SPATIAL DATA ANALYSIS WITH GIS:

AN INTRODUCTION TO APPLICATION IN THE SOCIAL SCIENCES

by

Luc Anselin
National Center for Geographic Information and Analysis
University of California
Santa Barbara, CA 93106

Technical Report 92-10

August 1992

ACKNOWLEDGEMENTS

The research of which this paper is an outgrowth was supported in part by grants SES 87-21875, SES 89-21385 and SES 88-10917 (to the National Center for Geographic Information and Analysis) from the U.S. National Science Foundation, and by a grant from the Environmental Systems Research Institute, Inc. (ESRI). The data were collected by Sheri Hudak and the GIS operations were carried out by Rusty Dodson, both of whose assistance is greatly appreciated.

A revised version of this paper is forthcoming as a chapter in Geographic Information Systems: A Handbook for the Social Sciences, edited by Carville Earle, Leonard Hochberg and David W. Miller (Oxford, Basil Blackwell).

SPATIAL DATA ANALYSIS WITH GIS: AN INTRODUCTION TO APPLICATION IN THE SOCIAL SCIENCES

INTRODUCTION

What is Special About Spatial Data?

An attention to location, spatial interaction, spatial structure and spatial processes lies at the heart of research in several subdisciplines in the social sciences. Empirical studies in these fields routinely employ data for which locational attributes (the "where") are an important source of information. Such data typically consist of one or a few crosssections of observations for either micro-units, such as households, store sites, settlements, or for aggregate spatial units, such as electoral districts, counties, states or even countries. Observations such as these, for which the absolute location and/or relative positioning (spatial arrangement) is taken into account are referred to as spatial data. In the social sciences, they have been utilized in a wide range of studies, such as archeological investigations of ancient settlement patterns (e.g., in Whitley and Clark, 1985, and Kvamme, 1990), sociological and anthropological studies of social networks (e.g., in White et al., 1981, and Doreian et al., 1984), demographic analyses of geographical trends in mortality and fertility (e.g., in Cook and Pocock, 1983, and Loftin and Ward, 1983), and political models of spatial patterns in international conflict and cooperation (e.g., in O'Loughlin, 1985, and O'Loughlin and Anselin, 1991). Furthermore, in urban and regional economics and regional science, spatial data are at the core of the field and are studied to model the spatial structure for a range of socioeconomic variables, such as unemployment rates (Bronars and Jansen, 1987), household consumer demand (Case, 1991), and prices for gasoline (Haining, 1984) or housing (Dubin, 1992).

The locational attributes of spatial data (i.e., for the settlements, households, regions, etc.) are formally expressed by means of the geometric features of points, lines or areal units (polygons) in a plane, or, less frequently, on a surface. This spatial referencing of observations is also the salient feature of a Geographic Information System (GIS), which makes it a natural tool to aid in the analysis of spatial data. I return to this issue in more detail below.

The crucial role of location for spatial data, both in an absolute sense (coordinates) and in a relative sense (spatial arrangement, distance) has major implications for the way in which they should be treated in statistical analysis, as discussed in detail in Anselin (1990a). Indeed, location gives rise to two classes of so-called *spatial effects*, *spatial dependence*, and *spatial heterogeneity*. The first, often also referred to as spatial autocorrelation or spatial association, follows directly from Tobler's (1979) *First Law of Geography*, according to which "everything is related to everything else, but near things are more related than distant things." As a consequence, similar values for a variable will tend to occur in nearby locations, leading to spatial clusters. For example, a high crime neighborhood in an inner city will often be surrounded by other high crime areas, or a low income county in a remote region may be neighboring other low income counties. This spatial clustering implies that many samples of geographical data will no longer satisfy the usual statistical assumption of independence of observations.

A major consequence of the dependence in a spatial sample is that statistical inference will not be as efficient as for an independent sample of the same size. In other words, the dependence leads to a loss of information.[1] Roughly speaking, and everything else being the same, this will be reflected in larger, variances for estimates, lower significance levels in tests of hypotheses and a poorer fit for models estimated with data from dependent samples, compared to independent samples of the same size. I will refer to this aspect of spatial dependence in the rest of the paper as a nuisance. The loss in efficiency may be remedied by increasing the sample size or by designing a sampling scheme that spaces observations such that their interaction is negligible. Alternatively, it may be taken into account by means of specialized statistical methods. In this paper, I will focus on the latter. When spatial dependence is considered to be a nuisance, one only wants to make sure that the interpretation of the results of a statistical analysis are valid. One is thus not really interested in the source of the spatial association, i.e., in the form of the spatial interaction, the characteristics of the spatial structure, or the shape of the spatial and/or social processes that led to the dependence. When the latter is the main concern, I will use the term *substantive* spatial dependence instead.

The second type of spatial effect, *spatial heterogeneity*, pertains to the spatial or regional differentiation which follows from the intrinsic uniqueness of each location. This is a special case of the general problem of structural instability. As is well known, in order to draw conclusions with a degree of general validity from the study of a spatial sample, it is necessary that this sample represents some type of equilibrium. In the analysis of cross-sectional data in the social sciences this assumption is typically made. However, this assumption is considered with respect to the time dimension only, and systematic instability or structural variation that may be exhibited across different locations in space is mostly ignored. Such spatial heterogeneity may be evidenced in various aspects

[1] Note that these statements hold for positive spatial dependence only. Some authors have argued that in the case of negative spatial dependence the sample actually contains more information than an independent sample of the same size. The concept of spatial autocorrelation is further discussed in a later section.

of the statistical analysis: it may occur in the form of different distributions holding for spatial, subsets of the data, or more simply, in the form of different means, variances or other parameter values between the subsets. I will refer to discrete changes over the landscape, such as a difference in mean or variance between inner city and suburb, or between northern and southern states as *spatial regimes*, where each regime corresponds to a well-defined subset of locations. Alternatively, I will call a continuous variation with location *spatial drift*. This would be the case if the parameters of a distribution vary in a smooth fashion with location, for example, when their mean follows a polynomial expression in the x and y coordinates (this is referred to as a trend surface). As is the case for spatial dependence, spatial heterogeneity can also be considered either as a nuisance or as substantive heterogeneity.

Spatial Data Analysis

In Anselin and Griffith (1988), it is shown in some detail how the results of data analyses may become invalid if spatial dependence and/or spatial heterogeneity are ignored. Consequently, specialized techniques must be used instead of those that follow the standard assumptions of independence and homogeneity. By now, a large body of such techniques has been developed, which appears in the literature under the rubrics of spatial statistics, geostatistics, or spatial econometrics. The differences between these "fields" are subtle and to some extent semantic. Spatial statistics is typically considered to be the most general of the three, with geostatistics focused on applications in the physical (geological) sciences, and spatial econometrics finding application in economic modeling.

A useful taxonomy for spatial data analysis was recently suggested by Cressie (1991). He distinguishes between three broad classes of spatial data and identifies a set of specialized techniques for each. Crressie's taxonomy consists of *lattice data* (discrete variation over space, with observations associated with regular or irregular areal units), *geostatistical data* (observations associated with a continuous variation over space, typically in function of distance), and *point patterns* (occurrences of events at locations in space). In the remainder of this paper, I will focus exclusively on the first category (lattice data), due to space limitations, but also because I have found it to be the most appropriate perspective for applications in the social sciences that utilize GIS. I chose not to discuss geostatistics, since the requirement of continuous variation with distance in an isotropic space is typically not satisfied by spatial samples in the social sciences. Such samples are mostly limited to data for areal units, which are often defined in a rather arbitrary fashion, making an assumption of continuity tenuous at best. Recent reviews of geostatistical techniques can be found in Davis (1986), Isaaks and Srivastava (1989), Webster and Oliver (1990), and Cressie (1991). In contrast to the geostatistical data viewpoint, point patterns represent a very appropriate perspective for the study of many phenomena in the social sciences, such as the analysis of the spatial arrangement of settlements, of store locations, occurences of crime, infectuous diseases, etc. I elected not to discuss them in this paper because their study does not require much in terms of the functionality of a GIS, once the coordinates of the locations have been determined. A very readable introduction to point pattern analysis is given in Boots and Getis (1988) and Upton and Fingleton (1985). More advanced treatments can be found in Getis and Boots (1978), Ripley (1981) and Diggle (1983), as well as in Cressie (1991).

Unfortunately, the need for specialized spatial data analysis techniques is not commonly appreciated in empirical work, as illustrated by an analysis of the contents of recent journal issues in regional science and urban economics in Anselin and Hudak (1992).[2] Over 200 empirical articles were reviewed, of which slightly more than one fifth employed spatial data, roughly evenly divided between purely cross-sectional and pooled cross-section and time series data. Of those, only one considered spatial dependence in a rigorous fashion. This absence of a strong dissemination of the methodological findings to the practice of empirical research is often attributed to the lack of operational software for spatial data analysis, e.g., as argued in Haining (1989, p. 201). While this may have been the case in the past, several recent efforts have added features for spatial analysis to many existing statistical and econometric software packages, in the form of macros and special subroutines. A small number of dedicated spatial data analysis software packages have become available as well, which should greatly facilitate the use of these techniques by a wider range of social scientists.

GIS and Spatial Data Analysis

A linkage between GIS and spatial data analysis is considered to be an important aspect in the development of GIS into a research tool to explore and analyze spatial relationships. The limited availability of advanced analytical capabilities in commercial GIS packages is by now a familiar complaint in the research literature, going back to Goodchild (1987), and several calls for a closer integration between spatial analysis and GIS have been formulated (e.g., Openshaw, 1990). In the past few years, this has resulted in considerable research activity in this area, as evidenced by an increasing number of review articles, conceptual outlines, and guides for practical implementation of the Linkage, e.g., in Anselin and Getis (1992), Bailey (1992), Fischer and Nijkamp (1992), Goodchild et al. (1992), and Anselin, Dodson and Hudak (1992).

[2] The content analysis pertained to all articles that appeared over the three year period 1988-1991 in the Journal of Urban Economics, Regional Science and Urban Economics and the Journal of Regional Science.

Simply put, the power of a GIS as an aid in spatial data analysis lies in its georelational. data base structure, i.e., in the combination of value information and locational. information. The link between these two allows for the fast computation of various characteristics of the spatial arrangement of the data, such as the contiguity structure between observations, which are essential inputs into spatial data analysis. The GIS also provides a flexible means to "create new data," i.e., to transform data between different spatial scales of observation, and to carry out aggregation, partitioning, interpolation, overlay and buffering operations. Of course, such "data" is nothing but the result of computations, themselves based on particular algorithms that often use parameter estimates and model calibrations obtained by statistical means. The powerful display capabilities contained in a GIS also provide excellent tools for the visualization of the results of statistical analyses.

A Short Guide to the Literature

The treatment of spatial data analysis from the lattice data perspective focuses on two main issues: testing for the presence of spatial association, and the estimation of regression models that incorporate spatial effects. Most of the introductory level writings in the field only deal with the former. Examples are selected materials in textbooks on "statistics for geographers," such as in Ebdon (1985) and Griffith and Amrhein (1991), and the small pedagogic volumes devoted to the topic of "spatial autocorrelation" by Griffith (1987) and Odland (1988). A more technical treatment of these issues can be found in the classic works of Cliff and Ord (1973, 1981) and in Upton and Fingleton (1985). In addition to dealing with spatial autocorrelation, these texts also cover several aspects of spatial regression modeling.

A more specific focus on spatial effects in regression analysis can be found in Haining (1990) at the intermediate level, and in Anselin (1988a), Griffith (1988a), and also Cressie (1991) at the more advanced level. An extensive discussion of operational implementation issues, including extensive listings of software code, is given in Anselin and Hudak (1992), and Anselin and Griffith (1993).

A good overview of current research issues in spatial statistics can be found in a publication by the Panel on Spatial Statistics and Image Processing of the National Research Council (NRC, 1991). Other recent sources that include both methodological discussions and empirical applications relevant for spatial data analysis in the social sciences are the volume edited by Griffith (1990), and a special issue of the journal *Regional Science and Urban Economics* that is devoted to "Space and Applied Econometrics" (Anselin, 1992a).

Empirical Illustration

In order to present the technical materials in this paper in more concrete terms, I will illustrate several of the methods discussed in the following sections with a simple empirical example. I chose to present a partial replication of a study by Ormrod (1990) on the adoption of air conditioners and food freezers in the U.S. in the 1950s, which uses data for a cross-section of the 48 contiguous states.[3] Only the analysis of food freezers will be replicated. I selected this study for two main reasons. Firstly, it illustrates the type of empirical analysis where the focus of interest is on spatial context and spatial interaction, which is typical of many spatial investigations in the social sciences. Secondly, it demonstrates how the careful inclusion of the proper spatial variables into a model may provide important additional insights.

Ormrod's main thesis is that "local context," in the sense of "a number of state characteristics ... capable of influencing the local acceptance ... [is] significantly associated [with the rate of adoption of an innovation]" (Ormrod, 1990, p. 120). In other words, he stresses the importance of the receptiveness factor over the flow of information in determining the rate of adoption of the innovations in question. He argues that the universal knowledge and availability of these consumer appliances "allowed local contexts to become dominant in shaping the patterns of acceptance" (Ormrod, 1990, p. 120). Ormrod's study is based on a linear regression analysis, but he does not take into account spatial effects. He regresses, the degree of adoption of food freezers in each state (FREEZ) on population density (DENSITY), percent of population classified as rural farm population (RURAL), and median family income (INCOME), with all observations taken from the 1960 U.S. census. The values for these variables, reconstructed from the sources listed in the Ormrod article are listed in Table 1. In addition, Table I also includes the x and y coordinates for the centroid of each state (X and Y), and the values for two dummy variables (FREDUM and WEST), which were not used in the original study, but which are employed in the spatial analysis presented in this paper.

All computations were carried out by means of SpaceStat (Anselin, 1992b), a software package for spatial data analysis available from the National Center for Geographic Information and Analysis at Santa Barbara, CA. The figures were produced in the *Arc/Info* GIS (Environmental Systems Research Institute, Inc.).

[3] The Ormrod article does not include the data needed for this analysis, but does contain good sources so that the study could be easily replicated. All data used are from the 1962 County and City Data Book. For more details on the data, see Hudak (1992).

Outline of the Paper

In the remainder of the paper I review various aspects of spatial data analysis in some detail. I focus on those characteristics that are most relevant for applications in the social sciences. In addition, whenever possible, I outline how the data analysis can be carried out more effectively by making use of a GIS.

I start by further considering the integration of spatial data analysis with GIS. Next, I outline some important issues and concepts related to spatial arrangement and spatial scale, and specifically deal with the role of the spatial weights matrix. This is followed by a review of various measures of spatial autocorrelation, a brief taxonomy of spatial process models, and a discussion of spatial dependence as a nuisance and substantive spatial dependence, in the context of linear regression modeling. Next, I briefly go over several so-called spatial regression models that are particularly suited for use with a GIS. I close with an evaluation of the importance of spatial effects in the study of adoption rates of home freezers.

Some of the material in this paper is similar in nature to other introductions to various aspects of spatial data analysis that appeared earlier in Anselin and O'Loughlin (1992), O'Loughlin and Anselin (1992), and Anselin (1992c, 1992d). Parts have also been adopted from Anselin and Getis (1992) and Anselin, Dodson and Hudak (1992).

INTEGRATING SPATIAL DATA ANALYSIS WITH GIS

Spatial Analysis in GIS

Traditionally, geographic information systems are considered to perform four basic functions on spatial data: *input*, *storage*, *analysis* and *output* (Goodchild, 1987). Of these, analysis has so far received least attention in existing commercial systems. Typically, a variety of map description and manipulation functions are defined by commercial vendors as being "spatial analysis," but this has little to do with the usual interpretation of the concept in the academic world. There, the emphasis is on spatial modeling and statistical analysis, e.g., as illustrated in the review in Fischer and Nijkamp (1992).

The analysis function of a GIS was further divided into four components in Anselin and Getis (1992). These consist of *selection* (sampling of data from the data base), *manipulation* (partitioning, aggregation, overlay, buffering, and interpolation), *exploration* and *confirmation*. The first two of these are typical features of existing GIS, while the exploratory and confirmatory analysis functions are only beginning to be included, and only in a very limited way. In Anselin, Dodson and Hudak (1992), the first two components are referred to as a *GIS module* and the last two as a *data analysis module*, in order to reflect the typical division of labor in existing systems. This is also the perspective taken in much of the literature on the nature of the linkage between GIS and spatial analysis, which is often approached in terms of a linkage between two different software systems. A number of taxonomies of this integration have been suggested, such as a classification into close or loose coupling (Goodchild et al., 1992), or into encompassing and modular (Anselin and Getis, 1992).

Taxonomy of Linkages

A useful taxonomy of the linkage between GIS and spatial data analysis should be based on the difference in functionality needed in the process of spatial analysis, as argued in Anselin, Dodson and Hudak (1992). In practice, what counts is which types of information contained in the GIS module can be most effectively transferred to the spatial analysis module, and what kinds of results of the data analysis are most meaningfully moved to the GIS.

In Anselin, Dodson and Hudak (1992), a simple taxonomy was suggested, based on the number of times information is moved between the GIS and the data analysis modules. The first category in this classification is *one-directional* or *static integration*. In this form of linkage, there is no feedback between the modules. In other words, after an item is moved from the GIS to the data analysis module, or the other way around, there is no return necessary. This is the most common manner in which the linkage between the two has been implemented to date. For example, the interaction can originate in the data analysis, e.g., when the results of a non-spatial regression analysis, such as predicted values or residuals, are displayed by means of a GIS. However, no earlier transfer of information from the GIS to the data analysis has occured. Similarly, the interaction can originate in the GIS, e.g., when information on the spatial arrangement of the observations is transferred to the data analysis for the construction of spatial weights matrices and computation of global measures of spatial autocorrelation. Again, no return is needed, since these global measures can easily be reported in a simple table or bar chart, without a need to resort to the more sophisticated display capabilities of the GIS.

The second category in the classification is *bi-directional integration*. In this form of linkage, there is a two-way interaction between the GIS and the data analysis. Thus, after some information is moved from the GIS to the data analysis, something is returned to the GIS, or vice versa. For example, this will be the case whenever the results of a location- specific spatial statistic must be

displayed by means of the GIS. Before the spatial statistic can be computed, information on the spatial arrangement of the observations must first have been transferred from the GIS to the data analysis.

The final category is *dynamic integration*. In this form of linkage, information moves back and forth between the GIS and the data analysis. These multiple feedbacks are obtained from the interaction of the analyst with the results. In other words, this last category comes closest to a fully interactive spatial data analysis, combining simple data queries with exploratory and confirmatory analysis and visualization.

SPATIAL ARRANGEMENT AND SPATIAL SCALE

Spatial Weights Matrix

In addition to their absolute location, an important aspect of observations in spatial data analysis is their relative positioning, or spatial arrangement. An often unstated assumption is that this spatial arrangement is directly related to the interaction between units of observation. In other words, the spatial arrangement in and of itself is considered to be an important determinant of the spatial interaction revealed in measures of spatial autocorrelation or spatial association.

For each data point, a relevant "neigborhood" is defined as those locations surrounding it that are considered to interact with it. The values at those locations are thus expected to influence the observed values at the data point. The determination of the set of neighbors is not without some degree of arbitrariness. In a strict sense, it is the operationalization of the First Law of Geography mentioned earlier, and neighbors are those units that share a border (simple contiguity for areal units), or are within a given critical distance of each other (for point data). Consequently, two spatial units are either neighbors or are not, hence the use of the *term binary contiguity*. However, in a broader interpretation, the purely geometric properties (e.g., euclidean distance) are less important and the meaning of "contiguity" is generalized to any measure of potential spatial interaction between two units. Typically, this potential interaction is identified with the relative length of the common border, or specified in function of inverse distance (e.g., as in the gravity model of spatial interaction), or of a prior pattern of interaction (e.g., migration flows).

Formally, the membership of observations in the neighborhood set for each location is expressed by means of a square contiguity or spatial weights matrix (W), of dimension equal to the number of observations (N), in which each row and matching column correspond to an observation pair I,j. The elements w_{ij} of the weights matrix W take on a non-zero value (1 for a binary matrix, but any other positive value for general weights) when observations i and j are considered to be neighbors, and a zero value otherwise. By convention, the diagonal elements of the weights matrix, w_{ii}, are set to zero. Also, for ease of interpretation, the weights matrix is often standardized such that the elements of a row sum to one. In other words, each element w_{ij} of W is divided by its row sum, $\Sigma j\ w_{ij}$ (with the summation over all column elements j) to yield values between 0 and 1. Since the row sum in row i (for w_{ij}) and row j (for wji) are not necessarily the same, the resulting row-standardized weights matrix is likely to become asymmetric, even though the original matrix may have been symmetric. In the calculation of several spatial statistics, this complicates computational matters considerably.

In essence, the spatial weights matrix summarizes the topology of the data set in' graph theoretic terms (nodes and arcs). For a binary matrix (with 0-1 elements), higher order contiguity can be defined as well. This is done in a recursive manner, in the sense that a unit is considered to be contiguous of a higher order to a given location if it is first order contiguous to a unit that is contiguous of the next lower order (and not already contiguous of a lower order, to avoid circularity). In other words, units that are second order contiguous to a location must be first order contiguous to the first order contiguous ones. Higher order contiguity thus results in bands of observations around a given location being included in the neighborhood set, at increasing distances from the location. In the example data set of the U.S. states, both Nevada and Arizona are first order contiguous to California, since they share a common border. In turn, Utah is second order contiguous to California, since it borders Nevada, which is first order contiguous. However, Arizona, which also borders Nevada, is not second order contiguous, since it already is first order contiguous to California. For further details on the structure and characteristics of spatial weights matrices considered as graphs, see Cliff and Ord (198 1), Anselin (1988a), Griffith (1988a) and also Blommestein (1985).,

Obtaining Spatial Weights with a GIS

In practice, for applications with a small number of areal units of observation, simple contiguity can be found from a visual inspection of boundaries on a map, say a map of the 48 U.S. states. However, for most medium and large size data sets, e.g., of 100 or more spatial units, this becomes impractical, since it too easily leads to inaccuracies. In those instances, the only practical way to obtain the contiguity structure for the data is by means of a GIS. The way in which such information is contained in the GIS depends on the underlying data model and corresponding data structure, i.e., on the way in which the features of the continuous spatial reality are "discretisized" in order to be handled by computing devices and stored in the spatial data, base (Goodchild, 1992). Most existing commercial GIS emphasize either a vector or a raster data structure. In a vector-based GIS, the adjacency information is typically

stored explicitly and may be accessed by means of a simple query. For example, in *Arc/Info*, the left-polygon and right-polygon on an arc represent the polygon adjacency, and the length of the polygon border (perimeter) is automatically stored as an item in the feature attribute table for each polygon (ESRI, 1991a). This information can be transformed into a spatial weights matrix in a straightforward manner, as illustrated in Kehris (1990), Can (1992), Ding and Fotheringham (1992), and Anselin, Hudak and Dodson (1993), among others. On the other hand, in a raster data structure, adjacencies are stored implicitly and cannot be retrieved directly. In order to obtain this information for so-called raster polygons (i.e., contiguous groups of like-valued cells), the grid cell values of neighboring cells must be compared explicitly by means of a rather slow search process, as illustrated in Anselin, Hudak and Dodson (1993). Even when data are originally associated with points (e.g., settlements in archaeology), and not with areal units, they can be converted to the latter by means of a number of tessellations, such as Thiessen polygons.[4] The spatial arrangement of the resulting tiles can then be used to derive a contiguity matrix, although this may not always be the best measure of potential interaction.

When the spatial weights matrix must be computed by means of distance measures between points (e.g., for the distance-based indicators of spatial association discussed later), a meaningful point can be associated with each areal unit in a fairly easy way. Most vector-based GIS store a unique identifying point automatically with each polygon (e.g., a label point in *Arc/Info*). If the coordinates of this point would not be appropriate for the analysis at hand, other points (e.g., the locations of state capitals) can easily be introduced explicitly, or formal properties of each areal unit, such as a centroid, can be computed. The distance between "areal units" can then be obtained as a distance between their corresponding centroids or other meaningful points. For technical details on the practical implementation of these operations, see Anselin, Hudak and Dodson (1993).

The vast computational power of a GIS may easily create a false impression of precision in the derivation of the spatial weights for a data set. Digitizing errors may lead to the creation of spurious contiguities, and other GIS operations, such as the conversion between a raster and a vector format (and vice versa) affect the constructed weights matrix (see Anselin, Dodson and Hudak, 1992). Clearly, by means of a GIS, a large number of weights matrices can be derived for the same spatial layout or map. It should always be kept in mind that the choice of the weights matrix is an important determinant of the results of any spatial statistical analysis. In practice, it is therefore useful to experiment with a few different matrices and to assess the sensitivity of the statistical inference to the choice of the matrix.

To illustrate the concept of spatial weights, two forms of first order binary contiguity for the 48 U.S. states are listed in summary form in Table 2, as well as second and third order contiguity. For each state, there are four lines in the table. In the first, the sequence numbers are listed for the states that have their centroids within a critical distance of 6 digitizing units from the centroid of the state in question. Similarly, the second line contains the sequence numbers for the states that share a common border with it. The third and fourth lines show the second and third order contiguous states, in the same format.[5]

A closer look at Table 2 reveals quite a few differences between the two definitions of first order contiguity. These differences can also be quantified in the form of various measures of connectedness (see Anselin, 1988a). For example, it turns out that the distance-based matrix has a higher degree of connectivity: it has a higher percentage of nonzero weights (12.7% vs. 9.5%) and a higher average number of links per observation (6.0 vs. 4.5). The most connected state according to this definition is Pennsylvania (#36), with 12 state centroids within its critical distance band. In contrast, using shared borders as the criterion, Missouri (#23) and Tennessee (#40) are most connected, with each 8 first order contiguous states. Least connected states, with one "neighbor" each, are California (#4), Florida (#8) and Texas (#41), using the distance criterion, and Maine (#17), using the border criterion. The degree of connectivity increases in the second and third order contiguity matrices, as indicated by the percentage nonzero cells (respectively 15.6% and 19.0%) and by the average number of links per observation (respectively 7.3 and 8.9). The most connected state is Missouri (#23), in terms of second order contiguity, and Kansas (#14), in terms of third order contiguity. The least connected states for these two contiguities are respectively Maine (#17) and New Hampshire (#27).

Spatially Lagged Variable

An easy way to compare the value at a location to that of its "neighbors" (as defined by a spatial weights matrix) is to compute a so-called spatially lagged variable, or *spatial lag*. Such a spatial lag is a weighted average of the values in neighboring

---

[4] See Upton and Fingleton (1985), for an extensive discussion of tessellations in spatial data analysis.

[5] First order contiguity was derived from the information stored in *Arc/Info's* Arc Attribute Table, using Arc Macro Language (AML) commands for the digitized coverage of the 48 U.S. states used in all figures in this paper. The centroids were also computed by means of an *Arc/Info* AML routine. Note that neither the upper peninsula of Michigan nor Long Island were included in the computation of the centroids for Michigan and New York. A listing of the AML commands used for these computations may be found in Anselin, Hudak and Dodson (1993). The spatial weights matrices and higher order contiguities were computed in *SpaceStat*, using the information on adjacencies and state centroids from files created by the AML routines. Note that the actual value of the cut-off distance (6) is meaningless, since it is expressed in digitizing units. Of course, in a GIS, these digitizing units can be easily converted into longitude and latitude, and the distance between centroids could be expressed in miles or kilometers.

locations. More precisely, if an observation on a variable x at location i is represented by x_i, then its spatial lag is $\Sigma_j w_{ij}x_j$, i.e., the surn of the product of each observation in the data set (x_j) with its corresponding weight from the i-th row of the spatial weights matrix (w_{ij}). In matrix notation, the spatial lags for all observations can be expressed as the matrix product Wx (where x is a N by 1 vector containing all observations). When the spatial weights matrix is in row-standardized form, the spatial lag corresponds to a form of spatial smoothing, which facilitates its interpretation. A spatial lag is used in many tests for spatial association and is an essential element in the spatial process models discussed later in the paper.

The spatial lags for the observations on home freezer adoption in the 48 states are listed in Table 3, for both types of first order contiguity spatial weights (distance based and border based). Since the original contiguity weights were binary, all elements of each row in the row-standardized form will be equal to 1 divided by the number of contiguities. For example, Table 2 lists New Mexico (#29) and Utah (#42) as neighbors for Arizona (#2) according to the distance criterion (first line in the table for the second state). Using the corresponding row-standardized weights matrix gives a weight of 1/2 to the values of both states, which are respectively 23.4 and 26.0. Hence, the spatial lag for Arizona is found in the second column in of Table 3 as (23.4 + 26.0)/2 or 24.7. Note that California only has one neighbor using this criterion, i.e., Nevada (#26), since the centroids of California and respectively Oregon and Arizona are more than 6 digitizing units apart. As a result, the spatial lag for California is the value observed in Nevada, 2 1. 1, compared a value of 23.9 when an average of the three states is used (in the third column of Table 3). For row-standardized weights, the mean of the original variable and its spatial lag are the same (e.g., 22.0 for the FREEZ variable in the example). However, the spatial smoothing can have a considerable variance reduction effect. For example, the variance of 68.3 for the original freezer variable is reduced to 47.7 for the distance based spatial lag, and to 49.7 for the shared border spatial lag.

Using a procedure suggested in Anselin, Dodson and Hudak (1992), a bivariate map of the variable and its spatial lag can be constructed in a GIS to provide a visual indication of local non-stationarities in the pattern of spatial association. As illustrated in Figure 1, a circle with diameter proportional to the value of freezer adoption rate observed in a state is superimposed on the matching polygon. In each circle, the top part corresponds to the original variable, and the bottom part to the spatial lag. This simple graph provides an intuitive insight into the magnitude of values surrounding a spatial unit, relative to the observed value at the spatial unit. This is similar to the logic behind some statistics for spatial association, such as the Gi statistic discussed later in the paper. In this respect, a symbolic map such as Figure 1 is vastly superior to a choropleth map, since many familiar problems of data simplification are avoided (e.g., the choice of categories to represent continuous variables). Figure 1 reveals a strong spatial clustering of high adoption rates in the northern great plains states, such as Montana, Wyoming, North and South Dakota, and a cluster of very low adoption rates in the North East An almost horizontal dividing line between the two shares in the circle indicates a great similarity between the value in the state and the weighted average of the values in neighboring states. This strong spatial relationship is evident at the high end of the scale for South Dakota and Nebraska, among others, and at the low end of the scale for New Hampshire and Connecticut. A considerable inequality between the pie shares illustrates a pattern of negative autocorrelation, i.e., either large values surrounded by smaller ones, as for North Dakota and Idaho, or small values surrounded by larger ones, as for Illinois and New Jersey. A more rigorous assessment of these patterns must be carried out by means of the tests for spatial association discussed later in the paper.

The Problem of Spatial Scale

The data model and corresponding data structure in a GIS determine the spatial unit of observation that can be used for the statistical analysis. For example, in a raster data structure, the unit of analysis is the grid cell and all points within the grid are assumed to take on the same value. This is an implicit form of spatial sampling. Clearly, if the size and location of the grid cell do not match the spatial arrangement of values in the underlying spatial process, various types of misspecification may result. Similar problems follow from the choice of points, lines and polygons that will be represented in a vector data structure. It is important to keep in mind how this sampling process structures the database, and how it may affect the analysis that follows. In the social sciences, the choice of sampling unit is often not under the control of the analyst. Instead, it is typically determined by various administrative agencies or by nongovernmental data gathering organizations, whose criteria for delineating spatial units are not necessarily inspired by sound scientific principles.

Once the basic unit of analysis is chosen for the spatial data in the GIS, a wide range of manipulations can be carried out that may change the spatial scale, such as aggregation, overlay and interpolation. This may give the impression that GIS operations and the analysis associated with them are not sensitive to the choice of scale. However, this is clearly not the case, as documented in the literature on the so-called *modifiable areal unit problem* (MAUP), or *ecological fallacy* problem. Some well-known studies on this topic, e.g., by Openshaw and Taylor (1979) and Arbia (1989) have demonstrated how the correlative and autocorrelative relationship between' variables changes with different spatial scales. This is no surprise, since one could hardly expect that the process of aggregation would be neutral with respect to spatial scale. Since the results of many statistical tests will be partially affected by the chosen scale, its selection is an important decision to which a great deal of attention must be given. One would expect that substantive social theory would be a good guide in this respect, although in purely exploratory analyses there is no such guidance. The importance of the modifiable areal unit problem should not be exaggerated however. It often is not obvious whether this problem is indeed an

artifact of a particular data set, as is typically argued, or instead should be attributed to the use of an improper model and/or technique, as suggested by Tobler (1989).

SPATIAL AUTOCORRELATION

The Concept of Spatial Autocorrelation

A visual interpretation of the spatial clustering in freezer adoption, such as is portrayed for the northern and western great plains states in Figure 1 is insufficient to assess whether this pattern is indeed "significant," or instead is purely a random occurrence. Tests for spatial autocorrelation are designed to quantify the extent of clustering and to allow for statistical inference. A number of such statistics have been suggested. The common principle that underlies them is the comparison of the value of the statistic for a particular data set to its distribution under the null hypothesis of "no spatial autocorrelation." Such a null hypothesis implies that space does not matter, or, in other words, that the assignment of values to particular locations is irrelevant. Hence, it is only the values that provide information to the analyst, and "where" they occur does not add any insight. In contrast, under the alternative hypothesis of *spatial autocorrelation* (spatial dependence, spatial association), the interest focuses on instances where large values are systematically surrounded by other large values, or where small values are surrounded by other small values, or where large values are surrounded by small values (and vice versa). The former two are referred to as positive spatial autocorrelation, the latter as negative spatial autocorrelation. *Positive spatial autocorrelation* implies a spatial clustering of similar values, but *negative spatial autocorrelation* is a qualitatively very different concept, in that it implies a checkerboard pattern of values.[6]

Tests for spatial autocorrelation for a single variable in a cross-sectional data set (a single data layer or coverage) are based on the magnitude of an indicator that combines the value observed at each location with the values at neighboring locations (i.e., the spatial lags). Basically, the tests are measures of the similarity between association *in value* (covariance, correlation, or difference) and *association in space* (contiguity). Spatial autocorrelation is considered to be present when the statistic for a particular map pattern takes on an extreme value, compared to what would be expected under the null hypothesis of no spatial autocorrelation. Clearly, the exact interpretation of what is "extreme" will depend on the distribution of the test statistic under the null hypothesis, and on the chosen level of the Type I error, i.e., on the critical value for a given significance level.

There are three main approaches to determine the distribution of a test for spatial autocorrelation under the null hypothesis. The first, and most commonly taken assumption is that the data follow an uncorrelated *normal* distribution. Based on the properties of this distribution, the moments of the statistic under the null hypothesis can be derived analytically. Moreover, by resorting to an appropriate central limit theorem, the statistic itself can be shown to tend to a normal distribution as well. The second approach is non-parametric and exploits the interpretation of the null hypothesis as being non-spatial. In other words, each observation can be assumed to occur with equal probability at all locations. This approach is referred to as the *randomization* assumption. The moments of the statistic can often be derived analytically as wen, or, if this is not possible, they can be approximated to an acceptable degree of precision. They are typically different from the moments computed under the normality assumption. In most instances, the statistic under the randomization assumption also tends to a normal distribution asymptotically. The third approach is quite different. Rather than basing inference on the analytical properties of an asymptotic distribution, the data themselves are used to construct an artificial reference distribution. This is referred to as the *permutation* approach. Using the same rationale as in the randomization approach, all permutations of observed values over the locations should be equally likely under the null hypothesis of no spatial association. Thus, by resampling the data over the locations, i.e., by allocating the same set of observations randomly to the different locations, and by computing the statistic for each allocation, an artificial reference distribution can be created. The degree of "extremeness" of the statistic for the observed pattern can then be assessed by comparing it to the frequency distribution of the random permutations. A simple rule of thumb can be based on a so-called pseudo significance level. This is computed as (T+1)/(M+1), where T is the number of values in the reference distribution that are equal to or more extreme than the observed statistic, and M is the number of permutations that are carried out (typically, for ease of interpretation, M is taken to be 99 or 999). A small value of this pseudo significance level is an indication of an extreme pattern of clustering compared to what would be expected under the null hypothesis.

Traditional Measures of Spatial Autocorrelation

Probably the two most commonly used measures for spatial autocorrelation are *Moran's I* statistic (Moran, 1948) and *Geary's c* statistic (Geary, 1954).[7] These tests indicate the degree of spatial association as reflected in the data set as a whole. They both necessitate the choice of a spatial weights matrix. While Moran's I is based on cross products to measure value association, Geary's c employs squared differences.

---

[6] Strictly speaking, spatial dependence, spatial autocorrelation and spatial association are not identical concepts. However, for the purposes of this paper, I will not dwell on the distinctions between them (for a technical discussion, see Anselin, 1988a).

[7] See also Cliff and Ord (1973) for an extensive discussion of these statistics.

Formally, Moran's I for N observations on a variable x, with observation x_i at location i, is expressed as:

$$I = (N/S_0) \sum_i \sum_j w_{ij}(x_i-\mu)(x_j-\mu) / \sum_i (x_i-\mu)^2$$

where μ is the mean of the x variable, w_{ij} are the elements of the spatial weights matrix, and S_0 is a normalizing factor equal to the sum of the elements of the weights matrix:

$$S_0 = \sum_i \sum_j w_{ij}.$$

For a row-standardized spatial weights matrix, which is the preferred way to implement this test, the normalizing factor S_o equals N (since each row sums to 1), and the statistic simplifies to a ratio of a spatial cross product to a variance:

$$I^* = \sum_i \sum_j w_{ij}(x_i-\mu)(x_j-\mu) / \sum_i (x_i-\mu)^2$$

or, in matrix notation:

$$I^* = (x-\mu)'W(x-\mu) / (x-\mu)'(x-\mu).$$

Geary's c statistic is expressed as:

$$c == (N-1)/2S_0 \{ \sum_i \sum_j w_{ij}(x_i-x_j)^2 / \sum_i(x_i-\mu)^2$$

in the same notation as above.

Inference for the two statistics is based on one of the three assumptions outlined before: the normal assumption, the randomization assumption, or the permutation assumption. Instead of reporting the statistics as such, which are not easy to interpret, a standardized z-value is often given instead. This standardized z-value is obtained by subtracting the expected value for the statistic, according to one of the three assumptions, and dividing the result by the corresponding standard deviation. Under the normal and randomization assumptions, the resulting z-values can be compared to a table of standard normal variates to assess significance. Detailed expressions for the first two moments of the statistics, under the normal and randomization assumptions are given in Cliff and Ord (1973, 1981), and will not be reproduced here. When the permutation approach is used, a standardized z-value can also be computed, based on the mean and standard deviation of the reference distribution. This may facilitate the intuitive interpretation of the statistics, but it should not be thought of as a standard normal variate. The permutation approach is a useful option when the variables in the data set are clearly non-normal. However, since it only uses the observations at hand, it does not have the degree of generality associated with the other two approaches. A value of Moran's I that is larger than its theoretical mean of –1/N-1, or, equivalently, a positive z-value, points to positive spatial autocorrelation. In contrast, for Geary's c, positive spatial autocorrelation is indicated by a value smaller than its mean of 1, or by a negative z-value.

The results for Moran's I and Geary's c for the four variables in the example are listed in Tables 4 and 5. Each table shows the statistic with the associated z-value in parentheses below it. The statistics are computed for four different weights matrices, all in row-standardized form: the distance-based contiguity (DISTANCE_l in the table) and first to third order contiguity from the shared border criterion (CONTIG_1 through CONTIG_3).[8] The indication given for the two first order conguity matrices (DISTANCE_1 and CONTIG_1) is very similar and points to a high and significant degree of positive spatial autocorrelation for all four variables (all statistics are significant at $p<0.001$). For example, for FREEZ, Moran's I is 0.69 (z-value 7.62) with the distance based weights and 0.72 (z-value 7.66) for shared border contiguity, which confirms the intuitive impression given by the patterning in Figure 1. Similarly, Geary's c for the freezer variable is 0.29 (z-value -6.97) for the distance weights, which is virtually the same as 0.28 (z-value -7.03) for the border contiguity. Note that the sign of the z-values is positive for Moran's I and negative for Geary's c, while both indicate positive spatial autocorrelation.

[8] The z-values are based on the randomization assumption, since neither the rural nor the density variable turn out to follow a normal distribution, as indicated by a test in SpaceStat. For the other two variables, the normal assumption would be equally valid.

A listing of spatial autocorrelation statistics for increasing orders of contiguity, such as CONTIG_1 to CONTIG_3 in the tables is referred to as a spatial correlogram. This provides insight into the range over which the spatial clustering occurs. When certain (restrictive) assumptions are satisfied, a spatial correlograrn may also indicate the type of spatial process that is likely to have generated the spatial pattern. For example, a spatial autoregressive process (discussed in more detail in the next section) will be reflected in decreasing autocorrelation with higher order contiguities, often switching from positive to negative measures at high orders. Both FREEZ and DENSITY remain highly autocorrelated up to order 3 (although with slightly decreasing significance). For example, the z-values associated with Moran's I for FREEZ go from 7.66 to 6.39 down to 3.43, and those associated with Geary's c go from -7.03 to -5.91 down to -3.53. In contrast, the statistics for RURAL and INCOME are no longer significant beyond second order contiguity.

Generalized Measures of Spatial Association

Moran's I and Geary's c have been shown to be special cases of a general crossproduct statistic. Such an index is often referred to as a *gamma index* (Γ), and indicates the association between two matrices of similarity (or dissimilarity) for a set of objects. The gamma index is an example of combinatorial data analysis, as outlined in Hubert (1985, 1987), and applied to measures of spatial autocorrelation in Hubert et al. (1981, 1985), among others. The index is based on two (or more) matrices that show the similarity between items. In spatial data applications, one of the matrices will be a contiguity or distance matrix. Indeed, each element in such a matrix indicates the locational similarity between the object corresponding to the row (i) and the object corresponding to the column (j). Value association can be expressed in the same manner, such that each element in the matrix reflects the similarity between the object in the row and the object in the column. For example, such similarity may be formalized such that each element in the matrix, say b_{ij}, is the product of the two values, $b_{ij}=x_i x_j$, or its squared difference, $b_{ij}= (x_i-x_j)^2$, respectively as in Moran's I and Geary's c.

More precisely, the gamma index of association consists of the sum of all the cross products between the corresponding elements of each matrix:

$$\Gamma = \sum_i \sum_j a_{ij} b_{ij}$$

where a_{ij} and b_{ij} are the two measures of similarity between objects i and j, contained in the square matrices A and B, of dimension N by N. Clearly, a Moran-like index can be constructed by using a spatial weights matrix for the first and the product $x_i x_j$ for the elements of the second matrix. A similar approach can be taken for a wide range of measures of spatial association suggested in the literature, such as Tjostheim's index, join count and directional statistics (see Hubert 1987, for an overview). The significance of a Γ index can be assessed by means of an asymptotic approximation, or can be based on the permutation approach outlined before. The latter is the preferred strategy.

The value for the Γ statistic itself is not very meaningful, since it is scale dependent. Therefore, a large value for Γ does not necessarily mean strong association. Instead, a standardized z-value is often reported. A positive z-value indicates an association between the two matrices that is stronger than expected, i.e., that large values tend to occur in the same cells of both matrices. Negative z-values point to the reverse, i.e., an association between the elements of the two matrices that is less than expected.

A wide range of statistics for spatial association can be expressed in the form of a Γ index. One such test, which is useful to deal with categorical data, is the so-called join count statistic for binary variables (see Cliff and Ord, 1973, for details). In this test, the values of 1 and 0 are associated with two colors on a choropleth map (e.g., black and white). The join count is the number of times that a join or contiguity on the map corresponds to two like colors (or two different colors). Formally, a so-called blackblack or BB join count test can be represented as the product of x_i times x_j, where x is either 1 or 0. Only when the two observations both take on the value of 1 (B) will the product be one. Similarly, a so-called white-white or WW join count test for the clustering of the 0 (W) values can be included in form of a product of $(1-x_i)$ times $(1-x_j)$. Only when both x_i and x_j are 0 will the product equal one. This value similarity will be included in the join count only if the i-j pair also corresponds to a contiguity, i.e.., to a value of 1 for the element w_{ij} in a binary contiguity matrix.

The BB join count statistic can be expressed as:

$$BB = \sum_i \sum_j (x_i x_j) w_{ij}$$

which follows the general formulation of a gamma index. Similarly, a WW join count index can be expressed as:

$$WW = \sum_i \sum_j \{(1-x_i)(1-x_j)\} w_{ij}$$

Actually, both indices will equal two times the number of joins counted, since each contiguity is included twice (once as $_{wij}$ and once as w_{ji}). To illustrate this measure in the example, I created a dummy variable, FREDUM in Table 1, which takes on a value of 1 for freezer adoption rates above the median (21.0), and zero otherwise. The spatial pattern for this new variable is illustrated in Figure 2. The question now becomes whether the visual impression of strong spatial clustering in this pattern can be confirmed by a significant statistic for spatial autocorrelation. A BB join count test with the first order contiguity matrix based on shared borders yields a Γ index value of 80, with a corresponding z-value of 3.72, which is evidence of highly significant positive spatial autocorrelation. However, when the distance based contiguity is used as the spatial weights matrix, the index drops to 58, with a z-value of -0.96, which is clearly not significant. The discrepancy between the qualitative indication given by the two tests illustrates how sensitive the measures of spatial autocorrelation can be to the choice of the weights matrix. In this particular instance, the combination of the cut-off value for the binary variable (which was arbitrary in this example) and the longer range of the distance based contiguity measure results in more states with a different value for FREDUM being included as contiguous, and thus will tend to yield a lowerjoin count. A WW join count for the distance based contiguity yields a value of 140 (z-value of 5.61), which is highly sigrificant. This indicates a clustering of the lack of adoption. Note that in Moran's I and Geary's c clusters of both high and low values contribute to the indication of spatial autocorrelation, while this is not the case in the join count. The latter allows for a finer distinction between association of low and high values. Hence the apparent conflict between the join count results for the distance based weights and the results in Tables 4 and 5.

Distance-Based Measures of Spatial Association

A slightly different perspective on measuring spatial association is offered by a series of tests recently suggested by Getis and Ord (1992). Their *distance or G statistics* are computed by defining a set of neighbors for each location as those observations that fall within a critical distance (d) from the location. As mentioned earlier, a series of such spatial weights matrices can be constructed for different values of the critical distance. For notational purposes, the elements of the weights matrix are expressed in function of the critical distance, as $w_{ij}(d)$. One limitation of the distance-based measures is that they are only applicable to positive observations.

Getis and Ord suggest two types of statistics. The so-called G(d) statistic is similar to the traditional measures of spatial autocorrelation in that it assesses a global pattern of clustering, summarized into one value. Formally, the statistic, for a chosen critical distance d, G(d), is defined as:

$$G(d) = \sum_i \sum_j w_{ij}(d) x_i x_j \; / \; \sum_i \sum_j x_i x_j$$

where x_i is the value observed at location i, and $w_{ij}(d)$ stands for an element of the symmetric (unstandardized) spatial weights matrix for distance d. This statistic is similar to Moran's I in the numerator, but differs in the denominator. Its significance is assessed by means of a standardized z-value, obtained in the usual fashion. 'Me mean and variance for the G(d) statistic are computed under a randomization assumption and the zvalue can be shown to tend to a standard normal variate in the limit (see Getis and Ord, 1992, for detailed derivations).

The interpretation of the G(d) statistic is somewhat different from that of Moran's I. A positive and significant value indicates spatial clustering of high values, while a negative and significant value would point to a spatial clustering of small values. This contrasts with the interpretation of the traditional measures, where positive spatial autocorrelation refers to the clustering of either large or small values, and where negative spatial autocorrelation is a totally different concept.

The G statistics for spatial association for the variables in the example are reported in Table 6. The first line shows the values for the statistic (with z-values in parentheses below) using the same distance based contiguity matrix as in the other illustrations (but not row-standardized). For FREEZ, the z-value fo -2.30 is significant, but negative, indicating the spatial clustering of the lack of adoption, rather than that of the adoption of the innovation. This is in line with the findings for the BB and WW statistics above. In contrast, the z-value of 8.89 for DENSITY strongly points to the spatial clustering of high values. Neither RURAL nor INCOME show a significant pattern of association, conflicting with the indications given by Moran's I and Geary'c in Tables 4 and 5.[9] While not specifically designed with such applications in mind, a G statistic can also be computed for the usual contiguity matrix based on shared borders. The second line in Table 6 shows the resulting values for the variables in the example. For FREEZ, DENSITY and RURAL, a strong indication of positive spatial association is given, with z-values of respectively 3.55, 3.59, and 4.98. This is in

[9] A conflicting indication between Moran's I and Geary's c on the one hand, and the G(d) statistic on the other hand is not unusual, e.g., as illustrated in Anselin (1992c). For a more detailed discussion of this aspect of the tests, see Getis and Ord (1992).

accordance with the findings in Tables 4 and 5. However, for the INCOME variable, no such association can be found, since the z-value of -0.08 is clearly insignificant. Again, this apparent contradiction may be explained by the fact that the G statistic only values the spatial association of high or low values, but not of mixtures of the two.'

A second type of statistic suggested by Getis and Ord (1992) is a measure of spatial association for each individual spatial unit For each observation, i, the so-called Gi and Gi* statistics indicate the extent to which that location is surrounded by high values or low values for the variable under consideration. Again, this is associated with a particular distance band d. Formally, the $G_i(d)$ statistic for observation i and distance d can be expressed as:

$$G_i(d) = \sum_j w_{ij}(d)x_j / \sum_j x_j$$

with the summation in j exclusive of i. In other words, the value at the observation itself is not included in the statistic. This index is simply a ratio of the sum of the values in the surrounding locations to the sum of the values in the data set as a whole (exclusive of the location under consideration). In contrast, in the so-called G_j* statistic, the value at i is included:

$$G_i^*(d) = \sum_j w_{ij}(d)x_j / \sum_j x_j$$

where the summation is now over all locations, inclusive of i. The $G_i(d)$ statistic should be interpreted as a measure of clustering of like values around a location, irrespective of the value at that location. In contrast, the G_i*(d) statistic includes the value at the location within the measure of clustering and is thus more in accordance with the usual interpretation. The Gi*(d) statistic allows for the decomposition of a global measure of spatial association into its contributing factors, by location. It is thus particularly suitable to detect potential non-stationarities in a spatial data set, e.g., when the spatial clustering is concentrated in one subregion of the data only. A global measure of spatial association, such as Moran's 1, Geary's c, or the G(d) statistic, will fail to detect such a pattern.

Inference for the $G_i(d)$ and G_i*(d) statistics is again based on the computation of a standardized z-value, which asymptotically tends to a normal distribution. The expressions of the moments for the statistics are given in Getis and Ord (1992), and will not be reproduced here. The z-values for the G_i and G_i* indices computed for FREEZ are listed in Table 7. Similar to the interpretation of the global G statistics, a positive z indicates a spatial clustering of high values, while a negative z indicates a spatial clustering of low values. This is further illustrated in Figure 3, where states with significant G_i* statistics for FREEZ are identified by means of triangles, pointing upward for a positive z-value of the statistic and downward for a negative z-value. The three sizes of the triangles in Figure 3 correspond to significance levels of less than 0.001, less than 0.01 and less than 0.05 respectively.

The table and figure confirm the earlier findings that the global indications of spatial association are obtained from a mixture of strong clustering of the very high and the very low rates of adoption, the former in the western and northern plain states, the latter in the North East. For example, the top 5 states in terms of Gi* are South Dakota (z-value of 3.59), Minnesota (3.58), North Dakota (3.49), Montana (3.33) and Idaho (3.17), while the most negative ones are found in Pennsylvania (-3.81), New Jersey (3.78), New Hampshire (-3.73), Vermont (-3.73) and Rhode Island (-3.73). Since the pattern of adoption in the mid- and southwestern states is more scattered (see also Figure 1), the Gi* statistics for those states tend not to be significant.

Measuring Spatial Association with a GIS

The location-specific Gi and Gi* statistics form a good example of so-called twodirectional integration between a GIS module and a spatial data analysis module. In order to compute the statistics, a series of weights matrices must be constructed, based on point locations and the distances between them. This is easily carried out in a GIS, as outlined earlier in this paper. After the statistics are computed in the data analysis module, they can be graphically presented in the GIS, e.g., by means of triangle symbols, as in Figure 3. Moreover, the statistics for different distance bands can be combined on one map, thereby providing further insight into the sensitivity of the indices to the choice of the spatial range of influence.

This two-way interaction is not needed for the global measures of spatial association, such as Moran's I, Geary's c, or the G-statistic. While these indices still need to obtain information from the GIS in order to construct a spatial weights matrix, the results can easily be presented in the form of a table, simple graph or bar chart. The sophisticated graphics contained in the GIS can of course achieve this as well, but it is not crucial to the integration.

A more interesting integration between the GIS and spatial data analysis is the so-called *windowing* of measures of spatial assocation. As outlined in Anselin, Dodson and Hudak (1992), any number of measures of spatial autocorrelation can be computed for

a subset of the data, which is selected by moving a window or lasso over it in the GIS. In principle, this would allow for the instantaneous computation of several measures, for different weights matrices and different subsets of the data, which could be displayed simultaneously in windows on the screen. While a seamless implementation of this approach has not yet been achieved, in its current form it is an initial step towards a truly dynamic exploratory analysis of spatial data.[10]

SPATIAL PROCESS MODELS

Spatial Autoregressive and Spatial Moving Average Processes

The indication of a significant pattern of spatial clustering given by a test for spatial autocorrelation is only an initial step in the analysis of spatial data. Such an indication shows that the observations are more clustered than they would be under a random assignment, but it does not explain why such clustering occurs, nor which factors determine its shape and strength. In other words, the alternative hypothesis of "spatial autocorrelation" is too vague to be very useful in the construction of theory. A concept that formalizes the way in which the spatial association is generated, is that of a spatial process, or spatial stochastic process.[11] Roughly speaking, such a process expresses how observations at each location depend on values at neighboring locations, i.e., on the spatial lags.

Analogous to the terminology used in time series analysis (see, e.g., Box and Jenkins, 1976), spatial stochastic processes can be classified as *spatial autoregressive* (SAR) or *spatial moving average* (SMA) processes. The combination of the two, with an autoregressive process for the spatial association in the dependent variable, and a moving average process for the errors, is called a spatial autoregressive, moving average process (SARMA). For ease of interpretation, I will limit the discussion to linear processes, although spatial processes may be nonlinear as well.

In its simplest form, a spatial autoregressive process can be expressed in matrix notation as:

$$y = \rho W y + \varepsilon$$

where y is a N by 1 vector of observations on the dependent variable,[12] Wy is the corresponding spatial lag for weights matrix W, p is a spatial autoregressive parameter and ε is a N by 1 vector of error terms. For an individual observation at location i, the corresponding formulation is:

$$y_i = \rho(\sum_j w_{ij} y_j) + \varepsilon_i$$

In such a first-order (or pure) spatial autoregressive process, the observed magnitudes are simply a weighted average of neighboring values. When higher order contiguity is well defined, e.g., for binary contiguity on a regular lattice, a higher order spatial autoregressive process, say of order s, can be specified as:

$$y = \rho_1 W^{(1)} y + \rho_2 W^{(2)} y + \ldots + \rho_s W^{(s)} y + \varepsilon$$

where the $W^{(1)}$, $W^{(2)}$, ..., $W^{(s)}$ are the weights matrices corresponding to each order of contiguity and the ρ_1, ρ_2, ..., ρ_s are the matching autoregressive coefficients.

A first order spatial moving average process can be expressed as:

$$y = \lambda W \varepsilon + \varepsilon$$

---

[10] For further discussion of the use of dynamic graphics for spatial data analysis in a GIS, see also Haslett et al. (1990,1991).

[11] A technical treatment of this concept is beyond the scope of the current paper. For more details, see Anselin (1988a), Bennett (1979) and Cressie (199 1).

[12] Since there is no constant term in this expression, the y variables must be in deviations from the mean, in order to ensure that there is no systematic bias in the error term.

or, for location i, as:

$$y_i = \lambda(\Sigma_j\ w_{ij}\varepsilon_j) + \varepsilon_i$$

with λ as the moving average coefficient, and the other notation as before. In contrast with the SAR model, here the spatial lag pertains to the error term. A moving average process is often interpreted as the combination of an "innovation" or "shock" unique to each location, the ε_i, with a weighted average from neighboring locations, i.e., the term $\lambda(\Sigma_j\ w_{ij}\varepsilon_j)$. Similar to the autoregressive case, higher order spatial moving average models are easily specified as an extension.

Using the same notation as before, a mixed spatial autoregressive moving average process, say of order s in the SAR part and of order q in the SMA part, can then be specified as:

$$y = \rho_1 W^{(1)}y + \rho_2 W^{(2)}y + ...+ \rho_s W^{(s)}y + \lambda_q W^{(q)}\varepsilon + ...+ \lambda_2 W^{(2)}\varepsilon + \lambda_1 W^{(1)}\varepsilon + \varepsilon$$

In and of itself, a spatial process has no true explanatory power, since it only relates the observed value of a variable in each location to its magnitudes at other locations. In other words, a variable is "explained" by itself. In time series analysis, such stochastic processes are useful for forecasting, since they only utilize information on past occurrences for the variable of interest. There is thus no need to construct elaborate models that include other variables. Similarly, in spatial analysis, spatial process models may be used to produce interpolated values for locations that were not part of the original data set, e.g., to interpolate missing values, without having to resort to information other than the variable of interest.[13] However, the scope of application of this approach is much more limited than in time series analysis. A serious obstacle is the specification of the spatial weights matrix, which must either pertain to regular lattice units (grids), or must be expressed in function of an observable variable (e.g., distance) in order to be useful in the interpolation process.[14]

Spatial process models have been extended to incorporate the time dimension as well, in so-called space time autoregressive and moving average models, or STARMA. Their application has been primarily for forecasting purposes in meteorology and water resources modeling, but much less in the social sciences. Moreover, their implementation is limited by the difficulty to identify the nature of the process from autocorrelation and partial autocorrelation coefficients. For a more detailed review of these issues, see Bennett (1979), Hooper and Hewings (1981), Pfeifer and Deutsch (1980a, b, c) and Stoffer (1986), among others.

Mixed Regressive Spatial Autoregressive Processes

An alternative to the pure spatial process specification is the inclusion of “exogenous" explanatory variables in addition to the spatial lag terms. This is sometimes referred to as a *mixed regressive-spatial autoregressive* specification. In its simplest form, with only one order of contiguity used for the spatial lag terms, it is expressed in matrix notation as:

$$y = \rho Wy + X\beta + \varepsilon$$

where the X is a N by K matrix with observations on the K exogenous explanatory variables (including a constant term), with matching coefficient vector β (K by 1), and the other notation is as before. Such a model combines the spatial dependence in the form of a spatial lag term with the usual linear explanation of a dependent variable by a set of explanatory variables. It is thus more useful than the pure spatial process in terms of generating theoretical insights.

A mixed regressive-spatial autoregressive specification can be interpreted in three different ways. In a first perspective, the main focus is on the spatial dependence. One is interested in finding out how the variable y relates to its value in surrounding locations (the spatial lag), while controlling for the influence of other explanatory variables. This is often useful, since the impression of spatial autocorrelation as given by a simple univariate test may disappear when other controlling variables are introduced. This is the case

[13] For a review of the missing value problem in spatial data analysis, see, e.g., Bennett, Haining and Griffith (1984); Haining, Griffith and Bennett (1984); and Griffith, Bennett and Haining (1989).

[14] In geostatistics, the spatial dependence in a spatial process model is expressed in function of distance, hence its use as a predictive devise in the practice of kriging. See Cressie (1991) and the references listed in the introduction for more details.

when the spatial pattern in a dependent variable is actually due to the spatial pattern of other variables with which it is strongly correlated.

$$y - \rho Wy = X\beta + \varepsilon$$

or as:

$$(I-\rho W)y = X\beta + \varepsilon$$

A second interpretation is when the interest is really in -the relation between the explanatory variables X and the dependent variable, after the spatial effect has been controlled for. This is often referred to as spatial filtering or spatial screening (Getis, 1990), and can be formally expressed as:

where the matrix I-pW is referred to as a spatial filter, with I as an identity matrix of dimension N by N, and the other notation as before. In other words, once the spatial effect has been filtered out on the left hand side of the expression, the correct relationship with the other explanatory variables can be assessed.

Finally, the model can be viewed in its nonlinear form, as a description of the mean or expected values for the dependent variable in function of the exogenous explanatory variables (i.e., as a reduced form):

$$E[y] = E[(I-\rho W)^{-1}X\beta] + E[(I-\rho W)^{-1}\varepsilon]$$

or

$$E[y] = (I-\rho W)^{-1}X\beta$$

Here, the matrix inverse is the inverse of the spatial filter I-pW, and, as usual, the mean of the error term is taken to be zero. This nonlinear form clearly illustrates how the expected value of the dependent variable at each location is a function not only of explanatory variables at that location, but of the explanatory variables at all other locations as well.

SPATIAL DEPENDENCE AS A NUISANCE

Spatial Dependence in Regression Residuals

The error term in a regression can be considered to contain all ignored elements. If some of these show a significant spatial pattern, it will be reflected in a spatial pattern for the error terms. Hence, the indication of such a spatial pattern in regression residuals may lead to the discovery of additional variables that should be included in the model. Spatial autocorrelation in the error terms also violates one of the basic assumptions of ordinary least squares (OLS) estimation in linear regression analysis, i.e., the assumption of uncorrelated errors. When the spatial dependence is ignored, the OLS estimates will be inefficient, the t- and F-statistics for tests of significance will be biased, and the R^2 measure of fit will be misleading. In other words, the statistical interpretation of the regression model will be wrong. However, the OLS estimates themselves remain unbiased, contrary to what is sometimes suggested in the literature.[15]

When spatial dependence is present in the error term of a regression, I refer to it as a nuisance. The main objective is to detect it and to correct for it, if found. In social science applications using cross-sectional data, spatial error dependence is likely to be a problem, in part due to the poor choice of spatial units (see the discussion of spatial scale earlier), but also due to the predominance of spatial interaction and spatial externalities. In practice, this aspect of regression modeling is typically ignored, since the focus of attention in cross-sectional data tends to be on heteroskedasticity (i.e., unequal error variance). However, given that the consequences of ignored spatial error dependence are at least as serious as those of ignored heteroskedasticity, it should be standard practice to test regression residuals in cross-sectional studies for the former as well.

[15] For a more technical discussion of the effect of ignored spatial autocorrelation in the error terms on OLS estimates, see Anselin (1988a), and Anselin and Griffith (1988).

The dependence in the error term can be specified as a spatial process, either as a spatial autoregressive or as a spatial moving average process. It turns out that most tests for spatial error autocorrelation are the same for either form. However, typically, a spatial autoregressive specification is implemented. Using standard matrix notation to express the regression model, the spatial dependence in the errors, is specified as:

$$y = X\beta + \varepsilon$$

with

$$\varepsilon = \lambda W\varepsilon + \xi$$

where y is a N by 1 vector of observations on the dependent variable, X is a N by K matrix of observations on explanatory variables (including a constant term), with matching regression coefficients in the vector β, and ε is a N by 1 vector of error terms. The latter are assumed to follow a spatial autoregressive process, with coefficient λ and a white noise error ξ.[16] It is easy to show that the variance matrix for the error term ε is no longer the homoskedastic and uncorrelated $\sigma^2 I$ (with I as an identity matrix and σ^2 as the error variance), but instead becomes:

$$E[\varepsilon\varepsilon'] = \sigma^2[(I-\lambda W)'(I-\lambda W)]^{-1}$$

with E as the expected value operator, σ^2 as the error variance, and the other notation as before. This is the most common specification of spatial dependence in the error terms, although other forms have been suggested as well, e.g., where the dependence is a direct function of the distance between observations.[17]

Testing for Spatial Dependence in Regression Residuals

There are currently three types of diagnostics available to test for spatial dependence in the residuals of a linear regression: an extension of Moran's I statistic to regression residuals (Cliff and Ord, 1972), a test based on the Lagrange Multiplier principle (Burridge, 1980), and a specification robust approach (Kelejian and Robinson, 1992). All three tests are derived from the results of an ordinary least squares estimation in a standard regression model, i.e., in:

$$y = X\beta + \varepsilon$$

where y is a N by 1 vector of observations on the dependent variable, X is a N by K matrix of observations on the explanatory variables, with matching K by 1 coefficient vector β, and ε is a N by 1 vector of error terms. Since the error terms are unobservable, all tests use the least squares residuals instead.

Moran's I for regression residuals is a straightforward application of the statistic discussed earlier. Its form is (in matrix notation):

$I = (N/So)\ e'We/e'e$

where e is a N by 1 vector of residuals, W is the spatial weights matrix and S_o is the same normalizing factor as before (i.e., the sum of all weights). For a row standardized weights matrix, which is the preferred way to implement this test, it reduces to:

$I^* = e'We/e'e$

This form is very similar to the familiar Durbin-Watson test for serial error correlation in the time domain (see Anselin, 1988a, for a more technical discussion). Inference for Moran's I test is obtained in the usual fashion: a standardized z-value is constructed, based on the theoretical expected value and standard deviation,of the statistic under the null hypothesis of no spatial autocorrelation. These moments are somewhat more complex than the ones for the general version of Moran's I, due to the nature of

[16] In what follows, I will use 2, as a symbol for the autoregressive coefficient that pertains to an error term, and p for the coefficient that pertains to a spatially lagged dependent variable.

[17] For a more elaborate discussion, see Anselin (1988a).

regression residuals. The complete expressions are given in Cliff and Ord (1981) and will not be reported here. The resulting z-value tends asymptotically to a standard normal distribution.

The *Lagrange Multiplier* (LM) or score test for spatial error autocorrelation is very similar in form to Moran's 1. The main difference between the two lies in the normalizing factor that is used. The expression for the LM statistic is:

$$LM_{ERR} = (e'We/\sigma^2)^2/tr(W'W+W^2)$$

where tr stands for the matrix trace operation (sum of the elements of the diagonal), and where σ^2 is replaced by its estimate, e'e/N. This is not the usual unbiased estimate (the sum of squared residuals divided by the degrees of freedom), but instead it is the consistent estimate from maximum likelihood estimation (see the section on estimation below). The numerator in this test is equivalent to the square of N times Moran's I. The LMERR statistic is asymptotically distributed as a χ^2 variate with one degree of freedom.

In order for the asymptotic distributions to hold for Moran's I and the LM test, the error term must be distributed as a normal variate. This is not required for the recently suggested test by Kelejian and Robinson (1992). This test is also applicable to both linear and nonlinear regression, and it necessitates less information about the exact form of the spatial weights matrix. The statistic is computed from an auxiliary regression of cross products of residuals on cross products of the explanatory variables (collected in a matrix Z with P colunms). The cross products are for all pairs of observations for which a nonzero correlation is postulated, with each pair only entered once, for a total of hn pairs. Using γ for the coefficient vector in this auxiliary regression, and α for the resulting residual vector, the statistic is:

$$KR = (\gamma'Z'Z\gamma)/(\alpha'\alpha/h_N)$$

Similar to the LM statistic, the KR diagnostic is also a, large sample test and asymptotically follows a χ^2 distribution with P degrees of freedom.

Of the three tests, Moran's I is the most familiar one. However, as Anselin and Rey (1991) have shown in a large number of Monte Carlo simulations, it is unreliable when other forms of misspecification are present, such as non-normality and heteroskedasticity.[18] In fact, Moran's I often indicates significant spatial error autocorrelation when none is actually present. In contrast, the LM test is fairly reliable, especially when used in conjunction with its counterpart for a spatial lag, which is discussed in the section on substantive spatial dependence. A weakness of both Moran's I and the LM test is their reliance on the assumption of normality, which is often not satisfied in practice. In such instances, the KR test may be a useful alternative.

As pointed out in the introduction to this paper, Ormrod's (1990) study on the diffusion of home freezers was based on a regression analysis of the FREEZ variable on DENSITY, RURAL and INCOME. No diagnostics are reported in the paper, neither for spatial effects nor for other sources of misspecification. The OLS estimates for this model are listed in the first column of Table 8 (under the heading OLS), with the t-values for the estimates in parentheses below them. The values for, the estimates are close enough to those reported in Ormrod (1990, p. 119) so that any small differences are likely to be due to rounding, or to slight discrepancies in the data used. At first sight, this is a regression that fits the data well, with an adjusted R^2 of 0.78, and with all three explanatory variables highly significant, with the expected sign.

Upon closer examination, the spatial pattern of regression residuals, e.g., as portrayed in Figure 4, suggests a potential problem with spatial autocorrelation. The results of a more rigorous assessment of this issue are shown in Table 9, where the Moran's I, LM and KR tests are listed for three orders of contiguity, based on the shared border concept (all three matrices are row-standardized). For the, first order of contiguity, the three tests indicate a strong pattern of positive spatial autocorrelation, significant for a Type I error smaller than 0.01 (Moran's z value of 4.38, LM error of 10.77, and KR of 17.75). For the second order, none of the statistics are significant, while for the third order of contiguity, there is a weak indication of a negative spatial autocorrelation (but not for p<0.0 1). This pattern of decreasing values for the autocorrelation coefficient with higher orders of contiguity indicates that a spatial autoregressive process in the error term may be a more appropriate specification.[19]

---

[18] See also Anselin and Griffith (1988), and Anselin (1990b), for a discussion of the interaction between various types of misspecification and their effect on diagnostics for spatial effects in regression analysis.

[19] Other diagnostics for regression misspecifications did not reveal any problems. There is a slight degree of multicollinearity, with a condition number of 19.3 (typically, 20 or 30 or more is considered to be a problem, see Belsley et al., 1980), but neither the

Estimating Models with Spatial Autoregressive Errors

The spatial error model is a special case of a so-called non-spherical error model, i.e., a regression specification for which the assumptions of homoskedastic (constant variance) and uncorrelated errors are not satisfied. It is easy to show that if the autoregressive coefficient λ were known, the regression coefficients in the model (the β) could be estimated by means of OLS in a model with spatially filtered variables. This is similar to the rationale behind the procedure of Weighted Least Squares (WLS) developed for many cases of nonspherical errors.

After some manipulation, the spatial error model can be shown to be equivalent to:

$$y - \lambda Wy = X\beta - \lambda WX\beta + \xi$$

or, to:

$$(I-\lambda W)y = (I-\lambda W)X\beta + \xi$$

where $(I-\lambda W)$ is the spatial filter, ξ is a well-behaved (uncorrelated homoskedastic) error, and the other notation is as before. In other words, after computing the spatially filtered dependent and explanatory variables y^* and X^*, defined as in:

$$y^* = y - \lambda Wy$$

$$X^* = X - \lambda WX$$

the estimates for β are found by applying OLS to a regression of y* on X*. This approach is referred to as *Generalized Least Squares* (GLS).

Unfortunately, in practice, the autoregressive coefficient is not known and must be estimated jointly with the other parameters. This is similar to the problem encountered for serial autocorrelation in the time domain. However, the two-step and iterative estimation methods that have been suggested for the time series case, such as the familiar Cochrane-Orcutt approach, are not applicable to spatial data.[20] Instead, estimation must be based on the principle of *maximum likelihood*.

A technical discussion of maximum likelihood (ML) estimation is beyond the scope of the current paper. The classic reference for ML estimation in spatial models is Ord (1975), but extensive treatments can also be found in Cliff and Ord (1981), Anselin (1988a), Griffith (1988a), and Anselin and Hudak (1992), among others. Roughly speaking, maximum likelihood estimation consists of finding those parameter values for which the joint probability of the observed dependent variable (y), conditional upon the parameters, is maximized. In order to implement this approach, a distribution must first be assumed, and a formal expression for the joint likelihood of the observations must be derived. In the estimation of spatial autoregressive models, the point of departure is a normal distribution for the unobservable error term E, in order to derive the distribution of the observable dependent variable. In this process, a so-called Jacobian term is used, which corrects for the autocorrelation in the error terms. In the spatial error model this Jacobian is the determinant of (IAW), i.e., the determinant of the spatial filter. The presence of the Jacobian term in the likelihood function considerably complicates the computation of the ML estimates, and its efficient treatment has been the subject of much research (see, e.g., Griffith, 1988a).

The likelihood function for the spatial error model must be maximized with respect to the parameters λ, β, and σ^2. Typically, this maximization of the likelihood function (or, equivalently, of its logarithm) requires a process of nonlinear optimization. In the

---

Kiefer-Salmon (1983) test for non-normality (value of 2.4, distributed as X2 with 2 degrees of freedom, with $p<0.30$), nor a Breusch-Pagan (1979) test for heteroskedasticity (value of 1.14, distributed as X2 with 3 degrees of freedom, with $p<0.77$) were significant.

[20] For a more extensive technical discussion of the relative merits of the various estimators suggested in the literature, see Anselin (1988a, Chapter 8).

spatial error model, this is somewhat simplified, since the estimates for both β and σ^2 can be expressed analytically in function of the autoregressive parameter λ. When this is taken into account, the maximization of the resulting so-called concentrated likelihood function requires only a search over a single parameter (λ).

The application of the maximum likelihood principle has a number of important implications for the resulting estimates. The properties of these estimates, such as their consistency and asymptotic efficiency, are based on sample sizes that tend to infinity, and they may not hold in realistic finite samples. In other words, even though the ML estimates are in principle superior to OLS, this may not be the case in small samples.[21] Also, the ML estimates are based on the assumption of normality, which is often violated in practice. Furthermore, the traditional measures of fit, such as the R^2, are no longer valid and specialized indices, such as the maximized likelihood and criteria based on information theory must be used instead (see Anselin, 1988b). In sum, even though the ML estimation is the methodologically correct approach for this model, its results may have to be interpreted with caution, especially when any of the underlying assumptions are not satisfied.

The outcome of a maximum likelihood estimation of the spatial error model for the freezer example is listed in the second column of Table 8, under the heading ERROR. The autoregressive coefficient X is positive (0.637) and highly significant (asymptotic tvalue of 5.3 1), confirming the earlier indication given by the diagnostics in Table 9. The estimates for the regression coefficients all decrease in absolute value, relative to the OLS estimates. While this change is not very large in absolute terms, in relative terms it ranges from 15.6% for RURAL to 25.7% for INCOME (the relative change for DENSITY is 23.1%). This may indicate that the error model may not be the appropriate specification. Indeed, since OLS is an unbiased estimator for this model, the difference between the coefficient estimates obtained by OLS and ML should not be large. More importantly however, the ranking of coefficients in terms of significance is altered. This is an aspect of the model that received considerable attention in the original study. The introduction of the spatial autoregressive error term changes the order of importance of the INCOME (asymptotic t-value of 3.37 vs. 5.75 in OLS) and DENSITY (-3.66 vs. -5.04 in OLS) variables. This change in significance is typical of an implementation of the spatial error model, since the main effect of the ML estimation is on the variance of the estimates. In some instances, this may even change variables from being significant to not being significant, and vice versa.

The maximum likelihood estimation of the spatial error model also yields two additional tests on the significance of the autoregressive coefficient. A so-called *Wald test* on this coefficient is simply the square of the reported asymptotic t-test, $(5.31)^2$ or 28.2. This statistic is distributed asymptotically as a χ^2 variate with one degree of freedom. The other asymptotic test is a so-called *Likelihood Ratio* test, which is obtained by comparing the maximized likelihoods of the spatial error model and of the standard regression model. The statistic equals twice the difference between the log-likelihoods and is also distributed as a χ^2 variate with one degree of freedom. For the freezer example, it is 2(-125.2+131.5), or 12.6. The Lagrange Multiplier, Wald and Likelihood Ratio tests are asymptotically equivalent, but typically conform to the following inequality in finite samples:

$$W \geq LR \geq LM$$

It is quite possible that this inequality would lead to an indication of significant spatial error autocorrelation according to the Wald or LR, but not the LM criterion. Since it is necessary to carry out a maximum likelihood estimation to obtain the W and LR statistics, but only OLS to obtain LM statistic, the latter is likely to be used most often to decide on the presence of spatial error autocorrelation. It should be kept in mind that its indication may tend to be on the conservative side, i.e., it may not reject a null hypothesis of no spatial autocorrelation when the other tests would reject it. In the freezer example, this is not the case, and the three test results conform to the inequality, since:

$$28.2\ (W) > 12.6\ (LR) > 10.8\ (LM)$$

Outliers

Another approach towards finding spatial patterns in regression residuals can be based on an analysis of the outliers, i.e., of the extreme residuals. This perspective lends itself extremely well to an integrated approach between a GIS and spatial data analysis module, as illustrated in Anselin, Dodson and Hudak (1992). The display of extreme residuals can form the basis for a number of traditional diagnostics, e.g., of the type outlined in Belsley et al. (1980) and Cook and Weisberg (1982), and applied to spatial models in Haining (1990). However, such diagnostics ignore spatial patterns. Instead, the extreme residuals can be converted to simple dummy variables, whose spatial pattern can then be analyzed by means of join count statistics, as in Anselin, Dodson and Hudak (1992), or by means of specially developed measures of clustering, as in Nass and Garfinkle (1992). The latter use an indication of significant clustering of outliers as a way to detect new variables to add to the model specification. Again, this is a process that is very suitable for incorporation into a GIS, e.g., to match the patterns of selected variables in a data set to the spatial pattern of the outliers with an overlay procedure.

[21] See Florax and Folmer (1992), for a recent discussion of this aspect.

Implementation

As already pointed out in the introduction to this paper, none of the popular commercial statistical and econometric software packages currently include facilities to carry out the tests for spatial error autocorrelation, or to estimate models with spatially autocorrelated errors. As illustrated in Anselin and Hudak (1992) for the econometric packages *Gauss, Splus, Shazam, Rats* and *Limdep*, it is fairly straightforward to implement the tests, once a spatial weights matrix has been constructed. The basic requirement to carry this out is that a statistical package include a way to efficiently compute matrix products. Other matrix manipulations, such as a transpose and a trace make matters easier, but are not essential. Most sophisticated econometric software includes these facilities.

In order to obtain maximum likelihood estimates for the spatial error model, the software must include nonlinear estimation as well as a way to obtain the eigenvalues, determinant and inverse of a matrix. As shown in Anselin and Hudak (1992), most packages can be manipulated into carrying out the proper nonlinear optimization, but the resulting estimates are not always totally reliable.[22] Griffith (1988b) and Griffith et al. (1990) provide similar illustrations of how the macro programming facilities in the packages *Minitab* and *SAS* can be used to estimate spatial error models. In Bivand (1992) the same is shown for the package *Systat*.

SUBSTANTIVE SPATIAL DEPENDENCE

Spatial Dependence in the Form of a Spatial Lag

A measure of spatial autocorrelation, such as Moran's I, may indicate true spatial dependence, which remains present even after the influence of explanatory variables has been taken into account. I refer to this as substantive spatial dependence, to contrast it with the autocorrelation in the error term discussed in the previous sections. Substantive spatial dependence implies that the value of a variable observed at each location is truly jointly determined with that at other locations. In social science research, it is not useful to interpret this in the tradition of geographic determinism, but instead it should be seen as an indication of a behavioral, political or economic process characterized by significant spatial externalities. In the freezer example, it would imply that the adopters of home freezers are influenced by the actual decisions of their neighbors, and not just by the fact that the "context," as described by the explanatory variables (e.g., INCOME, DENSITY and RURAL) is similar in neighboring states. Of course, one would still have to discover the social, economic and political mechanisms through which such externalities are realized. The usefulness of the spatial lag model is that it allows a clear distinction between spatial similarity in the dependent variable and spatial similarity in the explanatory variables. Simple, univariate descriptive statistics of spatial autocorrelation do not provide this information.

Formally, substantive spatial dependence can be expressed as a mixed regressive spatial autoregressive process:

$$y = \rho Wy + X\beta + \varepsilon$$

where, as before, Wy is the spatial lag and X is a matrix of observations on the explanatory variables. The vector of error terms ε is assumed to follow a normal distribution that satisfies the classic properties (uncorrelated and homoskedastic).

In contrast to what holds in the spatial error case, ignoring spatial dependence in the form of a spatial lag will yield biased estimates. This is similar to the specification problems caused by omitting any significant variable in regression analysis. In addition, even when a spatial lag is included in the model, the ordinary least squares estimator is biased and inconsistent. The reason for this is the simultaneity between the lagged dependent variable and the error term, similar to what happens with endogenous variables in models of simultaneous equations. For the properties of OLS to hold, all explanatory variables should be uncorrelated with the error term. The correlatedness between the Wy and F, violates this basic assumption.[23] Consequently, the mixed regressive spatial autoregressive model requires specialized estimation techniques, to which I return below.

Testing for an Omitted Spatial Lag

In Anselin (1988c), a Lagrange Multiplier test for a spatially lagged dependent variable was suggested, based on the principles for testing for an omitted explanatory variable. However, due to the simultaneity of the spatial dependence, the standard approach, based on a simple auxiliary regression, does not hold and a special derivation is needed.[24] As in the spatial error model, the

---

[22] A detailed listing of source code is contained in Anselin, Hudak and Dodson (1993).

[23] For a detailed technical discussion, see Anselin (1988a, Chapter 6).

[24] See Godfrey (1988) for an extensive review of the standard approach towards misspecification tests based on the Lagrange Multiplier principle.

Lagrange Multiplier test for a spatial lag proceeds from the results of the OLS estimation in a standard regression model, i.e., a model without a spatial lag. Formally, the test is:

$$LM_{LAG} = \{e'Wy/\sigma^2\}^2/\{(WXb)'MWXb/\sigma^2 + tr(W'W+W^2)\}$$

where $M = I - X(X'X)^{-1}X'$, b is a vector with OLS estimates for the regression coefficients, and the rest of the notation is the same as before. This statistic asymptotically follows a χ^2 distribution with one degree of freedom. The numerator in the expression is again very similar to Moran's I, except that the cross product is now between the residual vector and a spatial lag in the dependent variable Wy, rather than a lag in the residuals. The first part of the denominator term may seem quite formidable at first sight. However, (WXb)'MWXb is a residual sum of squares in a regression with WXb as the dependent variable and X as the explanatory variables. In other words, this is found in a regression of the spatial lag of the predicted values from the OLS estimation on the explanatory variables. As illustrated in Anselin and Hudak (1992), the test can be easily implemented in the econometric software packages *Gauss, Splus, Limdep, Rats* and *Shazam*, using this auxiliary regression.

In the example, the LM lag statistic yields a value of 14.7 for first, 4.9 for second and 0.2 for third order contiguity (with a row-standardized matrix). The result for first order contiguity is highly significant ($p<0.001$), but the others are not. It compares to a value of 10.8 for the LM error test, which is also significant, but slightly less so. This type of situation, where both error and lag dependence are indicated by the LM tests, is very common in practice. A rule of thumb suggested in Anselin and Rey (1991), based on a large number of simulation experiments, is to consider the alternative model that achieves the highest value on the LM statistic. In this case, the rule would point to the lag model as the most likely alternative. The difficulty in distinguishing between the spatial error and the spatial lag model is a direct result of the similarity between the two specifications, to which I return below. Both alternative models will lead to OLS residuals that are spatially correlated. As a result, a test such as, Moran's I, which cannot distinguish between the two alternatives, is not very useful as a guide in the search for a proper model specification.

Estimation of the Spatial Lag Model

The presence of the spatially lagged dependent variable on the right hand side of the regression equation is similar to the inclusion of an endogenous variable in systems of simultaneous equations. Two general approaches exist to deal with the resulting conr,lation between the Wy term and the error. One is based on the maximum likelihood principle, similar to the estimation of the spatial error model. A normal distribution is assumed for the error term and the corresponding likelihood function is derived. As in the spatial error model, a crucial role is played by the Jacobian term, i.e., the determinant of the spatial filter I-pW. The estimates for the β and σ^2 coefficients can again be expressed in function of the spatial autoregressive parameter ρ, and the maximum of the resulting nonlinear concentrated likelihood function can be found by means of a straightforward search. As before, the properties of the ML estimates are asymptotic in nature and may not always hold in small samples. Many technical details are the same as for the spatial error model, and I will not repeat them here. The implementation of this approach in the econometric packages *Gauss, Splus, Shazam, Rats* and *Limdep* is illustrated in Anselin and Hudak (1992).[25]

Another approach towards estimating the spatial lag model can be based on the *instrumental variable* principle. This is equivalent to two stage least squares (2SLS) estimation in systems of simultaneous equations. The correlation between the spatial lag and the error term is controlled for by replacing the former with an appropriate instrument, i.e., a variable that is highly correlated with the spatial lag, but uncorrelated with the error term. The choice of the proper instrument is a major problem in the practical implementation of this approach. On the other hand, it is easy to carry out in practice, since two stage least squares is a common feature of commercial econometric software. Since there are insufficient variables available to construct a good instrument in the freezer example, I will not illustrate this approach here.[26]

The results of a maximum likelihood estimation of the spatial lag model in the example are listed in the third column of Table 8 (under the heading LAG). The spatial autoregressive coefficient takes on a value of 0.408, which is highly significant (asymptotic t-value of 3.97). The decrease in absolute value of the other coefficients, relative to the OLS results is even more pronounced than for the spatial error case. Specifically, for DENSITY (from -0.0173 to -0.0103, or 40.4%) and INCOME (from 4.433 to 2.779, or 37.3%), these changes illustrate the bias of the OLS estimates when the significant spatial lag is ignored. However, in contrast to the results in the spatial error model, the relative significance of the explanatory variables remains the same as for OLS, but with the spatial lag being more significant than both INCOME and DENSITY. The value of the maximized likelihood of -124.3 for the spatial lag model

[25] For derivations of the estimators and their asymptotic variance matrix, see Ord (1975), Cliff and Ord (1981), and Anselin (1988a).

[26] For a more in-depth discussion of the instrumental variable and other robust approaches towards estimating the spatial lag model, such as a bootstrap estimator, see Anselin (1990c).

is the highest of the three specifications. This confirms the indication given by the LM lag statistic relative to the LM error statistic. Similar to the spatial error case, a Wald and Likelihood Ratio statistic for spatial dependence can be derived from the ML estimates. In the example, these amount to respectively 15.8 and 14.5. Note that the latter is slightly smaller than the LM test of 14.7, thus violating the expected inequality between the three asymptotic statistics. This is not uncommon in practice, and usually points to the presence of other specification errors, such as non-normality, heteroskedasticity, or, more importantly, to the wrong choice of the spatial weights matrix.[27] In addition, for the particular model under consideration, one other potential problem could be the truncated nature of the dependent variable (percentage adoption), which may have to be taken into account explicitly by means of a probit or logit model.[28]

Spatial Lag or Spatial Error Model?

In practice, the distinction between a spatial error model and a spatial lag model is often difficult Even though the interpretation of the two specifications is fundamentally different (i.e., substantive vs nuisance), they are closely related in formal terms. Indeed, the filtered version of the spatial error model used above:

$$y - \lambda Wy = X\beta - \lambda WX\beta + \xi$$

can also be written as:

$$y = \lambda Wy + X\beta - \lambda WX\beta + \xi$$

The latter is a special case of a spatial lag model, where in addition to the usual set of explanatory variables (contained in the matrix X), the spatially lagged explanatory variables WX are included as well. This model is referred to in the literature as the *spatial Durbin* or *common factor* model. The equivalence between a spatial error model and a spatial lag specification with X and WX as explanatory variables is only satisfied when a set of nonlinear constraints hold for the coefficients of the model. These constraints are referred to as the common factor hypothesis, in analogy with a similar approach used in time series analysis (Hendry and Mizon, 1978). Specifically, the negative of the product of the spatial autoregressive coefficient (λ, the coefficient of Wy) with each regression coefficient associated with the unlagged X (β) should equal the matching coefficient of the lagged WX ($-\lambda\beta$).

An easy way to check whether the common factor hypothesis is satisfied is to estimate an unconstrained spatial lag model of the form above (i.e., without imposing the nonlinear constraints), and to compare its likelihood to that of the spatial error model. The two likelihoods can be combined in a so-called Likelihood Ratio test for the common factor hypothesis. In the freezer example, using the same first order contiguity matrix as before, the result of this test is 3.66, which is clearly not significant for a χ^2 variate with three degrees of freedom ($p<0.30$). In other words, the common factor hypothesis cannot be rejected, and there is no logical inconsistency in the spatial error specification. If the hypothesis had been rejected, this would have been evidence that the spatial error model is the wrong one, or that the weights matrix was poorly chosen.[29]

The common factor hypothesis is one of two tests than can be used to aid in distinguishing between the spatial error and spatial lag model. The other test does not exploit a formal equivalence such as the spatial Durbin model, but it focuses on the extent to which a spatial lag has eliminated the spatial autocorrelation in the errors. In Anselin (1988c), such a diagnostic was formulated as a Lagrange Multiplier test, which is carried out for the residuals of a maximum likelihood estimation of the spatial lag model. A highly significant statistic would indicate that the spatial lag model has not eliminated all evidence of spatial autocorrelation. This often means that the spatial error model is more appropriate, or that the spatial weights matrix was poorly chosen. In the freezer example, there is no indication of any remaining error autocorrelation.[30]

Choice of Weights Matrix

The choice of the spatial weights matrix is an important aspect in the specification search for a spatial model. There are two distinct approaches to this, one based on hypothesis tests, the other on measures of fit. According to the first approach, the comparison of the appropriateness of different spatial weights matrices is formalized as a test on non-nested hypotheses. A number of such tests have been developed, but their useftilness for spatial models has been rather limited, due to the lack of power of the tests, and because

[27] A Lagrange Multiplier test for heteroskedasticity in the spatial lag model is suggested in Anselin (1988a). For the current model, it yields a value of 3.63, which is not significant for a χ^2 variate with 3 degrees of freedom ($p<0.30$).

[28] For a recent example of the application of a spatial lag specification in a probit model, see Case (1992).

[29] For a technical discussion, see Burridge (1981), Bivand (1984), and Anselin (1989a).

[30] The statistic is asymptotically distributed as a χ^2 variate with one degree of freedom. Its value in the example, using the first order contiguity matrix, is 0.86, which is clearly not significant ($p<0.35$).

it is necessary to carry out maximum likelihood estimation before they can be implemented (see Anselin, 1986 and 1988b, for a review).

A much easier way to compare the various weights matrices is to assess the fit of the models by means of their likelihood, and to carefully check all diagnostics in each model. In addition to the test on the common factor hypothesis and a test on remaining error autocorrelation, an often insightful indicator is the extent to which the expected inequality between LM, LR and W statistics is satisfied. The 'best' model will have the highest likelihood and should pass all diagnostics without problems.

To illustrate this point, the results of the estimation of the spatial lag model for the FREEZER variable, with four different spatial weights are listed in Table 10. The first column is the same as for the spatial lag model shown in Table 8, and is based on a first order contiguity weights matrix. The other three columns correspond to different distance-based weights matrices. The first (DISTANCE_1) is the binary matrix used before in the computation of the G statistics. The second (DISTANCE_2) is a gravity-like specification, using the squared inverse distance, but for the same range as specified in the binary weights matrix. In other words, up to the critical distance the spatial weight takes a value of $1/(d_{ij})^2$, but beyond the critical distance it becomes 0. In the third distance-based matrix (DISTANCE_3), there is no cut-off, and the full squared inverse distance matrix is used.

The best fit is found for the gravity specification with a cut-off distance (DISTANCE_2), which yields a log likelihood of -122.3 (compared to -122.9 for the binary distance, - 124.3 for simple contiguity and - 125.5 for continuous gravity). All four weights result in highly significant statistics for spatial autocorrelation, as shown in the bottom part of the table. However, only for the two gravity specifications is the expected order $W \geq LR \geq LM$ satisfied, although barely so for the full matrix (respectively with 25.0 > 18.4 > 16.7 for DISTANCE_2, and 18.6 > 12.1 > 12.0 for DISTANCE_3). Since a spatial autoregressive process for freezer adoption is likely to follow from a pattern of spatial interaction between households, the superiority of the gravity specification has some theoretical foundation as well.

A common characteristic of the use of the weights matrix in the models considered so far is that its elements are treated as if they were known a priori. In essence, the spatial autoregressive parameter is only a scaling factor which is applied to a matrix of N times N- 1 coefficients. Clearly, there is insufficient information in a sample of N observations to estimate all of these, but some additional flexibility may be introduced, e.g., by means of Bayesian approaches, or by making the weights a nonlinear function with several additional parameters. A detailed discussion of this issue is beyond the scope of this paper. Nevertheless, it is worthwhile to point out that the spatial weights approach is seriously limited in terms of the range of specifications for the spatial interaction that can be incorporated into the models.[31]

OTHER SPATIAL REGRESSION MODELS

Spatial Regression Models

A limited number of statistical capabilities have recently been added to some commercial GIS, primarily in raster-based systems, such as *Idrisi* (Eastman, 1992), and the GRID module of *Arc/Info* (ESRI, 1991b). In terms of regression analysis, the models supported are often referred to as *spatial regression* models. While such models do not incorporate spatial dependence as discussed up to this point in the paper, they include various spatial aspects in their specification, such as locational characteristics. They can be thought of as dealing with spatial heterogeneity, either in the form of *spatial drift* (continuous variation over space) or *as spatial regimes* (discrete variation over space).

The most common among the spatial regression specifications is the so-called *trend surface* model. In a trend surface, the spatial drift in a variable is explained by a polynomial in the coordinates of the observations.[32] Since the coordinates of a grid cell are easily obtained in a raster-based GIS, trend surface analysis can be implemented in a straightforward fashion. Other forms of spatial regression models are: *spatial analysis of variance* (ANOVA), where an indicator variable corresponds to a particular subregion of the data; *spatial regime* analysis, where the regression coefficients vary between subregions of the data; and the *spatial expansion* model, where the regression coefficients follow a trend surface or other continous function (i.e., to model spatial drift). While these three specifications are currently not yet incorporated in the regression commands of commercial GIS, they are particularly, suited to be analyzed in conjunction with a GIS, as outlined in more detail in Anselin, Dodson and Hudak (1992).

In the remainder of this section, I will briefly illustrate each of the spatial regression models by means of the home freezer adoption example. Since these models are simply special cases of standard regression specifications, they can be estimated and analyzed by means of existing software. However, as discussed in the previous two sections, they may also be subject to the effects of

[31] For further discussion, see Anselin (1988a), and Bolduc et al. (1992).

[32] Trend surface analysis is carried out by means of the command Trend in both Idilsi and Arcl/Info's GRID module.

spatial dependence, either in the error term or in the form of a spatial lag. Of course, if this is the case, the specialized techniques discussed before will be needed.

Trend Surface Analysis

As mentioned, a trend surface model is a special regression model that has as its explanatory variables the elements of a polynomial in the coordinates of the observations, say x and y. These coordinates correspond to specific data points or grid cells, as in a raster-based GIS, or represent meaningful points associated with an areal unit, such as the centroid of a polygon in a vector-based GIS. For example, a second order trend surface would have the following formal specification:

$$z = \alpha + \beta_1 x + \beta_2 x + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy + \varepsilon$$

where, in contrast to the notational convention used elsewhere in the paper, the dependent variable is represented by the symbol z. The explanatory variables x and y correspond the coordinates, with α as the constant and the βi as the regression coefficient for each term in the polynomial. As before, ε is a random error term. Except for the particular choice of explanatory variables, the trend surface model is otherwise treated exactly like any other regression specification.

A trend surface is particularly useful to filter out large scale spatial trends in order to focus on smaller scale variation, i.e., on the residuals. Another common application of this model is to obtain spatial interpolations. Since the trend surface is only a function of the absolute location, predicted values can be computed for any pair of coordinates. This is commonly exploited to produce a smoothed surface. Clearly, such a surface can be easily visualized in a GIS, or may be used to produce interpolated values for addition to a georelational data base.

A common problem with the estimation of a trend surface is the high degree of multicollinearity. This is due to the strong functional relation between the various terms in the polynomial. As a consequence, the indication of significance (t-values) and fit (R^2) may be suspect. It is generally not a very good idea to use a trend surface for more than simple smoothing, filtering or interpolation of data.[33]

The results of a second order trend surface fitted to the FREEZER variable are listed in Table 11. The coefficient estimates are given, as well as two t-statistics for significance, one based on the standard OLS approach, the other on a robust Jacknife approach (MacKinnon and White, 1985; Anselin, 1990c). The latter is used to correct for the presence of heteroskedasticity.[34] The fit of the model, achieving an adjusted R2 of 0.71 is only slightly worse than that of the original specification (compare to an adjusted R^2 of 0.78 in Table 8). There is a strong indication of a quadratic trend in the x (westeast) dimension, with the terms in x and x^2 highly significant, for both OLS (t-values of 5.46 and -6.20 respectively) and robust (t-values of 6.19 and -5.07 respectively) measures. The interaction term in xy is significant for OLS (t-value of -2.41), but not for the Jacknife (-1.52). In general, these t-values should be interpreted with some caution, given the high degree of multicollinearity, illustrated by a condition number of 88.7, well above the acceptable levels of 20 to 30 (Belsley et al., 1980). Interestingly, there is no indication of any trend in the north-south dimension.

A final point illustrated by this model is the unreliability of Moran's I statistic. The Lagrange Multiplier tests for error and lag autocorrelation are not significant for any of the four weights matrices used so far (simple contiguity and the three distance-based weights), but Moran's I is significant for the first order contiguity (z-value of 2.07) and for the inverse distance matrix with a cut-off (z-value of 2.16).[35] In those instances, it is likely to pick up the heteroskedasticity in the model, rather than the spatial autocorrelation, illustrating a point raised in the simulations of Anselin and Rey (1991).

Spatial Analysis of Variance

In many instances in exploratory spatial data analysis, the interest focuses on the extent to which the mean for a variable differs significantly between spatial subsets of the data. This is assessed by means of analysis of variance (ANOVA). In ANOVA terminology, a test is formulated on the difference between the means of a variable subject to a number of different

---

[33] For a further discussion of some technical issues encountered in trend surface modeling, see Haining (1987, 1990), and Unwin and Wrigley (1987a, 1987b), among others.

[34] A Breusch-Pagan (1979) test dw uses the x and y as heteroskedastic variables yields a value of 7.27, which is significant at $p<0.03$ for a χ^2 variate with 2 degrees of freedom.

[35] LM error results are respectively 0.41, 0.01, 0.77 and 0.56 (for the weights CONTIG_1, DISTANCE_1, DISTANCE_2 and DISTANCE_3 used in Table 10), and LM lag results are 1.64, 0.52, 2.62, and 2.00, neither of which is significant at p=0.05.

"treatments."[36] I refer to this as *spatial ANOVA*, in the sense that the treatments are spatial and refer to subregions of the data set (e.g., east-west, north-south, core-periphery). Classic ANOVA is based on a number of simplifying assumptions that may not be satisfied in practice. Two important ones are similar to the assumptions of homoskedasticity and uncorrelatedness in a regression model. The easiest way to take such complications into account is to treat the ANOVA as a regression model with dummy variables (indicator variables). Each treatment (or subregion) is associated with a particular indicator variable. If the model includes a constant term, then the overall mean corresponds to the value of that constant term, and the coefficient of each indicator variable measures the difference between each subset and the overall mean. Alternatively, when the model does not include a constant term, the coefficient of the indicator variable measures the mean for each subset individually. Since the spatial ANOVA takes the form of a regression model, diagnostics for spatial autocorrrelation can be implemented, and an ANOVA model with a spatial error or spatial lag can be estimated as well.[37]

A casual look at the spatial pattern of freezer adoption in Figure 1, or at the Gi* statistics in Figure 3 would suggest a major difference between western and eastern states in the U.S. In order to assess this more rigorously, I created an indicator variable, WEST, which takes on a value of 1 for all states with the x coordinate of their centroid (in Table 1) less than 29.0 (the variable WEST is also listed in Table 1). Such a designation of a spatial subset can easily be carried out by means of a window or lasso operation in a GIS (see Anselin, Dodson and Hudak, 1992), or even as a simple query in a georelational data base. The results for a spatial ANOVA of the east-west dichotomy of FREEZER are given in Table 12. The constant term in the regression shows an overall mean adoption of 17.3%, and a differential for the western states of 10.3% (hence, the western states achieve a mean of 27.6%). This differential is highly significant, as indicated by a t-value of 5.33 ($p<0.001$).

The simple ANOVA model explains 37% of the variance in the FREEZER variable. A Breusch-Pagan statistic of 3.77 (for a χ^2 variate with one degree of freedom) indicates a slight degree of heteroskedasticity ($p<0.06$), but more importantly, both LM error and LM lag statistics are highly significant, e.g., taking on values of 20.5 and 24.4 respectively for the inverse distance weights matrix with a cut-off (DISTANCE_2).[38] Since LM lag is considerably higher than LM error, a model with the former is estimated. The results are listed in the second column of Table 12. The spatial autoregressive coefficient of 0.73 is highly significant (t-value of 8.23), confirming the indication given by the test. Note that the coefficients of the constant and WEST dummy are no longer comparable to the values obtained by OLS, but should be interpreted as the mean and regional differential for the spatially filtered fi-eezer adoption rates, i.e., after the spatial autoregressive filter is applied (using 0.73 as the coefficient). After this spatial screening process, the regional differential of 2.96 (relative to an overall filtered mean of 4.63) is much less proriounced, achieving a t-statistic of 2.03, which is only slightly significant ($p<0.05$ compared to $p<0.001$ for OLS). In other words, the original indication of two clearly different spatial regimes is much lessened after the spatial dependence in the data is taken into account. Similarly, the heteroskedasticity in the model is removed by the spatial filter.[39]

Spatial Regimes of Structural Change

Neither the simple spatial ANOVA nor the trend surface model are very meaningful in their own right as final explanations of the patterns observed in the data. However, they highlight the importance of spatial drift (trend surface) and spatial regimes (ANOVA) and may indicate that the incorporation of one of these forms of spatial heterogeneity into the original specification may be useful.

Spatial regimes are formalized as varying coefficients between spatial subsets of the data. In other words, different spatial regimes can be distinguished, across which the relation between dependent and explanatory variables in the regression shows a significant variation. Such variation may be expressed in a very simple manner in the form of a change in "intercept," by means of a dummy variable for each of the spatial regimes (less one, if a constant term is included in the regression). Formally, the observations are classified into two (or more) subsets according to the value taken by an indicator variable, say r:

---

[36] A classic reference on ANOVA is Scheffe (1959).

[37] For further discussion of spatial ANOVA, see Griffith (1978, 1992), Legendre et al. (1990), and the example in Anselin (1992d). An illustration of the integration of spatial ANOVA within a GIS is given in Anselin, Dodson and Hudak (1992).

[38] For this weights matrix, Moran's I is 0.507, with a highly significant z-value of 5.19. The statistics for the other three weights matrices were highly significant as well. However, since the best fit is consistently obtained for the DISTANCE_2 matrix, I will no longer report results for the other weights.

[39] The LM statistic for heteroskedasticity in the spatial lag model (Anselin, 1988a) takes on a valueof 1.24, which is clearly not significant for a χ^2 variate with one degree of freedom ($P<0.27$).

$$y_1 = \alpha_1 + X_1\beta + \varepsilon_1 \quad \text{for } r = 0$$

$$y_2 = \alpha_2 + X_2\beta + \varepsilon_2 \quad \text{for } r = 1$$

or

$$y = \alpha + X\beta + (\alpha_2 - \alpha)r + \varepsilon$$

Note that a spatial ANOVA regression follows the same specification, except that no other explanatory variables are included besides the dummy variable r. A test of significance on the coefficient of this dummy variable indicates the extent to which the constant terms are different between the two subsets in the data.

A slightly more complex form of regional differentiation occurs when both the constant terms ((α_1 and α_2) as well as the slope terms (β_1 and β_2) take on different values, as in:

$$y_1 = \alpha_1 + X_1\beta_1 + \varepsilon_1 \quad \text{for } r = 0$$

$$y_2 = \alpha_2 + X_2\beta_2 + \varepsilon_2 \quad \text{for } r = 1$$

In the extreme, a different coefficient can be specified for each observation, by means of approaches such as spatial adaptive filtering (Foster and Gorr, 1986).

The strength of the regional differentiation can be assessed by means of the familiar Chow test on the stability of regression coefficients, and its extension for spatial process models (i.e., a spatial Chow test, suggested in Anselin, 1990b). The null hypothesis in such a test is:

$$H_0: \alpha_1 = \alpha_2 \text{ and } \beta_1 = \beta_2$$

Rejection of this null hypothesis thus implies that all coefficients taken jointly differ significantly between the spatial subsets in the data. Alternatively, such a test may be carried out for each coefficient individually. Of course, the model itself is treated as a standard regression model, and spatial effects may be incorporated as outlined before.

The implementation of spatial regimes for the regression coefficients in the freezer adoption model is illustrated in Tables 13 and 14. In Table 13, estimates are presented for an "extended" model that includes the dummy variable WEST as one of the explanatory variables. The results of the OLS estimation are listed in the first column. The coefficient of the dummy variable is significant (t value of 2.81, $p<0.01$), and indicates an adoption rate in the west which is higher than the national average by 3.7%, even after all the other explanatory variables have been taken into account. All the coefficients of these variables remain significant, but they decrease in absolute value relative to the original model. However, several diagnostics (not reported in the table) indicate some problems with this specification. A Breusch-Pagan statistic of 10.88 is significant with $p<0.03$ for a χ^2 variate with 4 degrees of freedom, pointing to heteroskedasticity. In addition, both LM error (12.63) and LM lag (11.73) statistics, using the inverse distance weights matrix with a cut-off (DISTANCE-2) are highly significant ($p<0.001$). The estimates for the alternative models are given in Table 13 as well, respectively in the second column for the error model, and in the third column for the lag model. The two are marginally different in terms of fit, with a slight edge for the spatial error model (log likelihood of -121.5 vs -121.6). However, both the LM test and the t-value in the model indicate a higher significance for the error model (note that the squared correlation, though intuitively appealing, is a misleading measure of fit). Both models eliminate the heteroskedasticity from the original specification.[40]

An interesting difference between the two specifications is that the coefficient of the WEST dummy is no longer significant in the spatial lag model (1.467 with a t-value of 1.18), while it remains significant in the spatial error model (3.902 with a t-value of 2.61). More pronounced than what was found for the spatial ANOVA, the spatial filtering implied by the lag model effectively

[40] The respective LM statistics for heteroskedasticity, which are distributed as X2 with four degrees of freedom, are 5.36 ($p<0.25$) in the error model, and 5.90 ($p<0.21$) in the lag model.

removes the indication of spatial regimes in the mean (i.e., for the constant term), after the influence of the other explanatory variables is taken into account.

The estimates for a specification where all regression coefficients are different between east and west are given in Table 14. The first column lists the results for OLS estimation. Except for the constant term, all coefficients are strongly significant in both regimes, and the model fit (adjusted R^2 of 0.82) is a slight improvement over the original specification (adjusted R^2 of 0.78) and the extended model (adjusted R^2 of 0.81). A Chow test on the joint equality of all coefficients between subsets (including the constant term) yields a value of 3.70, which for an F(4,40) variate is significant at 0.0 1. However, when limited to each individual coefficient, a significant difference can only be found for the DENSITY variable.[41] Again, various diagnostics (not listed in the table) indicate a number of specification problems. A Breusch-Pagan test for groupwise heteroskedasticity, based on the WEST variable (i.e., with a different error variance in each of the subsets) yields a value of 9.05, which is highly significant for a χ^2 variate with one degree of freedom ($p<0.01$). Similarly, both LM error and LM lag are significant (for DISTANCE-2), with values of respectively 6.11 4KO.01) and 9.91 ($p<0.003$).

The estimates for the spatial lag and groupwise heteroskedastic specifications are listed in the second and third columns of Table 14. The latter are the same as for OLS, but the standard deviations and t-values (and also the likelihood) differ when the heteroskedasticity is taken into account. The spatial autoregressive parameter is strongly significant (0.365, with a t-value of 3.41). In contrast to the filtering effect of the lag in the spatial ANOVA and in the extended model, here all coefficients (except the constant term) remain significant in both regimes. The same is true for the heteroskedastic model. The error variance corresponding to each regime is respectively 4.46 in the east and 17. 10 in the west, clearly illustrating the presence of heteroskedasticity. An asymptotic Chow-like test on the difference between regimes is very significant for both lag and heteroskedastic models, yielding values of respectively 18.67 and 17.81, for a χ^2 variate with 4 degrees of freedom (significant at $p<0.001$). However, there is a clear distinction in the tests on the individual coefficients. In the spatial lag model, the coefficients of DENSITY and INCOME are found to be strongly different (test values of 15.01 and 11.63 respectively, both significant at $p<0.001$ for a χ^2 variate with one degree of freedom), and the coefficient of RURAL turns out to be weakly so (3.57, significant at $p<0.06$). In contrast, the results for the heteroskedastic model are the same as for OLS, with test values of 4.28 ($p<0.03$) for DENSITY, 0.04 ($p<0.85$) for RURAL and 0.05 ($p<0.82$) for INCOME. While there is a slight edge for the heteroskedastic model in terms of likelihood (-118.8 vs. -119.0), both models still suffer from specification problems. On the one hand, a test for groupwise heteroskedasticity in the spatial lag model yields 9.05, which is significant at $p<0.01$ for a χ^2 variate with one degree of freedom. On the other hand, a LM test for a spatial lag in the h~teroskedastic model (see Anselin, 1988a) gives a value of 9.61, also significant at $p<0.01$ for a χ^2 variate with one degree of freedom.[42] Clearly, a more satisfactory model should combine the two, i.e., should have a spatial lag and groupwise heteroskedasticity, but its implementation is beyond the scope of the current paper.

Spatial Expansion Model

The expansion method, originally suggested by Casetti (1972) for applications in geography is a way to deal with continuous drift in the regression coefficients. In its early forms, the approach was primarily a way to implement a form of varying coefficients, where the variation is in function of a set of auxiliary variables. Recently, it has been extended to become a general framework for model development (e.g., Casetti, 1986; Casetti and Jones, 1988; Jones and Casetti, 1992).

A straightforward implementation of a spatial expansion model is to specify a trend surface for each regression coefficient. For example, a second order trend surface for a coefficient β_k, associated with an explanatory variable z_k would be defined as:

$$\beta_k = \gamma_0 + \gamma_1 x + \gamma_2 y + \gamma_3 x^2 + \gamma_4 y^2 + \gamma_5 xy$$

where x and y are the coordinates for the observations. A substitution of this expression in the original specification (the initial model, in the terminology of the expansion method) would then yield five additional regression coefficients, i.e.,,γ_1 through γ_5 in:

[41] The statistic yields 5.88, which for an F(1,40) variate is significant at $p<0.02$. The values for RURAL and INCOME are respectively 0.02 and 0.05, which are clearly not significant ($p<0.90$).

[42] A LM test for spatial error autocorrelation in the heteroskcedastic model (see Anselin, 1988a) does not give a strong indication (2.53 for a χ^2 variate with one degree of freedom, $p<0.1$ 1).

$$\beta_k z_k = (\gamma_0 + \gamma_1 x + \gamma_2 y + \gamma_3 x^2 + \gamma_4 y^2 + \gamma_5 xy) z_k$$

or

$$\beta_k z_k = \gamma_0 z_k + \gamma_1 (x z_k) + \gamma_2 (y z_k) + \gamma_3 (x^2 z_k) + \gamma_4 (y^2 z_k) + \gamma_5 (xy z_k)$$

The resulting model can be estimated by means of standard regression techniques, and spatial effects may be incorporated as discussed before.

A linear expansion of the coefficients in the original freezer adoption example is illustrated in Table 15. The explanatory variables are listed on the left hand side and the expansion variable are shown at the top. The first column of estimates gives the results for the non-varying coefficients (e.g., the γo in the equation above), the second column gives the estimates for the coefficients of the original variable multiplied by the x coordinate of the observations, and the third column does the same for the y coordinate. Of the initial coefficients, only the one for RURAL remains significant (t value of 2.37, $p<0.03$). In contrast, only two expanded coefficients are not or only weakly significant: INCOME times X (t value of -1.95, $p<0.06$) and RURAL times Y (t value of 1.11, $p<0.28$). All the others are strongly significant. The signs of the coefficients show an interesting spatial pattern: the coefficients of RURAL and INCOME both decrease from west to east (smallest X are for the west), and from north to south (smallest Y are for the south), while the coefficient of DENSITY increases along the same dimensions. Clearly, more complex spatial patterns could be discovered by means of higher order expansions.

The spatial expansion model has by far the best fit of all specifications considered in this paper, with an adjusted R_2 of 0.90 and a likelihood of -100.9. It also fails to suffer from either heteroskedasticiy or spatial dependence.[43] However, a major problem with this approach is the high degree of multicollinearity, as evidenced by a condition number of 73.8, which is way beyond the acceptable level. As a consequence, the indication of significance and the measure of fit may have to be interpreted with caution.[44]

CONCLUSION: DO SPATIAL EFFECTS MATTER?

How do the substantive conclusions of Ormrod's original paper stand up after the introduction of spatial effects in this paper, or, in other words, do the spatial effects matter? In Table 16, I summarize some salient features of the original model, as well as of seven of the various spatial specifications implemented here: the original model with a spatial lag (DISTANCE_2. in Table 10); the extended model, with a dummy variable for WEST (Table 13); the extended model with spatial error dependence (Table 13); the three specifications for spatial regimes (OLS, LAG and HET in Table 14); and the spatial expansion model (Table 15).

I consider four aspects of the models: evidence of ignored spatial dependence (in the error, or in the form of a spatial lag); evidence of ignored heteroskedasticity; goodness-of-fit according to the log likelihood; and goodness-of-fit according to the Akaike information criterion, or AIC. The latter corrects the likelihood for the number of explanatory variables that are included in the model, similar to an adjusted R^2 As a result, the AIC represents a compromise between parsimony, and fit, and is a more reliable indicator when comparing models with different numbers of explanatory variables, as is the case here. The model with the lowest AIC is best.[45]

From the information in Table 16, it is clear that the model suggested by Ormrod was a rather preliminary specification: it suffers from both types of ignored spatial dependence, and has by far the poorest fit of all eight models (likelihood of -131.5 and AIC of 271.0). Three of the alternative models pass both diagnostics for spatial dependence and heteroskedasticity: the original model with a spatial lag, the extended model with a spatial autoregressive error, and the expansion model. Of these, the expansion model has by far the best fit, according to AIC (239.8) as well as likelihood (109.9). The extended error model is next best, placing second on the AIC (253.0) and fourth on the likelihood (- 121.5) criteria. The lag model for the original specification has a somewhat lower fit, rating fourth on AIC (254.7) and fifth on likelihood (-122.3). As mentioned above, the models with spatial regimes and either a spatial lag or heteroskedasticity should really be combined, and may then achieve a slight improvement in fit. As it stands, the latter rates third in terms of AIC (253.5), while the former rates fifth (255.9). In sum, it seems that the three most promising specfications would

[43] A Breusch-Pagan test based on the original variables in the model yields 2.73, which is clearly not significant for a χ^2 variate with three degrees of freedom ($p<0.44$). For an inverse distance weights matrix with a cut-off, the z-value for Moran's I is 1.3 1, the LM error statistic is 0.003 and the LM lag statistic is 1.42, neither of which is significant. The same holds for the other weights matrices as well.

[44] The multicollinearity intrinsic to the construction of the expansion method may be remedied by the use of principal components, as outlined in Casetti and Jones (1988).

[45] For a technical discussion of the use of AIC in spatial modelling, see Anselin (1988b).

be the expansion model, the extended model with a spatial autoregressive error, and a spatial lag model with two coefficient regimes as well as groupwise heteroskedasticity.

The substantive interpretation of these three types of models is quite different. In the expansion model, there is no role for spatial interaction, between households in different states to explain the adoption rates: only the "true" explanatory variables

DENSITY, RURAL and INCOME matter, as in the original Ormrod article. However, quite in contrast with Ormrod's assertion, the strength of this relationship shows a definite spatial pattern and is by no means constant over the landscape. Why this pattern is such a regular function of absolute location, and what possible socio-economic mechanisms may be behind it remains to be determined. In the extended spatial error model, there is spatial interaction, but it occurs only in the form of a spatial spill-over in unknown ignored variables. There is also a clear dichotomy between the average adoption rate in the west and east, but a reason for this, or a possible additional explanatory variable cannot be readily suggested. Finally, in the model that would combine a spatial lag, spatial regimes and groupwise heteroskedasticity, there is a role for explicit spatial interaction (following a truncated gravity specification), but the strength of the explanatory power of the other variables varies between west and east. Similarly, the variance of the error, or the importance of the ignored influences is clearly different between the two regions. Again, this is only a partial explanation, since the mechanisms that elicit such a clear dichotomy between west and east would still need to be discovered.

The upshot of this exploration into spatial regression analysis is that a richness of spatial patterns was found, which was ignored in the original paper. As a consequence, the substantive interpretation offered in the Ormrod article was shown to be incomplete, and to some extent also incorrect. The spatial patterns that are discovered by means of the data analysis techniques, and their visualization by means of a GIS raise new questions about the social mechanisms behind the phenomenon studied. The combination of spatial data analysis and GIS thus provides insights that remain hidden in a traditional analysis, and therefore should be an important tool in social science research.

REFERENCES

Anselin, Luc (1986). Non-nested tests on the weight structure in spatial autoregressive models: some Monte Carlo results, Journal of Regional Science 26, 267-84.

Anselin, Luc (1988a). Spatial Econometrics, Methods and Models (Dordrecht, Kluwer Academic).

Anselin, Luc (1988b). Model validation in spatial econometrics: a review and evaluation of alternative approaches, International Regional Science Review 11, 279-316.

Anselin, Luc (1988c). Lagrange Multiplier test diagnostics for spatial dependence and spatial heterogeneity, Geographical Analysis 20, 1-17.

Anselin, Luc (1990a). What is special about spatial data? Alternative perspectives on spatial data analysis. In DA. Griffith (Ed.), Spatial Statistics, Past, Present and Future, pp. 63-77 (Ann Arbor, ML, Institute of Mathematical Geography).

Anselin, Luc (1990b). Spatial dependence and spatial structural instability in applied regression analysis, Journal of Regional Science 30, 185-207.

Anselin, Luc (1990c). Some robust approaches to testing and estimation in spatial econometrics, Regional Science and Urban Economics 20, 141-63.

Anselin, Luc (1992a). Space and applied econometrics: introduction, Regional Science and Urban Economics 22, 307-16.

Anselin, Luc (1992b).. SpaceStat, a program for the statistical analysis of spatial data (Santa Barbara, CA, National Center for Geographic Information and Analysis).

Anselin, Luc (1992c). Discrete space autoregressive models. In M.F. Goodchild, B. Parks and L.T. Steyaert (Eds.), GIS and Environmental Modeling (Oxford, Oxford University Press) (in press).

Anselin, Luc (1992d). SpaceStat tutorial: a workbook for using SpaceStat in the analysis of spatial data (Santa Barbara, CA, National Center for Geographic Information and Analysis).

Anselin, Luc and Arthur Getis (1992). Spatial statistical analysis and geographic information systems, The Annals of Regional Science 26, 19-33.

Anselin, Luc and Daniel A. Griffith (1988). Do spatial effects really matter in regression analysis? Papers, Regional Science Association 65, 11-34.

Anselin, Luc and Daniel A. Griffith (1993). Operational Methods of Spatial Data Analysis (Oxford, Oxford University Press) (forthcoming).

Anselin, Luc and Sheri Hudak (1992). Spatial econometrics in practice: a review of software options, Regional Science and Urban Economics 22, 509-36.

Anselin, Luc and John O'Loughlin (1992). Geography of international conflict and cooperation: spatial dependence and regional context in Africa. In MD. Ward (FA), The New Geopolitics, pp. 39-75 (London, Gordon and Breach).

Anselin, Luc and Serge Rey (1991). Properties of tests for spatial dependence in linear regression models, Geographical Analysis 23, 112-3 1.

Anselin, Luc, Rustin Dodson and Sheri Hudak (1992). Linking GIS and spatial data analysis in practice. Paper Presented at the 32nd European Congress of the Regional Science Association International, Brussels, Belgium.

Anselin, Luc, Sheri Hudak and Rustin Dodson (1993). Spatial data analysis and GIS: interfacing GIS and econometric software (Santa Barbara, CA, National Center for Geographic Information and Analysis) (forthcoming).

Arbia, G. (1989). Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems (Dordrecht, Kluwer Academic).

Bailey, Trevor (1992). Statistical analysis and geographic information systems: a review of the potential and progress in the state of the art. EGIS 92, Proceedings of the Third European Conference on Geographic Information Systems, pp. 186-203 (Utrwht, EGIS Foundation).

Belsley, D., E. Kuh and R. Welsch (1980). Regression Diagnostics, Identifying Influential Data and Sources of Collinearity (New York, Wiley).

Bennett, R. (1979). Spatial Time Series (London, Pion).

Bennett, R., R. Haining and D. Griffith (1984). The problem of missing data on spatial surfaces, Annals, Association of American Geographers 74, 138-56.

Bivand, R. (1984). Regression modelling with spatial dependence: an application of some class selection and estimation methods, Geographical Analysis 16, 25-37.

Bivand, R. (1992). Systat compatible software for modelling spatial dependence among observations, Computers and Geosciences (forthcoming).

Blommestein, Hans (1985). Elimination of circular routes in spatial dynamic regression equations, Regional Science and Urban Economics 13, 251-70.

Bolduc, D., R. Laferri&e and G. Santarossa (1992). Spatial autoregressive error components in travel flow models, Regional Science and Urban Economics 22, 371-85.

Boots, Barry N. and Arthur Getis (1988). Point Pattern Analysis (Newbury Park, Sage Publications).

Box, G. and G. Jenkins (1976). Time Series Analysis, Forecasting and Control (San Francisco, Holden Day).

Breusch, T. and A. Pagan (1979). A simple test for heteroskedasticity and random coefficient variation, Econometrica 47, 1287-94.

Bronars, S. and D. Jansen (1987). The geographic distribution of unemployment rates in the US, Journal of Econometrics 36, 251-79.
Burridge, P. (1980). On the Cliff-Ord test for spatial autocorrelation, Journal of the Royal Statistical Society B 42, 107-8.
Burridge, P. (1981). Testing for a common factor in a spatial autoregressive model, Environment and Planning A 13,795-800.
Can, Ayse (1992). Residential quality assessment: alternative approaches using GIS, The Annals of Regional Science 26, 97-110.
Case, Anne (1991). Spatial patterns in household demand, Econometrica 59,953-965.
Case, Anne (1992). Neighborhood influence and technological change, Regional Science and Urban Economics 22,491-508.
Casetti, Emilio (1972). Generating models by the expansion method: applications to geographical research, Geographical Analysis 4, 81-9 1.
Casetti, Emilio (1986). The dual expansion method: an application for evaluating the effects of population growth on development, IEEE Transactions on Systems, Man, and Cybernetics SMC- 16, 29-39.
Casetti, Emilio and J.P. Jones (1988). Spatial parameter variation by orthogonal trend surface expansions: an application to the analysis of welfare program participation rates, Social Science Research 16, 285300.
Cliff, A.D. and J.K. Ord (1972). Testing for spatial autocorrelation among regression residuals, Geographical Analysis 4, 267-84.
Cliff, A.D. and JK Ord (1973). Spatial Autocorrelation (London, Pion).
Cliff, AD. and J.K. Ord (198 1). Spatial Processes, Models and Applications (London, Pion).
Cook, D. and S. P&ock (1983). Multiple regression in geographical mortality studies, with allowance for spatially correlated effors, Biometrics 39, 361-71.
Cook, R., D. and S. Weisberg (1982). Residuals and Influence in Regression (New York, Chapmann and Hall).
Cressie, Noel (199 1). Statisticsfor Spatial Data (New York, Wiley).
Davis, John C. (1986). Statistics and Data Analysis in Geology (2nd Ed) (New York, Wiley).
Diggle, P. (1983). Statistical Analysis of Spatial Point Patterns (New York, Academic Press).
Ding, Yuemin and A. Stewart Fotheringham (1992). The integration of spatial analysis and GIS, Computers, Environment and Urban Systems 16, 3-19.
Doreian, P., K. Teuter and C-H Wang (1984). Network autocorrelation models, Sociological Methods and Research 13, 155-200.
Dubin, Robin A. (1992). Spatial autocorrelation and neighborhood quality, Regional Science and Urban Economics 22,433-52.
Eastman, Ron (1992). Idrisi, Version 4.0 (Worcester, IU, Clark University, Graduate School of Geography).
Ebdon, David (1985). Statistics in Geography (2nd Ed) (Oxford, Basil Black-well).
Environmental Systems Research Institute, Inc. (1991a). Arc/Info Data Model, Concepts and Key Terms (Redlands, CA, ESRI).
Environmental Systems Research Institute, Inc. (1991b). Cell-based modeling with GRID: analysis, display and management (Redlands, CA, ESRI).
Fischer, Manfred M. and Peter Nijkamp (1992). Geographic information systems and spatial analysis, The Annals of Regional Science 26, 3-17.
Florax, Raymond and Henk Folmer (1992). Specification and estimation of spatial linear regression models: Monte Carlo evaluation of pre-test estimators, Regional Science and Urban Economics 22, 405-32.
Foster, S.A. and W.L. Gorr (1986). An adaptive filter for estimating spatially-varying parameters: application to modeling police hours in response to calls for service, Management Science 32, 878-89.
Geary, R. (1954). The contiguity ratio and statistical mapping, The Incorporated Statistician 5, 11545.
Getis, Arthur (1990). Screening for spatial dependence in regression analysis, Papers, Regional Science Association 69, 69-8 1.
Getis, Arthur and Barry N. Boots (1978). Models of Spatial Processes (Cambridge, Cambridge University Press).
Getis, Arthur and J. Keith Ord (1992). The analysis of spatial association by use of distance statistics, Geographical Analysis 24, 189-206.
Godfrey, L.G. (1988). Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches (Cambridge, Cambridge University Press).
Goodchild, Michael F. (1987). A spatial analytical perspective on geographical information systems, International Journal of Geographical Information Systems 1, 327-34.
Goodchild, Michael F. (1992). Geographical data modeling, Computers and Geosciences (in press).
Goodchild, M.F., R. Haining and S. Wise (1992). Integrating GIS and spatial data analysis: problems and possibilities, International Journal of Geographic Information Systems 6 (in press).
Griffith, Daniel A- (1978). A spatially adjusted ANOVA model, Geographical Analysis 10, 296-301.
Griffith, Daniel A. (1987). Spatial Autocorrelation, A Primer (Washington, D.C., Association of American Geographers).
Griffith, Daniel A. (1988a). Advanced Spatial Statistics (Dordrecht, Kluwer Academic).
Griffith, Daniel A. (1988b). Estimating spatial autoregressive model parameters with commercial statistical packages, Geographical Analysis 20, 176-86.
Griffith, Daniel A. (1990). Spatial Statistics, Past, Present and Future (Ann Arbor, Ml., Institute of Mathematical Geography).
Griffith, Daniel A. (1992). A spatially adjusted N-way ANOVA model, Regional Science and Urban Economics 22, 347-69.
Griffith, Daniel A. and Carl G. Amrhein (1991). Statistical Analysis for Geographers (Englewood Cliffs, NJ, Prentice-Hall).
Griffith, D., R. Bennett and R. Haining (1989). Statistical analysis of spatial data in the presence of missing observations: a methodological guide and an application to urban census data, Environment and Planning A 21, 1511-23.

Griffith, D., R. Lewis, B.Li, I. Vasiliev, S. McKnight and X Yang (1990). Developing Minitab software for spatial statistical analysis: a tool for education and research, The Operational Geographer 8, 28-33.
Haining, Robert (1984). Testing a spatial interacting-markets hypothesis, Review of Economics and Statistics 66, 576-83.
Haining, Robert (1987). Trend surface analysis with regional and local scales of variation, with an application to areal survey data, Technometrics 29, 461-9.
Haining, Robert (1989). Geography and spatial statistics: current positions, future developments. In Bill Macmillan (Ed.), Remodelling Geography, pp. 191-203 (Oxford, Basil Blackwell).
Haining, Robert (1990). Spatial Data Analysis in the Social and Environmental Sciences (Cambridge, Cambridge University Press).
Haining, R., D. Griffith and R. Bennett (1994). A statistical approach to the problem of missing data using a first order markov model, Professional Geographer 36, 338-48.
Haslett, J., G. Wills, A. Unwin (1990). SPIDER - an interactive statistical tool for the analysis of spatially distributed data, International Journal of Geographical Information Systems 4, 285-96.
Haslea, J., R. Bradley, P. Craig, A- Unwin and G. Wills (1991). Dynamic graphics for exploring spatial data with application to locating global and local anomalies, The American Statistician 45, 234-42.
Hendry, D. and G. Mizon (1978). Serial correlation as a convenient simplification, not a nuisance: a comment on a study of the demand for money by the Bank of England, Econornic Journal 88, 549-563.
Hooper, P., and G. Hewings (1981). Some properties of space-time processes, Geographical Analysis 13, 203-223.
Hubert, L. (1985). Combinatorial data analysis: association and partial association, Psychometrica 50, 449-67.
Hubert, L. (1987). Assignment Methods in Combinatorial Data Analysis (New York: Marcel Dekker).
Hubert, L., R. Golledge and C.M. Costanzo (1981). Generalized procedures for evaluating spatial autocorrelation, Geographical Analysis 13, 224-33.
Hubert L., R. Golledge, C.M. Costanzo and N. Gale (1985). Measuring association between spatially defined variables: an alternative procedure, Geographical Analysis 17, 36-46.
Hudak, Sheri (1992). Spatial Econometrics in Practice. Unpublished Master's Thesis, Department of Geography, University of California, Santa Barbara.
Isaaks, Edward H. and R. Mohan Srivastava (1989). An Introduction to Applied Geostatistics (Oxford, Oxford University Press).
Jones, J.P. and E. Casetti (1992). Applications of the Expansion Method (London, Routledge).
Kehris, E. (1990). Spatial autocorrelation statistics in Arc/Info. North West Regional Research Laboratory, Research Report No. 16, Lancaster University.
Kelejian, Harry and Dennis P. Robinson (1992). Spatial autocorrelation: a new computationally simple test with an application to per capita county police expenditures, Regional Science and Urban Economics 22,317-31.
Kiefer, N. and IVL Salmon (1983). Testing normality in econometric models, Economics Letters 11, 123-8.
Kvamme, Kenneth, L. (1990). Spatial autocorrelation and the classic Maya collapse revisited: Refined techniques and new conclusions, Journal of Archaelogical Science 17, 197-207.
Legendre, P., N. Oden, R. Sokal, A. Vaudour, J. Kim (1990). Approximate analysis of variance of spatially autocorrelated regional data, Journal of Classification 7, 53-75.
Loftin, C. and S. Ward (1983). A spatial autocorrelation model of the effects of population density on fertility, American Sociological Review 48, 121-8.
MacKinnon, J. and H. White (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties, Journal of Econometrics 29, 305-25.
Moran, P. (1948). The interpretation of statistical maps, Journal of the Royal Statistical Society B 10, 24351.
Nass, C. and D. Garfinkle (1992). Localized autocorrelation diagnostic statistic (LADS) for spatial models: conceptualization, utilization, and computation, Regional Science and Urban Economics 22, 333-46.
National Research Council (1991). Spatial Statistics and Digital Image Analysis (Washington, D.C., National Academy Press).
Odland, John (1988)., Spatial Autocorrelation (Newbury Park, Sage Publications).
O'Loughlin, John (1986). Spatial models of international conflicts: extending current theories of war behavior, Annals, Association of American Geographers 76, 63-80.
O'Loughlin, John and Luc Anselin (1991). Bringing geography back to the study of international relations: spatial dependence and regional context in Africa, 1966-1978, International Interactions 17, 29-61.
O'Loughlin, John and Luc Anselin (1992). Geography of international conflict and cooperation: theory and methods. In M.D. Ward (Ed.), The New Geopolitics, pp. 11-38 (London, Gordon and Breach).
Openshaw, Stan (1990). Spatial analysis and geographical information systems: a review of progress and possibilities. In H. Scholten and J. Stillwell (Eds.), Geographical Information Systems for Urban and Regional Planning, pp. 153-63 (Dordrecht, Kluwer Academic).
Openshaw, S. and P. Taylor (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In N. Wrigley and R. Bennett (Eds.), Statistical Applications in the Spatial Sciences, pp. 127-44 (London, Pion).
Ord, J. Keith (1975). Estimation methods for models of spatial interaction, Journal of the American Statistical Association 70,120-6.
Ormrod, Richard R. (1990). Local context and innovation diffusion in a well-connected world, Economic Geography 66, 109-122.

Pfeifer, P. and S. Deutsch (1980a). A STARIMA model-building procedure with applications to description and regional forecasting, Transactions of the Institute of British Geographers 5, 330-49.
Pfeifer, P. and S. Deutsch (1980b). A three-stage iterative procedure for space-time modeling, Technometrics 22, 35-47.
Pfeifer, P. and S. Deutsch (1980c). Identification and interpretation of first order space-time ARIMA models, Technometrics 22,397-408.
Ripley, B. (198 1). Spatial Statistics (New York, Wiley).
Scheffe, H. IL (1959). The Analysis of Variance (New York, Wiley).
Stoffer, D. (1986). Estimation and identification of space-time ARMAX models in the presence of missing data, Journal of the American Statistical Association 81, 762-72.
Tobler, Waldo (1979). Cellular geography. In S. Gale and G. Olsson (Eds.), Philosophy in Geography, pp. 379-86 (Dordrecht, Reidel).
Tobler, Waldo (1989). Frame independent analysis. In M. Goodchild and S. Gopal (Eds.), The Accuracy of Spatial Databases, pp. 115-22 (London, Taylor and Francis).
Unwin, D. and N. Wrigley (1987a). Control point distribution in trend surface modelling revisited: an application of the concept of leverage, Transactions, Institute of British Geographers 12, 147-60.
Unwin, D. and N. Wrigley (1987b). Towards a general theory of control point distribution effects in trend surface models, Computers in Geosciences 13, 351-5.
Upton, G. and B. Fingleton (1985). Spatial Data Analysis by Example (Vol. 1) (New York, Wiley).
Webster, R. and MA. Oliver (1990). Statistical Methods in Soil and Land Resource Survey (Oxford, Oxford University Press).
White, D., M. Burton and M. Dow (1981). Sexual division of labor in African agriculture: a network autocorrelation analysis, American Anthropologist 84, 824-49.
Whitley, D.S. and W.A.V. Clark (1985). Spatial autocorrelation tests and the Classic Maya collapse: methods and inferences, Journal of Archaelogical Science 12, 377-95.

FIGURE 1: Freezer Variable and its Spatial Lag

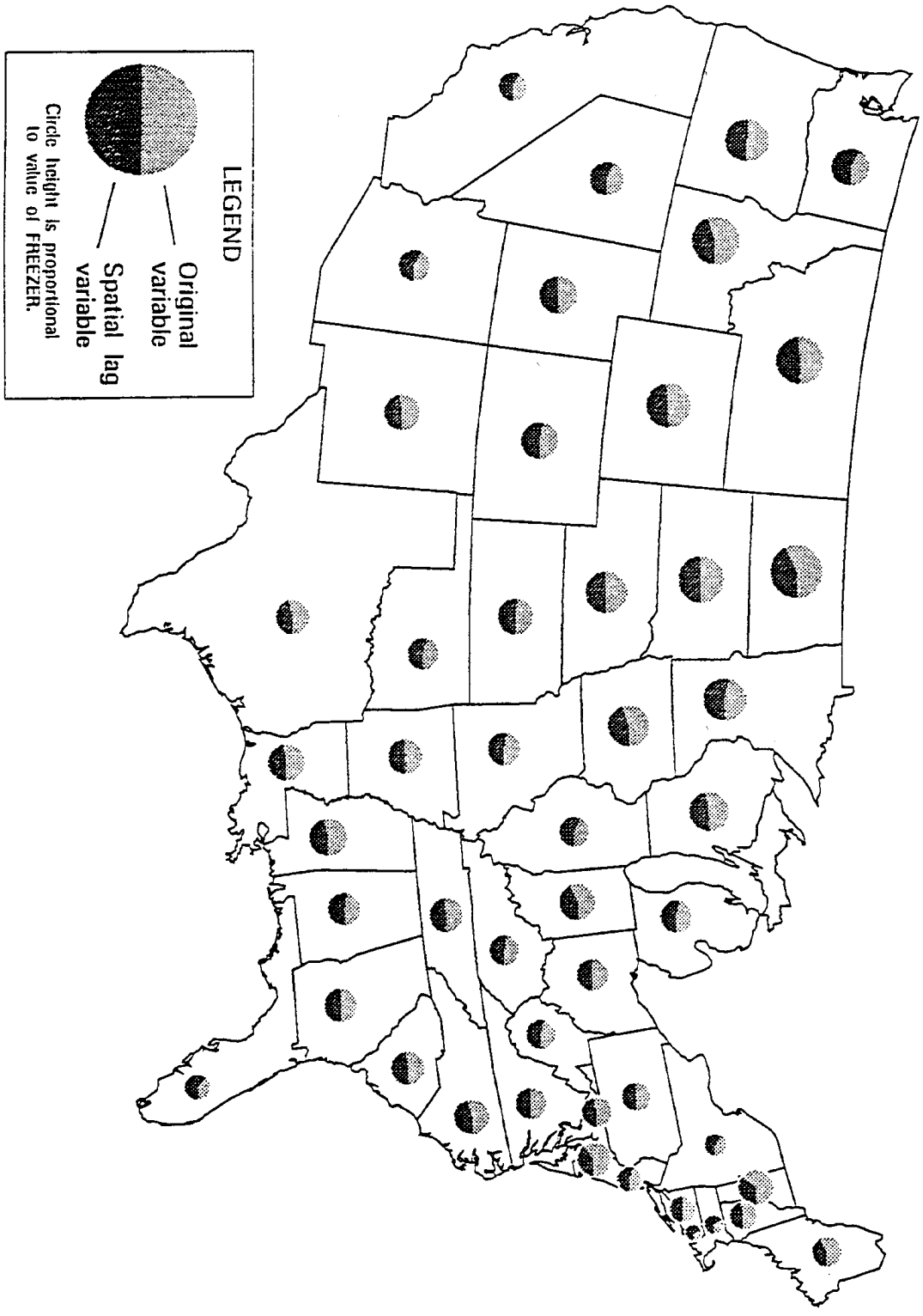
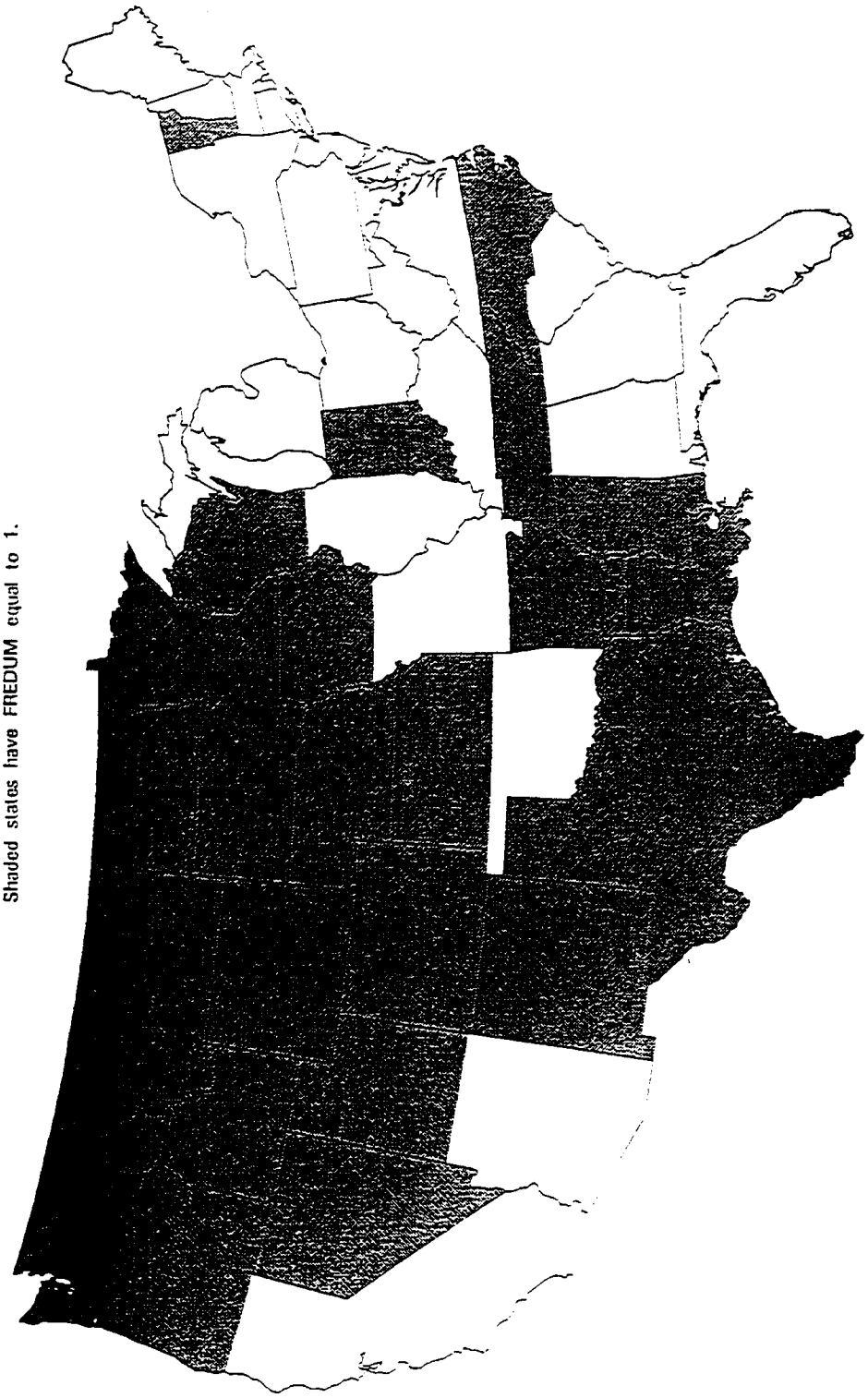


FIGURE 2: Freezer Adoption as a Dummy Variable



Shaded states have FREDUM equal to 1.

FIGURE 3: G_i^* Statistics of Spatial Association

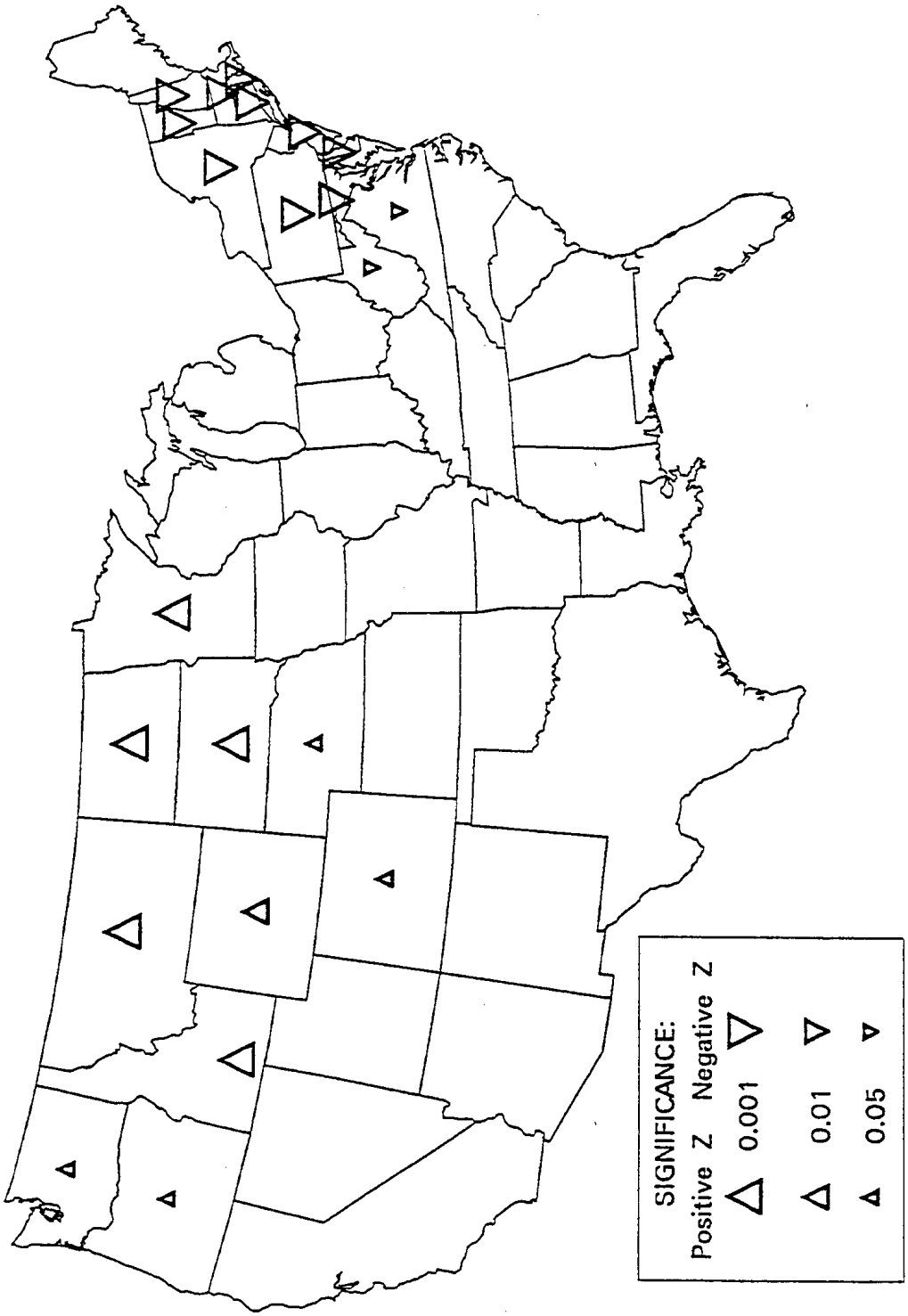


FIGURE 4: Regression Residuals

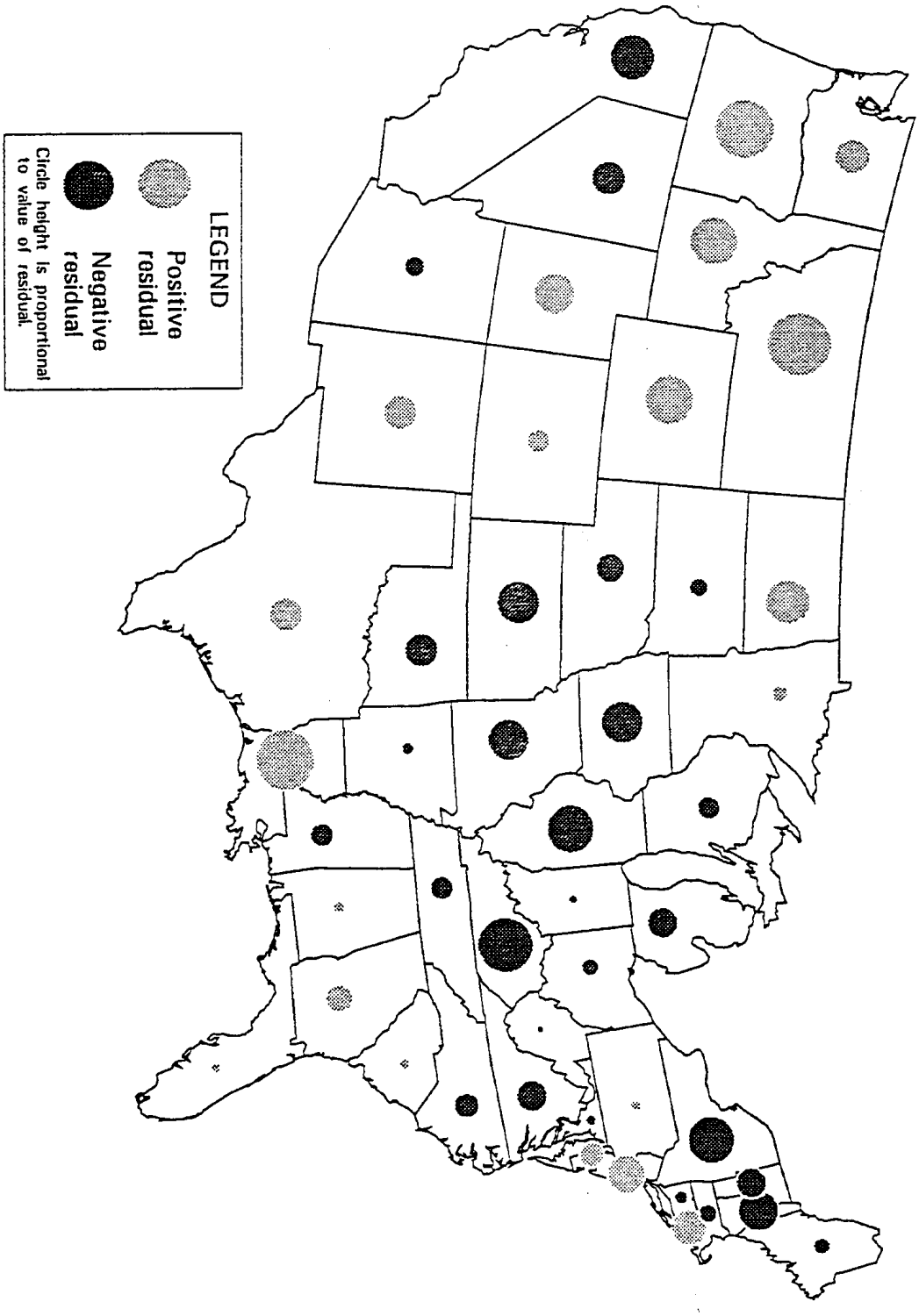


TABLE 1. Data for the Empirical Illustration

STATE	NO.	FREEZ	DENSITY	RURAL	INCOME	X	Y	FREDUM	WEST
AL	1	19.3	64.0	12.3	3.937	33.52	11.21	0	0
AZ	2	18.0	11.5	3.8	5.568	10.76	13.64	0	1
AR	3	21.6	34.0	18.6	3.184	28.24	13.22	1	1
CA	4	16.0	100.4	2.1	6.726	4.56	18.27	0	1
CO	5	24.7	16.9	7.3	5.780	16.83	18.15	1	1
CT	6	13.1	517.5	1.0	6.887	44.06	23.04	0	0
DE	7	20.9	225.6	4.9	6.197	42.46	19.62	0	0
FL	8	12.7	91.3	2.1	4.722	38.21	7.16	0	0
GA	9	20.1	67.7	10.3	4.208	36.66	11.41	0	0
ID	10	38.7	8.1	19.9	5.259	10.29	25.17	1	1
IL	11	16.8	180.3	5.6	6.566	30.74	19.12	0	0
IN	12	24.1	128.9	10.4	5.798	33.21	19.17	1	0
IA	13	30.6	49.2	24.0	5.069	27.05	21.23	1	1
KS	14	23.2	26.6	14.7	5.295	22.95	17.21	1	1
KY	15	17.1	76.2	18.0	4.051	34.35	16.63	0	0
LA	16	24.1	72.2	7.2	4.272	28.79	8.99	1	1
ME	17	16.1	31.3	5.0	4.873	45.68	27.80	0	0
MD	18	16.7	314.0	3.6	6.309	41.42	19.39	0	0
MA	19	7.3	654.5	0.7	6.272	44.60	23.91	0	0
MI	20	19.2	137.2	5.6	6.256	34.12	23.30	0	0
MN	21	31.8	42.7	17.2	5.573	26.31	25.89	1	1
MS	22	24.9	46.1	24.9	2.884	30.89	10.93	1	0
MO	23	20.2	62.5	12.5	5.127	28.06	17.09	0	1
MT	24	38.8	4.6	15.6	5.403	14.62	27.47	1	1
NE	25	30.5	18.4	21.9	4.862	21.86	20.65	1	1
NV	26	21.1	2.6	3.5	6.736	7.51	20.03	1	1
NH	27	13.7	67.3	3.1	5.636	44.40	25.49	0	0
NJ	28	11.7	806.7	0.8	6.786	42.85	21.11	0	0
NM	29	23.4	7.8	6.1	5.371	15.80	13.09	1	1
NY	30	9.6	350.1	1.9	6.371	41.43	24.00	0	0
NC	31	21.9	92.9	17.7	3.956	39.83	15.19	1	0
ND	32	46.0	9.1	32.3	4.530	21.61	27.25	1	1
OH	33	18.6	236.9	5.4	6.171	36.09	19.96	0	0
OK	34	18.5	33.8	11.1	4.620	23.66	13.95	0	1
OR	35	32.5	18.4	7.8	5.892	5.55	25.79	1	1
PA	36	15.3	251.5	3.1	5.719	40.12	21.31	0	0
RI	37	5.5	812.4	0.5	5.589	44.95	23.34	0	0
SC	38	20.7	78.7	14.7	3.821	38.79	13.19	0	0
SD	39	37.0	8.9	30.2	4.251	21.66	23.91	1	1
TN	40	21.0	85.4	16.4	3.949	33.63	14.66	1	0
TX	41	21.8	36.5	7.2	4.884	21.85	9.40	1	1
UT	42	26.0	10.8	4.9	5.899	11.68	19.19	1	1
VT	43	21.3	42.0	12.5	4.890	43.45	25.70	1	0
VA	44	18.2	99.6	10.0	4.964	39.89	17.43	0	0
WA	45	26.4	42.8	5.7	6.225	6.62	29.46	1	1
WV	46	16.0	77.3	6.5	4.572	38.20	18.41	0	0
WI	47	27.8	72.2	14.0	5.926	29.74	24.18	1	0
WY	48	35.1	3.4	13.0	5.877	15.67	22.79	1	1

TABLE 2. Contiguity Structures for 48 U.S. States

State		Contiguities (sequence numbers)
1	AL	3 9 15 16 22 38 40
		8 9 22 40
		3 15 16 23 31 38 44
		11 12 13 14 18 25 33 34 41 46
2	AZ	29 42
		4 5 26 29 42
		10 14 25 34 35 41 48
		3 13 16 23 24 39 45
3	AR	1 16 22 23 34 40
		16 22 23 34 40 41
		1 5 9 11 13 14 15 25 29 31 44
		2 8 12 18 21 33 38 39 42 46 47 48
4	CA	26
		2 26 35
		5 10 29 42 45
		14 24 25 34 41 48
5	CO	25 29 42 48
		2 14 25 29 34 42 48
		3 4 10 13 23 24 26 39 41
		11 15 16 21 22 32 35 40 45 47
6	CT	7 17 18 19 27 28 30 36 37 43
		19 30 37
		27 28 36 43
		7 17 18 33 46
7	DE	6 18 19 28 30 31 36 37 44 46
		18 28 36
		30 33 44 46
		6 12 15 19 20 31 40 43
8	FL	9
		1 9
		22 31 38 40
		3 15 16 23 44
9	GA	1 8 15 22 31 38 40
		1 8 31 38 40
		3 15 22 23 44
		11 12 13 14 16 18 25 33 34 41 46
10	ID	24 26 35 45 48
		24 26 35 42 45 48
		2 4 5 25 29 32 39
		13 14 21 23 34 41
11	IL	12 13 15 20 23 33 40 47
		12 13 15 23 47
		3 14 20 21 25 33 34 39 40 44 46
		1 5 9 16 18 22 24 29 31 32 36 41 48
12	IN	11 15 20 23 33 40 46
		11 15 20 33
		13 23 36 40 44 46 47
		1 3 7 9 14 18 21 22 25 28 30 31 34 39

TABLE 2 (continued)

13	IA	11 14 21 23 25 47
		11 21 23 25 39 47
		3 5 12 14 15 20 24 32 34 40 48
		1 2 9 10 16 22 29 31 33 41 42 44 46
14	KS	13 23 25 34
		5 23 25 34
		2 3 11 13 15 29 39 40 41 42 48
		1 4 9 10 12 16 21 22 24 26 31 32 33 44 46 47
15	KY	1 9 11 12 31 33 38 40 44 46
		11 12 23 33 40 44 46
		1 3 9 13 14 18 20 22 25 31 34 36 47
		5 7 8 16 21 28 29 30 38 39 41 48
16	LA	1 3 22
		3 22 41
		1 23 29 34 40
		2 5 8 9 11 13 14 15 25 31 42 44
17	ME	6 19 27 30 37 43
		27
		19 43
		6 30 37
18	MD	6 7 19 28 30 31 33 36 37 44 46
		7 36 44 46
		15 28 30 31 33 40
		1 3 6 9 11 12 19 20 22 23 38 43
19	MA	6 7 17 18 27 28 30 36 37 43
		6 27 30 37 43
		17 28 36
		7 18 33 46
20	MI	11 12 33 47
		12 33 47
		11 13 15 21 36 46
		7 18 23 25 28 30 32 39 40 44
21	MN	13 32 39 47
		13 32 39 47
		11 20 23 24 25 48
		3 5 10 12 14 15 33 34 40 42
22	MS	1 3 9 16 40
		1 3 16 40
		8 9 15 23 31 34 41 44
		5 11 12 13 14 18 25 29 33 38 46
23	MO	3 11 12 13 14 34
		3 11 13 14 15 25 34 40
		1 5 9 12 16 21 22 29 31 33 39 41 44 46 47 48
		2 8 10 18 20 24 32 36 38 42
24	MT	10 48
		10 32 39 48
		5 13 21 25 26 35 42 45
		2 4 11 14 23 29 34 47

TABLE 2 (continued)

25	NE	5 13 14 39
		5 13 14 23 39 48
		2 3 10 11 15 21 24 29 32 34 40 42 47
		1 4 9 12 16 20 22 26 31 33 35 41 44 45 46
26	NV	4 10 42
		2 4 10 35 42
		5 24 29 45 48
		14 25 32 34 39 41
27	NH	6 17 19 28 30 36 37 43
		17 19 43
		6 30 37
		28 36
28	NJ	6 7 18 19 27 30 36 37 43 44 46
		7 30 36
		6 18 19 33 43 46
		12 15 20 27 37 44
29	NM	2 5
		2 5 34 41 42
		3 4 10 14 16 23 25 26 48
		11 13 15 22 24 35 39 40 45
30	NY	6 7 17 18 19 27 28 36 37 43
		6 19 28 36 43
		7 18 27 33 37 46
		12 15 17 20 44
31	N	C 7 9 15 18 38 44 46
		9 38 40 44
		1 3 8 15 18 22 23 46
		7 11 12 13 14 16 25 33 34 36 41
32	ND	21 39
		21 24 39
		10 13 25 47 48
		5 11 14 20 23 26 35 42 45
33	OH	11 12 15 18 20 36 40 44 46
		12 15 20 36 46
		7 11 18 23 28 30 40 44 47
		1 3 6 9 13 14 19 21 22 25 31 34 43
34	OK	3 14 23 41
		3 5 14 23 29 41
		2 11 13 15 16 22 25 40 42 48
		1 4 9 10 12 21 24 26 31 33 39 44 46 47
35	OR	10 45
		4 10 26 45
		2 24 42 48
		5 25 29 32 39
36	PA	6 7 18 19 27 28 30 33 37 43 44 46
		7 18 28 30 33 46
		6 12 15 19 20 43 44
		11 23 27 31 37 40 47

TABLE 2 (continued)

37	RI	6 7 17 18 19 27 28 30 36 43
		6 19
		27 30 43
		17 28 36
38	SC	1 9 15 31 40 44 46
		9 31
		1 8 40 44
		3 15 18 22 23 46
39	SD	21 25 32
		13 21 24 25 32 48
		5 10 11 14 23 42 47
		2 3 12 15 20 26 29 34 35 40 45
40	TN	1 3 9 11 12 15 22 33 38 46
		1 3 9 15 22 23 31 44
		8 11 12 13 14 16 18 25 33 34 38 41 46
		5 7 20 21 29 36 39 47 48
41	TX	34
		3 16 29 34
		2 5 14 22 23 40 42
		1 4 9 10 11 13 15 25 26 31 44 48
42	UT	2 5 26 48
		2 5 10 26 29 48
		4 14 24 25 34 35 39 41 45
		3 13 16 21 23 32
43	VT	6 17 19 27 28 30 36 37
		19 27 30
		6 17 28 36 37
		7 18 33 46
44	VA	7 15 18 28 31 33 36 38 46
		15 18 31 40 46
		1 3 7 9 11 12 22 23 33 36 38
		8 13 14 16 20 25 28 30 34 41 47
45	WA	10 35
		10 35
		4 24 26 42 48
		2 5 25 29 32 39
46	WV	7 12 15 18 28 31 33 36 38 40 44
		15 18 33 36 44
		7 11 12 20 23 28 30 31 40
		1 3 6 9 13 14 19 22 25 34 38 43 47
47	WI	11 13 20 21
		11 13 20 21
		12 15 23 25 32 33 39
		3 5 14 24 34 36 40 44 46 48
48	WY	5 10 24 42
		5 10 24 25 39 42
		2 13 14 21 23 26 29 32 34 35 45
		3 4 11 15 40 41 47

First row is distance-based contiguity (DISTANCE_1), second through fourth rows are first throught third order of contiguity (CONTIG_1 through CONTIG_3)

TABLE 3. Spatially Lagged Variables

STATE	NO	FREEZ	CO-FREEZ	DI-FREEZ
AL	1	19.3	19.7	21.4
AZ	2	18.0	22.2	24.7
AR	3	21.6	21.8	21.3
CA	4	16.0	23.9	21.1
CO	5	24.7	25.0	28.8
CT	6	13.1	7.5	13.8
DE	7	20.9	14.6	13.5
FL	8	12.7	19.7	20.1
GA	9	20.1	19.1	19.7
ID	10	38.7	30.0	30.8
IL	11	16.8	24.0	22.3
IN	12	24.1	17.9	18.4
IA	13	30.6	27.4	25.1
KS	14	23.2	23.5	25.0
KY	15	17.1	19.3	19.7
LA	16	24.1	22.8	21.9
ME	17	16.1	13.7	11.8
MD	18	16.7	17.6	14.4
MA	19	7.3	12.6	14.4
MI	20	19.2	23.5	21.8
MN	21	31.8	35.4	35.4
MS	22	24.9	21.5	21.2
MO	23	20.2	22.4	22.5
MT	24	38.8	39.2	36.9
NE	25	30.5	28.5	28.9
NV	26	21.1	26.2	26.9
NH	27	13.7	14.9	12.5
NJ	28	11.7	15.3	14.3
NM	29	23.4	21.8	21.4
NY	30	9.6	13.7	14.2
NC	31	21.9	20.0	18.5
ND	32	46.0	35.9	34.4
OH	33	18.6	18.3	18.3
OK	34	18.5	22.5	21.7
OR	35	32.5	25.6	32.6
PA	36	15.3	15.6	14.4
RI	37	5.5	10.2	14.6
SC	38	20.7	21.0	19.1
SD	39	37.0	35.5	36.1
TN	40	21.0	20.4	19.9
TX	41	21.8	21.9	18.5
UT	42	26.0	26.8	24.7
VT	43	21.3	10.2	11.5
VA	44	18.2	18.5	17.7
WA	45	26.4	35.6	35.6
WV	46	16.0	17.2	18.7
WI	47	27.8	24.6	24.6
WY	48	35.1	32.6	32.1

TABLE 4. Moran's I Statistic of Spatial Autocorrelation

Weight	FREEZ	DENSITY	RURAL	INCOME
DISTANCE_1	0.693	0.441	0.539	0.590
	(7.62)	(5.22)	(5.95)	(6.44)
CONTIG_1	0.719	0.609	0.532	0.587
	(7.66)	(6.92)	(5.71)	(6.23)
CONTIG_2	0.451	0.317	0.266	0.340
	(6.39)	(4.85)	(3.88)	(4.83)
CONTIG_3	0.204	0.227	-0.02	-0.051
	(3.43)	(4.00)	(0.02)	(-0.44)

z-values in parentheses

TABLE 5. Geary's c Statistic of Spatial Autocorrelation

Weight	FREEZ	DENSITY	RURAL	INCOME
DISTANCE_1	0.291	0.589	0.433	0.416
	(-6.97)	(-3.61)	(-5.63)	(-5.91)
CONTIG_1	0.280	0.340	0.461	0.404
	(-7.03)	(-5.92)	(-5.29)	(-5.94)
CONTIG_2	0.492	0.634	0.679	0.636
	(-5.91)	(-3.51)	(-3.79)	(-4.47)
CONTIG_3	0.719	0.673	1.021	1.025
	(-3.53)	(-3.26)	(0.27)	(0.33)

z-values in parentheses

TABLE 6. G Statistic of Spatial Association

Weight	FREEZ	DENSITY	RURAL	INCOME
DISTANCE_1	0.109	0.444	0.137	0.131
	(-2.30)	(8.89)	(0.65)	(1.22)
CONTIG_1	0.109	0.180	0.140	0.095
	(3.55)	(3.59)	(4.98)	(-0.08)

z-values in parentheses

TABLE 7. G_i and G_i^* statistics

NO	STATE	$Z(G_i)$	$Z(G_i^*)$
1	AL	-0.235	-0.331
2	AZ	0.450	0.102
3	AR	-0.207	-0.213
4	CA	-0.123	-0.595
5	CO	1.690	1.683
6	CT	-3.578	-3.730
7	DE	-3.583	-3.524
8	FL	-0.254	-0.967
9	GA	-0.807	-0.846
10	ID	2.681	3.169
11	IL	0.082	-0.112
12	IN	-1.201	-1.064
13	IA	1.019	1.320
14	KS	0.739	0.738
15	KY	-1.028	-1.156
16	LA	-0.003	0.121
17	ME	-3.238	-3.298
18	MD	-3.486	-3.556
19	MA	-3.460	-3.730
20	MI	-0.056	-0.196
21	MN	3.412	3.578
22	MS	-0.199	-0.050
23	MO	0.136	0.052
24	MT	2.737	3.329
25	NE	1.770	2.039
26	NV	1.037	0.865
27	NH	-3.599	-3.734
28	NJ	-3.599	-3.776
29	NM	-0.105	0.009
30	NY	-3.503	-3.730
31	NC	-1.177	-1.128
32	ND	2.442	3.486
33	OH	-1.503	-1.571
34	OK	-0.091	-0.264
35	OR	1.879	2.258
36	PA	-3.712	-3.811
37	RI	-3.430	-3.730
38	SC	-0.997	-1.003
39	SD	3.173	3.585
40	TN	-0.884	-0.893
41	TX	-0.419	-0.319
42	UT	0.700	0.845
43	VT	-3.851	-3.734
44	VA	-1.750	-1.822
45	WA	2.353	2.258
46	WV	-1.523	-1.664
47	WI	0.680	0.919
48	WY	2.632	3.018

TABLE 8. Model Estimates

Variable	OLS	ERROR	LAG
W_FREEZ			0.408
			(3.97)
CONSTANT	-8.856	-1.529	-7.364
	(-1.87)	(-0.25)	(-1.92)
DENSITY	-0.0173	-0.0133	-0.0103
	(-5.04)	(-3.66)	(-3.20)
RURAL	0.932	0.787	0.676
	(9.35)	(6.89)	(6.47)
INCOME	4.433	3.292	2.779
	(5.75)	(3.37)	(3.83)
Lambda		0.637	
		(5.31)	
R2	0.78	0.79	0.85
Lik	-131.5	-125.2	-124.3

t-values and asymptotic t-values in parentheses
R2 is adjusted R2 for OLS and
squared correlation between predicted
and observed for ERROR and LAG models
Lik is maximized log likelihood

TABLE 9. Diagnostics for Spatial Error Autocorrelation

	CONTIG_1	CONTIG_2	CONTIG_3
Moran's I	0.335	0.042	-0.17.
Moran's z	4.38	1.20	-2.29
LM error	10.77	0.27	5.42
KR	17.75	1.89	11.42

Critical values for LM error (1 d.f.) are 3.84 for p=0.05
and 6.63 for p=0.01
Critical values for KR (4 d.f.) are 9.49 for p=0.05
and 13.28 for p=0.01

TABLE 10. Choice of Weights Matrix in Spatial Lag Model

Variable	CONTIG_1	DISTANCE_1	DISTANCE_2	DISTANCE_3
W_FREEZ	0.408	0.424	0.457	0.527
	(3.97)	(4.57)	(5.00)	(4.31)
CONSTANT	-7.364	-7.180	-7.797	-11.886
	(-1.92)	(-1.93)	(-2.15)	(-2.94)
DENSITY	-0.0103	-0.0110	-0.0092	-0.0081
	(-3.20)	(-3.70)	(-3.01)	(-2.20)
RURAL	0.676	0.669	0.649	0.747
	(6.47)	(6.61)	(6.50)	(7.91)
INCOME	2.779	2.712	2.689	3.008
	(3.83)	(3.90)	(4.01)	(4.29)
R2	0.85	0.86	0.87	0.85
Lik	-124.3	-122.9	-122.3	-125.5
LM	14.67	17.37	16.67	12.00
LR	14.46	17.24	18.36	12.10
W	15.78	20.90	25.03	18.56

R2 is squared correlation between predicted and observed
Lik is maximized log likelihood
LM, LR, and W are tests on the spatial lag
Critical values for LM, LR and W (1 d.f.) are 3.84 for p=0.05
and 6.63 for p=0.01

TABLE 11. Trend Surface Regression

Variable	Coefficient	t-OLS	t-Jacknife
CONSTANT	-4.739	-0.41	-0.48
X	1.933	5.46	6.19
Y	0.306	0.32	0.31
X2	-0.033	-6.20	-5.07
Y2	0.031	1.48	1.02
XY	-0.027	-2.41	-1.52
R2	0.71		
Lik	-136.8		

R2 is adjusted R2, Lik is maximized log likelihood

TABLE 12. Spatial Analysis of Variance

Variable	OLS	LAG
W_FREEZ		0.729
		(8.23)
CONSTANT	17.292	4.628
	(13.3)	(2.56)
WEST	10.253	2.963
	(5.33)	(2.03)
R2	0.37	0.73
Lik	-157.9	-143.6

R2 is adjusted R2 for OLS and
squared correlation between predicted
and observed for LAG model
Lik is maximized log likelihood

TABLE 13. Estimates in Extended Model

Variable	OLS	ERROR	LAG
W_FREEZ			0.399
			(3.76)
CONSTANT	-5.897	-4.374	-6.757
	(-1.30)	(-0.81)	(-1.82)
DENSITY	-0.0125	-0.0119	-0.0084
	(-3.46)	(-3.30)	(-2.65)
RURAL	0.850	0.813	0.653
	(8.75)	(7.97)	(6.55)
INCOME	3.596	3.378	2.580
	(4.64)	(3.80)	(3.79)
WEST	3.693	3.902	1.467
	(2.81)	(2.61)	(1.18)
Lambda		0.539	
		(4.13)	
R2	0.81	0.83	0.87
Lik	-127.5	-121.5	-121.6

R2 is adjusted R2 for OLS and
squared correlation between predicted
and observed for ERROR and LAG models
Lik is maximized log likelihood

TABLE 14. Spatial Regimes

Variable	OLS	LAG	HET
W_FREEZ		0.365	
		(3.41)	
EAST			
CONSTANT	-2.755	-6.274	-2.755
	(-0.41)	(-1.15)	(-0.68)
DENSITY	-0.0117	-0.0084	-0.0117
	(-2.94)	(-2.50)	(-4.89)
RURAL	0.773	0.645	0.773
	(4.17)	(4.08)	(6.92)
INCOME	3.086	2.616	3.086
	(2.88)	(2.94)	(4.79)
WEST			
CONSTANT	1.362	0.284	1.362
	(0.18)	(0.05)	(0.16)
DENSITY	-0.0885	-0.0685	-0.0885
	(-2.82)	(-2.70)	(-2.39)
RURAL	0.807	0.613	0.807
	(6.99)	(5.57)	(5.93)
INCOME	3.435	2.121	3.435
	(2.95)	(2.16)	(2.51)
R2	0.82	0.88	0.85
Lik	-124.0	-119.0	-118.8

R2 is adjusted R2 for OLS and
squared correlation between predicted
and observed for LAG and HET models
Lik is maximized log likelihood

TABLE 15. Spatial Expansion Model

Variable	Initial	X	Y
CONSTANT	7.917		
	(1.43)		
DENSITY	0.0233	0.0020	-0.0054
	(0.68)	(2.67)	(-3.86)
RURAL	0.861	-0.017	0.012
	(2.37)	(-2.31)	(1.11)
INCOME	0.999	-0.024	0.070
	(0.84)	(-1.95)	(2.36)
R2	0.90		
Lik	-109.9		

R2 is adjusted R2, Lik is maximized log likelihood
t-values are in parentheses

TABLE 16. Comparison of Model Characteristics

Model	DEP	HET	LIK	AIC
Original	Lag, Err	No	-131.5 (8)	271.0 (8)
Orig. Lag	No	No	-122.3 (5)	254.7 (4)
Extended	Err, Lag	Yes	-127.5 (7)	264.9 (7)
Ext. Error	No	No	-121.5 (4)	253.0 (2)
Regimes	Lag, Err	Yes	-124.0 (6)	263.9 (6)
Regimes Lag	No	Yes	-119.0 (3)	255.9 (5)
Regimes Het	Lag	N/A	-118.8 (2)	253.5 (3)
Expansion	No	No	-109.9 (1)	239.8 (1)

DEP is spatial dependence, HET is heteroskedasticity
LIK is log likelihood, AIC is Akaike Information Criterion
ranks are in parentheses