\sum_{k=1}^{r} \beta_k x_{ik} y_i + \sum_{j \in N_i} \gamma_j y_i y_j\right\}}{1 + \exp\left\{\sum_{k=0}^{r} \beta_k x_{ik} + \sum_{j \in N_i} \gamma_j y_j\right\}}$$

Since *y* takes the value 1 if disease is present, the log of the odds of disease being present is:

$$\text{logit}(p_i \mid y_j, \text{ for } j \neq i) = \sum_{k=0}^{r} \beta_k x_{ik} + \sum_{j \in N_i} \gamma_j y_j \ ,$$

which looks just like the standard logistic model (1) with the addition of terms for disease in neighboring quadrats (Y_j). This model has flexibility in that neighbors may be defined in any way that makes sense, the observations do not necessarily need to be taken on a rectangular lattice, and covariates (X_{ik}) can be incorporated. The covariate terms in the autologistic model are not necessarily required to enter in a linear fashion. The general form of the autologistic model is

$$\text{logit}(p_i \mid y_j, \text{ for all other sites } j \neq i) = \alpha_i y_i + \sum_{j \in N_i} \gamma_{ij} y_i y_j \qquad (2)$$

Besag (1974). The parameters β_i and γ_{ij} can be different for every node, *i*. Consequently, the parameter α_i may be a linear combination of covariates, as in model (2), or α_i may be some nonlinear function of covariates, $f(x_i)$. Furthermore, the dependence parameters γ_{ij} can differ for different neighbors $j \in N_i$, which allows different amounts of dependence in different directions. The dependence structure is not completely general, however, being limited to pairwise dependence among sites. A completely general covariance structure would include three-term products of neighbors, four-term products, and so on. Hence the covariance structure of the autologistic model may be too restrictive to fit a particular set of data well.

In logistic regression models the regression coefficients contain information about the effects of changes in the covariate values on the odds of disease. In model (1) the log odds of disease in a quadrat appears as a function of *k* regressor variables, X_1, ..., X_k. If X_1 increases by one unit while the other variables remain constant, the log odds of disease increases by β_1 units. The odds ratio for an increase of one unit is defined as

$$\text{odds ratio} = \frac{\text{odds of disease if } X_1 = x + 1}{\text{odds of disease if } X_1 = x} \; .$$

Thus the odds ratio is $e^{\beta 1}$. If the odds ratio = 1.5 for increasing X_1 by one unit, the odds of disease increase 50% for every unit increase in X_1.

The method of choice for estimating parameters of the ordinary logistic regression model is maximum likelihood. In the autologistic model for spatially correlated responses the observations are not independent and it is not possible to write the likelihood function in closed form. Besag (1975) coined the term "pseudolikelihood" (PL) for the function that would be the likelihood were the data independent,

$$PL(\beta, \gamma \mid Y) = \prod_{i=1}^{n} (p_i \mid y_j, \text{ for } j \neq i).$$

Maximization of the pseudolikelihood function as though it were a true likelihood function is simple to implement because it just involves fitting the autologistic model using standard software for ordinary logistic regression, such as SAS® PROC LOGISTIC (Strauss 1992).

Before implementing the pseudo likelihood method, it is necessary to first create a variable for each type of neighbor. For instance, if your model includes a term for adjacent quadrats within the rows (W), and another term for adjacent quadrats one row apart (A), then two new variables containing these disease values must be created. These new variables are most easily constructed using a matrix programming language such as SAS® IML. First create an array of disease values representing the actual field. Then shift the array in the desired directions to create a new array of neighbor values for each distance and direction, and finally output these neighbor values to your data set.

The pseudolikelihood estimates have good statistical properties when the spatial dependence is not too large (Strauss 1992, Geyer 1991). Be aware, however, that the standard errors of regression coefficients from standard logistic regression software are not correct for the autologistic model. Hence, Wald-type tests of hypotheses cannot be constructed using the output from PROC LOGISTIC and, furthermore, likelihood ratio-type tests should not be constructed using pseudolikelihood values. In ordinary logistic regression, large sample likelihood ratio tests are constructed by maximizing the likelihood function under two competing models, one nested within the other, and comparing twice the difference between the logarithms of the likelihoods to a quantile of the Chi-square distribution. If pseudolikelihood values are substituted for true likelihood values in this procedure, the resulting statistic has a distribution far from Chi-square (Graham 1994), so should be avoided.

A new method of estimation, Monte Carlo maximum likelihood, has recently been developed which gives good estimates of the parameters even when the correlations among neighboring plots are extremely high (Geyer 1991), and which provides standard errors and tests of hypotheses (Graham 1994). At the current time, there are still practical difficulties in implementing this method and no commercial software for Monte Carlo maximum likelihood estimation exists. In this paper all parameters have been estimated using the maximum pseudolikelihood method.

3. APPLICATION TO PHYTOPHTHORA EPIDEMIC

The pathogen *Phytophthora capsici* Leonian causes lesions on the crown, stem, and leaves of bell pepper, and rapidly causes the plant to die. Ristaino et al. (1993, 1994) and Larkin *et al.* (1994) described the spatial pattern of *Phytophthora* epidemics in six naturally infested commercial bell pepper fields. Ristaino *et al.* found that the spatial correlations of disease incidence down rows and across rows contain information about whether the disease is spread by water, by root-to-root contact, or aerially. They demonstrated that disease generally tends to form longer clusters along rows than across rows, and from this concluded that movement of surface water within furrows is important in spreading inoculum. The spatial and temporal order in which wilt occurs and stem and crown lesions develop provide further clues to the methods of dispersal (Ristaino *et al.* 1994). It appears that root infections spreading to the crown are the most frequent paths of infection. Stem, leaf, and fruit infections were much rarer or nonexistent, indicating that splash dispersal was not as important for spreading inoculum. Larkin *et al.* (1994) presented spatial correlograms and crosscorrelograms for disease severity and soil variables in the Phytophthora study. The correlograms show that the distance over which quadrats are correlated increases steadily over time and reveal some correspondence between disease and the soil variables.

Soil water content measurements and leaf disk assays of soil pathogen population were collected for two of the naturally infested commercial bell pepper fields that were studied by Ristaino *et al.* (described in detail in Ristaino *et al.* 1993). Each field was a square lattice of 20 rows by 20 quadrats with 2 to 3 bell pepper plants per quadrat. The response variable of interest was presence or absence of disease in a quadrat. If any plant was wilted or dead or had lesions on stem, crown, or leaves, disease was considered to be present in the quadrat. Disease presence or absence was recorded for each quadrat on 9 dates throughout the growing season, from 6/16/92 to 8/5/92. Soil water content (%) was measured in each quadrat of field 1 on 7/2/92 and field 2 on 6/22/92, number of leaf disks colonized (out of five) was counted in each quadrat on two dates: 6/29/92 and 7/29/92 for field 1 and 6/19/92 and 8/5/92 for field 2.

For one of the fields (field 2) the initial pattern of disease closely matched the soil water content pattern and disease developed in areas where population density of Phytophthora in the soil was high (Fig. 2). For the other field (field 1) no such patterns are obvious from the maps of soil water content and number of leaf disks colonized (Fig. 1). The patterns of soil water content were quite different in these two fields. Field 2 had a distinct wet corner and disease was present in most of the quadrats in this corner from the first sampling date. Field 1 was wetter overall (mean water content = 10.8%, compared to field 2 mean water content = 8.8%) but more homogeneous with wet quadrats dispersed throughout the field (field 1 std. dev. = 1.82, field 2 std. dev. = 2.39).

As a preliminary step we fit the logistic model

$$\text{MODEL1:}\quad \text{logit}(p_{ij}) = \beta_0 + \beta_1 M_{ij} + \beta_2 L1_{ij} + \beta_3 L2_{ij}\ , \qquad (3)$$

to the data from each quadrat for the last sampling date, where M = soil water content, L1 = number of leaf disks colonized in June, L2=number of leaf disks colonized in late July or early August, and the subscripts i and j indicate row and quadrat respectively. All models were fit to the inner 16 × 16 lattice of 256 quadrats, so that models involving adjacent quadrats and quadrats two spaces away could be accommodated. In each field one quadrat had a measured water content value greater than 25%, which was far from the distribution of remaining water content values, so was omitted from all regression and correlation computations, and 4 or 5 water content values were missing altogether. The preliminary model (MODEL1) ignores the spatial nature of the data and is used to check whether spatial correlations exist in disease incidence among neighboring quadrats after accounting for the effects of soil water content and pathogen population in the soil.

Before doing any regression, disease incidence showed fairly high correlations (r = .47 and r = .55 in fields 1 and 2, respectively) between adjacent quadrats within a row on the last sampling date. The lag one autocovariance between adjacent quadrats within a row is computed as the covariance between the "tail" and "head" variables, where the tail variable is the response in quadrats 3 through 17 and the head variable is the response in quadrats 4 through 18.

$$C(1) = \frac{1}{16\cdot 15}\sum_{i=3}^{18}\sum_{j=3}^{17} y_{ij}y_{i,j+1} - \left(\frac{1}{16\cdot 15}\sum_{i=3}^{18}\sum_{j=3}^{17} y_{ij}\right)\left(\frac{1}{16\cdot 15}\sum_{i=3}^{18}\sum_{j=4}^{18} y_{ij}\right),$$

The spatial autocorrelation is computed by dividing C(1) by the product of the "head" and "tail" standard deviations. After fitting the logistic model for the effects of soil water content and the leaf disk assays, ignoring spatial correlations (MODEL1), the Pearson residuals showed correlations between neighboring quadrats within a row of .44 and .35 for fields 1 and 2, respectively (Figure 3). The reduction in correlation from .55 to .35 in field 2 indicates that a large part of the spatial correlation in disease incidence may be attributable to the environmental variables soil water content and *Phytophthora* population level. In field 1 the correlation between adjacent quadrats does not appear to be related to these variables. Pearson residuals,

$$\chi_{ij} = \frac{y_{ij} - \hat{p}_{ij}}{\sqrt{\hat{p}_{ij}(1 - \hat{p}_{ij})}},$$

are standardized differences between the observed response (Y = 1 for disease present, Y = 0 otherwise) and the predicted probability of disease.

In some settings spatial correlations can be completely eliminated by regression on covariates. In the present application, however, disease is actually spread from one plant to another, so it is likely that, even after taking the soil variables into account, the disease status of the neighboring quadrats would be an important predictor of disease presence. After verifying that the Pearson residuals were not free of spatial correlation, terms for disease in adjacent quadrats were added to MODEL1 to form a second order autologistic model.

$$\text{MODEL2: } \operatorname{logit}(p_{ij}) = \beta_0+\beta_1 M_{ij}+\beta_2 L1_{ij}+\beta_3 L2_{ij}+\beta_4 W_{ij}+\beta_5 A_{ij}+\beta_6 D_{ij1}+\beta_7 D_{ij2}, \quad (4)$$

In this model $W_{ij} = y_{i,j} + y_{i,j+1}$, the number of diseased quadrats of the two adjacent quadrats within the same row, A_{ij} = number of diseased quadrats of the two adjacent quadrats in neighboring rows, D_{ij1} = number of diseased quadrats of the two diagonal quadrats in the (1,1) and (-1,-1) direction, and D_{ij2} = number of diseased quadrats of the two diagonal quadrats in the (-1,1) and (1,-1) direction. The types of neighbors of site S_{ij} are diagrammed below; notice that the rows are numbered from right to left to match the actual field layout.

			Row	
		i + 1	*i*	*i* - 1
	j + 1	D_{ij1}	W_{ij}	D_{ij2}
Quadrat	*j*	A_{ij}	S_{ij}	A_{ij}
	j - 1	D_{ij2}	W_{ij1}	D_{ij1}

Including four separate terms for neighbors allows us to examine whether correlations across rows are as strong as those within rows, and whether there are any diagonal gradients in the field. If we thought that there was a gradient in a different direction, such as the (1,2) direction, we could add terms to the model to capture the expected pattern.

Finally, we fit a pure autologistic model without covariates to check whether predictions based on solely disease in the neighboring quadrats would be adequate.

$$\text{MODEL3: logit}(p_{ij}) = \beta_0+\beta_1 W_{ij}+\beta_2 A_{ij}+\beta_3 D_{ij1}+\beta_4 D_{ij2}. \qquad (5)$$

The estimated odds ratios ($e^{\hat{\beta}}$) and the number of quadrats misclassified for each fitted model for field 1 on the last sampling date are shown in Table 1. For any given quadrat, if $\hat{p}_{ij} > .5$,, disease was predicted to be present. The misclassification rate is the proportion of quadrats for which the predictions do not match the disease status actually observed.

In field 1, within-row effects were pronounced and some diagonal trend across the field was seen. The estimated odds of disease were over four times higher if one neighboring quadrat within the row was diseased than if the two neighbors were disease-free (MODEL 3, Table 1), with all other variables held constant. Disease in neighboring rows appeared to have little effect on disease presence (odds ratio $\exp\{\hat{\beta}\}$ approximately equal to one). The soil water content and leaf disk data did not appear to be helpful in predicting disease presence or absence, as the odds ratios for water content and leaf disk assays were all close to one. Fitting the autologistic model was successful in reducing the residual spatial autocorrelation (Fig. 3), but not completely successful in reproducing the spatial pattern of disease (Fig. 4). The percent of quadrats misclassified was 21 or 22%

regardless of whether the soil variables were included in the autologistic model or not (compare MODELs 2 and 3 in Table 1).

In field 2 there was a clear visual correspondence between the maps of soil water content and disease incidence (Fig. 2), with the southeast corner having both high soil water content and high disease incidence. The second order autologistic model with covariates (MODEL 2) was fitted to the field 2 data for 3 sampling dates: 6/25/92, 7/13/92, and 8/4/92 (Table 2). The relationship with water content showed up most strongly on 6/25/92, which was just 3 days after the soil water content measurements were taken, and decreased as time went on. The relationship between disease and the number of leaf disks colonized in the 6/19/92 assay was fairly strong early in the season and weak late in the season. For example, at the 6/25 sampling date, the estimated odds of disease if one leaf disk was colonized was 48% higher than if no leaf disks were colonized. At the end of the season, the leaf disk assay of 8/5/92 was a better predictor of disease than the early leaf disk assay, and the odds of disease were estimated to be about 60% higher if the number of leaf disks colonized was increased by one.

The most striking relationships in the autologistic models for field 2 are those between diagonally adjacent quadrats. Soil water content also shows a diagonal trend in this field (Figs 2, 5). The estimated relationships between the soil variables and disease are weaker than might be expected, and in particular, much weaker than the relationship among disease states in neighboring quadrats. When spatial correlations are ignored (MODEL1), the effects of water content and the leaf disk assays appear stronger (Table 3). However, MODEL1 is clearly not adequate because strong spatial correlations remain after regressing disease incidence on soil water content and the two leaf disk assays (Fig. 3). When terms for disease in neighboring quadrats are omitted from the model (MODEL1), the effects of soil water content and the leaf disk assays are biased upward. This type of bias in regression parameters may be expected when the response depends on disease in neighboring quadrats and the covariates are also spatially correlated. There is high correlation between soil water content levels in neighboring diagonal quadrats (Fig. 5) and there is a discernible relationship between water content in one quadrat and disease in the diagonal neighbors (Fig. 5).

MODEL2 does a better job of predicting disease for field 2 than for field 1, with a misclassification rate of 16%. If we explore further, and consider models that include soil water content and leaf disk assays of neighboring quadrats as well as in the current quadrat, and if we expand our definition of neighbors to include quadrats two spaces away, we can find some models that have misclassification rates of 10 to 12 percent. There are several models that do well, so it is difficult to choose among them, and it is difficult to interpret the coefficients in these more complicated models. The simplest model that gave a substantially improved misclassification rate included nine terms:

$$\text{logit }(\hat{p}_{ij}) = -4.48 + 0.66\text{M}_{\text{ij}} + 0.61\text{L2}_{\text{ij}} + 1.26\text{A}_{\text{ij}} + 2.16\text{D}_{\text{ij2}} + 0.74\text{W}_{\text{ij2}}$$

$$- 0.48\text{M}_{\text{ijA1}} - 0.49\text{M}_{\text{ijW2}} + 0.66\text{M}_{\text{ijA2}} + 0.74\text{L2}_{\text{ijD1}}, \qquad (6)$$

where W_{ij2} indicates disease two quadrats away within the same row, M_{ijA1} = water content in adjacent rows, M_{ijW2} = water content two quadrats away within the same row, M_{ijA2} = water content two rows away and L2_{ijD1} = leaf disk assay 2 in the diagonal (1,1) direction. Fitting this model reduced the misclassification rate to 12%. When the values of water content and leaf disk assays for neighboring quadrats are included in the model, the values of the coefficients of the soil covariates increase; however, the relationship between a covariate and disease cannot be inferred from one coefficient alone in these more complicated models. The instability in the coefficients for the soil variables is probably an indication that the autologistic model is not completely adequate.

4. EVALUATION OF FIT

The second order autologistic model was successful in decreasing the spatial autocorrelation in both fields studied. The maps of misclassified quadrats show that this model had difficulty predicting disease status at the boundary between diseased and disease-free areas. If the autologistic model does not fit a set of data well, it is possible that an autologistic model is appropriate, but that a higher order neighbor system is needed. Another possibility is that the assumption of pairwise dependence, which is basic to the autologistic model, is not adequate; *i.e.*, that disease in a quadrat depends on some products or more complicated functions of neighboring disease values. These deficiencies of the model can be checked by examination of the prediction errors. For example, in field 2 there were 42 misclassified quadrats after fitting the autologistic model with covariates (MODEL 2). Categorizing these according to the disease status of the neighbors two quadrats away reveals that 79% of them had one diseased neighbor two rows away (Table 4). However including the third order neighbors in the autologistic model did not improve the predictive ability of this model. The third order model with covariates misclassified 41 of the 253 quadrats, which was not noticeably better than the misclassification rate of MODEL2.

Crossvalidation provides another tool for evaluating the predictive ability of a model, for examining the stability of parameter estimates, and for checking for influential quadrats. For crossvalidation we refit the autologistic model 256 times, omitting one quadrat each time, and then predicted disease presence or absence for the omitted quadrat. If the model included neighboring disease or soil variables, the neighbor quadrats were also omitted from the model fitting. Thus, the dataset used to fit the model for prediction of a given quadrat was completely free of the point to be predicted. The numbers of quadrats misclassified by crossvalidation were very similar to the simple misclassification rates reported in Tables 1 and 3: 25% for field 1 MODEL3, 19% for field 2 MODEL2, and 10%

for the 9-term model (6) for field 2. The signs and general magnitudes of the regression coefficients did not vary substantially when any one quadrat was omitted from fitting the model. For example, in field 2 the diagonal (-1,1) direction still had the largest coefficients, followed by disease in the adjacent row neighbors; the coefficients for the other diagonal direction and the neighbors within rows were relatively small, and the second leaf disk assay appeared to be the most important of the soil covariates.

Collinearity among the regressor variables causes difficulty in estimation and interpretation of the coefficients in logistic regression models just as in linear regression models. In autologistic regression there is also the potential for the disease in neighbors in different directions to be highly correlated with each other. This has happened on the last sampling date in field 2. Visually, it appears that disease status runs in strips down rows; however the fitted model gives much higher odds of disease if an adjacent row is diseased than if an adjacent quadrat within the same row is diseased. Using the measure of association, $\hat{\gamma}$, which estimates the difference between the probabilities of concordance and discordance (Agresti 1990), the number of diseased neighbors one quadrat away within the row, $W_{ij} = (y_{i,j-1} + y_{i,j+1})$, is highly correlated with the number of diseased neighbors in every other direction: across rows $\hat{\gamma} = .85$, within the row two quadrats away $\hat{\gamma} = .87$, diagonal (1,1) direction $\hat{\gamma} = .84$, and diagonal (-1,1) direction $\hat{\gamma} = .87$. In contrast, for field 1 the corresponding correlations are much lower: across rows $\hat{\gamma} = .58$, two quadrats away within the row $\hat{\gamma} = .78$, diagonal (1,1) $\hat{\gamma} = .64$, and diagonal (-1,1) $\hat{\gamma} = .57$. Cross-classifying the numbers of diseased quadrats in different directions reveals that in field 2 it was very rare that a quadrat had two diseased neighbors within a row but no disease in the adjacent row quadrats, and vice versa (Table 5). Also, if a quadrat had two diseased within-row neighbors, it most often had 2 diseased neighbors in any other direction.

This collinearity does not show up in standard regression diagnostics, probably because of the categorical nature of the regressor variables. The result, however, is that the regression coefficients may not give meaningful estimates of the odds ratios. The model may still be a good predictive model, but the parameter estimates cannot stand alone. As a demonstration of the difficulty of fitting the second order autologistic model to the field 2 data and interpreting the coefficients, we refit MODEL2 using Monte Carlo maximum likelihood. For the MCML estimation missing water content values were replaced by the mean of their two within-row neighbors. The parameter estimates for the two estimation methods are quite different (Table 6), but the misclassification rates are nearly identical. Consequently, it would not be wise to conclude that the odds of disease depends more on neighbors in adjacent rows than within the row (or vice-versa) for this agricultural field.

5. SUMMARY

This paper has demonstrated the use of a logistic model for describing the spatial pattern of *Phytophthora* epidemics in bell pepper, while incorporating soil covariates and spatial correlation. The autologistic approach is intuitively reasonable. If a low-order model fits well, the parameter estimates give a concise summary of the factors affecting the odds of disease. If no low-order model fits, it may be that a logistic or probit model with another type of spatial correlation pattern would be appropriate (see *e.g.* Breslow and Clayton 1993). One aim of this paper, in addition to showing the autologistic model, has been to bring models that are specifically developed for binary data to the attention of researchers in the plant sciences. These models are not as familiar, and statistical theory and software for them tend not to be as well developed as for models for continuous response variables, but they are rapidly becoming more available.

For autologistic models specifically, there are many questions that remain to be answered. The method of estimation demonstrated (maximum pseudolikelihood) is simple to implement, but has two major practical deficiencies: 1) lack of standard errors for parameter estimates; and 2) lack of formal tests of hypotheses about the parameters. In the future, these two deficiencies can be remedied by use of Monte Carlo maximum likelihood estimation. The lack of standard errors is particularly important for interpreting the estimated coefficients. As a case in point, the estimates of odds ratios given in Section 3 can only be interpreted as guides to the direction and magnitude of changes in odds because there are no estimates of precision for them.

Practitioners also require methods for comparing the covariance structure imposed by the autologistic model with other models of spatial correlation. With Monte Carlo maximum likelihood, tests could be done to determine whether the pairwise-only dependence structure is adequate by adding products of neighbors to the autologistic model. This may not be, in the end, a practical way of detecting departures from the autologistic covariance structure; because the number of forms that departures could take is so large. We also need ways of determining whether neighboring values of the covariates should be incorporated into the model. It may be possible to adapt time series methods for identifying the number of lags of covariates to spatial problems.

Although no formal tests for aptness of the model are available for use with pseudolikelihood estimation, much can be learned from examination of the fitted model, the Pearson residuals, and the prediction errors. The following plots and tables are useful for studying whether the fitted model adequately captures the spatial correlation structure: 1) correlograms of the Pearson residuals; 2) crosstabulations of the number of misclassified quadrats according to numbers of diseased neighbors omitted from the model; and 3) maps of the misclassified quadrats. The maps of misclassified quadrats and the numbers of misclassified quadrats can also be compared to those for higher order models to determine whether the fitted neighborhood structure is large enough. Crossvalidation is another useful tool, and the quadrats misclassified in crossvalidation can be mapped and crosstabulated in the same ways as the simple misclassification errors. Finally, it is important to view the estimated odds ratios with a critical eye. If they seem unreasonable, examination of the covariates and the disease states of the neighbors via correlograms, descriptive plots and tables may shed light both on the relationships among the variables and on the reliability of the parameter estimates.

ACKNOWLEDGEMENTS

This research was funded in part by grant number 92-37303-7616 from the National Research Initiatives Competitive Grants Program of the USDA.

LITERATURE CITED

Agresti, A. 1990. *Categorical Data Analysis.* Wiley. New York. 558pp.

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

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of Royal Statistical Society B.* 36: 192-225.

Besag, J. 1975. Statistical analysis of non-lattice data. *The Statistician.* 24:179-195.

Besag, J. 1977. Some methods of statistical analysis for spatial data. *Bulletin of the International Statistical Institute Proceedings of the 41st Session*. Vol. XLVII, book 2:77-92.

Breslow, N. E. and Clayton, D. G. 1993. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*. 88:9-25.

Chadoeuf, J., Nandris, D., Geiger, J.P., Nicole, M., and Pierrat, J.C. 1992. Modelisation spatio-temporelle d'une epidemic par un processus de Gibbs: estimation et tests. *Biometrics.* 48: 1165-1175.

Cressie, N. A. C. 1991. *Statistics for Spatial Data.* John Wiley & Sons, New York.

Geyer, C. J. 1991. Markov chain Monte Carlo maximum likelihood. In *Computer Science and Statistics: Proc. 23rd Symp. Interface*. 156-163.

Graham, J. M. 1994. Monte Carlo Markov chain likelihood ratio test and Wald test for binary spatial lattice data. Abstract. Joint Statistical Meetings.

August 13-18, Toronto.

Larkin, R. P., Gumpertz, M.L., and Ristaino, J.B. 1995. Geostatistical analysis of Phytophthora epidemic development in commercial bell pepper fields. *Phytopathology*. To appear.

Lecoustre, R., Fargette, D., Fauquet, C. and de Reffye, P. 1989. Analysis and mapping of the spatial spread of african cassava mosaic virus using geostatistics and the kriging technique. *Phytopathology* 79: 913-920.

Reynolds, K. M., and Madden, L. V. 1988. Analysis of epidemics using spatio-temporal autocorrelation. *Phytopathology* 78: 240-246.

Ristaino, J. B., Larkin, R. P., and Campbell, C. L., 1993. Spatial and temporal dynamics of Phytophthora epidemics in commercial bell pepper fields. *Phytopathology* 83:1312-1320.

Ristaino, J. B., Larkin, R. P., and Campbell, C. L., 1994. Spatial dynamics of disease symptom expression during Phytophthora epidemics in bell pepper. *Phytopathology*. In press.

Smyth, G. K., Chakraborty, S., Clark, R. G., and Petitt, A. N. 1992. A stochastic model for anthracnose development in *Stylosanthes scabra. Phytopathology* 82: 1267-1272.

Strauss, D. 1992. The many faces of logistic regression. *The American Statistician* 46:321-326.

Table 1. Field 1 on 7/29/92. Estimated odds ratios ($e^{\hat{\beta}}$) and proportion of quadrats misclassified from fitted autologistic models, MODEL1, MODEL2, and MODEL3.

Model	Intercept	Soil Water Content	6/29 Leaf Disk	7/29 Leaf Disk	Within-Row Disease	Across Rows Disease	Diagonal (1,1)	Diagonal (-1,1)	Miss-Class
MODEL1	0.22	1.08	1.19	1.16					85/252
MODEL2	0.11	0.97	1.16	1.13	4.17	1.07	1.83	1.22	54/252
MODEL3	0.086				4.21	1.08	1.84	1.24	57/256

Table 2. Field 2. Estimated odds ratios ($e^{\hat{\beta}}$) from fitted MODEL2 for three sampling dates.

Date	Intercept	6/22/92 Water Content	6/19/92 Leaf Disk	8/5/92 Leaf Disk	Within-Row	Across-Rows	Diagonal (1,1)	Diagonal (-1,1)
6/25/92	0.002	1.35	1.48	1.14	1.29	1.30	3.41	2.21
7/13/92	0.01	1.14	1.42	1.12	2.31	0.57	2.59	4.09
8/04/92	0.01	1.12	1.19	1.63	1.40	3.02	1.25	4.48

Table 3. Field 2, last sampling date, 8/4/92. Estimated odds ratios, $e^{\hat{\beta}}$, and proportion of quadrats misclassified from fitted autologistic models, MODEL1, MODEL2, and MODEL3.

Model	Intercept	Soil Water Content	6/19 Leaf Disk	8/5 Leaf Disk	Within-Row Disease	Across-Rows Disease	Diagonal (1,1)	Diagonal (-1,1)	Misclass
MODEL1	0.019	1.40	1.41	1.97					63/253
MODEL2	0.011	1.12	1.19	1.63	1.40	3.02	1.25	4.48	42/253
MODEL3	0.039				1.57	4.55	1.23	4.77	42/256

Table 4. Field 2. MODEL2. Quadrats misclassified.

Number diseased two quadrats away within the row	Number of Diseased quadrats two rows away		
	0	1	2
0	3	11	3
1	1	8	1
2	1	14	0
total	5	33	4

Table 5. Field 2. 8/4/92. Cross-classification of numbers of diseased neighbors.

Within-	Across-Rows 1 row away			Within-row, 2 quadrats away			Diagonal (1,1)			Diagonal (-1,1)		
Row	0	1	2	0	1	2	0	1	2	0	1	2
0	90	30	3	100	21	2	88	32	3	89	34	0
1	16	28	14	19	22	17	14	34	10	17	30	11
2	1	31	43	4	16	55	4	27	44	2	28	45

Table 6. Field 2. 8/4/92. MODEL2 parameter estimates and number of quadrats misclassified for Monte Carlo maximum likelihood and maximum pseudolikelihood estimation. Standard errors for Monte Carlo maximum likelihood estimates are given in parentheses.

Method	Intercept	Soil Water Content	6/19 Leaf Disk	8/5 Leaf Disk	Within-Row Disease	Across-Rows Disease	Diago-nal (1,1)	Diago-nal (-1,1)	Mis-class (n=253)
PL	-4.52	0.068	0.23	0.45	0.53	1.18	1.48	0.14	42
MCML	-5.33	0.14	0.43	0.85	1.00	0.42	2.02	0.38	44
	(0.43)	(.074)	(.21)	(.27)	(.41)	(.45)	(.40)	(.36)	

Figure 1. Field 1 disease incidence, soil moisture, and leaf disk assays.

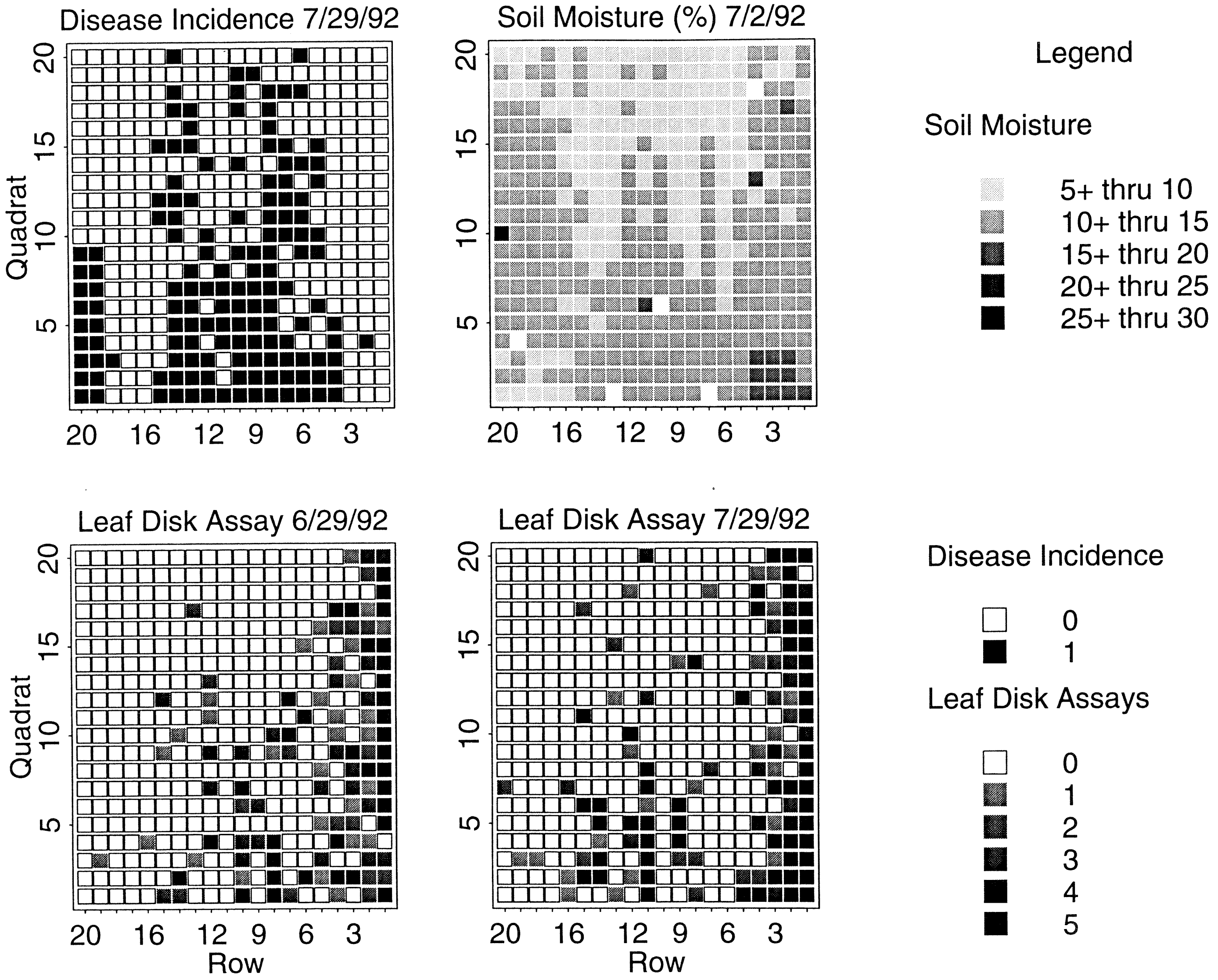


Figure 2. Field 2 disease incidence, soil moisture, and leaf disk assays.

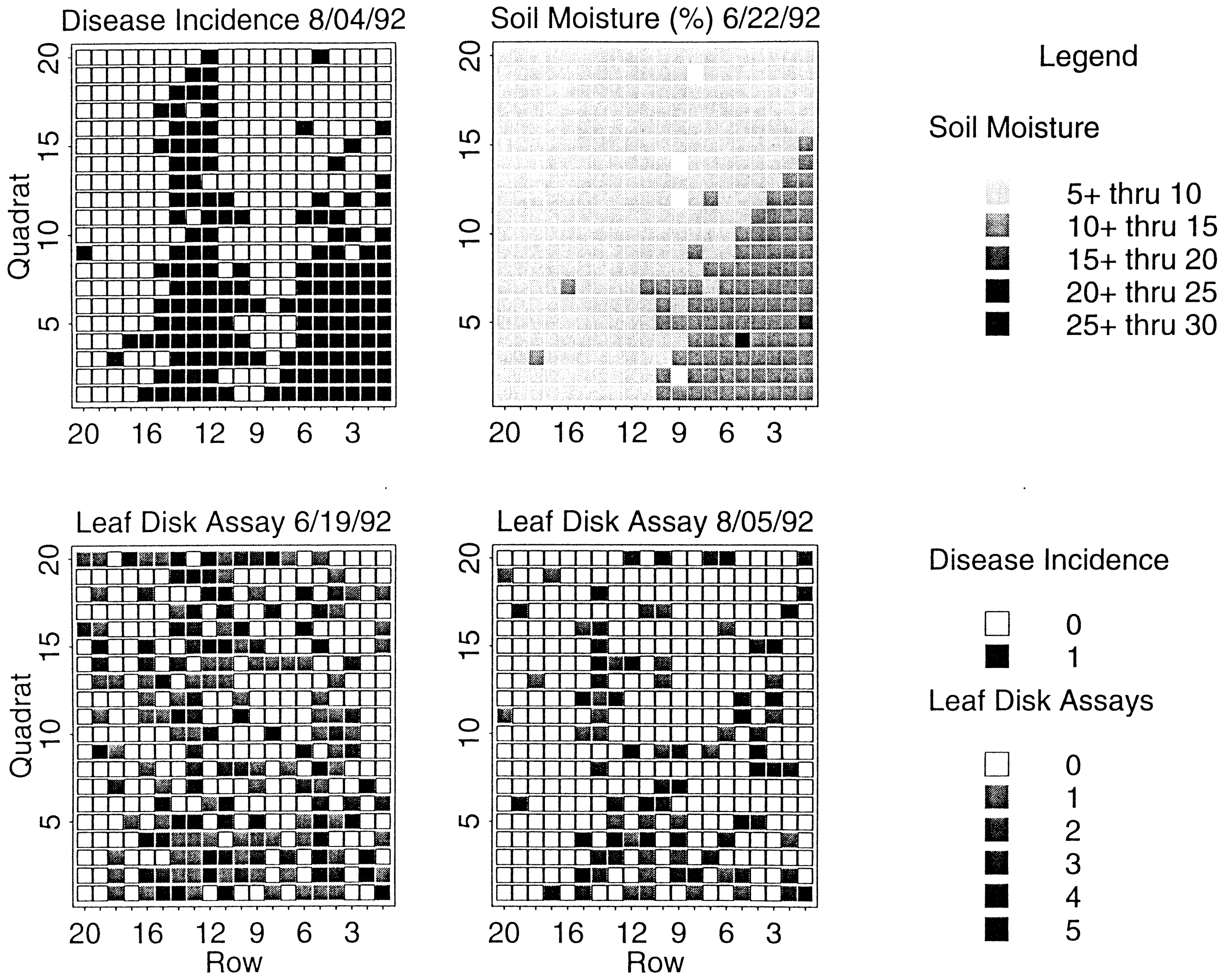


Figure 3. Correlograms of Pearson residuals.

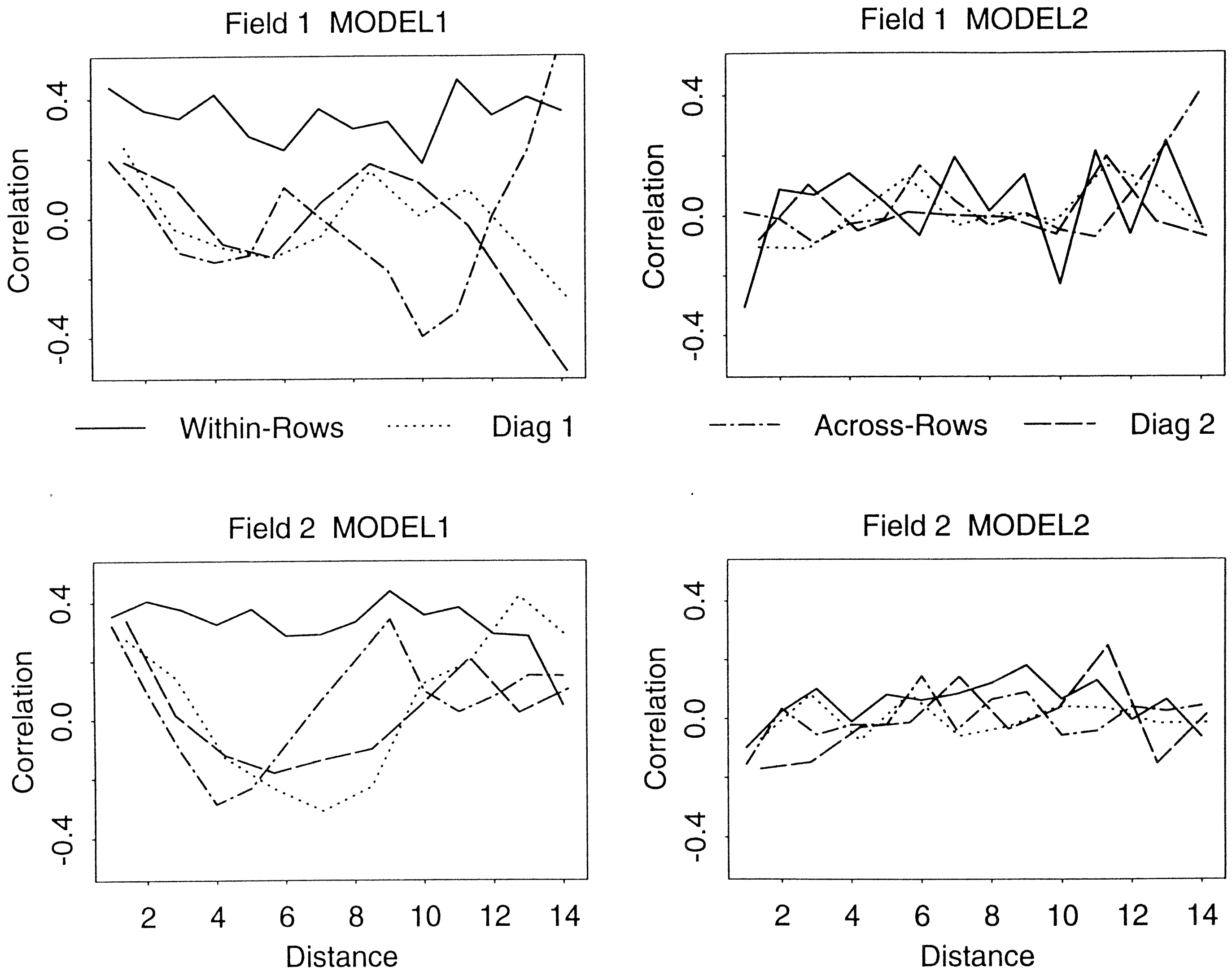


Figure 4. Misclassified quadrats after fitting MODEL2.

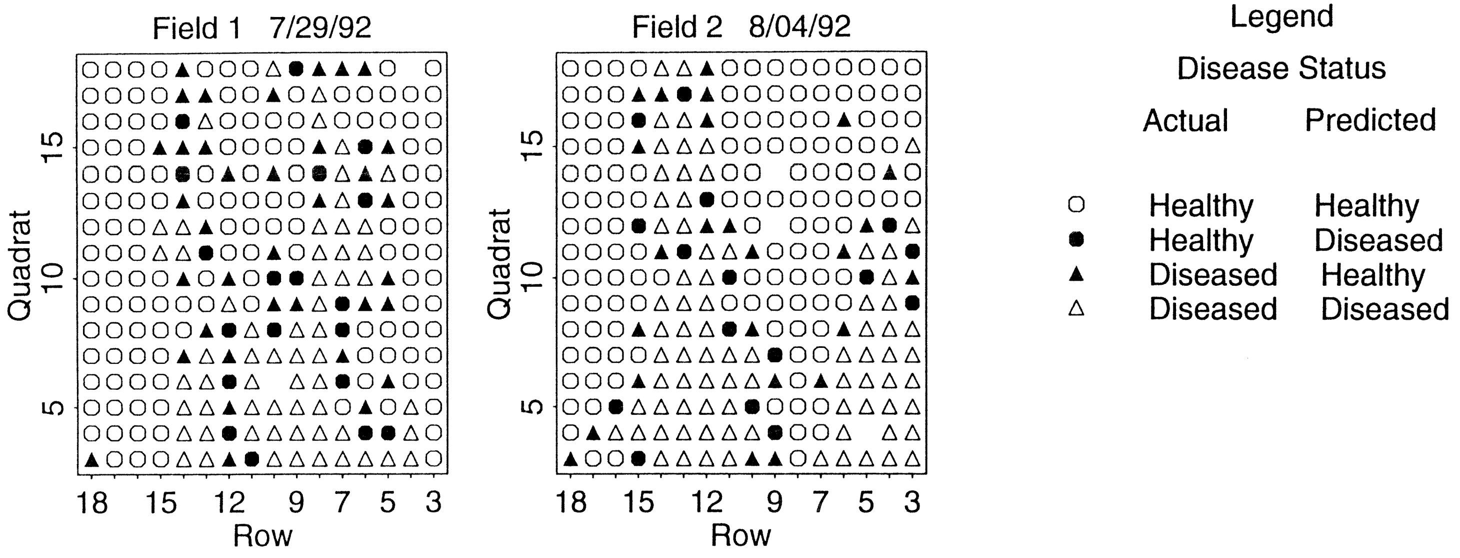
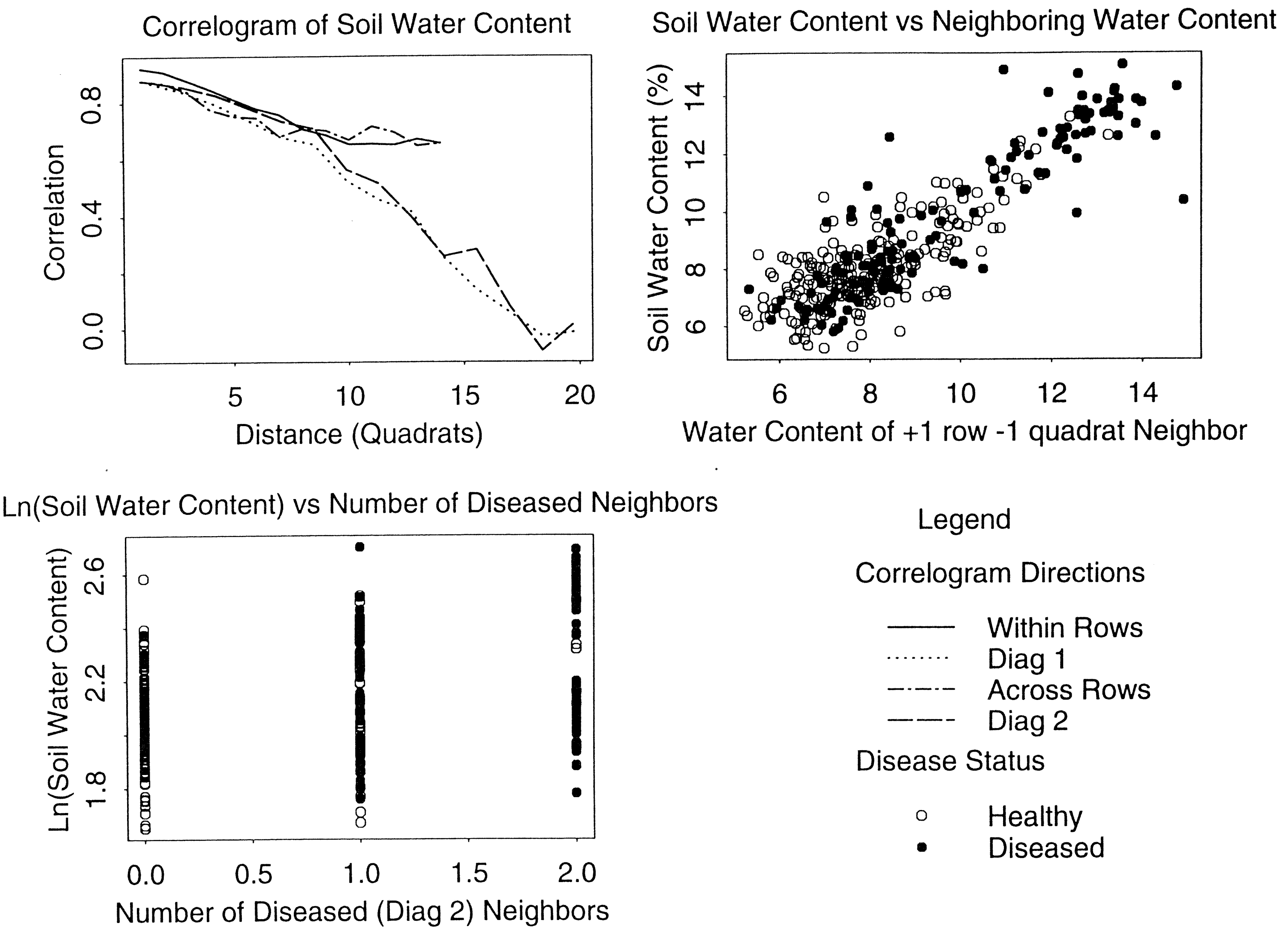


Figure 5. Field 2 Spatial Variability of Soil Water Content


Correlogram of Soil Water Content
Correlation
Distance (Quadrats)
Soil Water Content vs Neighboring Water Content
Soil Water Content (%)
Water Content of +1 row -1 quadrat Neighbor
Ln(Soil Water Content) vs Number of Diseased Neighbors
Ln(Soil Water Content)
Number of Diseased (Diag 2) Neighbors
Legend
Correlogram Directions
Within Rows
Diag 1
Across Rows
Diag 2
Disease Status
○ Healthy
● Diseased