Environment and Planning A 2004, volume 36, pages 1791 – 1811

DOI:10.1068/a36247

A spatial filtering specification for the autologistic model

Daniel A Griffith¶
Department of Geography, 144 Eggers Hall, Syracuse University, Syracuse, NY 13244-1020, USA
Received 8 October 2003; in revised form 31 December 2003

Abstract. The autologistic model describes binary correlated data; its spatial version describes georeferenced binary data exhibiting spatial dependence. The conventional specification of a spatial autologistic model involves difficult-to-nearly-impossible computations to ensure that appropriate sets of probabilities sum to 1. Work summarized here accounts for spatial autocorrelation by including latent map pattern components as covariates in a model specification. These components derive from the surface zonation scheme used to aggregate attribute data, to construct a geographic weights matrix, and to evaluate geographic variability. The illustrative data analysis is based upon field plot observations for the pathogen *Phytophthora capsici* that causes disease in pepper plants. Results are compared with pseudolikelihood and Markov chain Monte Carlo estimation techniques, both for the empirical example and for two simulation experiments associated with it. The principal finding is that synthetic map pattern variables, which are eigenvectors computed for a geographic weights matrix, furnish an alternative, successful way of capturing spatial dependency effects in the mean response term of a logistic regression model, avoiding altogether the need to use other than traditional standard techniques to estimate model parameters.

1 Introduction

Statistical techniques for analyzing correlated binary variables are not as plentiful as those for analyzing correlated continuous data. One exception is two-groups discriminant function analysis (DFA) using a 0/1 indicator variable as the response variable, because it can be rewritten as a standard regression problem. Another is autologistic regression. When binary spatial data are of interest—such as the spatial distribution of wildlife, the spatial pattern of a disease, or the geographic distribution of plant species—the first model that should come to mind for describing these data and their latent spatial autocorrelation is one whose specification captures spatial dependencies. This is exactly what appears in an autologistic specification.

Many practical questions are associated with how spatial dependency is incorporated into an autologistic specification. How is spatial autocorrelation quantified and portrayed? How does spatial autocorrelation impact the assessment of covariates included in an autologistic specification? How are standard errors (that is, statistical efficiency) and model description or prediction enhanced by accounting for spatial autocorrelation? And what improved understanding of a georeferenced binary variable is furnished by an autologistic specification? In addition, several serious technical questions are associated with an autologistic specification. Foremost, how can complications be handled that arise from an autologistic model's intractable normalizing factor (for one possible solution to this problem, see Pettitt et al, 2003); this normalizing factor ensures that the resulting probability mass function sums to 1. Initially this complication was circumvented by employing a specific pseudolikelihood estimation (PLE) procedure, which tends to be less efficient than maximum likelihood estimation because it still assumes independence (Besag, 1975). More recently this particular complication has been avoided through the use of Markov chain Monte Carlo

¶ Current address: Department of Geography and Regional Studies, University of Miami, PO Box 240867, Coral Gables, FL 33124-2221, USA; e-mail: dagriffith@miami.edu

(MCMC) maximum likelihood estimation procedures (for example, Gumpertz et al, 1997; Huffer and Wu, 1998), which allows efficiency to be regained at the expense of numerical intensity.

This paper departs from more conventional autologistic modeling efforts by focusing on specification of a mean response that forces the automodel spatial dependency parameter value to 0. It introduces the eigenvector filtering approach for spatial autologistic models by extending the spatial filtering concept promoted by Getis (1995), Griffith (2000a), and Getis and Griffith (2001). The eigenvector filtering approach is a nonparametric technique that removes the inherent spatial autocorrelation from generalized linear regression models by treating it as a missing variables (that is, first-order) effect. The aim of nonparametric spatial filtering is to control for spatial autocorrelation with a set of spatial proxy variables rather than to identify a global spatial autocorrelation parameter for a spatial process. As such, it utilizes the misspecification interpretation of spatial autocorrelation, which assumes that spatial autocorrelation is induced by missing exogenous variables, which themselves are spatially autocorrelated; spatial autocorrelation latent in the X covariates accounts for spatial autocorrelation in the dependent variable Y, effectively transferring spatial autocorrelation effects from Y, and hence the residuals, to the Xs. In doing so, spatial filtering enables standard software to be used to estimate generalized linear models, which rely on stochastically independent observations, with georeferenced data whether or not the X covariates are known and measured. Each eigenvector of a spatial filter, derivable from a geographic weights matrix, exhibits a distinctive spatial pattern and has associated with it a given spatial autocorrelation level. The eigenvector spatial filtering approach adds a minimally sufficient set of eigenvectors as proxy variables to a set of linear predictors, and in doing so eliminates spatial autocorrelation among the observations by inducing mutual independence. This methodology is much easier to implement than MCMC, and produces parameter estimates that are more efficient than those calculated with the maximum pseudolikelihood method. Furthermore, results permit a more detailed interpretation of spatial autocorrelation effects by allowing explicit visualization and decomposition of an initially hidden autocorrelation pattern in georeferenced binary data.

Griffith et al (1998) furnish one illuminating application of spatial filtering in the context of autologistic modeling. This present paper provides a comparison of results from this form of spatial filtering with those from MCMC and maximum pseudolikelihood in order to demonstrate the feasibility of spatial filtering methodology, enabling spatial scientists to obtain a better evaluation of findings such as those reported in Griffith et al (1998).

2 A brief history of the autologistic model

One of the first versions of the two-dimensional autologistic model was derived in 1941 from work by Ising, who developed its one-dimensional counterpart in 1925 and after whom the Ising model is named (McCoy and Wu, 1973). The Ising model is a simple model of magnetism, and one of the pillars of statistical mechanics. Each site on a two-dimensional lattice can take on one of two possible values (for example, 0 or 1), with the geographic distribution of site values displaying strong positive first-order spatial autocorrelation. The phrase autologistic model began being widely used in the spatial statistics literature following Besag (1972), who went on to introduce a range of possible autoprobability models (Besag, 1974). Besag characterizes this model, in its purely spatial form, as being "analogous to a classical logistic model ... except that here the explanatory variables are themselves observations on the process" (1974, page 201).

Bartlett immediately began using this terminology (for example, 1975; 1978), directly linking it to the Ising model in his discussions.

One of the earlier applications of the autologistic model in the geographic literature appears in Haining (1985), where spatial price competition is defined in terms of autologistic-model-generated probabilities, which in turn are used to generate regional price distributions. In more recent years an increasing number of other researchers have estimated autologistic models, sometimes addressing issues associated with using this class of model. For example, Meier et al (1993) apply a linear landscape version of the model to transect sampling to investigate the population size of St. Croix ground lizards; le Cessie and van Houwelingen (1994) discuss modeling correlated binary outcomes in a way that preserves the logistic property of the marginal response probabilities; Augustin et al (1996) use the Gibbs sampler and explicitly model spatial autocorrelation intrinsic in presence – absence species data; Wu and Huffer (1997) use an autologistic regression model for binary georeferenced data to describe the distribution of plant species; Dubin (1995; 1997) specifies a logit diffusion model for the probability of adoption; Hoeting et al (1999) apply an autologistic model with covariates to interpolate sparsely sampled georeferenced data over a region in order to construct maps of the likelihood of presence – absence of particular plant species; and Brownstein et al (2003) employ an autologistic model to analyze the spatial distribution of the black-legged tick, in order to assess human risk for Lyme disease in much of the United States. Regardless of order of spatial dependency specified, number of covariates included, or phenomenon studied, each of these studies analyzes spatially autocorrelated dichotomous variables.

Classical maximum likelihood estimation techniques cannot be used to estimate autologistic model parameters. Therefore, in addition to outlining the autologistic model, Besag (for example, 1974) describes two procedures for estimating its parameters that he labels pseudolikelihood and coding. His maximum pseudolikelihood method treats areal unit values as though they are conditionally independent, and is equivalent to maximum likelihood estimation when they are independent. A pseudolikelihood is specified as the product of conditional probability density functions for each areal unit, given neighboring areal units. For logistic regression, then, each areal unit value is regressed on a function of its surrounding areal unit values. This estimation procedure involves a trade-off between simplicity and statistical efficiency; efficiency is lost when dependent values are assumed to be independent. Meanwhile, Besag's coding scheme divides a set of areal units into (ideally two) subsets free of, for example, areal unit adjacencies for a first-order dependency structure. This allows the first-order Markov assumption to imply that values of each areal unit in a given subset are mutually independent. Then each value in one subset can be regressed on a function of its corresponding neighboring values in the other subset, with this estimation repeated by switching regressors and regressands. The resulting spatial autoregressive parameter estimates then can be averaged, as long as they are comparable. Preferring two subsets makes implementing this coding scheme practical for regular lattice data, and cumbersome for irregular lattice data. Regardless, this coding procedure essentially is equivalent to the first step in MCMC estimation of parameters for an autologistic model. Bartolucci and Besag (2002) propose a recursive algorithm as a more useful alternative to estimate the joint distribution of georeferenced binary values. But this new algorithm can be as numerically intensive as MCMC estimation. In contrast, Heagerty and Lele (1998) propose a computationally simple method for estimation and prediction using georeferenced binary data—whose spatial autocorrelation is described with a geostatistical semivariogram model—that is based upon pairwise likelihood contributions. And Albert and McShane (1995) propose a generalized estimating equations approach, again employing geostatistical semivariogram models to parameterize spatial dependency.

3 Logistic regression

Logistic regression (see Hosmer and Lemeshow, 2000) assumes independent Bernoulli outcomes, denoted by $Y_i = 0$ or 1, taken at locations $i = 1, 2, ..., n$, where these binary values can be described by a set of explanatory variables denoted by $\boldsymbol{X}_i$, a $1 \times (K+1)$ vector of K covariate values, and a 1 (for the intercept term), for location i. The probability of a 1 being realized for these data is given by

$$\mathrm{P}(Y_i = 1|\boldsymbol{X}_i) = \frac{\exp(\boldsymbol{X}_i \boldsymbol{\beta})}{1+\exp(\boldsymbol{X}_i \boldsymbol{\beta})}, \tag{1}$$

where $\boldsymbol{\beta}$ is the $(K+1) \times 1$ vector of nonredundant parameters, and where $\mathrm{P}(Y_i = 0|\boldsymbol{X}_i) = 1 - \mathrm{P}(Y_i = 1|\mathbf{X}_i)$. Equation (1) has been employed in geographic studies of dichotomous phenomenon (for example, Clark and Hosking, 1986; Wrigley, 1985). Its simplest form is for a constant probability across areal units: $\mathrm{P}(Y_i = 1|\boldsymbol{X}_i) = \mathrm{P}(Y_i = 1|\alpha) = \exp\alpha/(1+\exp\alpha)$, for some constant α (using popular bivariate regression notation, which in multiple regression notation is denoted by β_0), where $\mathrm{P}(Y_i = 1|\alpha) \rightarrow 0$ as $\alpha \rightarrow -\infty$, $\mathrm{P}(Y_i = 1|\alpha) \rightarrow 0.5$ as $\alpha \rightarrow 0$, and $\mathrm{P}(Y_i = 1|\alpha) \rightarrow 1$ as $\alpha \rightarrow \infty$.

3.1 Autologistic regression

Suppose the $n \times n$ 0/1 binary geographic connectivity or weights matrix $\mathbf{C}$ represents the geographic arrangement of data values. Accordingly, $c_{ij} = 1$ if two locations i and j are neighbors, and $c_{ij} = 0$ otherwise (note: $c_{ii} = 0$); matrix $\mathbf{C}$ contains n^2 0/1 values. Pairwise-only spatial dependence often is assumed when specifying such automodels in terms of matrix $\mathbf{C}$ (for an example employing more than pairwise cliques see Tjelmeland and Besag, 1998). Retaining this assumption, for a spatial autoregressive situation, the problem becomes one of estimating the parameters of the following probability function:

$$\mathrm{P}(Y_i = 1|\alpha_i, \boldsymbol{C}_i \boldsymbol{Y}) = \frac{\exp\left(\alpha_i + \rho \sum_{j=1}^{n} c_{ij} y_j\right)}{1+\exp\left(\alpha_i + \rho \sum_{j=1}^{n} c_{ij} y_j\right)}, \tag{2}$$

where Y_i denotes a random variable and y_i denotes an observed realization of the random variable, α_i is the parameter capturing large-scale variation (and hence could be specified in terms of vector $\boldsymbol{X}_i$), ρ is the spatial autocorrelation parameter, and $\boldsymbol{C}_i$ is the row vector of c_{ij} values for location i. Matrix $\mathbf{C}$ must be symmetric for identifiability reasons (Besag, 1974), and often is, but need not be, binary. The pure spatial autoregressive form of equation (2) is the extension of equation (1) for which α_i is the constant α:

$$\mathrm{P}(Y_i = 1|\alpha, \boldsymbol{C}_i \boldsymbol{Y}) = \frac{\exp\left(\alpha + \rho \sum_{j=1}^{n} c_{ij} y_j\right)}{1+\exp\left(\alpha + \rho \sum_{j=1}^{n} c_{ij} y_j\right)}.$$

In this situation, spatial autocorrelation may be measured with the join-count statistics (Cliff and Ord, 1981), denoted by BB if two ones are geographically nearby (that is, $c_{ij} = 1$), BW if a one and a zero are geographically nearby, and WW if two zeros are geographically nearby. These join-count statistics can be converted to z-scores by subtracting their respective means and dividing the resulting differences by their respective standard deviations, parameters that are established under a null hypothesis of zero spatial autocorrelation (see Cliff and Ord, 1981, pages 36–41).

The proposition promoted in this paper is that by including variables in matrix $\mathbf{X}$—the $n \times (K+1)$ concatenation of the n $\boldsymbol{X}_i$ vectors—that account for the spatial autocorrelation observed in the associated geographic distribution of binary values, the explicit autoregressive term in equation (2) can be dispensed with. In other words, spatial dependence effects are shifted from a small-scale variation term to the mean response term, resulting in ρ being forced to 0. This perspective contends that spatial autocorrelation appears in residuals because variables are missing from the mean response specification (for example, the geographic distribution of soil types or moisture content for agricultural yields). This shift can occur by introducing appropriate synthetic variables into matrix $\mathbf{X}$ that serve as surrogates for spatially autocorrelated missing variables. These synthetic variables are the eigenvectors of the following modified version of binary matrix $\mathbf{C}$:

$$\left(\mathbf{I} - \frac{\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}}{n}\right)\mathbf{C}\left(\mathbf{I} - \frac{\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}}{n}\right), \tag{3}$$

where $\mathbf{I}$ is the identity matrix, $\boldsymbol{1}$ is an $n \times 1$ vector of ones, T denotes the operation of matrix transpose, and n is the number of areal units. This particular matrix expression appears in the numerator of the widely used Moran coefficient (MC) index of spatial autocorrelation. Tiefelsdorf and Boots (1995) show that all of the eigenvalues of matrix expression (3) relate to specific MC values. One appealing property of equation (3) is that matrix $\mathbf{C}$ is constant for a given surface partitioning and adjacency definition, rendering the same set of eigenvectors for all attributes geographically distributed across the given surface partitioning (that is, zonation scheme).

In extending the findings of Tiefelsdorf and Boots (1995), and linking them to principal components analysis (PCA) (Griffith, 1984), the eigenvectors of expression (3) may be interpreted in the context of latent map pattern as follows. The first eigenvector, $\boldsymbol{E}_1$, of expression (3) is the set of numerical values that has the largest MC achievable by any possible set of numerical values, for the arrangement of locations given geographic connectivity matrix $\mathbf{C}$. The second eigenvector is the set of numerical values that has the largest achievable MC by any set of numerical values that is uncorrelated with $\boldsymbol{E}_1$. This sequential construction of eigenvectors continues through $\boldsymbol{E}_n$, which is the set of numerical values that has the largest negative MC achievable by any set of numerical values that is uncorrelated with the preceding $(n-1)$ eigenvectors.

Hence these n eigenvectors describe the full range of all possible mutually orthogonal map patterns, and may be interpreted as synthetic map variables that represent specific natures (that is, positive or negative) and degrees (for example, negligible, weak, moderate, strong) of potential spatial autocorrelation. This perspective also is alluded to by Switzer (2000), who is more concerned with efficient sampling issues, and Boots and Tiefelsdorf (2000), who discuss the construction of linear combinations of these eigenvectors in order to obtain any prespecified level of spatial autocorrelation. In the presence of positive spatial autocorrelation, then, an analysis can employ those eigenvectors depicting map patterns exhibiting consequential levels of positive spatial autocorrelation; operationally speaking, attention initially can be restricted to eigenvectors having $\mathrm{MC}/\mathrm{MC}_{\mathrm{extreme}} > 0.25$, where $\mathrm{MC}_{\mathrm{extreme}}$ denotes either the maximum ($\mathrm{MC}_{\max}$) or the minimum ($\mathrm{MC}_{\min}$) possible value of MC, because $\mathrm{MC}/\mathrm{MC}_{\mathrm{extreme}} = 0.25$ refers to situations in which 5%–10% of the information content[1]—measured in terms of the pseudo-R^2-value accompanying an autonormal model (see Griffith, 2003)—tends to be redundant because of the presence of nonzero spatial autocorrelation. If the autocorrelation detected with indices (for example, BB, BW, MC) is positive, then only eigenvectors for which $\mathrm{MC}/\mathrm{MC}_{\max} \geqslant 0.25$ initially would be considered; if the autocorrelation detected with indices is negative, then only eigenvectors for which $\mathrm{MC}/|\mathrm{MC}_{\min}| \leqslant 0.25$ (or, equivalently, $\mathrm{MC}/\mathrm{MC}_{\min} \geqslant 0.25$) initially would be considered. If overcorrection for spatial autocorrelation occurs, then this threshold value of 0.25 needs to be increased.

Restricting the set of eigenvectors over which a stepwise selection search is performed is sensible for several reasons. First, if a homogeneous process underlies a set of georeferenced data, then the nature of spatial autocorrelation components should not vary. A Moran scatterplot may help assess the feasibility of this characterization. Second, if a parsimonious set of eigenvectors is to be selected, then eigenvectors depicting near-zero spatial autocorrelation should be avoided, because they fail to capture any geographic information. Third, as n increases, the number of eigenfunctions increases, and hence the numerical intensity involved in executing a stepwise regression also increases. This issue is of paramount importance when dealing with remotely sensed data, where the number of eigenfunctions is in the hundred thousands, millions, or billions. Finally, selection experience with positive spatial autocorrelation analyses suggests that virtually all prominent eigenvectors are contained in the set of eigenfunctions whose eigenvalues constitute the top quartile.

Given the foregoing MC decomposition result, the research problem becomes one of determining whether or not expression (2) can be replaced by

$$\mathrm{P}(Y_i = 1|\mathbf{E}_{i,K}) = \frac{\exp(\alpha + \mathbf{E}_{i,K}\boldsymbol{\beta})}{1+\exp(\alpha + \mathbf{E}_{i,K}\boldsymbol{\beta})}, \tag{4}$$

where K denotes some subset of the n eigenvectors that has been chosen by supervised selection criteria (that is, $\mathbf{E}_{i,K}$ is an $n \times K$ matrix whose columns are the K selected eigenvectors), dispensing with the $\rho \sum c_{ij} y_j (j = 1, ..., n)$ term by shifting spatial dependence effects to the large-scale variation term represented by $\mathbf{E}_{i,K}\boldsymbol{\beta}$, forcing ρ to 0, and letting α_i be the constant α (that is, no covariates other than eigenvectors are included in the specification). A link between equations (2) and (4) may be gleaned from the following algebraic manipulations:

$$\mathrm{P}(\boldsymbol{Y} = \boldsymbol{1}|\mathbf{C}\boldsymbol{Y}) = \mathrm{P}(\boldsymbol{Y} = \boldsymbol{1}|\mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^{\mathrm{T}}\boldsymbol{Y}) = \mathrm{P}(\boldsymbol{Y} = \boldsymbol{1}|\mathbf{E}\boldsymbol{\delta}),$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues whose order is the same as the corresponding eigenvectors in matrix $\mathbf{E}$, and $\boldsymbol{\delta}$ is a vector of coefficients. A straightforward extension of equation (4) would be to include covariates in its specification; the version selected for study here avoids dealing with covariates in order to focus solely on the eigenvector spatial filter. The parameters α and $\boldsymbol{\beta}$ of equation (4) can be estimated with the method of maximum likelihood.

3.2 Eigenvector selection criteria

One difficulty associated with equation (4) is that n eigenvectors are extracted from expression (3). One of these eigenvectors is proportional to the vector $\boldsymbol{1}$, which is associated with the constant parameter α (the intercept of a model), leaving $(n-1)$ eigenvectors as candidates for describing latent spatial autocorrelation. Restricting attention to only those eigenvectors describing positive spatial autocorrelation, when

[1] Common variance of at least 5% is a standard rule-of-thumb threshold value employed in applied statistics to identify substantively meaningful relationships (for example, see Griffith and Amrhein, 1997, page 95).

latent spatial autocorrelation is positive, further reduces the candidate set. The symmetry of expression (3) ensures that the eigenvectors are orthogonal; the projection matrix $(\mathbf{I} - \boldsymbol{11}^{\mathrm{T}}/n)$ ensures that they are uncorrelated.[2] Unfortunately, these properties are lost to some degree in logistic regression analysis because of the weighting involved in parameter estimation (see the appendix); collinearity among the weighted eigenvectors causes some difficulty in estimation. The degree of this collinearity can be indexed by the square root of the ratio of the largest and smallest eigenvalues extracted from a correlation matrix (known in the statistics literature as the condition number). This value is 1 for the orthogonal case. For the numerical example presented in section 4, this index increases to roughly 2.1, indicating increased but not harmful multicollinearity (see Griffith and Amrhein, 1997, page 98).

Nevertheless, supervised stepwise selection of eigenvectors is a useful and effective approach to identifying the subset of eigenvectors that best describes latent spatial autocorrelation in a particular georeferenced binary variable. This procedure begins with only the intercept included in the logistic regression specification. Then, at each step an eigenvector is considered for addition to the model specification. The one that produces the greatest reduction in the log-likelihood function χ^2-test statistic is selected, but only if it produces at least a prespecified minimum reduction; this is the criterion used to establish statistical importance of an eigenvector. At each step all eigenvectors previously entered into the model specification are reassessed, with the possibility of removal of vectors added at an earlier step. The forward–backward stepwise procedure terminates automatically when some prespecified threshold values are encountered for entry and removal of all candidate eigenvectors. The final inclusion criterion may be determined by the MC value of the residuals or the join-count statistics for misclassified rounded off values of $\hat{Y}$—with a maximum likelihood criterion, $0 \leqslant \hat{Y} < 0.5$ becomes 0, and $0.5 \leqslant \hat{Y} \leqslant 1$ becomes 1—which should indicate an absence of spatial autocorrelation. The MC may be used because the predicted probability, $\hat{Y}$, is a continuous measure in the interval [0, 1], and its sampling distribution for the $(Y - \hat{Y})$ residuals can be constructed using permutations across the set of areal units. If substantive covariates belong in a model specification, they can be forced to be retained during the stepwise selection of eigenvectors; eigenvectors are synthetic and capture correlated map patterns, whereas substantive covariates have conceptual or theoretical meaning as well as latent spatial autocorrelation.

Supervision of this selection procedure involves monitoring reduction in the residual spatial autocorrelation. At each step the BB and BW z-scores (say z_{BB} and z_{BW}) for misclassified areal units (residuals) need to be calculated. Equation (4) produces a value between 0 and 1 (the estimated probability), that when rounded off gives a binary prediction, $\hat{Y}$, for the values of variable Y. The correlation coefficient, ϕ, can be calculated with these two measures. BB and BW statistics then can be calculated for the absolute value of the difference between these two values, $|\hat{Y} - Y|$; a 0 indicates a correct prediction whereas a 1 indicates an incorrect prediction. The stepwise procedure should be terminated when $|z_{\mathrm{BB}}|$ and $|z_{\mathrm{BW}}|$ suggest the presence of negligible or trace spatial autocorrelation. Of note is that this evaluation is approximate, as the sampling

[2] This situation differs from PCA in that here eigenvectors themselves are used as synthetic variables, rather than as coefficients for constructing linear combinations of a set of original variables. Thus, the numerator of a product moment correlation coefficient is

$$\boldsymbol{E}_k^{\mathrm{T}}\boldsymbol{E}_k - (\boldsymbol{1}^{\mathrm{T}}\boldsymbol{E}_k/n)(\boldsymbol{1}^{\mathrm{T}}\mathrm{E}k/n) = 0 - (\boldsymbol{1}^{\mathrm{T}}\boldsymbol{E}_k/n)(\boldsymbol{1}^{\mathrm{T}}\boldsymbol{E}_k/n)$$

because the eigenvectors are orthogonal. Hence this numerator equals 0 only if $\boldsymbol{1}^{\mathrm{T}}\boldsymbol{E}_k/n = 0$; in other words, the sum of the elements of eigenvector $\boldsymbol{E}_k$ must be zero. This result is guaranteed by the presence of a single eigenfunction for which $\lambda \equiv 0$ and $\boldsymbol{E} \equiv (1/n^{1/2})\boldsymbol{1}$, which also accounts for the intercept term in equation (4).

distribution of spatial autocorrelation in logistic regression residuals remains to be formalized.

Employing this second criterion of monitoring the residual spatial autocorrelation is advisable because otherwise the χ^2 criterion may allow too many eigenvectors to be added to the logistic equation, resulting in an overcorrection for nonzero spatial autocorrelation latent in variable *Y*.

4 Numerical results

The spatial filtering procedure is evaluated here both with real data and with simulated data in order to illustrate its general applicability.

The pathogen *Phytophthora capsici* causes disease in pepper plants. Data for the incidence of this disease—recorded as a 0 if a plant is healthy and a 1 if a plant is infected—in a North Carolina pepper field divided into a 20 × 20 grid of quadrats is reported in Graham (1994) and portrayed in figure 1(a). The neighborhood structure employed here has $c_{ij} = 1$ if two quadrats have a nonzero-length common boundary, and $c_{ij} = 0$ otherwise (that is, the rook's move chess analogy). Statistics for these data include 143 quadrats with diseased plants, join-count statistics of BB = 181 and BW = 185, and 189 candidate eigenvectors associated with positive spatial autocorrelation extracted from expression (3)—using the **C** matrix for this geographic landscape—depicting map patterns portraying positive spatial autocorrelation.

Two sets of simulated data (following Guyon, 1995, page 212) were generated based upon this empirical dataset. The first of these two datasets comprises 1000 simulated maps constructured with the logistic regression equation (4) specified with $K = 19$ selected eigenvectors; the coefficients employed for this equation appear in table 1 (over):

$$\begin{aligned}\hat{p}_i = {} & 1 - 1/\{1 + \exp[-1.5521 + 20.9094\boldsymbol{E}_{1,i} - 14.0700\boldsymbol{E}_{2,i} + 19.9043\boldsymbol{E}_{4,i} \\ & - 28.5770\boldsymbol{E}_{5,i} - 12.2360\boldsymbol{E}_{7,i} - 17.0488\boldsymbol{E}_{8,i} \\ & - 12.9671\boldsymbol{E}_{10,i} + 8.5895\boldsymbol{E}_{13,i} + 21.6357\boldsymbol{E}_{16,i} + 20.7751\boldsymbol{E}_{17,i} \\ & - 16.2489\boldsymbol{E}_{18,i} + 15.0228\boldsymbol{E}_{22,i} - 12.2855\boldsymbol{E}_{26,i} - 6.7825\boldsymbol{E}_{30,i} \\ & - 11.3520\boldsymbol{E}_{33,i} + 14.8709\boldsymbol{E}_{41,i} + 10.2964\boldsymbol{E}_{45,i} - 9.4756\boldsymbol{E}_{47,i} + 8.3727\boldsymbol{E}_{48,i}]\} \\ i = {} & 1, 2, ..., 400 . \qquad (5)\end{aligned}$$

In other words, by substituting the coefficients contained in table 1 into equation (4), a probability—given by equation (5)—can be calculated for each of the 400 quadrats into which the agricultural field was divided. Each of these probabilities represents the (population) chance of a 1 being randomly selected for the associated quadrat. Next, for each of the 1000 simulation replications, 400 independent random selections were made from a Bernoulli distribution using these 400 probabilities. This sampling was implemented with the IMSL RNBIN routine (Visual Numerics Inc., San Ramon, CA), which generates pseudorandom numbers from a binomial distribution. For agricultural field plot *i* for a given simulated map, a pseudorandom number was generated for a single Bernoulli trial having probability $\hat{p}_i$ given by equation (5). The seed selected to initiate routine RNBIN was calculated using the computer system clock. Independent selections can be made here because spatial autocorrelation is contained solely in the pattern of the probabilities, induced through the linear combination of eigenvectors substituted into equation (4). This procedure produced 1000 spatially autocorrelated binary maps.

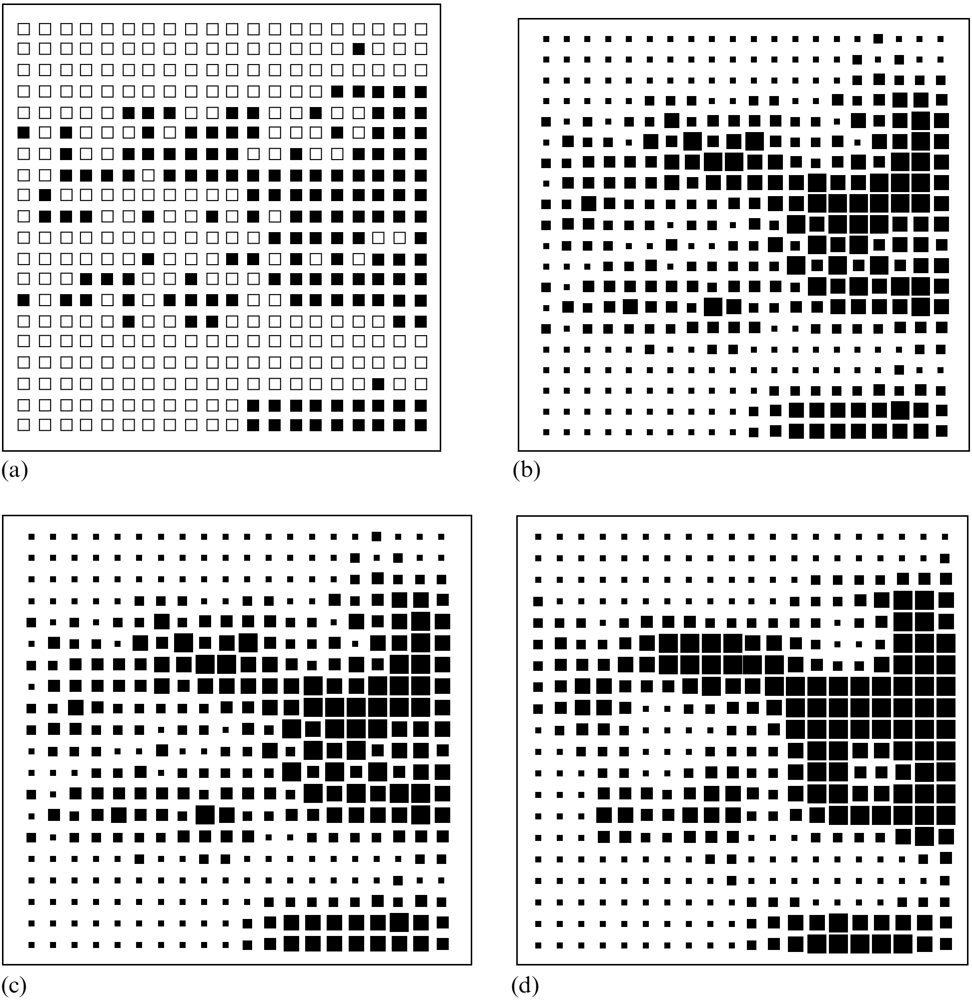

Figure 1. (a) Distribution of disease (solid black squares) across field plots. (b) Probability of disease, which is directly proportional to the size of a solid black square, rendered by conventional logistic regression (pseudolikelihood estimation). (c) Probability of disease, which is directly proportional to the size of a solid black square, rendered by MCMC – maximum likelihood estimation logistic regression. (d) Probability of disease, which is directly proportional to the size of a solid black square, rendered by the spatial filter model—equation (4).

The second dataset also comprises 1000 simulated maps, constructed using the logistic regression equation (2) and the MCMC[3] maximum likelihood parameter estimates appearing in table 2 (over):

$$p_{i,\tau} = \frac{\exp\left(-2.8242 + 1.4114 \sum_{j=1}^{n} c_{ij} y_{j,\tau-1}\right)}{1 + \exp\left(-2.8242 + 1.4114 \sum_{j=1}^{n} c_{ij} y_{j,\tau-1}\right)},$$

[3] MCMC convergence was monitored with convergence of the relative frequency of ones and of the BB join-count statistic.

Table 1. Spatially filtered logistic regression parameter estimation results based on a parsimonious set of eigenvectors selected from a restricted set ($\mathrm{MC}/\mathrm{MC}_{\max} > 0.5$) by simultaneously maximizing the χ^2 criterion and minimizing residual spatial autocorrelation.

Term	Empirical data		Simulated maps (r = 1000)			MCMC maps (r = 1000)		
	coefficient	SE	average coefficient	SE	SW	average coefficient	SE	SW
intercept	−1.5521	0.2514	−1.7410	0.2999	0.989***	−0.7433	0.4867	0.995***
$\boldsymbol{E}_1$	20.9094	3.6373	23.3828	4.4885	0.990***	−0.5603	9.0389	0.998
$\boldsymbol{E}_2$	−14.0700	3.4006	−15.7441	3.9273	0.998	0.1738	8.9761	0.999
$\boldsymbol{E}_4$	19.9043	3.6634	21.9133	4.2612	0.984***	8.6942	8.1449	0.999
$\boldsymbol{E}_5$	−28.5770	4.2431	−32.0469	5.3117	0.988***	0.4857	7.5655	0.996***
$\boldsymbol{E}_7$	−12.2360	4.3588	−13.6315	4.9120	0.991***	0.1047	7.1625	0.996**
$\boldsymbol{E}_8$	−17.0488	3.6469	−19.0721	4.1735	0.991***	−0.0718	6.5995	0.997*
$\boldsymbol{E}_{10}$	−12.9671	3.8687	−14.2414	4.7221	0.994***	−2.2549	6.3701	0.998
$\boldsymbol{E}_{13}$	8.5895	3.2588	9.5346	3.7312	0.994***	6.2305	5.6361	0.998
$\boldsymbol{E}_{16}$	21.6357	4.0994	24.2504	4.9459	0.985***	−0.2351	5.5465	0.999
$\boldsymbol{E}_{17}$	20.7751	3.8546	23.1638	4.5999	0.985***	0.0458	5.4090	0.998
$\boldsymbol{E}_{18}$	−16.2489	3.9902	−18.0229	4.7846	0.992***	−0.0880	5.0405	0.998
$\boldsymbol{E}_{22}$	15.0228	3.4241	16.8955	3.9003	0.996***	−0.1487	4.7311	0.999
$\boldsymbol{E}_{26}$	−12.2855	3.7770	−13.9481	4.4221	0.995***	−0.0970	4.4601	0.998
$\boldsymbol{E}_{30}$	−6.7825	3.0978	−7.6449	3.5205	0.998	4.5137	4.5934	0.997
$\boldsymbol{E}_{33}$	−11.3520	3.7003	−12.7405	4.2060	0.997*	−0.1558	4.2629	0.995***
$\boldsymbol{E}_{41}$	14.8709	3.3158	16.3687	3.7776	0.996**	0.0084	3.7087	0.998
$\boldsymbol{E}_{45}$	10.2964	3.3470	11.2270	3.8618	0.994***	−0.1318	3.7425	0.998
$\boldsymbol{E}_{47}$	−9.4756	3.3411	−10.5502	3.9203	0.995***	−0.1423	3.4684	0.999
$\boldsymbol{E}_{48}$	8.3727	3.4017	9.0807	3.8655	0.993***	−0.2311	3.4644	0.998

Note: signs are unimportant, as properties of eigenvectors do not change when multiplied by −1. Results here relate to those appearing for the MC threshold value of 0.51669, but with the marginal eigenvector (for which MC = 0.55977) manually removed after stepwise selection. r denotes the number of replications in a simulation. MC—Moran coefficient. MCMC—Markov chain Monte Carlo. SE—standard error. SW—Shapiro–Wilk test statistic.
*, **, ***, respectively, denote a significant difference from 1 at the 0.10, 0.05, and 0.01 level.

Table 2. Autologistic parameter estimates for the diseased pepper data.

Parameter	Maximum pseudolikelihood			MCMC–maximum likelihood	
	estimate	standard error		simulated	asymptotic standard error
		conventional	MCMC simulation (r = 100)		
α	−2.6657	0.2814	0.2417	−2.8242	0.1950
ρ	1.3297	0.1391	0.1485	1.4214	0.1056

Note: r denotes the number of replications in a simulation. MCMC—Markov chain Monte Carlo.

for iteration τ, yielding a conditional probability for each of the 400 quadrats into which the agricultural field was divided. The initial map ($\tau = 0$) was constructed with a sample of zeroes and ones drawn using the IMSL RNBIN routine and probabilities $p_{i,0} = p = 0.5$ ($i = 1, 2, ..., 400$). For each iteration and each agricultural field plot i, a pseudorandom number was generated for a single Bernoulli trial having probability $p_{i,\tau}$. The seed selected to initiate routine RNBIN was calculated by using the computer system clock. Each iterative pass over the map was done randomly, with the order of field plot iterative (that is, τ) updates being determined by IMSL routine IPER.

Iterations were performed until the Markov chain converged. In other words, each of these maps was constructed by using a Gibbs sampler, with the chain running for 1000 iterations; each of the final iteration results constitutes a single simulated binary map.

4.1 Parameter estimation results

Maximum pseudolikelihood estimates—conventional generalized linear model parameter estimates for equation (2)—were obtained using PROC LOGISTIC in SAS (SAS Inc., Cary, NC). MCMC – maximum likelihood estimates were obtained using PROC NLIN in SAS to estimate the score equation reported in Huffer and Wu (1998) and the log-likelihood equation reported in Graham (1994). The initial state was a map whose binary values were selected at random, each with a probability of 0.5, and whose external edge quadrats were assumed to take on the value of 0.[4] The maximum pseudolikelihood estimates were used to generate a Markov chain by using a Gibbs sampler. In other words, the parameter estimates $\hat{\alpha}$ and $\hat{\rho}$ are obtained with the pseudolikelihood method, which renders reasonable parameter but poor standard error estimates, and then used to compute

$$\hat{p}_{i,\tau+1} = \frac{\exp\left(\hat{\alpha} + \hat{\rho}\sum_{j=1}^{n} c_{ij}\, y_{j,\tau\rightarrow\tau+1}\right)}{1 + \exp\left(\hat{\alpha} + \hat{\rho}\sum_{j=1}^{n} c_{ij}\, y_{j,\tau\rightarrow\tau+1}\right)}, \qquad i = 1, 2, ..., n,$$

for each iteration $\tau + 1$ where $\tau \rightarrow \tau + 1$ denotes the transition updating from iteration τ to $\tau + 1$. Once calculated, each probability, $\hat{p}_{i,\tau+1}$, then is used to draw a random value from a Bernoulli distribution, which becomes the map value for location i at iteration $\tau + 1$. The first probabilities are $\hat{p}_{i,1}$ ($i = 1, 2, ..., n$), which are calculated by using the initial state map obtained with random sampling. Iteration $\tau + 1$ calculations begin by using map values for iteration τ, and gradually change to using map values for iteration $\tau + 1$ as they increasingly replace the values for iteration τ. Repeated application of these n marginal probability calculations, randomly permuting the set of locational indices $\{1, 2, ..., n\}$ at each iteration, produces the joint distribution for the n locations. The warm-up – burn-in portion of the chain was 5000 iterations, after which every fifth map was selected, yielding a total of 1000 maps in the chain, and during each iteration all 400 quadrats were visited in a new random order. With regard to the speed of convergence of the chain (that is, its mixing rate), deterministic sweepings (that is, updates across a map) appear to be more efficient than random sweepings when strong spatial autocorrelation prevails, whereas the converse holds when weak spatial autocorrelation is present (Peskun, 1973). Random permutations can be viewed as a compromise between these two possibilities. The two estimation functions were used in order to check that a single set of estimates was obtained. The maximum pseudolikelihood and MCMC – maximum likelihood estimates of the pair of parameters contained in equation (2), where the locational mean is a constant (that is, $\alpha_i = \alpha$), appear in table 2. The conventional standard errors are those reported by SAS for PROC LOGISTIC, the MCMC simulation-based standard errors are those given by the variation of pseudolikelihood estimates for 100 MCMC simulations generated with the empirical pseudolikelihood parameter estimates, and the asymptotic standard errors are based upon the Fisher information matrix presented in Huffer and Wu (1998).

[4] Setting beginning probabilities to the empirical value of 143/400 has no noticeable effect on the outcome, because the Markov chain loses its memory of initial conditions as it converges on a steady state.

A spatial filtering analysis can follow two approaches here. The first is to execute a stepwise selection of eigenvectors from the entire set of eigenvectors associated with positive spatial autocorrelation. The second is to execute a stepwise selection while restricting attention to those eigenvectors depicting prominent levels of positive spatial autocorrelation. This first approach—which is the first threshold of the second approach—ultimately resulted in the selection of 36 eigenvectors (see table 4 below for details of the explicit selection results). Results from this approach include: 32 misclassified quadrats, $\hat{\phi} = 0.82644$, $z_{\mathrm{BB}} = 5.1$ and $z_{\mathrm{BW}} = -0.7$. This BB join-count statistic suggests some overcorrection for effects of the latent spatial autocorrelation; a fewer number of eigenvectors produces a z_{BB} value that is closer to 0 (see table 4). Based upon achieving the minimum residual z_{BB} value with the fewest eigenvectors, the second approach resulted in restricting attention to those eigenvectors whose associated MC exceeds half of the maximum possible MC value, or $\frac{1}{2}\mathrm{MC}_{\max} = 1.02337/2 = 0.51169$ (which is equivalent to $\mathrm{MC}/\mathrm{MC}_{\max} = 0.5$). This result suggests that the threshold value of $\mathrm{MC}_{\max}/4$ (which is equivalent to $\mathrm{MC}/\mathrm{MC}_{\max} = 0.25$) proposed in section 3.1 may be too liberal. Stepwise results produced by these two approaches respectively appear in tables 3 and 4. Eigenvectors are denoted by $\boldsymbol{E}_k$, where k is the rank order number, with $k = 1$ denoting that eigenvector having the largest MC value, through $k = 189$ denoting

Table 3. Stepwise filtered logistic regression results when candidate eigenvectors are all those associated with positive spatial autocorrelation.

Ordered, selected eigenvector	Moran coefficient	Likelihood ratio χ^2	$\hat{\phi}$	Number of misclassified quadrats	z_{BB}	z_{BW}
			0	143	13.0	−4.3
$\boldsymbol{E}_5$	0.99463	33.65	0.25954	126	11.1	−3.6
$\boldsymbol{E}_1$	1.02337	63.91	0.44286	98	8.9	−2.7
$\boldsymbol{E}_{17}$	0.88875	84.17	0.36083	111	9.4	−2.9
$\boldsymbol{E}_8$	0.95530	101.09	0.44976	97	8.1	−2.4
$\boldsymbol{E}_{22}$	0.84859	116.11	0.50464	88	6.2	−1.6
$\boldsymbol{E}_{41}$	0.71272	131.01	0.49910	89	7.2	−1.9
$\boldsymbol{E}_4$	1.00317	146.13	0.50041	89	6.6	−1.7
$\boldsymbol{E}_{16}$	0.90625	163.63	0.47304	94	9.1	−2.5
$\boldsymbol{E}_{30}$	0.78971	178.13	0.52826	85	6.2	−1.6
$\boldsymbol{E}_{45}$	0.69522	189.17	0.59014	74	5.8	−0.1
$\boldsymbol{E}_2$	1.02337	197.82	0.62084	69	4.5	−1.0
$\boldsymbol{E}_{10}$	0.94925	207.25	0.65310	63	3.4	−0.7
$\boldsymbol{E}_{47}$	0.66648	217.53	0.65310	63	5.0	−1.0
$\boldsymbol{E}_{18}$	0.88875	228.27	0.68106	58	3.0	−0.1
$\boldsymbol{E}_{116}$	0.28882	238.27	0.72637	50	3.8	−0.6
$\boldsymbol{E}_{181}$	0.04905	246.48	0.69731	55	3.9	−0.6
$\boldsymbol{E}_{140}$	0.19228	253.49	0.70836	53	3.8	−0.6
$\boldsymbol{E}_{186}$	0.02874	259.87	0.73153	49	2.3	−0.2
$\boldsymbol{E}_{48}$	0.66648	266.94	0.74072	47	3.1	−0.4
$\boldsymbol{E}_{26}$	0.82068	273.21	0.74152	47	3.4	−0.6
$\boldsymbol{E}_{31}$	0.78360	278.78	0.74657	46	4.3	−0.8
$\boldsymbol{E}_{33}$	0.76609	284.49	0.75225	45	3.1	−0.4
$\boldsymbol{E}_7$	0.97713	290.52	0.75754	44	3.0	−0.4
$\boldsymbol{E}_{13}$	0.91335	298.43	0.75257	45	3.1	−0.4
(and $\boldsymbol{E}_{31}$ removed)		(and reduced to 296.21)				

The remaining sequential additions are: $\boldsymbol{E}_{62}$, $\boldsymbol{E}_{81}$, $\boldsymbol{E}_{42}$, $\boldsymbol{E}_{84}$, $\boldsymbol{E}_{76}$, $\boldsymbol{E}_{119}$, $\boldsymbol{E}_{23}$, $\boldsymbol{E}_{176}$, $\boldsymbol{E}_{69}$, $\boldsymbol{E}_{102}$ (and $\boldsymbol{E}_{45}$ removed), $\boldsymbol{E}_{32}$, $\boldsymbol{E}_{52}$ (and $\boldsymbol{E}_{31}$ reentered), $\boldsymbol{E}_{20}$ (and $\boldsymbol{E}_{10}$, $\boldsymbol{E}_{32}$, $\boldsymbol{E}_{62}$ removed), $\boldsymbol{E}_{38}$, $\boldsymbol{E}_{100}$ (and $\boldsymbol{E}_{32}$ reentered).

Table 4. Stepwise filtered logistic regression results for eigenvector subsets selected using a χ^2 maximization criterion, a 0.15 stepwise entry level of significance, and a 0.10 stepwise removal level of significance.

Moran coefficient threshold	Number of candidate eigenvectors	Number of eigenvectors selected	$\hat{\phi}$	Number of misclassified quadrats	z_{BB}	z_{BW}	$\frac{1}{2}(\lvert z_{\mathrm{BB}}\rvert + \lvert z_{\mathrm{BW}}\rvert)$
0.00	189	36	0.82644	32	5.1	−0.7	2.900
0.10	162	**26**	**0.77654**	**41**	**0.9**	**0.03**	**0.465**
0.15	148						
0.20	136	25	0.75550	45	2.0	−0.2	1.100
0.25	123						
0.30	115	26	0.75298	45	1.0	0.03	0.515
0.35	103						
0.40	92						
0.45	84						
0.50	76	21	0.73622	48	1.5	−0.1	0.800
0.51669 = $\frac{1}{2}\mathrm{MC}_{\max}/2$[a]	74	20	0.74152	47	0.9	0.04	0.470
0.55	67						
0.60	57	22	0.73219	49	3.3	−0.5	1.900
0.65	49						
0.70	42	21	0.63874	66	5.4	−1.2	3.300
0.75	36	11	0.58759	74	4.6	−1.1	2.850

Note: the optimal solution appears in bold.

[a] Results appearing in table 1 relate to the MC (Moran coefficient) threshold value of 0.51669, but with the marginal eigenvector (for which MC = 0.55977) manually removed after stepwise selection.

that eigenvector having the positive MC value closest to zero. The likelihood ratio criterion (that is, a χ^2-statistic given by -2 times the difference between the maximized value of the log-likelihood function before and after a variable has been entered into a logistic regression equation) is optimized by conventional stepwise logistic regression. Nonmonotonicity of other sequential results appearing in table 3 (for example, step 3 inclusion of $\boldsymbol{E}_{17}$) suggests the need to formulate selection criteria other than maximization of the log-likelihood function value.

Similar to the numerous stopping rules available for stepwise variable selection in conventional multiple linear regression, other approaches could be adopted here. One criterion might be to minimize the residual-based term $\frac{1}{2}(\lvert z_{\mathrm{BB}}\rvert + \lvert z_{\mathrm{BW}}\rvert)$. This criterion would favor a MC threshold value of 0.10 in table 4, which was constructed by sequentially reducing the candidate set of eigenvectors by incrementally increasing the minimum eigenvector MC threshold value. Another criterion comparable to the adjusted-R^2-criterion of linear regression would be to minimize the quantity

$$\hat{\phi}^2 - \frac{\text{number of eigenvectors} - 1}{n - \text{number of eigenvectors} + 1}\,(1 - \hat{\phi}^2)\,.$$

A trade-off also could be made between the number of selected eigenvectors and the number of misclassified areal units. Regardless, the second of the two approaches implemented here seems preferable, because it seems to minimize the chance of overcorrection for spatial autocorrelation.

The final filtered equation is based upon the set of common eigenvectors selected by the two stepwise approaches. This logistic regression equation contains 19 eigenvectors, and renders the following statistics: 48 misclassified quadrats, $\hat{\phi} = 0.73622$,

$z_{\mathrm{BB}} = 2.5$ and $z_{\mathrm{BW}} = -0.3$. Although slight but detectable spatial autocorrelation remains, z_{BB} and z_{BW} are dramatically closer to 0 than are their corresponding unfiltered data values of 13.0 and -4.3; Gumpertz et al (1997) report a similar failure to account for all spatial autocorrelation.

4.2 Pseudolikelihood simulation findings

Maximum pseudolikelihood estimation of parameters for each of the two simulated spatial datasets produced interesting properties. Results for those data simulated using the filtered logistic regression equation (4) include: the average number of ones is 142.7, which essentially is the same as the count of 143 ones for the empirical data; a join-count statistic of $\overline{\mathrm{BB}} = 194.8$, which is greater than the observed value of 181 but falls just within its 95% confidence interval—the accompanying Shapiro–Wilk (SW) test statistic calculated for the simulated sampling distribution of BB indicates that it conforms to a normal distribution; and, parameter estimate means of $\hat{\rho} = 1.1784$ and $\hat{\alpha} = -2.4660$, with standard errors of $s_{\hat{\rho}} = 0.1038$ and $s_{\hat{\alpha}} = 0.2171$—all of these values are noticeably less than their empirical counterparts (see table 2), and neither simulated sampling distribution adequately conforms to a normal distribution (the respective SW values are 0.994 and 0.989, both of which are highly significantly different from 1). Inspection of quantile plots for the frequency distributions of the simulated parameter estimates suggest that they basically are symmetric, with deviations occurring in their tails.

Results for those data simulated by generating MCMC maps using the autologistic regression equation (2) include: the average number of ones is 145.6, which is very close to the count of 143 ones for the empirical data; a join-count statistic of $\overline{\mathrm{BB}} = 185.7$, which essentially is the same as the observed value of 181; and maximum pseudolikelihood estimated parameter estimate arithmetic averages of $\hat{\alpha} = -2.8428$ and $\hat{\rho} = 1.4302$, with standard errors of $s_{\hat{\alpha}} = 0.2793$ and $s_{\hat{\rho}} = 0.1638$—these parameter estimates are almost identical to the MCMC–maximum likelihood ones for the empirical data, which were used to generate the Markov chains, and these standard errors are very similar to those for the empirical maximum pseudolikelihood estimates (see table 2). Again, neither simulated sampling distribution conforms to a normal distribution (the respective SW values are 0.992 and 0.992, both of which are highly significantly different from 1).

4.3 Spatial filtering simulation findings

Spatial filtering results appear in table 1. Eigenvector coefficient estimates computed for maps generated with equation (4) are comparable with their empirical counterparts, although in every case the coefficients calculated with the simulated maps are further from 0 than are their empirical counterparts. In addition, the accompanying simulation-based standard errors are greater than their empirical counterparts, with few simulated sampling distributions adequately conforming to a normal distribution.

Eigenvector coefficient estimates computed for MCMC-generated maps from equation (2) fail to capture spatial autocorrelation effects here, although most of their sampling distributions do conform to a normal curve. This finding is attributable to the combinatorial nature of spatial autocorrelation: many map patterns can be associated with a single global measure of spatial autocorrelation (for example, see Boots and Tiefelsdorf, 2000). Given this feature of spatial autocorrelation, the 74 eigenvectors identified in table 4 (for MC = 0.51669) describing prominent degrees of positive spatial autocorrelation were used to account for autocorrelation in the MCMC simulated maps. Results for these logistic regressions appear in table 5. These results differ from those reported in table 1 because the restricted candidate set of eigenvectors here is determined with a lower MC value. On average, 26 eigenvectors contribute to

Table 5. Summary of individual spatial filtering results for 100 MCMC (Markov chain Monte Carlo) generated maps.

Term for which b was calculated	Number of maps for which $\|b/s_b\| > 2$	$\bar{\hat{\phi}}$	Term for which b was calculated	Number of maps for which $\|b/s_b\| > 2$	$\bar{\hat{\phi}}$
intercept	28	0.792786	$\boldsymbol{E}_{37}$	29	0.860387
$\boldsymbol{E}_{1}$	26	0.713942	$\boldsymbol{E}_{38}$	29	0.724235
$\boldsymbol{E}_{2}$	24	0.752116	$\boldsymbol{E}_{39}$	31	0.734536
$\boldsymbol{E}_{3}$	30	0.811987	$\boldsymbol{E}_{40}$	30	0.693760
$\boldsymbol{E}_{4}$	27	0.704758	$\boldsymbol{E}_{41}$	20	0.784933
$\boldsymbol{E}_{5}$	20	0.703752	$\boldsymbol{E}_{42}$	16	0.756664
$\boldsymbol{E}_{6}$	47	0.733497	$\boldsymbol{E}_{43}$	39	0.748211
$\boldsymbol{E}_{7}$	23	0.704521	$\boldsymbol{E}_{44}$	24	0.779282
$\boldsymbol{E}_{8}$	21	0.774862	$\boldsymbol{E}_{45}$	33	0.702431
$\boldsymbol{E}_{9}$	17	0.719281	$\boldsymbol{E}_{46}$	13	0.686990
$\boldsymbol{E}_{10}$	23	0.732252	$\boldsymbol{E}_{47}$	23	0.708872
$\boldsymbol{E}_{11}$	23	0.780618	$\boldsymbol{E}_{48}$	19	0.790692
$\boldsymbol{E}_{12}$	20	0.745951	$\boldsymbol{E}_{49}$	14	0.769425
$\boldsymbol{E}_{13}$	19	0.762103	$\boldsymbol{E}_{50}$	38	0.788857
$\boldsymbol{E}_{14}$	21	0.754092	$\boldsymbol{E}_{51}$	17	0.764384
$\boldsymbol{E}_{15}$	26	0.758813	$\boldsymbol{E}_{52}$	21	0.860477
$\boldsymbol{E}_{16}$	32	0.683522	$\boldsymbol{E}_{53}$	16	0.752492
$\boldsymbol{E}_{17}$	27	0.706157	$\boldsymbol{E}_{54}$	18	0.680707
$\boldsymbol{E}_{18}$	21	0.796314	$\boldsymbol{E}_{55}$	27	0.763202
$\boldsymbol{E}_{19}$	30	0.794667	$\boldsymbol{E}_{56}$	30	0.611233
$\boldsymbol{E}_{20}$	26	0.735911	$\boldsymbol{E}_{57}$	23	0.655087
$\boldsymbol{E}_{21}$	27	0.813860	$\boldsymbol{E}_{58}$	26	0.789154
$\boldsymbol{E}_{22}$	26	0.749108	$\boldsymbol{E}_{59}$	22	0.809770
$\boldsymbol{E}_{23}$	29	0.768074	$\boldsymbol{E}_{60}$	26	0.804842
$\boldsymbol{E}_{24}$	20	0.652206	$\boldsymbol{E}_{61}$	28	0.717241
$\boldsymbol{E}_{25}$	31	0.839641	$\boldsymbol{E}_{62}$	26	0.684714
$\boldsymbol{E}_{26}$	25	0.560267	$\boldsymbol{E}_{63}$	27	0.829588
$\boldsymbol{E}_{27}$	31	0.627427	$\boldsymbol{E}_{64}$	25	0.764829
$\boldsymbol{E}_{28}$	24	0.706283	$\boldsymbol{E}_{65}$	22	0.763858
$\boldsymbol{E}_{29}$	33	0.679712	$\boldsymbol{E}_{66}$	35	0.660345
$\boldsymbol{E}_{30}$	19	0.829540	$\boldsymbol{E}_{67}$	31	0.818690
$\boldsymbol{E}_{31}$	21	0.631433	$\boldsymbol{E}_{68}$	18	0.782466
$\boldsymbol{E}_{32}$	37	0.714919	$\boldsymbol{E}_{69}$	27	0.828827
$\boldsymbol{E}_{33}$	18	0.607428	$\boldsymbol{E}_{70}$	31	0.732010
$\boldsymbol{E}_{34}$	29	0.713226	$\boldsymbol{E}_{71}$	31	0.621491
$\boldsymbol{E}_{35}$	25	0.725714	$\boldsymbol{E}_{72}$	25	0.832277
$\boldsymbol{E}_{36}$	35	0.815517	$\boldsymbol{E}_{73}$	31	0.741804
			$\boldsymbol{E}_{74}$	39	0.840654

describing latent spatial autocorrelation; the range is 13 to 47, with a standard error of 6.4. The distribution of these counts conforms closely to a normal distribution (SW = 0.978). On average, these descriptions render $\hat{\phi} = 0.7512$; the range is 0.5 to 0.99, with a standard error of 0.0657. The simulated sampling distribution of this correlation coefficient deviates markedly from a normal distribution (SW = 0.996, which is significantly different from 1 at the 0.01 level). Moreover, different subsets of eigenvectors furnish a mean response description of latent spatial autocorrelation, depending upon the specific map pattern of zeros and ones.

5 Further assessment of the spatial filtering model specification

The predictive performance of the spatial filtering model specification also was evaluated through cross-validation, whose tabular results appear in table 6, and whose graphical results appear in figure 2. Both assessment tools suggest that probabilities generated by the spatial filtering logistic regression model are very good; accompanying statistics include: 62 misclassified quadrats and $\hat{\phi} = 0.65829$.

An important feature differentiating the spatial filtering [that is, equation (4)] and conventional autologistic [equation (2)] models is the number of parameters, which are $(K+1)$ and 2, respectively. The eigenfunctions of matrix expression (3) and matrix $\mathbf{C}$ are nearly identical, once the principal eigenvalue has been replaced by 0 and its corresponding eigenvector has been replaced by $(1/n^{1/2})\boldsymbol{1}$ (Griffith, 2000b). Meanwhile, the autoregressive term associated with equation (2) can be rewritten as follows:

$$\rho\mathbf{C}\boldsymbol{Y} = \rho\mathbf{E}\Lambda\mathbf{E}^{\mathrm{T}}\boldsymbol{Y} = \rho\mathbf{E}\Lambda\mathbf{I}\mathbf{E}^{\mathrm{T}}\boldsymbol{Y} = \rho\mathbf{E}\Lambda(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}}\boldsymbol{Y} = \rho\mathbf{E}\Lambda\boldsymbol{b} = \mathbf{E}\boldsymbol{b}^{*},$$

where $\mathbf{E}\Lambda\mathbf{E}^{\mathrm{T}}$ is the eigenfunction decomposition of matrix $\mathbf{C}$, and the vector $(\mathbf{E}^{\mathrm{T}}\mathbf{E})^{-1}\mathbf{E}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{b}$ is the familiar vector of linear regression coefficients. In other words, the two-parameter spatial autoregressive specification includes all of the eigenvector coefficients, regardless of their distances from 0. The spatial filtering specification simply explicitly removes those eigenvectors whose coefficients are unimportant.

This equivalency also relates to the amount of redundant information represented by spatial autocorrelation. When $\rho = 0$, the effective sample size—the comparable number of independent observations—is n. But, as ρ increases, the effective sample size decreases to 1. The number of eigenvectors appearing in equation (4) potentially is proportional to the accompanying reduction in effective sample size.

The role of the geographic neighbor structure also is of concern here. Popular first-order definitions are the 'rook's' case employed in this paper (see section 4), and the 'queen's' case ($c_{ij} = 1$ if two locations share either a zero or a nonzero-length common boundary, and $c_{ij} = 0$ otherwise). For a regular square tessellation forming a rectangular region (for example, the field plots analyzed in this paper), the eigenvectors are

Table 6. Classification results produced by the different estimation procedures.

Estimation technique	Pseudolikelihood		MCMC– maximum likelihood		Spatial filter			
					estimated model		cross-validation	
Predicted presence (1) or absence (0)	0	1	0	1	0	1	0	1
Actual absence: 0	241	16	209	48	237	20	231	26
Actual presence: 1	62	81	26	117	28	115	36	107
Percentage correctly classified	80.5		81.5		88.0		84.5	
Residual Moran coefficient	−0.274		−0.291		−0.023		−0.023	
Approximate z_{MC} score	−7.5		−8.0		−0.5		na	
Residual Geary ratio	1.299		1.317		1.048		1.050	
Approximate z_{GR} score	8.0		8.5		1.3		na	

Note: z-scores are approximate, computed using 10 000 random permutations of the residuals; results from case to case were consistent with theory.
na denotes 'not applicable'.
[a] MCMC—Markov chain Monte Carlo.

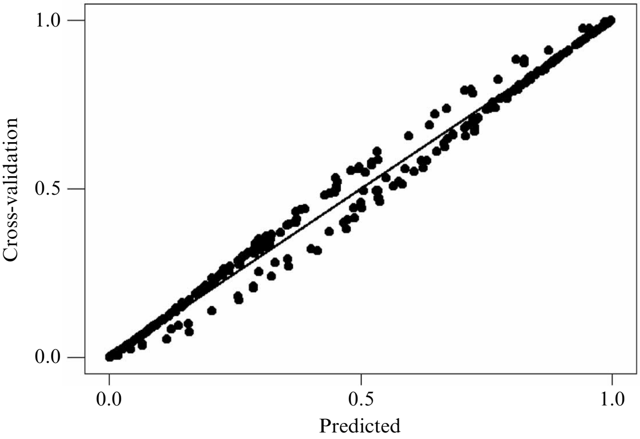

Figure 2. Scatterplot of the cross-validation probabilities versus the predicted probabilities from the fitted spatial filter logistic regression model.

analytically known and are the same for both of these neighbor definitions (Griffith, 2003). In addition, because

$$(\mathbf{I} - \rho\mathbf{C})^K = \mathbf{E}(\mathbf{I} - \rho\Lambda)^K\mathbf{E}^{\mathrm{T}} ,$$

and

$$\left(\mathbf{I} - \frac{\mathit{\mathbf{11}}^{\mathrm{T}}}{n}\right)\mathbf{C}^K\left(\mathbf{I} - \frac{\mathit{\mathbf{11}}^{\mathrm{T}}}{n}\right) = \left(\mathbf{I} - \frac{\mathit{\mathbf{11}}^{\mathrm{T}}}{n}\right)\mathbf{E}\Lambda^K\mathbf{E}^{\mathrm{T}}\left(\mathbf{I} - \frac{\mathit{\mathbf{11}}^{\mathrm{T}}}{n}\right) ,$$

if certain types of higher order autoregressive models are based upon powers of matrix **C**—for example, replacing a conditional spatial autoregression with a simultaneous spatial autoregression specification—then the eigenvectors remain the same. The spatial dependency adjustments that accompany these higher order models convert to changes in estimated regression coefficients for the spatial filter specification, and changes in weights applied to eigenvectors used to calculate MC values. In other words, more than likely different eigenvectors will be selected for constructing matrix $\mathbf{E}_k$. As with the first-order and second-order spatial autoregressive models, except for the special case of the principal eigenvector of matrix **C**, the eigenvectors of matrix $(\mathbf{I} - \mathit{\mathbf{11}}^{\mathrm{T}}/n)\mathbf{C}^{\mathrm{K}}(\mathbf{I} - \mathit{\mathbf{11}}^{\mathrm{T}}/n)$ asymptotically converge on those of matrix $(\mathbf{I} - \rho\mathbf{C})^K$. Therefore, a single spatial filtering specification competes with a wide range of autoregressive model specifications. But this result is lost when higher order autoregressive models are not simply powers of matrix **C**. Discussions of these cases are found in Anselin and Smirnov (1996), in Tiefelsdorf (2000, pages 32–33), and in Haining (2003, pages 79–87).

6 Conclusions

The spatial filtering methodology outlined in this paper furnishes an alternative pseudolikelihood procedure—correlated georeferenced binary data are assumed to be conditionally independent, given selected geographic weights matrix eigenvectors—that successfully captures spatial dependency effects in the mean response term of a logistic regression model, avoiding the complication of an intractable normalizing factor. It is easily implemented with standard logistic regression and PCA software. This filtering focuses on the particular form of positive spatial autocorrelation latent in a given georeferenced dataset. In contrast to findings of Huffer and Wu (1998), many of the simulated sampling distributions for estimated parameters inspected here failed to conform to a normal distribution (for the most part, they are symmetric with heavy tails).

Classification results for the particular empirical example examined here appear in table 6. Probabilities calculated with Besag's pseudolikelihood estimation procedure perform better when classifying the absence of disease. Probabilities calculated with MCMC – maximum likelihood estimation perform better when classifying the presence of disease. In contrast, probabilities calculated with the spatial filtering equation are as good as the pseudolikelihood-based ones for absence of disease, and are as good as the MCMC – maximum likelihood-based ones for presence of disease. The percentages of correctly classified field plots are roughly the same for the pseudolikelihood and MCMC – maximum likelihood results, but somewhat better for the spatial filter results. Further, the residuals (binary value minus predicted probability) produced by both the pseudolikelihood and MCMC – maximum likelihood results appear to contain non-negligible negative spatial autocorrelation. These levels of spatial autocorrelation are in contrast with the weak-to-moderate level (MC $= 0.473$) detected in the original binary data. In contrast, only a trace amount of spatial autocorrelation appears to remain in the spatial filter residuals. Finally, the maps of predicted probabilities and their accompanying residuals appear in figures 1(b) – 1(d). All three procedures capture the basic map pattern exhibited by the disease. Pseudolikelihood appears to do the poorest job of prediction in the core area of disease concentration, whereas spatial filtering appears to do the best job. The linear combination of 19 eigenvectors used in the spatial filtering contains moderate-to-strong spatial autocorrelation, having an MC value of 0.917. The map pattern portrayed by this composite variable can be used as a clue to search for missing substantive variables (for example, soil type, soil – water tension) that should be included in equation (4) as replacements for some or all of the selected eigenvectors; such covariates also would reduce the magnitude of $\hat{\rho}$ in the equation (2) specification. Similar findings are reported in Griffith (2003, pages 76 – 80, 115 – 116) for a spatial analysis of the presence – absence of West Nile virus by state (an irregular lattice) in the United States.

Results of this type of spatial filtering analysis offer useful insights into the spatial process under study. The empirical example analyzed here is the diffusion of a particular disease over a geographic landscape, a spatially autocorrelated phenomenon by its very nature. Based on the ϕ correlation coefficient, the pseudolikelihood results suggest that roughly 25% of the variance in presence – absence of the pepper plant disease is locationally redundant. MCMC results, which statistically are more efficient, suggest that about 33% is redundant information. Spatial filtering suggests that about 50% is redundant information. In addition, both pseudolikelihood and MCMC findings suggest that the number of ones on the map should be approximately 50% of the total (that is, $\hat{\alpha} \approx -2\hat{\beta}$). But the number of ones is only about 36%. Although MCMC indicates the presence of slightly stronger spatial autocorrelation, neither estimator has an upper bound. In contrast, with an $\mathrm{MC}/\mathrm{MC}_{\max}$ value of 0.896, the spatial filter specification suggests a well-structured pattern, which is visible in figure 1(a); and the index has an upper limit of 1. Finally, both the pseudolikelihood and MCMC procedures appear[5] to have overcorrected for spatial autocorrelation (see table 6); these methods are based upon $\rho\mathbf{C}$, which uses all the eigenvectors of matrix $\mathbf{C}$ (see section 5). Because spatial filtering involves a judicious selection of a subset of the eigenvectors of matrix $\mathbf{C}$, it enables overcorrection for spatial autocorrelation to be better controlled. In other words, the map appearing in figure 1(d) should better portray the spatial process outcome than do the maps in figures 1(b) and 1(c).

[5] Z-scores reported in table 6 are approximate. The spatial autocorrelation sampling distribution theory for residuals from a logistic regression remains undeveloped.

Findings reported here are important for spatial statistics because the autologistic specification for binary data naturally lends itself to geographic studies of disease and species. These findings also complement those reported in Griffith (2002) for the auto-Poisson model. The synthetic variate constructed as a linear combination of eigenvectors yields a map that furnishes clues to help search for missing substantive variables. In doing so, it serendipitously removes biasing effects of ignoring autocorrelation latent in georeferenced binary data. And, because each eigenvector depicts an orthogonal map pattern, with increasing fragmentation of the attribute surface as the associated eigenvalue decreases, articulating connections between spatial filtering and local statistics (for example, local indicators of spatial autocorrelation) and local models (for example, geographically weighted regression) should be a fruitful avenue for future research.

Acknowledgements. This research was supported by the National Science Foundation, research grant number BCS-9905213.

References

Albert P, McShane L, 1995, "A generalized estimating equations approach for spatially correlated binary data: applications to the analysis of neuroimaging data" *Biometrics* **51** 627 – 638

Anselin L, Smirnov O, 1996, "Efficient algorithms for constructing proper higher order spatial lag operators" *Journal of Regional Science* **36** 67 – 89

Augustin N, Mugglestone M, Buckland S, 1996, "An autologistic model for the spatial distribution of wildlife" *Journal of Applied Ecology* **33** 339 – 347

Bartlett M, 1975 *The Statistical Analysis of Spatial Pattern* (Chapman and Hall, London)

Bartlett M, 1978 *An Introduction to Stochastic Processes* 3rd edition (Cambridge University Press, Cambridge)

Bartolucci F, Besag J, 2002, "A recursive algorithm for Markov random fields" *Biometrika* **89** 724 – 730

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

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

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

Boots B, Tiefelsdorf M, 2000, "Global and local spatial autocorrelation in bounded regular tesellations" *Journal of Geographical Systems* **2** 319 – 348

Brownstein J, Holford T, Fish D, 2003, "A climate-based model predicts the spatial distribution of Lyme disease vector Ixodes scapularis in the United States" *Environmental Health Perspectives* **111** 1152 – 1157

Clark W A V, Hosking P L, 1986 *Statistical Methods for Geographers* (Wiley, New York)

Cliff A, Ord J, 1981 *Spatial Processes* (Pion, London)

Dubin R, 1995, "Estimating logit models with spatial dependence", in *New Directions in Spatial Econometrics* Eds L Anselin, R Florax (Springer, New York) pp 229 – 242

Dubin R, 1997, "A note on the estimation of spatial logit models" *Geographical Systems* **4** 181 – 193

Fox J, 1997 *Applied Regression Analysis, Linear Models, and Related Methods* (Sage, Thousand Oaks, CA)

Getis A, 1995, "Spatial filtering in a regression framework", in *New Directions in Spatial Econometrics* Eds K Anselin, R Florax (Springer, New York) pp 172 – 185

Getis A, Griffith D, 2001, "Comparative spatial filtering in regression analysis" *Geographical Analysis* **34** 130 – 140

Graham J, 1994, "Monte Carlo Markov chain likelihood ratio test and Wald test for binary spatial lattice data", mimeo, Department of Statistics, North Carolina State University, Raleigh, NC

Griffith D, 1984, "Measuring the arrangement property of a system of areal units generated by partitioning a planar surface", in *Recent Developments in Spatial Analysis: Methodology, Measurement, Models* Eds G Bahrenberg, M Fischer, P Nijkamp (Gower, Aldershot, Hants) pp 191 – 200

Griffith D, 2000a, "A linear regression solution to the spatial autocorrelation problem" *Journal of Geographical Systems* **2** 141 – 156

Griffith D, 2000b, "Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses" *Linear Algebra and Its Applications* **321** 95 – 112

Griffith D, 2002, "A spatial filtering specification for the auto-Poisson model" *Statistics and Probability Letters* **58** 245 – 251

Griffith D, 2003 *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding through Theory and Scientific Visualization* (Springer, Berlin)

Griffith D, Amrhein C, 1997 *Multivariate Statistical Analysis for Geographers* (Prentice-Hall, Englewood Cliffs, NJ)

Griffith D, Doyle P, Wheeler D, Johnson D, 1998, "A tale of two swaths: urban childhood blood-lead levels across Syracuse, New York" *Annals of the Association of American Geographers* **88** 640 – 665

Gumpertz M, Graham J, Ristaino J, 1997, "Autologistic model of spatial pattern of Phytophthora epidemic in bell pepper: effects of soil variables on disease presence" *Journal of Agricultural, Biological, and Environmental Statistics* **2** 131 – 156

Guyon X, 1995 *Random Fields on a Network: Modeling, Statistics and Applications* (Springer, Berlin)

Haining R, 1985, "The spatial structure of competition and equilibrium price dispersion" *Geographical Analysis* **17** 231 – 242

Haining R, 2003 *Spatial Data Analysis: Theory and Practice* (Cambridge University Press, New York)

Heagerty P, Lele S, 1998, "A composite likelihood approach to binary spatial data" *Journal of the American Statistical Association* **93** 1099 – 1111

Hoeting J, Leecaster M, Bowden D, 1999, "An improved model for spatially correlated binary responses", TR 9719, Department of Statistics, Colorado State University, Fort Collins, CO

Hosmer D, Lemeshow S, 2000 *Applied Logistic Regression* 2nd edition (John Wiley, New York)

Huffer F, Wu H, 1998, "Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species" *Biometrics* **54** 509 – 524

le Cessie S, van Houwelingen J, 1994, "Logistic regression for correlated binary data" *Applied Statistics* **43** 95 – 108

McCoy B, Wu T, 1973 *The Two-dimensional Ising Model* (Harvard University Press, Cambridge, MA)

Meier A, Nobel R, Rathbun S, 1993, "Population status and notes on the biology and behavior of the St. Croix ground lizard on Green Cay (St. Croix, U.S. Virgin Islands)" *Caribbean Journal of Science* **29** 147 – 152

Peskun P, 1973, "Optimum Monte Carlo sampling using Markov chains" *Biometrika* **60** 607 – 612

Pettitt A, Friel N, Reeves R, 2003, "Efficient calculation of the normalizing constant of the autologistic and related models on the cylinder and lattice" *Journal of the Royal Statistical Society, Series B* **65** 236 – 246

Switzer P, 2000, "Multiple simulation of spatial fields", in *Accuracy 2000* Eds G Heuvelink, M Lemmens (Delft University Press, Delft) pp 629 – 635

Tiefelsdorf M, 2000 *Modelling Spatial Processes — The Identification and Analysis of Spatial Relationships in Regression Residuals by Means of Moran's I* (Springer, Berlin)

Tiefelsdorf M, Boots B, 1995, "The exact distribution of Moran's *I*" *Environment and Planning A* **27** 985 – 999

Tjelmeland H, Besag J, 1998, "Markov random fields with higher-order interactions" *Scandinavian Journal of Statistics* **25** 415 – 433

Wrigley N, 1985 *Categorical Data Analysis for Geographers and Environmental Scientists* (Longman, Harlow, Essex)

Wu H, Huffer R, 1997, "Modeling the distribution of plant species using the autologistic regression model" *Environmental and Ecological Statistics* **4** 49 – 64

Appendix

Loss of eigenvector orthogonality in logistic regression

Suppose

$$\mathrm{P}(Y_i|\boldsymbol{X}_i) = p_i = \frac{\exp(\alpha+\boldsymbol{X}_i\boldsymbol{\beta})}{1+\exp(\alpha+\boldsymbol{X}_i\boldsymbol{\beta})}, \qquad i = 1, 2, ..., n,$$

where $\boldsymbol{X}_i$ is a $1\times K$ vector of K predictor variables. Let $\boldsymbol{Y}$ be the $n\times 1$ vector of observed binary 1/0 values. Then the log-likelihood function is given by

$$\boldsymbol{Y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}+\alpha\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{1}-\sum_{i=1}^{n}\ln[1+\exp(\alpha+\boldsymbol{X}_i\boldsymbol{\beta})]\,. \tag{A1}$$

Maximizing this function can be done efficiently with a quasi-Newton method, which approximates its appropriate second-order derivatives in terms of squared residuals.

The first term in equation (A1) reveals that some relief from multicollinearity problems may be realized when the predictor variables contained in matrix $\mathbf{X}$ are orthogonal. Maximization is with respect to the elements of vector $\boldsymbol{\beta}$ and scalar α, and orthogonality of matrix $\mathbf{X}$ results in an absence of covariation among the partial derivatives with respect to the individual β_j that would be attributable to the $\boldsymbol{Y}^{\mathrm{T}}\mathbf{X}$ term. When the mean of each X_j variable is 0, there also is an absence of covariation among the partial derivates that would be attributable to vector $\boldsymbol{1}$. But the partial derivative of the third term in expression (A1) with respect to β_j is

$$\frac{\sum_{i=1}^{n}x_{ij}\exp(\alpha+\boldsymbol{X}_i\boldsymbol{\beta})}{1+\exp(\alpha+\boldsymbol{X}_i\boldsymbol{\beta})},$$

resulting in covariations among $\ln X_j$ and X_k $(j\neq k)$. This covariation is present regardless of the orthogonality of matrix $\mathbf{X}$.

A more explicit illustration of how the orthogonality of matrix $\mathbf{X}$ is corrupted in logistic regression is furnished by considering an asymptotic equivalence to the maximization of equation (A1). The variance of population probabability parameter p_i is given by p_i $(1-p_i)$, indicating heterogeneous variance when p_i is not constant (that is, variation across the n observations). This feature of logistic regression is exploited in a second criterion that can be used for parameter estimation, namely minimizing the weighted sum of squares quantity

$$\sum_{i=1}^{n}\frac{(y_i-\hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)}, \tag{A2}$$

where p_i is given above. Expression (A2) reveals that orthogonality not only is lost through the nonlinearity of the logistic function [for example, through the third term in equation (A1)], but also through adjusting for heteroscedasticity with the weights $\hat{p}_i(1-\hat{p}_i)$, which are not included in the orthogonalization of the original predictor variables represented by matrix $\mathbf{X}$. The algebra of this situation can be studied in terms of iteratively reweighted least squares (for example, see Fox, 1997).

Therefore, if matrix $\mathbf{X}$ is the eigenvectors $\boldsymbol{E}_K$, collinearity complications are introduced because these eigenvectors retain their orthogonality only for simple linear models.