PREAMBLE

What we call progress is the exchange of one Nuisance for another Nuisance.

Havelock Ellis

A brief compendium of spatial regression model types is presented by Haining. In his discussion of this topic, "firmer" models are exchanged for "soft" models, "relative" mathematical space is exchanged for "absolute" space, and geo–referenced data complications are exchanged for traditional independent data complications. The purpose of this paper is to outline a class of models that seems to successfully handle redundant information contained in data arising from the locational positions of observations. What new idiosyncrasies, nuances, or subtleties of data will become problematic during analysis with the utilization of these alternate perspectives? A transcending issue beginning to emerge through the confusion surrounding development of spatial regression models appears to be even more fundamental than its old counterpart issue of multicollinearity that troubles traditional regression models. Dorean echoes many of these same sentiments, while raising questions concerning spatial regression model implementation.

The Editor

Models in Human Geography: Problems in Specifying, Estimating, and Validating Models for Spatial Data

Robert P. Haining*

Department of Geography, The University of Sheffield, Sheffield S10 2TN, England.

Overview: The use of the linear regression model for analysing relationships between a set of predictor variables and a response variable when the data refer to areal units, raises a number of distinctive issues. These issues include: specification of the regression model to allow for possible "spillover" effects; how to get good estimates of the spatial parameters. Further, there may be a separate set of concerns that derive from the nature of the data and the spatial distribution of certain types of values. The paper examines these problems and discusses ways of handling them. We conclude with two short examples.

1. Introduction

This paper is concerned with the problem of accounting for variation in some attribute (a response or dependent variable), in terms of a set of other attributes (explanatory, predictor or independent variables) where measurements are taken at "locations" (point sites distributed across a map or areas that partition a map).

For all its many perceived conceptual shortcomings as a model for variable relationships, the regression model is still widely used to treat questions of this type. Denoting the response variable as Y and the predictor variables as $X_1, \ldots, X_k$ the model is specified by a linear equation of the form

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \xi \tag{1.1}$$

where the βs are unknown parameters and the ξs are statistical errors (or disturbances). The data to fit equation (1.1) form an n–by–$(k+1)$ array, where n is the number of cases and $(y_i, x_{1,i}, \ldots, x_{k,i})$ is the vector of observations for case i. Given these data, equation (1.1) may be re–written as

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i} + \xi_i \quad (i = 1, \ldots, n), \tag{1.2}$$

where the ξs are assumed to be normally and independently distributed with mean zero and constant variance σ^2 $[\xi_i \sim NID(0, \sigma^2)]$.

If these assumptions are satisfied, least squares provides the best linear unbiased estimator for the unknown parameters. Given the data, let $\hat{\boldsymbol{\beta}}$ denote the least squares estimate for $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^T$, where the superscript T denotes the matrix operation of transpose. Then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y}), \tag{1.3}$$

* Part of the work for this paper was carried out while the author was in receipt of a Nuffield Social Science Research Fellowship.

where $\mathbf{X}$ is the n–by–$(k+1)$ matrix containing data on the predictor variables, with the i^th row given by $(1, x_{1,i}, \ldots, x_{k,i})$, and $\mathbf{Y}^T = (y_1, \ldots, y_n)$. Further, letting $\hat{\boldsymbol{\xi}} = \mathbf{e}$

$$\hat{\sigma}^2 = \mathbf{e}^T\mathbf{e}/(n-k-1), \tag{1.4}$$

where $\mathbf{e} = (e_1, \ldots, e_n)^T$ is the vector of least squares residuals, with $e_i = y_i - (\hat{\beta}_0 + \cdots + \hat{\beta}_k x_{k,i})$.

The variation accounted for by the linear combination of predictor variables is referred to as the explained variation, and the unallocated portion (associated with the residuals) is the unexplained variation, in variate Y. The steps that are followed in fitting a model such as equation (1) are:

(a) identification of a model (selection of predictor variables and specification of relationships),

(b) estimation of the unknown parameters,

(c) assessment of goodness–of–fit (residual analysis), and

(d) perhaps modification of the current model (new assumptions, data transformation), followed by a return to step (b).

This procedure cycles until a satisfactory model emerges, meaning a model where the percentage of explained variation is as high as possible and the residuals are well behaved in the sense of satisfying model assumptions. The initial model [at Step (a)] is often referred to as a "soft" model, which is made "firmer" by the cycles of fitting and model assessment. In addition to ensuring that the least squares assumptions are met (and applying remedial action if they are not), further problems may arise depending upon the nature of the data.

In this paper we examine the use of the regression model for describing relationships between variables measured across a set of locations on a map. In Section 2 we consider generalizations of equation (1.1) that reflect the spatial ordering of the data. Then in Section 3 we examine the implications for least squares parameter estimation. In Section 4 attributes of the spatial system that influence regression analysis are described. In particular we discuss how boundaries should be handled, the treatment of order relations between the set of locations and the influence of the surface partitioning. We also consider problems that arise when trying to assess the influence of individual cases and extreme values (outliers) on model fit. The last section briefly discusses two data sets in order to exemplify some of the issues raised in this paper.

2. Spatial Regression Models

The regression model defined by equations (1.1) and (1.2) disregards the geographic location of the n cases. Each case is treated as a distinct event. In specifying the regression model, the value of the response variable at any location is assumed only to be a function of values of the predictor variables at that *same* location (this accords with what sometimes is referred to as an "absolute" or "container" conceptualization of space). The location of each case only plays a significant role at the assessment stage of analysis. The errors in the regression model (1.2) are required, by assumption, to be independent. One aspect of model evaluation, therefore, consists of checking the residuals for evidence of pattern (spatial autocorrelation). Within an "absolutist" representation of space the presence of residual pattern is taken to

imply that important variables have been omitted, or that the functional relationship has been misspecified. In the former case new variables are sought that will eliminate the residual autocorrelation while in the latter case data transformations are used.

But space is not a series of separate, disconnected (independent) "boxes" or "containers," and the influence of events need not be restricted to the locations where they occur. The level of a response variable at a location may reflect the levels of predictor variables at other locations, and indeed a response variable at one location may act as a predictor variable for another location. Such considerations reflect the fact that events in space are not "parcelled-up." If these influences are present, then they may need to be taken into consideration when specifying a regression model. Several situations are presented next where such issues arise.

2.1. The spatially lagged response variable model.

In a lagged response variable model, the response variable at location j may act as a predictor variable at other locations. For example,

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + \rho \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ij} Y_j + \xi_i, \tag{2.1}$$

where ρ is an unknown parameter and the set $\{w_{ij}\}$ denotes a prior weighting scheme that may reflect, for instance, the distance between locations i and j. Often the influence of Y_j on Y_i is assumed to decrease as this distance increases.

Consider a set of retail sites scattered across a large urban area, with each site selling a more or less identical product. In such cases the market of each seller may be closely linked to neighboring markets, with the degree of interdependence lessening rather quickly with distance until this dependency becomes zero. The result of such local competitive interactions is a network of intricately interwoven markets—a "chain linking" (Chamberlin, 1956, p. 103) that is likely to be strongly influenced by the underlying movements of consumers within the city. If the retailers are gasoline retailers strung out along a road, for example, and if one retailer reduces his price, then it is probable that the nearest competitor also will drop his price. Such considerations lead to the specification of price models of the form given by equation (2.1), where $X_1, \ldots, X_k$ measure site effects (such as site quality, range of automotive services, brand type) and the set of weights $\{w_{ij}\}$ describes the structure of inter–site competition (Haining, 1986). The price charged at site i (namely Y_i) is in part dependent on prices charged in surrounding local retail sites, since the level of demand at site i is not only dependent on prices at i, but also on the level of prices at site i relative to prices at other sites with which i competes.

As an additional example, consider the problem of modeling variation in the total income accruing to the residents of a number of towns and cities distributed over a region. Let $\mathbf{Y}$ denote the vector of total income for the residents of the n places. Then

$$\mathbf{Y} = \mathbf{X} + \mathbf{C},$$

where $\mathbf{X}$ denotes the vector of exogenous income (deriving from export earnings, investment, government outlays), and $\mathbf{C}$ denotes the vector of endogenous income (local consumption by community residents). Assuming

$$\mathbf{C} = c\mathbf{Y}, \tag{2.2}$$

where the scalar c is the income creating local propensity to consume, then

$$\mathbf{Y} = (1 - c)^{-1}\mathbf{X}.$$

The vector $\mathbf{X}$ can be further decomposed into income earned from long distance (extra-regional) income transfers $(\mathbf{X}_1)$, and income earned from short distance (intra-regional) transfers $\mathbf{X}_2$). Thus

$$\mathbf{Y} = (1 - c)^{-1}(\mathbf{X}_1 + \mathbf{X}_2).$$

The vector $\mathbf{X}_2$ includes income accruing to each community arising from consumption expenditures by non-residents. Haining (1987) suggests that if equation (2.2) is reasonable in terms of intra-community consumption expenditure, then

$$\mathbf{X}_2 = \mathbf{\Omega Y},$$

where $\mathbf{\Omega}$ is an n-by-n matrix with diagonal values equal to zero, and non-negative off-diagonal entries that reflect the structure of inter-community movements for the purposes of purchasing consumer goods. Given a hierarchical ordering of the urban places, the non-zero elements of matrix $\mathbf{\Omega}$ can be specified using central place arguments. Accordingly the model becomes

$$\mathbf{Y} = (1 - c)^{-1}(\mathbf{X}_1 + \mathbf{\Omega Y}),$$

where again the response variable at location i may appear as a predictor variable at other locations. Unlike the price model, where interactions are reciprocal (Y_i is a function of Y_j **and** vice versa, so that matrix $\mathbf{W} = \{w_{ij}\}$ contains non-zero values above and below the main diagonal), central place principles suggest that interaction will be directional (consumers in low-order centers spending in high-order centers, but not vice versa), so that matrix $\mathbf{\Omega}$ will be an upper- (or lower-) triangular matrix.

2.2. The spatially lagged predictor variable model.

The simplest form of this model may be stated as

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + \tau \sum_{j=1}^{n} w_{ij} X_{r,j} + \xi_i, \tag{2.3}$$

where the set of weights $\{w_{ij}\}$ are as before, and τ is an unknown parameter. Variable X_r is usually a member of the set $\{X_1, \ldots, X_k\}$.

Models of this type have arisen in the study of the housing market, and in particular the modeling of spatial variation in house prices. House price is a function of structural characteristics of the house, the location of the house with respect to the city center, and characteristics of the area in which the house is situated (including environmental and demographic characteristics). Furthermore, depending upon the scale of the areal units, the characteristics of neighboring areas also may be significant. Hence, a house located in a desirable residential area that is adjoined by other desirable residential areas will tend to have a higher price than an equivalent house located in an equally desirable residential area, but where some of the adjoining residential areas are of lower status. Anas and Eum (1984, p. 105) remark that "the spillovers among neighboring and otherwise substitutable sub-markets can be taken into account to specify models in which market information from other submarkets becomes capitalized into housing prices." The estimate of each coefficient implicitly measures the price of that attribute.

2.3. The spatially correlated error model.

Equation (1.1) is modified here, yielding

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + \nu_i \tag{2.4}$$

where now ν_i are the statistical errors, with non–zero covariances $E(\nu_i, \nu_j) \neq 0$ for some i and j $(i \neq j)$. Therefore $E[\boldsymbol{\nu}\boldsymbol{\nu}^T] = \sigma^2\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is a matrix having some non–zero entries in its upper– and lower–triangles. Traditionally, in geography, matrix $\boldsymbol{\Sigma}$ has been modelled as a first–order simultaneous autoregressive scheme, such that

$$\boldsymbol{\Sigma} = [(\mathbf{I} - \rho\mathbf{W})^T(\mathbf{I} - \rho\mathbf{W})]^{-1},$$

where $\mathbf{I}$ is the n–by–n identity matrix, $\mathbf{W}$ is an n–by–n matrix of given weights $\{w_{ij}\}$ reflecting order relations on the map, and ρ is the (unknown) autoregressive parameter. However, many other models could be used.

In experimental situations (*e. g.*, agricultural uniformity trials), where the set of predictor variables and their levels are determined by the experiment, a correlated errors model is a natural choice when residuals are found to be correlated. In this context attention focuses on the choice of error model. In non–experimental situations the justification for this model is less clear cut, since the set of relevant predictor variables is not defined. In such cases it is usual to consider fitting a model such as equation (2.4), if the residuals are found to be correlated and if

(a) no further variables can be identified, or

(b) data are not available on other variables that might be significant, or

(c) adding further variables to the model does not remove this property of the residuals.

Residual correlation may be present because of the influence of a large number of variables that are difficult to specify, but that together display spatial persistence or continuity. Omission of variables representing these influences is responsible for the correlation detected in the residuals. Rather than attempt to model such influences, which might prove very difficult, a model such as equation (2.4) enables "safer" inference to be made with respect to those variables that can be included in the model. If such effects display smooth variation (such as trend), then replacing the error model by some order of polynomial trend surface may be preferable.

Loftin and Ward (1983) use a correlated errors model in examining the effects of population density on fertility rates in areas of Chicago. Such a modification apparently enables better estimates and safer inferences to be made on the influence of the included predictor variable. Further details are discussed in, for example, Cliff and Ord (1981), and Upton and Fingleton (1985).

3. Parameter Estimation

Estimation procedures for each of the three models presented in Section 2 will be discussed now, in turn.

3.1. The spatially lagged response variable model.

Using matrix notation, equation (2.1) may be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{Y} + \boldsymbol{\xi},$$

where matrix $\mathbf{W} = \{w_{ij}\}$. By re–arranging terms this expression becomes

$$(\mathbf{I} - \rho\mathbf{W})\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}, \tag{3.1}$$

and letting matrix $\mathbf{A} = (\mathbf{I} - \rho\mathbf{W})$ and matrix $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, substitutions into equations (1.3) and (1.4) yield

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\hat{\mathbf{A}}\mathbf{Y}), \tag{3.2}$$

$$\hat{\sigma}^2 = \mathbf{Y}^T\hat{\mathbf{A}}^T\mathbf{M}\hat{\mathbf{A}}\mathbf{Y}/(n-k-2), \tag{3.3}$$

where matrix $\hat{\mathbf{A}}$ denotes that the parameter ρ has been estimated (the problem of estimating ρ will be considered later), so that another degree of freedom is lost.

3.2. The spatially lagged predictor variable model.

Using matrix notation, equation (2.3) becomes

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \tau\mathbf{W}\mathbf{X}_r + \boldsymbol{\xi},$$

where vector $\mathbf{X}_r$ denotes one of the columns of matrix $\mathbf{X}$ (but not the first column). Again by re–arranging terms this equation may be expressed as

$$[\hat{\boldsymbol{\beta}}^T : \hat{\tau}]^T = (\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{Y}),$$

where the vertical dots symbol, :, denotes matrix partitioning, matrix $\mathbf{Z} = [\mathbf{X} : \mathbf{W}\mathbf{X}_r]$, and

$$\sigma^2 = \mathbf{e}^T\mathbf{e}/(n-k-2),$$

where $\mathbf{e} = \mathbf{Y} - (\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\tau}\mathbf{W}\mathbf{X}_r)$.

The regression parameters $\boldsymbol{\beta}$ and τ are estimated simultaneously. If variable $\mathbf{X}_r$ is spatially correlated, then $(\mathbf{X}^T\mathbf{X})^{-1}$ may be unstable (since in extreme cases vectors $\mathbf{X}_r$ and $\mathbf{W}\mathbf{X}_r$ may be linearly dependent, causing matrix $\mathbf{X}^T\mathbf{X}$ to be singular). This numerical problem produces inflated estimates of the parameter estimator variances, giving misleading or erroneous inferences.

3.3. The spatially correlated error model.

As noted in Section 2, a commonly used regression model with correlated errors is given by the pair of equations

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\nu} \\ \nu &= \rho\mathbf{W}\nu + \boldsymbol{\xi} \end{aligned} \tag{3.4}$$

so that $\boldsymbol{\nu} \sim MVN(0, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (\mathbf{A}^T\mathbf{A})^{-1}$. By substituting the second of these equations into the first one, and algebraically manipulating the result,

$$(\mathbf{I} - \rho\mathbf{W})\mathbf{Y} = (\mathbf{I} - \rho\mathbf{W})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi} \tag{3.5}$$

Then by substitution into equations (1.3) and (1.4),

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Y}), \tag{3.6}$$

$$\hat{\sigma}^2 = \mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{u}/(n-k-2), \tag{3.7}$$

where vector $\mathbf{u} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}^{-1} = (\hat{\mathbf{A}}^T\hat{\mathbf{A}}) = (\mathbf{I} - \hat{\rho}\mathbf{W})^T(\mathbf{I} - \hat{\rho}\mathbf{W})$.

Intuitively speaking, it appears that the general effect of including $\hat{\boldsymbol{\Sigma}}^{-1}$ in the estimate of $\hat{\boldsymbol{\beta}}$ is to downweight those observations with high spatial correlation. The presence of spatial correlation means that the information content of an observation is partially duplicated by those other observations (usually nearby) with which it is strongly correlated. Therefore a natural approach to this problem is to reduce the influence of such data duplication in the model fit. Those observations that are to be most strongly downweighted depend upon how matrix $\mathbf{W}$ is specified; often (because of the way matrix $\mathbf{W}$ is specified in applications) they tend to be observations associated with the highly connected interior sites of a spatial partition, particularly if these interior areas also are small relative to areal units closer to the boundary of the region.

3.4. Properties of the spatial parameter.

In the case of the lagged response and correlated error models, there is the additional spatial parameter, ρ, to be estimated. Estimation of this parameter could be avoided by evaluating the regression equation for different values of ρ contained in the permissible range which is, in fact, very restricted since $1/\eta_{\min} < \rho < 1/\eta\max$, where $\eta\max$ and $\eta_{\min}$ respectively are the largest and smallest eigenvalues of matrix $\mathbf{W}$. This identifies the sensitivity of $\hat{\boldsymbol{\beta}}$ to values of $\hat{\rho}$, and it is usually $\hat{\boldsymbol{\beta}}$ we are most interested in. As an extension of this strategy, a grid search can be conducted in order to find the minimum residual sum of squares for either equation (3.1) or equation (3.5).

Estimation of ρ in the case of the lagged response variable model is described in, amongst other sources, Upton and Fingleton (1985). The maximum likelihood estimate of ρ is obtained by minimizing (with respect to ρ)

$$(n/2) \star LN(\hat{\sigma}^2|\mathbf{A}|^{-2/n}), \tag{3.8}$$

where $\hat{\sigma}^2$ is given by equation (3.3), replacing the degrees of freedom value $n-k-2$ by the sample size n, and the pair of parallel lines, $|\bullet|$, denotes the determinant of a matrix. One

should note that, by substituting equation (3.3) into equation (3.8), the estimate of ρ does not depend upon $\boldsymbol{\beta}$, so that once $\hat{\rho}$ is obtained $\hat{\boldsymbol{\beta}}$ can be estimated. The estimation of ρ in the spatially correlated error model (with autoregressive errors) is more complicated. Again expression (3.8) is minimized, but $\hat{\sigma}^2$ is now given by equation (3.7), replacing $n-k-2$ by n, which depends upon $\hat{\boldsymbol{\beta}}$. A recommended estimation procedure here is to start with an initial estimate of $\boldsymbol{\beta}$ [*e. g.*, equation (1.3)], estimate from expression (3.8), then evaluate equation (3.6), and iteratively repeat these last two steps until convergence occurs. Mardia and Marshall (1984) discuss other possibilities, including computation of the standard error of $\hat{\rho}$. It is usually necessary to write a separate routine to estimate ρ, which includes evaluation of a matrix determinant. These problems currently limit the size of data sets that can be handled easily.

4. Regression Problems Associated with Spatial Data

In this section we examine a variety of problems encountered in the fitting of regression models that arise in one form or another from the spatial context of the data. The problems raised here fall into three generic groups, namely

(a) problems associated with representing the spatial distribution of observations—these usually call for modeling assumptions that cannot be directly tested;

(b) problems associated with the surface partitioning (in the case where observations are areal aggregates or densities) and which give rise to problems for least squares estimation; and,

(c) data problems that arise either as a consequence of the areal units system or independently of it.

4.1. The spatial distribution of observations.

The fitting of each of the models described in Sections 2 and 3 require an explicit, and largely *a priori*, representation of the areal units system to which the observations refer. This representation has two aspects to it: first defining order relationships between the sites or areas; and, second treating the boundary of the study area. For the most part any representation constitutes a set of modeling assumptions that are largely untestable (for instance, there is rarely sufficient information to estimate the elements of matrix $\mathbf{W}$). Since they are untestable, the sensitivity of results to different assumptions should be examined both as part of parameter estimation, and as assessment of fit (*e. g.*, inspecting whether or not the residuals are better behaved under one set of assumptions than under another).

4.1.1. Order relationships.

Order relationships across the map are specified by the matrix $\mathbf{W}$. This specification involves two separate decisions: (i) which sites or areas should be considered joined, and (ii) what weights should be attached to the joins. The first decision identifies which entries in matrix $\mathbf{W}$ are non–zero, whereas the second decision enables a particular value to be attached to each element w_{ij}. These issues are discussed at length in Cliff and Ord (1981). Table 1(a) identifies some of the main criteria used to select a join structure. The use of proximity criteria seems most appropriate where inter–site connections are not limited to special transport networks, whereas the use of interaction criteria seems most appropriate

where such a network does exist. Some level of flow existing between all pairs of sites or areas may necessitate introducing a cut off level, or alternatively an analysis of the system of flows, in order to identify the key linkages in the system (Holmes and Haggett, 1977). In addition, order relations might be hierarchical and directional [as on a central place lattice (Haining, 1987)] or discontinuous (for example, neighbors might be areas with stations on a railway track, if the analyst is looking at the spread of a rumor or an infectious disease). Table 1(b) identifies possible weighting schemes. Again one should note that these weighting schemes can be standardized by setting row sums to 1. This aspect of model specification clearly involves many *ad hoc* decisions, which implies that a need exists for assessing the sensitivity of results to *plausible* alternative definitions. The specification of matrix $\mathbf{W}$ has a direct effect on the fit of the model, since it enters into the estimation of $\boldsymbol{\beta}$ and σ^2 in all the models outlined in Sections 2 and 3, and in the case of the spatially correlated errors model it influences the relative downweighting of observations.

4.1.2. Boundary effects.

Boundary effects are likely to be more serious for the analysis of map data than for the analysis of time series data. In the time series case border effects are of order $1/n$, where n is the length of the observed series. In the case of an $n = N \star M$ rectangular lattice there are usually at least $2N + 2M - 4$ border sites, although the number of border sites depends on the specification of order relationships on the map.

Suppose the study area is not naturally bounded (for example it is a subarea of a much larger region). Here the analyst must consider how to model boundary effects. Consider the problem of modeling county death rates in Pennsylvania due to the spread of an infectious disease. The border counties of Pennsylvania will be influenced by death rates in immediately adjacent counties in New York, Ohio, West Virginia, Maryland, Delaware, and New Jersey. If these influences are simply ignored, the overlooked effects may distort the fit of the model. Border county residuals may be inflated (relative to non–border county residuals) because these external influences will be felt most strongly close to the border of the study area.

In the case of a model such as equation (2.1), if death rates are available in non Pennsylvanian counties, such values at the boundary could be included as additional exogenous predictor variables. Where such boundary information is not available, options include shrinking the study area (which would seem wasteful of data), or assuming values for the (non–observed) boundary counties. Fixed values could be assigned that preserved gradients at the boundary in some sense. Any selection would tend to be arbitrary, and the sensitivity of results to the choices made probably should be assessed.

Regardless of how boundary values are treated, the problem remains of estimating the spatial parameters in models such as equations (2.1) and (2.4). The issues are far from straightforward [see Ord (1981), Künsch (1983), Martin (1987), Griffith (1988)]. However, if the primary interest is in estimating the regression parameter vector $\boldsymbol{\beta}$ rather than the spatial parameter (ρ), then these problems may be rather less serious than they appear from the arguments put forward in the literature cited above. If some adjustment is thought necessary, since residuals might be larger for those cases in the study region close to the boundary, might not robust or resistant regression be considered? Alternatively, an *a priori* weighting might be considered in which observations close to the boundary are downweighted in the fit of the regression model. Different levels of downweighting could be tried and the

TABLE 1
SPECIFICATION OF **W**: DEFINING JOINS AND WEIGHTS

(a) Joins

Proximity:

i. Distance: each site/area is linked to all other sites/areas within a specified distance.

ii. Nearest neighbors: each site is linked to its k ($k = 1, 2, 3, \ldots$) nearest neighbor(s).

iii. Gabriel graphs: "any two sites A and B are said to be contiguous if and only if all other sites are outside the $A - B$ circle, that is the circle on whose circumference A and B are on opposite points" (Matula and Sokal, 1980).

iv. Delaunay triangulation: all sites that share a common border in a Dirichlet partitioning of the area are joined. Where the sites refer to areas that already partition the map, then the joins may be based upon whether the areas have a boundary in common.

Interaction:

All sites/areas between which there is a flow (measured, for example, by traffic movement, telephone calls, or person to person contact).

(b) Weights

Binary:

$w_{ij} = 1$ if areas i and j are joined; $w_{ij} = 0$ otherwise

Inverse distance:

$w_{ij} = d_{ij}^{-\gamma}$ $(\gamma > 0)$, where d_{ij} is the distance separating areas i to j

Exponential:

$w_{ij} = \exp(-d_{ij}^{\gamma})$

Boundary length:

$w_{ij} = (l_{ij}/l_i)^{\tau}$, where l_{ij} is the length of the common boundary between areas i and j, l_i is the perimeter of the border of area i, and τ is a constant.

Boundary and distance:

$w_{ij} = (l_{ij}/l_i)^{\tau} d_{ij}^{\gamma}$

results compared. As posed here, however, perhaps the most fundamental question concerns the sensitivity of $\hat{\boldsymbol{\beta}}$ to $\hat{\rho}$, and the influence of boundary assumptions on the estimation of ρ.

4.2. The surface partitioning.

Spatial data often refer to aggregate or density attributes of subareas into which the study area has been partitioned. The results of fitting a regression model to such data will be sensitive to the particular partition involved. Partitions that lump together individual micro–level units (*e. g.*, households), which are alike with respect to the important predictor variables, are generally considered better than those that lump together individual units that are unalike, simply because the level of the predictor variable then will be more representative of the area to which it refers. However such ideal partitions do not usually arise in practice, and there may be several "plausible" alternative partitions; ideally the sensitivity of results to these alternatives should be assessed.

Often the size of areal units within any surface partitioning varies with respect to either geometric size or total number of objects captured. If the response variable is a density measure (*e. g.*, with respect to the population size), then it is often the case that some density measures are taken with respect to subareas with large populations while others relate to subareas with small populations. If Y is a density variable derived from equally variable units, then we might expect

$$\mathrm{Var}(Y_i) = \sigma^2/n_i, \quad n_i > 0,$$

where n_i is a measure of the size (*e. g.*, total population) of areal unit i. This result is attributable to the law of large numbers, since the average is taken over more individual units. It follows that the errors are heteroscedastic:

$$\mathrm{Var}(\boldsymbol{\xi}_i) = \sigma^2/n_i,$$

and thus

$$E[\boldsymbol{\xi}\boldsymbol{\xi}^T] = \sigma^2\mathbf{P},$$

where $\mathbf{P}$ is a diagonal matrix with element $p_{ii} = 1/n_i$. The weighted least squares estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^T\mathbf{P}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{P}^{-1}\mathbf{Y}), \text{ and}$$
$$\hat{\sigma}^2_w = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w)^T\mathbf{P}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w)/(n-k-1).$$

Table 2 shows equivalent weighted estimators for the three models of Sections 2 and 3. The parameter ρ may be estimated as before. The determinant $|\mathbf{A}|$ is unchanged if $\mathbf{P}$ does not depend upon ρ, but $\hat{\sigma}^2$ is now given by $\hat{\sigma}^2_w$.

4.3. Data problems.

Certain data problems arise that often have nothing to do with the spatial nature of the data *per se*. These include such problems as multicollinearity (which renders parameter estimates unreliable), excessive numbers of predictor variables (which makes efficient analysis difficult), missing and unreliable values, and outliers. The latter group of problems relate to specific data values. However, even if these latter problems arise independently of the areal units system, how they are dealt with (*e. g.*, estimation of missing values, adjustments of the fit to the presence of extreme values) might be affected by where the problem observations are located on the map. If a missing value is near the boundary, for example, interpolation of its

TABLE 2
WEIGHTED LEAST SQUARES ESTIMATION RESULTS
FOR THE THREE REGRESSION MODELS

1. The regression model with spatially correlated errors:

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^T\hat{\mathbf{A}}^T\mathbf{P}^{-1}\hat{\mathbf{A}}\mathbf{X})^{-1}(\mathbf{X}^T\hat{\mathbf{A}}^T\mathbf{P}^{-1}\hat{\mathbf{A}}\mathbf{Y})$$
$$\hat{\sigma}_w^2 = \mathbf{u}^T\hat{\mathbf{A}}^T\mathbf{P}^{-1}\hat{\mathbf{A}}\mathbf{u}/(n-k-2)$$

2. The regression model with lagged response variable:

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^T\mathbf{P}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{P}^{-1}\hat{\mathbf{A}}\mathbf{Y})$$
$$\hat{\sigma}_w^2 = (\hat{\mathbf{A}}\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w)^T\mathbf{P}^{-1}(\hat{\mathbf{A}}\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w)/(n-k-2)$$

3. The regression model with lagged predictor variables:

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{Z}^T\mathbf{P}^{-1}\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{P}^{-1}\mathbf{Y})$$
$$\hat{\sigma}_w^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w - \hat{\tau}\mathbf{W}\mathbf{X}_\tau)^T\mathbf{P}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w - \hat{\tau}\mathbf{W}\mathbf{X}_\tau)/(n-k-2)$$

value might be more difficult than if it is near the center. Similarly, several missing values may be more difficult to estimate if they are clustered rather than scattered.

In regression analysis it is often of particular interest to assess the sensitivity of the model fit to individual cases, particularly cases with extreme values either in the response variable or in one or more of the predictor variables. In the standard regression model, individual cases can be deleted, one at a time, and the model refit. But with the models of Section 2, deletion of individual cases alters the order relationships between the sites, and creates internal boundaries within the study area if the cases refer to areal units. So procedures that are exact for the standard regression model no longer apply. Martin (1984) gives the estimator for vector $\boldsymbol{\beta}$, for the case of a general Gaussian correlated errors model, when one or more cases are (treated as) missing, and although implementation of the procedure taking each of the n cases in turn might be lengthy, his results can be used to develop an appropriate check on the sensitivity of $\hat{\boldsymbol{\beta}}$ to individual observations. The results also can be used to assess the sensitivity of estimates of $\hat{\boldsymbol{\beta}}$ to individual observations in the case of a general matrix $\boldsymbol{\Sigma}$.

The spatial distribution of extreme values may need to be considered, particularly whether they are scattered or clustered. Suppose the distribution of extreme regression residuals in equation (3.4) (defined by vector $\mathbf{u}$) is clustered. This might have a greater impact on the estimate of ρ than if the extreme values are scattered. A plot of the elements of vector $\mathbf{u}$ against the corresponding elements in vector $\mathbf{Wu}$ may highlight this problem. The potentially important consideration is the influence of the *distribution* of extreme values on the estimation of ρ, and thence on the estimation of $\boldsymbol{\beta}$ (since vector $\hat{\boldsymbol{\beta}}$ is a function of $\hat{\rho}$). Methods for resistant estimation of regression parameters by iterative downweighting of observations based on the frequency distribution of residuals are available [see Hoaglin,

Mosteller and Tukey (1983, 1985); see Besag (1981) for a spatial application]. Resistant estimators for $\boldsymbol{\beta}$ are of the general form

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{Q}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Q}\mathbf{Y}),$$

where matrix $\mathbf{Q}$ downweights observations with large residuals. The estimation is often performed by iterated weighted least squares, where the residuals at one iteration step are used to specify the elements of matrix $\mathbf{Q}$ at the next iteration step through a selected "weighting function." This procedure is distinguished from the weighting scheme discussed in Section 4 ("surface partitioning"), where the weights are specified *a priori*, as a function of areal unit attributes, and are not subsequently re–estimated. However, in the case of the regression models discussed in Sections 2 and 3, there is the further problem of acquiring resistant estimates of the spatial parameter ρ, a problem that has not as yet received much attention in the literature. Indeed many of the problems raised in this section have not yet received detailed consideration.

5. Two Case Studies

We conclude this discussion by briefly examining two applications that exemplify some of the points made in earlier sections of this paper.

5.1. Standardized Mortality Rates (SMRs) for areas of Glasgow.

Cancer data are available on SMR's for 87 community medicine areas (CMAs) in Glasgow (1981/82). An SMR is obtained for any area by dividing the observed number of deaths (O_i) by the expected number (E_i), given the age and sex composition of the area, and multiplying by 100. Data also are available on 15 relevant social and economic variables for the areas that are to act as predictors for the SMR data.

Scatterplots of the SMR values against the predictor variables suggest that the relationships are more linear and have better spread properties if the SMR data are subjected to a logarithmic transformation. Table 3 summarizes the fit of the best fitting model selected by a stepwise regression procedure.

TABLE 3
SUMMARY OF THE FIT OF THE REGRESSION MODEL
TO THE SMR DATA FOR GLASGOW

$$\mathrm{LN}(\widehat{\mathrm{SMR}}) = 4.23 + 0.014X_1 + 0.017X_2$$

$$R^2 = 67.9\%, \qquad \hat{\sigma} = 0.100$$

Values of t–statistics corresponding respectively to terms in the regression model containing variables X_1, X_2:

3.95, 13.07

X_1 = % of population that are pensioners living alone.

X_2 = % of population in Social Classes 4 and 5.

The residuals from the model show evidence of pattern with generally higher values near the center of the city, declining out towards the suburbs. A Moran test for residual autocorrelation, using a binary connectivity matrix **W** in which CMAs that share a common boundary are defined as being joined, leads to a rejection of the null hypothesis of no pattern, at the 10% level of significance. However, a correlated errors model has proven to be an unsatisfactory data descriptor in this example. The regression model was augmented with a second–order trend surface. The fit of this model is summarized in Table 4. Although the R^2 does not improve substantially, its increase is significant, and the residuals are better behaved (no residual autocorrelation is detected). The trend surface model peaks at the city center and declines towards the suburbs.

TABLE 4
SUMMARY OF THE FIT OF THE REGRESSION MODEL
WITH SECOND ORDER TREND COEFFICIENTS
TO THE SMR DATA FOR GLASGOW

$$\mathrm{LN}(\widehat{\mathrm{SMR}}) = 4.27 + 0.010X_1 + 0.017X_2 - 0.165X_E + 0.202X_N$$
$$-0.610X_E^2 - 0.710X_N^2 + 1.296X_EX_N$$
$$R^2 = 73.7\%, \qquad \hat{\sigma} = 0.093$$

Values of t–statistics corresponding respectively to terms in the regression model containing variables X_1, X_2, X_E, X_N, X_E^2, X_N^2, X_EX_N:

2.55, 12.27, −0.38, 0.65, −1.65, −2.35, 3.24

X_E and X_N are trend surface co–ordinates.

X_1 and X_2 are defined as in Table 3.

There is evidence in the residuals of an inverse relationship between residual variance and the observed number of deaths. Pocock *et al.* (1981) have argued that this should be expected, and have shown that observations should be weighted, with areas having a small number of observed deaths being downweighted. In the notation of Section 4, they suggest that

$$p_{ii} = 1 + 1/(\sigma^2 O_i). \tag{5.1}$$

In addition, high leverage values have been noted for three of the suburban CMAs, which are of large areal extent and hence, when represented by their centroids for purposes of fitting the trend surface component of the model, isolated from the rest of the map. Leverage effects in fitting trend surface models have been discussed in Unwin and Wrigley (1987). The best fit model derived from reanalyzing the data and downweighting observations using equation (5.1) are reported in Table 5. The effect of deleting the three CMA's with high leverages resulted in variable X_1 ceasing to be significant in the fit (failed to reject $H_0 : \beta_1 = 0$). Finally, Table 6 reports the results of using a resistant R–estimator (Li, 1985, p. 331) to fit the regression model. This uses a rank–based criterion. The table provides evidence of the resistance of the fit to the few large residuals.

TABLE 5
SUMMARY OF THE FIT OF THE REGRESSION MODEL
USING MATRIX **P** TO DOWNWEIGHT THE OBSERVATIONS

$$\mathrm{LN}(\widehat{\mathrm{SMR}}) = 4.26 + 0.009X_1 + 0.016X_2 - 0.142X_E + 0.242X_N$$
$$-0.588X_E^2 - 0.704X_N^2 + 1.210X_E \star X_N$$
$$\hat{\sigma} = 0.093$$

Values of t–statistics corresponding respectively to terms in the regression model containing variables X_1, X_2, X_E, X_N, X_E^2, X_N^2, $X_E \star X_N$:

$$2.49,\ \ 12.27,\ \ -0.32,\ \ 0.71,\ \ -1.57,\ \ -2.19,\ \ 3.03$$

X_E and X_N are defined in Table 4.

X_1 and X_2 are defined as in Table 3.

TABLE 6
R–RESISTANT REGRESSION

$$\mathrm{LN}(\widehat{\mathrm{SMR}}) = 4.34 + 0.009X_1 + 0.017X_2 - 0.375X_E + 0.135X_N$$
$$-0.461X_E^2 - 0.682X_N^2 + 1.373X_E \star X_N$$
$$R^2 = 46.0\%, \quad \hat{\sigma} = 0.087$$

Values of standard errors corresponding respectively to terms in the regression model containing variables X_1, X_2, X_E, X_N, X_E^2, X_N^2, $X_E \star X_N$:

$$0.003,\ \ 0.001,\ \ 0.412,\ \ 0.294,\ \ 0.347,\ \ 0.284,\ \ 0.375$$

X_E and X_N are defined in Table 4.

X_1 and X_2 are defined as in Table 3.

An interesting feature of this analysis is the presence of an "inner city" factor as an added risk factor that is in addition to the usual class and age variables. Such a factor might be associated either with the environmental characteristics of the inner cities or the characteristics of the inner city population (related to, for example, diet, exercise, higher levels of stress and overcrowding).

5.2. Agricultural consumption and accessibility.

Cliff and Ord (1981, pp. 209 and 237) report the results of an analysis of spatial variation in the percentage, in value terms, of the gross agricultural output of each county in Ireland consumed by itself (Y) as a function of a measure of county accessibility in terms of the arterial road network (X). Table 7 reports the results of the least squares regression fit. The residuals show evidence of spatial autocorrelation under three different order definitions

of matrix **W** (the binary matrix is constructed by setting $w_{ij} = 1$ if counties i and j share a common boundary, zero otherwise; the matrices are standardized by setting row sums to 1).

TABLE 7
ORDINARY LEAST SQUARES ANALYSIS OF IRISH DATA

$$\hat{Y} = -8.44 + 0.0053X_1$$

Values of t–statistics corresponding in order to the two terms in the regression model:

$$-2.65, \quad 7.42$$

$$R^2 = 69.7\%, \quad R^2(\text{adjusted}) = 68.4\%$$

$$n = 26, \quad \text{residual variance} = 13.58$$

Lag correlations computed for the residuals:

lag	0	1	2	3
correlation	1.000	0.387	−0.089	−0.218
number of pairs		58	93	94

Autocorrelation tests on the residuals:

Matrix type:	GMC	E[GMC]	standard deviation [GMC]	standard normal deviate
binary	1.726	−0.238	0.535	3.67
standardized binary	0.315	−0.057	0.126	2.95
weighted	0.429	−0.057	0.146	3.32

Source: Cliff and Ord, 1981, p. 230.

We consider the effects of adding spatially lagged forms of the original variables in order to deal with the problem of residual spatial autocorrelation. Added variable plots have been used in order to determine whether vector **WX** or vector **WY**, or both, should be added to the model, as well as what form of matrix **W** provides the best fit (Haining, 1990). The evidence of these plots, irrespective of how matrix **W** is constructed, is that adding vector **WY** is preferable to adding vector **WX**; if vector **WX** is added then vector **WY** also should be added, whereas if **WY** is added then **WX** is not needed.

Table 8 reports the results of regression model fitting with vector **WX** and then with **WY**. The evidence here confirms the superiority of including vector **WY** rather than **WX**. One should note that vectors **X** and **WX** are correlated, raising the problem of multicollinearity in the fit of the lagged predictor variable model. Table 9 reports the results of fitting a regression model with a spatially correlated errors model. Three error models have been tried: a simultaneous autoregressive (SAR) scheme, a moving average (MA) scheme (parameter θ), and a conditional autoregressive (CAR) scheme (parameter δ). Details of these models are summarized in Cliff and Ord (1981), Upton and Fingleton (1985) and

Ripley (1981). The first error model provides the best fit, although it is not quite as good as that obtained by the lagged response variable model. On the other hand, the spatial error model is probably more substantively justifiable than the lagged response variable model for this data set.

TABLE 8
FITTING DIFFERENT SPATIAL REGRESSION MODELS TO THE IRISH DATA

(a) Lagged predictor variable:

$$\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{X} + \tau\mathbf{WX} + \boldsymbol{\xi}$$

	Binary Matrix **W**	Standardized Binary Matrix **W**	Weighted Matrix **W** (Cliff & Ord, 1981, p. 230)
$\hat{\beta}_0$	−14.13	−23.97	−20.59
	(−3.55)	(−5.24)	(−4.22)
$\hat{\beta}_1$	0.0056	0.0026	0.0030
	(8.25)	(3.13)	(3.25)
$\hat{\tau}$	0.0002	0.0063	0.0050
	(2.15)	(4.05)	(3.02)
R^2	74.7%	82.3%	78.3%
adjusted R^2	72.5%	80.7%	76.4%
$r_{x,wx}$	0.06	0.57	0.58
error variance	11.80	8.28	10.16

(b) Lagged response variable:

$$\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{X} + \rho\mathbf{WY} + \boldsymbol{\xi}$$

	Standardized Binary Matrix **W**	Weighted Matrix **W** (Cliff & Ord, 1981, p. 230)
$\hat{\beta}_0$	−6.24	−6.71
	(−3.10)	(−3.21)
$\hat{\beta}_1$	0.0024	0.0028
	(4.38)	(5.11)
$\hat{\rho}$	0.731	0.646
	(6.38)	(5.13)
R^2	87.2%	86.2%
error variance	5.25	5.67

Figures in parentheses under the coefficient estimates are t–statistic values; $r_{x,wx}$ is the Pearson correlation between variable vectors **X** and **WX**; these results agree with those given in Anselin (1988) and Bivand (1984).

Finally we report the fitting of a robust form of the regression model with SAR errors (Table 10) using Tukey's biweight function to downweight the influence of large residuals (**e**) in the estimation of $\boldsymbol{\beta}$. Outliers also will affect the estimation of ρ, but after computing vector **u** at the first iteration and inspecting the plot of (**u**, **Wu**) a non–robust estimator

TABLE 9
FITTING A REGRESSION MODEL WITH SPATIALLY CORRELATED ERRORS TO THE IRISH DATA

(a) SAR errors model:

$$E[\mathbf{u}\mathbf{u}^{\mathbf{T}}] = \sigma^{\mathbf{2}}[(\mathbf{I} - \rho\mathbf{W})^{T}(\mathbf{I} - \rho\mathbf{W})]^{-1}$$

	Binary Matrix **W**	Standardized Binary Matrix **W**	Weighted Matrix **W** (Cliff & Ord, 1981, p. 230)
$\hat{\beta}_0$	1.155	4.670	1.359
	(0.33)	(1.04)	(0.36)
$\hat{\beta}_1$	0.0032	0.0024	0.0030
	(5.84)	(37.79)	(4.74)
$\hat{\rho}$	0.177 *	0.843 +	0.780 +
	(14.29)	(9.44)	(7.06)
R^2	87.0%	85.7%	86.1%
error variance	5.36	5.89	5.73

(b) Other spatial error models:

	Moving Average		Conditional
	Invertible Range	Unrestricted	Autoregressive
$\hat{\beta}_0$	−1.290	1.766	−3.725
$\hat{\beta}_1$	0.0037	0.0034	0.0041
$\hat{\theta}$	0.193	0.944	***
$\hat{\delta}$	***	***	0.184
R^2	78.3%	80.4%	81.9%
error variance	8.92	8.07	7.47

* denotes that the maximum value is 0.194.

+ denotes that the maximum value is 1.00.

Figures in parentheses under the coefficient estimates are t–statistic values.

for ρ was chosen that employed the usual expression (16) (indeed an R–estimator fit of **u** on **Wu** gives an estimate for ρ very close to the maximum likelihood estimate). The results reported here are for $B = cS$ where $c = 6$ and S is the median absolute deviation of the residuals (Li, 1985, p. 293). The downweighting is strongest for western counties. It happens that these are the counties for which X values are most unreliable because of the way the index was constructed (Cliff and Ord, 1981, p. 207).

TABLE 10
ROBUST ESTIMATION OF EQUATION (3.4) USING TUKEY'S BI-WEIGHT

$$\hat{Y} = 1.9941 + 0.0028X$$

$$\hat{\rho} = 0.794, \text{ error variance } (\hat{\sigma}^2) = 5.68, \; R^2 = 86.2\%$$

Final set of Tukey weights (in alphabetical order across rows)

0.989	0.962	0.918	0.866	0.971	0.980	0.959	0.908	0.997
0.972	0.999	0.821	0.991	0.970	0.999	0.829	0.847	0.999
0.990	0.873	0.893	0.928	0.990	0.973	0.970	0.999	

MAD = 1.715, B = 10.293

6. Conclusions

Social scientists are often accused of selecting models that derive too much from theory and too little from data properties. A "firm" model is specified on the basis of some theoretical argument (less kindly put, some "preconceived" idea) and attention then focuses on fitting the models and, at best, making comparisons with a small range of alternative models.

In developing "spatialized" forms of equation (1.1) in Section 1, the aim is to broaden the range of possible models that may be considered "soft" or "firm," depending upon the substantive context and the stage of data analysis reached. In discussing data related problems in Section 4, the aim was to draw attention to the sorts of fitting and assessment issues that often prove endemic to regression modeling with spatial data and to suggest some possible lines of treatment.

The development of interactive statistical packages with a range of graphical data inspection options should encourage closer inspection of data properties. Unfortunately, however, some of the procedures needed to fit the "extended" range of spatial regression models described here are yet to be made widely available in easy–to–use packages.

7. References

Anas, A., and S. Eum. (1984) Hedonic analysis of a housing market in disequilibrium. *Journal Urban Economics*, **15**: 87-106.

Anselin, L. (1988) Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis*, **88**: 1-17.

Besag, J. (1981) On resistant techniques and statistical analysis. *Biometrika*, **68**: 463-469.

Bivand, R. (1984) Regression modeling with spatial dependence: an application of some class selection and estimation methods. *Geographical Analysis*, **16**: 25-38.

Chamberlin, E. (1956) *The Theory of Monopolistic Competition*. London: Oxford University Press.

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

Griffith, D. (1988) A reply to R. Martin's 'Some comments on correction techniques for

boundary effects and missing value techniques'. *Geographical Analysis*, **20**: 70-75.

Haining, R. (1986) Intra urban retail price competition: corporate and neighbourhood aspects of spatial price variation. In *Spatial Pricing and Differentiated Markets*, edited by G. Norman, pp. 144-164. London: Pion, London Papers in Regional Science #16.

Haining, R. (1987) Small area aggregate income models; theory and methods with an application to urban and rural income data for Pennsylvania. *Regional Studies*, **21**: 519-530.

Haining, R. (1990) The use of added variable plots in regression modelling with spatial data. *The Professional Geographer*, (to appear).

Hoaglin, D., F. Mosteller, and J. W. Tukey (1983) *Understanding Robust and Exploratory Data Analysis*. New York: Wiley.

Hoaglin, D., F. Mosteller, and J. Tukey (eds.) (1985) *Exploring Data Tables, Trends and Shapes*. New York: Wiley.

Holmes, J., and P. Haggett (1977) Graph theory interpretation of flow matrices: a note on maximization procedures for identifying significant links. *Geographical Analysis*, **9**: 388-399.

Künsch, H. (1983) Approximations to the maximum likelihood equations for some Gaussian random fields. *Scandinavian Journal of Statistics*, **10**: 239-246.

Li, G. (1985) Robust regression. In *Exploring Data Tables, Trends and Shapes*, edited by Hoaglin, D., F. Mosteller, and J. Tukey, pp. 281-343. New York: Wiley.

Loftin, C., and S. Ward (1983) A spatial autocorrelation model of the effects of population density on fertility. *American Sociological Review*, **48**: 121-8.

Mardia, K., and R. Marshall (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, **71**: 135-146.

Martin, R. (1984) Exact maximum likelihood for incomplete data from a correlated Gaussian process. *Communications in Statistics A—Theory and Methods*, **13**: 1275-1288.

Martin, R. (1987) Some comments on correction techniques for boundary effects and missing value techniques. *Geographical Analysis*, **19**: 273-282.

Matula, D., and R. Sokal (1980) Properties of Gabriel graphs relevant to geographic variation and the clustering of points in the plane. *Geographical Analysis*, **12**: 205-222.

Ord, J. (1981) Towards a theory of spatial statistics: a comment. *Geographical Analysis*, **13**: 86-91.

Pocock, J., D. Cook, and A. Beresford (1981) Regression of area mortality rates on explanatory variables: what weighting is appropriate? *Applied Statistics*, **30**: 286-296.

Ripley, B. (1981) *Spatial Statistics*. New York: Wiley.

Unwin, D., and N. Wrigley (1987) Control point distribution in trend surface modelling revisited: an application of the concept of leverage. *Transactions of the Institute of British Geographers*, **12**: 147-160.

Upton, G., and B. Fingleton (1985) *Spatial Data Analysis by Example*, Volume 1: Point pattern and quantitative data. New York: Wiley.

DISCUSSION

"Models in human geography: Problems in specifying, estimating and validating models for spatial data"

by Robert P. Haining

Haining considers the following three kinds of model: (i) those with a spatially lagged response variable; (ii) those with a spatially lagged predictor variable; and, (iii) those with spatially correlated disturbance terms. Using his notion, with $\mathbf{A} = \mathbf{I} - \rho\mathbf{W}$, the presence of $\mathbf{ln}|\mathbf{A}|$ in the log–likelihood function complicates the estimation of the parameters for the first and third classes of models. With an emphasis on maximum likelihood methods, only a relatively small set of areas (making up a region) can be considered, as Haining notes. In turn, the smaller the number of cases (areas), the greater the vulnerability of the estimated regression parameters to problems stemming from the presence of high leverage data points and outliers. This vulnerability can be tackled on two fronts, namely (i) modifying software and data structures in order to analyze larger data sets, and (ii) paying close attention to diagnostic signals (of the presence of problems). Both are crucial. As the first strategy does little for small systems, it is necessary to take regression diagnostics very seriously and to look closely at robust regression methods. Haining's discussion of these issues for spatially distributed variables is particularly welcome.

Theories, models and statistical methods.

He notes that "social scientists are often accused of selecting models that derive too much from theory and too little from data properties." Excluding theorists who disavow any systematic examination of empirical evidence, the problem for social scientists is not that they choose models based on theories, but that they choose models from a superficial examination of data properties. This "examination" usually excludes consideration of the diagnostic procedures discussed by Haining. There is, however, another basis for the superficial consideration of data that is rooted in the modeling cycle outlined by him. As described, the process of moving from a "soft" model to a "firmer" model capitalizes on chance. It is not clear why making "the percentage of explained variation as high as possible" is of any real value in assessing the utility of an estimated model. Certainly, we must have well behaved residuals, but this criterion can be invoked without slavish adherence to the maxim of maximizing $\mathbf{R^2}$. A satisfactory model may "emerge" but, at most, it is a specification of a model that can be assessed with a different data set. Of course, such a model has a better chance of serving further tests if the problems discussed by Haining are addressed in a (modified–only modest attention given to $\mathbf{R^2}$) modeling cycle.

Specifying interdependence.

The specification of the weights matrix $\mathbf{W}$ is crucial. Leaving it out is problematic when linear models are specified and it is known that the data points are interdependent. Haining's Table 1 is helpful in laying out *some* of the possible specifications of $\mathbf{W}$. Many specifications of $\mathbf{W}$ in terms of joins and weights remain little more than guesses about the processes generating interdependencies. Until we know more about these processes, specifications of $\mathbf{W}$ will remain individual guesses or customary behavior. Even so, doing something with regard

to $\mathbf{W}$ for models where the disturbance terms are autocorrelated may be of some value for "safer" inference. For this class of models, the interdependence is a technical problem. However, for models with a spatially lagged response variable, the specification of $\mathbf{W}$ must be substantive, as it is an explicit part of the theoretical statement directly of interest.

Another important issue concerns the match, if any, between the level of aggregation of the data and the spatial scale of the phenomenon under study. Intuitively, it is unlikely that a social process with a spatial scale defined in terms of local neighborhoods can be captured in data aggregated to the ward, postal ZIP Code, or city levels. The larger units are likely to contain many diverse and distinct neighborhoods that have been grouped together in the (usually implicit) aggregation. At the other extreme, a process with a spatial scale at the county level will be modeled, at best, inefficiently with data assembled at the local neighborhood level.

Spatially lagged predictor variable models.

The simplest form, as stated by Haining, of such a model is as follows:

$$\mathbf{y}_i = \beta_0 + \beta_1 \mathbf{X}_{1,i} + \cdots + \beta_R \mathbf{X}_{R,i} + \tau \sum_{j=1}^{n} \mathbf{W}_{ij} \mathbf{X}_{\tau,j} + \epsilon_i.$$

Attention can be focused on the estimation of $[\beta : \tau]$. Haining writes, "if variable $\mathbf{X}_\tau$ is spatially correlated, the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ may be unstable (since in extreme cases vectors $\mathbf{X}_\tau$ and $\mathbf{W}\mathbf{X}_\tau$ may be linearly dependent, causing $\mathbf{X}'\mathbf{X}$ to be singular)." A *worse* situation would be an undiagnosed near singularity that would lead to inflated standard errors and compromised inference. (For an *exact* singularity, estimation would breakdown and diagnosis would be straightforward.) Why include both $\mathbf{X}_\tau$ and $\mathbf{W}\mathbf{X}_\tau$ in the specification of the model? If $\mathbf{X}_\tau$ is spatially autocorrelated, then $\mathbf{X}_{\tau,i}$, would be given by the weighted sum of the values of $\mathbf{X}_\tau$ for the areas, j, having non–zero $\mathbf{W}_{ij}$. The data value for $\mathbf{X}_{\tau,i}$ would include little new information beyond that contained in

$$\sum_{j=1}^{n} \mathbf{W}_{ij} \mathbf{X}_{\tau,j}$$

Use of $\mathbf{X}_\tau$ and $\mathbf{W}\mathbf{X}_\tau$ seems certain to generate collinearities and, as it is a specification problem, recourse to reduced rank methods seems premature. (Of course if $\mathbf{X}$ is not of full rank for a set of measured variables, then techniques like ridge regression *may* be of some value.)

The possibility of τ being a vector may merit further consideration. If one $\mathbf{X}_\tau$ is autocorrelated, then it is possible that other $\mathbf{X}$'s are autocorrelated, too. Further, each $\mathbf{X}_\tau$ may have its own $\mathbf{W}_\tau$ regime. This may be introducing an identification and specification nightmare, but there is no reason (other than simplicity) for assuming only one $\mathbf{X}_\tau$ is spatially autocorrelated. Coupled spatial processes with distinct autocorrelation regimes seem quite reasonable.

Data problems.

Haining is correct in directing our attention towards data problems. Missing and unreliable data values are a major problem, as are outliers. Influential data points can be included here also. In addition to the statistical problems discussed by Haining, there are many database management issues. Techniques for re–estimating a specified equation, when data points are dropped one at a time, must rest on an adequate database management system. Not only are $\mathbf{Y}$ and $\mathbf{X}$ changed when one (or more) observation(s) is(are) removed, but $\mathbf{W}$ also is changed. As Haining notes, deletion of a data point does create internal boundaries and it changes geometric relationships between areas. These are serious technical and substantive issues that can only be addressed if there is in place a sophisticated and flexible database management system that can handle dropped cases and the accompanying implied changes for matrix $\mathbf{W}$.

Haining's discussion uses a variety of alternative estimation procedures (*e. g.*, resistant regression methods) given a specific problem has been identified. This is useful, but many researchers will be left unsatisfied if they can neither implement nor have an adequate database management system to support such statistical procedures. It took a decade before regression diagnostics were widely available in the standard statistical software. As even fewer analysts grapple with interdependent data points, it may take even longer to have widely available software to support generalized autocorrelation modeling. Of course, if the importance of grappling with interdependent data points is recognized more widely, the (badly needed) software will become available for general use sooner. Haining has helped to push us in that direction.

Patrick Doreian, University of Pittsburgh