9-2020

Spatial Autocorrelation

John Odland

The Web Book of Regional Science

Sponsored by

Spatial Autocorrelation

By

John Odland

Scientific Geography

Series Editor:

Grant Ian Thrall

Sage Publications: 1988
Web Book Version: September, 2020

Web Series Editor: Randall Jackson
Director, Regional Research Institute
West Virginia University

<This page blank>

The Web Book of Regional Science is offered as a service to the regional research community in an effort to make a wide range of reference and instructional materials freely available online. Roughly three dozen books and monographs have been published as Web Books of Regional Science. These texts covering diverse subjects such as regional networks, land use, migration, and regional specialization, include descriptions of many of the basic concepts, analytical tools, and policy issues important to regional science. The Web Book was launched in 1999 by Scott Loveridge, who was then the director of the Regional Research Institute at West Virginia University. The director of the Institute, currently Randall Jackson, serves as the Series editor.

When citing this book, please include the following:

Odland, J. (1988). *Spatial Autocorrelation.* Reprint. Edited by Grant Ian Thrall. WVU Research Repository, 2020.

<This page blank>

<This page blank>

Contents

<This page blank>

SERIES EDITOR'S INTRODUCTION

Serial independence of data is assumed in many statistics, including ordinary least squares (regression analysis). If the order in which measurements are drawn from a population is unimportant, then the value that a measurement takes on should not be influenced by the position of the measurement in the draw. For some classes of problems, however, a measurement's value may be dependent upon the position in the draw, such as with time-series analyses where the value of measurement depends upon the value of other measurements one or more time periods previous. If interdependence exists between the values of the measurements (or their disturbance terms) serial autocorrelation is said to exist and the analysis may then be subject to error, including yielding biased estimates of the regression coefficients.

Now take the autocorrelation problem one more step. Say that in addition to the dependence being one directional and back in time, let the value of the measurement also be dependent upon values that measurements take on "down the road time." And then, in addition to serial dependence being behind and ahead, also let there be dependence to the right and left and every point of the compass in between. This, then is the problem of spatial autocorrelation.

Spatial autocorrelation statistics detect the presence of interdependence between data at neighboring locations and derive the effect upon the values of the measurements. Spatial autocorrelation statistics are the basic statistics for all data capable of being mapped.

Professor John Odland in this volume explains how to calculate and apply spatial autocorrelation statistics and illustrates how to use this class of statistics when the research hypotheses require data where the original spatial pattern of the measurements is retained.

Spatial autocorrelation statistics are important in dealing with the special problems that arise when other statistical models are applied to data that can be mapped, for such data often do not fulfill the conditions of serial independence. Spatial autocorrelation statistics measure the amount that the measurements depart from the requirements of independence.

Professor Odland demonstrates how spatial autocorrelation statistics can be used in diagnosing and correcting problems that arise when common statistical methods such as regression analysis are applied to spatially arrayed measurements. The statistics for spatial autocorrelation are also extended to the analysis of patterns in space and time, a development that makes it possible to investigate hypotheses about the processes that transform spatial patterns.

Because spatial analysis has long been central to the study of geography, geographers such as Professor Odland have been at the center of the research on spatial autocorrelation. At the same time, it may not be possible to delimit those disciplines that would find this material of use from those that would not other than by the state of evolution of the particular science, awareness of this class of statistics, and personality of the researchers. This is because though space may not be the central paradigm of a discipline, measurements that they make may be explicitly spatial. Hence, like time series analysis, spatial autocorrelation has no disciplinary boundaries.

This volume will be particularly useful to those disciplines that regularly draw upon the geographers' maps, such as geology, epidemiology, biological and ecological sciences, anthropology and archaeology, sociology, urban and regional sciences including city planning, and both human and physical geography.

-Grant Ian Thrall
Series Editor

<This page blank>

1 SPATIAL ANALYSIS AND STATISTICAL INFERENCE

The analysis of spatial distributions and the processes that produce and alter them is a central theme in geographic research and this volume is concerned with statistical methods for analyzing spatial distributions by measuring and testing for spatial autocorrelation. Spatial autocorrelation exists whenever a variable exhibits a regular pattern over space in which its values at a set of locations depend on values of the same variable at other locations. Spatial autocorrelation is present, for example, when similar values cluster together on a map. Spatial autocorrelation statistics make it possible to use formal statistical procedures to measure the dependence among nearby values in a spatial distribution, test hypotheses about geographically distributed variables, and develop statistical models of spatial patterns.

These methods provide ways of investigating the organization and structuring of phenomena over space--a concern that is so pervasive in geographic research that it has been identified as one of the central themes that unifies geography as a discipline and distinguishes it from other fields of study (Morrill, 1983). Even if a concern for spatial pattern and spatial organization is not, in itself, sufficient to define geography, an interest in some kind of spatial pattern is nearly universal in geographic research, and it is often necessary for geographers to measure spatial patterns in some numerical way or to test hypotheses that are formalized as statements about patterns or regularities in space. Analyses of spatial patterns are also important in other fields, including ecology (Sokal & Oden, 1978b), archaeology (Hodder & Orton, 1976), epidemiology (Mayer, 1983), sociology (Dorien, 1981), and geology (Agterburg, 1970). Consequently, the development of statistical methods for analyzing spatial data constitutes a very important step in the development of geography. The methods for analyzing spatial autocorrelation that are presented in this volume have been developed by geographers and others, mainly in the last twenty years and especially by Cliff and Ord (1973, 1981a), and they are among the most useful and general of these methods for spatial analysis.

The analysis of spatial patterns is characterized by the fact that the data can be arrayed as some kind of map and analyses of spatial autocorrelation can be thought of as formal statistical investigations of the patterns on maps. Spatial data include information on the absolute or relative locations of phenomena, so that the data are at least capable of being mapped, and spatial autocorrelation statistics provide a means of employing the general methods of statistical inference in order to test hypotheses about map patterns. Maps are capable of expressing enormously complex and varied information and autocorrelation statistics do not provide a complete set of methods for examining all the possible questions we might ask about map patterns. They can, however, be used to test useful hypotheses, such as the hypothesis that the values of some variable are randomly distributed over space or, alternatively, that the values follow some regular structure or distribution over space such as clustering.

Statements such as "this set of values is randomly distributed in space" are not, of course, very interesting in themselves but these kinds of statements are capable of being tested by means of formal statistical procedures and one aspect of the research process is to express interesting geographic questions in such a way that they can be tested and examined on the basis of statistical theory. This makes it necessary to formulate statements that are meaningful in terms of both geography and statistics and this has required some development in both fields since most statistical methods have not been designed for the special problems of spatial analysis. The material in this volume includes (1) discussions of how meaningful statements about geographic patterns can be expressed in forms that are suitable for analysis using autocorrelation statistics, (2) the problems that must be resolved in order to apply statistical methods to spatial data, and (3) the procedures that can be used to measure spatial autocorrelation and test hypotheses in particular situations.

1.1 Autocorrelation Statistics

Autocorrelation statistics are basic descriptive statistics for any data that are ordered in a sequence (Griffith, 1984b) because they provide basic information about the ordering of the data that is not available from other

AUTHOR'S NOTE: Most of this volume was prepared while I was a visitor in the Geography Department of the University of California at Santa Barbara. I am most grateful to the members of that department for their hospitality and support. I am also grateful to Arthur Getis and Gary Gaile. They both read the entire manuscript and provided a great deal of helpful advice. I am solely responsible for errors and omissions that remain.

descriptive statistics such as the mean and variance. When data are mapped, the map contains not only information about the values of variables but also information about how those values are arranged in space. Autocorrelation statistics provide summary information about this arrangement. Like the mean and variance, they do not express all the information about the data, but they do provide a numerical summary in a form that is useful for statistical testing. Autocorrelation statistics are not limited to spatial data but may be calculated for any data that are arrayed in a sequence. There is, for example, an extensive literature on autocorrelation in time series that deals with related problems although the calculation and interpretation of spatial autocorrelation statistics involves some difficulties that are not present for time series.

Autocorrelation statistics are functions of the same data values that are used to calculate other descriptive statistics, but they are also functions of the arrangement of those values in a sequence. In order to calculate an autocorrelation statistic, it is necessary to express that arrangement in some numerical fashion. The arrangement is expressed by some function that assigns values to pairs of locations in the sequence in order to represent their location with respect to one another. This function is called a "lag function" in time series analysis but we will use the term *weighting function* for spatial data because "lag function" implies a uniform relation with time or distance that is not always possible for spatial data.

A spatial weighting function is nothing more than a set of rules for assigning values to pairs of places in a way that represents their arrangement in space. The result of applying such a function to a map of n regions will be a set of numerical "weights" that express the relative locations of the regions on the map. For example, the spatial weighting function for a set of n regions might specify that the weight for a pair of regions i and j is one if i and j are neighbors and zero otherwise. (The weights that would express the location of a region with respect to itself, W_{ii}, are usually specified as zero.)

The set of weights is then combined with the data values to produce a function of both types of values, such as

$$\Sigma\Sigma w_{ij}(x_i - \bar{x})(x_j - \bar{x})$$

where the double summation indicates summation over all pairs of regions; w_{ij} is the spatial weight for the pair of regions i and j; x_i and x_j are their data values; and $\bar{x}$ is the mean for the entire sequence. This particular function is a spatial autocovariance. An autocovariance differs from an ordinary covariance because it is defined for lagged observations of a single sequenced variable, rather than joint observations of two variables, and this is a spatial autocovariance because the w_{ij} weight the individual cross-products $(x_i - \bar{x})(x_j - \bar{x})$ according to the relative locations of the regions i and j. Only the cross-products for neighboring values are used to calculate the spatial autocovariances if we adopt the zero-one weighting function suggested above, but other weighting functions are possible.

The spatial autocovariance measures the relation among nearby values of x_i where the meaning of "nearby" is specified by the w_{ij}, but a more useful spatial autocorrelation statistic, called Moran's I, can be produced by standardizing the spatial autocovariance. This statistic is

$$I = \frac{n}{\Sigma\Sigma w_{ij}} \frac{\Sigma\Sigma w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\Sigma(x_i - \bar{x})^2}$$

where n is the number of regions and the double summation indicates summation over all pairs of regions. Moran's I is merely the spatial autocovariance, standardized by two terms; the variance of the data series $\Sigma(x_i - \bar{x})^2$, which depends on the x_i values but is invariant with their arrangement; and $n/\Sigma\Sigma w_{ij}$. This second term says about as much concerning the arrangement of regions on a map as any single numerical value can. It is a measure of connectivity for the set of regions. Its value could change if the map of the regions were rearranged, but this value will not change with changes in the x_i.

Moran's I has an expected value of $-[1/(n-1)]$. The calculated value of I should equal this expectation, within the limits of statistical significance, if the x_i are independent of the values of x_i at neighboring locations. Values of I that exceed $[I/(n-l)]$ indicate *positive spatial autocorrelation* in which values of x_i tend to be similar to neighboring values. Values of I below the expectation indicate *negative spatial autocorrelation* in which neighboring values are not independent but tend to be dissimilar. Notice that $-[1/(n-1)]$ approaches zero, which is the expectation for an ordinary correlation coefficient, as the number of regions becomes large.

Moran's I is made up of the same components that define any correlation statistic, a measure of covariation (the autocovariance) and measures of total variation (the variance and the connectivity measure). Alternative statistics for spatial autocorrelation can be constructed by choosing different measures of covariation. For example, if we use the sum of squared differences between pairs of data values as a measure of covariation instead of the autocovariance and standardize in a slightly different way we have

$$c = \frac{n-1}{2\Sigma\Sigma w_{ij}} \frac{\Sigma\Sigma w_{ij}(x_i - x_j)^2}{\Sigma(x_i - \bar{x})^2}$$

This statistic is Geary's c, an alternative statistic for spatial autocorrelation that has a strong resemblance to the Durbin-Watson statistic that is widely used to test for autocorrelation in time series. Geary's c has an expectation of one for independence among neighboring values.

Spatial autocorrelation statistics can also be calculated for data that fall into categories instead of assuming continuous numerical values. These statistics are known as join-count statistics because they are calculated by counting the numbers of occurrences of like and unlike categories in adjoining locations.

The values of Moran's I and Geary's c depend on the w_{ij} and these are specified by the spatial weighting function that an investigator chooses. The freedom to choose alternative weighting functions introduces ambiguity into investigations of spatial autocorrelation but it also introduces flexibility because it becomes possible to use the statistics for alternative weighting functions to compare alternative hypotheses about how the data are organized in space. For example, the zero-one weighting function introduced above represents a hypothesis that the data for adjacent regions are related but ignores the actual distances among regions. A hypothesis that the values depend on the distances among regions rather than their adjacency could be represented by a weighting function that makes each w_{ij} the inverse of the distance between regions. These weighting functions represent slightly different hypotheses about the arrangement of the data over a set of regions and any number of weighting functions may be defined to represent different hypotheses. The possibility of defining alternative spatial weighting functions to test competing hypotheses makes autocorrelation statistics a very useful means of investigating spatial organization along with spatial pattern. The selection of spatial weighting functions is discussed at greater length in the next chapter.

1.2 Autocorrelation Statistics as General Cross-Product Statistics

The I and c statistics presented above are not the only statistics that could be used to measure spatial autocorrelation. In fact, the I and c statistics are only two special cases of a more general approach to measuring spatial autocorrelation by means of general cross-product statistics (Cliff & Ord, 1981a, pp. 22-24; Hubert, Gollege, & Costanzo, 1981; Upton & Fingleton, 1985, pp. 154-158). General cross-product statistics provide a means of comparing information on the proximity of locations with information on some other variable measured for the locations. They work by calculating the sum of cross-products for the two kinds of information and can be written in the general form:

$$\Gamma = \Sigma\Sigma W_{ij} Y_{ij}$$

where the double summation indicates summation over all pairs of locations. In the case of Moran's I the values for proximity, W_{ij}, are merely the weights, w_{ij} and the elements of Y_{ij} are the cross-product terms for each pair of regions, $(x_i - \bar{x})(x_j - \bar{x})$. The sum of their products is weighted by the constant term

$$\frac{n}{\Sigma\Sigma w_{ij} \Sigma(x_i - \bar{x})^2}$$

to produce Moran's I. A test for the statistical significance of Moran's I is nothing more than a test for the independence of the spatial proximity measure $W_{ij} = w_{ij}$ and the similarity measure $Y_{ij} = (x_i - \bar{x})(x_j - \bar{x})$.

Generalized cross-product statistics provide a very broad framework or statistical testing (Hubert & Gollege, 1982a, 1982b; Hubert et al., 1981). In fact the framework is not confined to autocorrelation statistics since this type of statistic could also be calculated for nonspatial variables. Many statistics other than Moran's I and Geary's c could be defined to make a general comparison between the spatial proximity of a set of locations and their similarity in terms of another variable. In order to test the hypothesis of independence, however, it is necessary to establish the distributional properties of the test statistics. That is, it is necessary to know the form of the probability distributions for the statistics, and to have information on their moments, including their variances, in order to test the hypothesis that the x_i are independent of location.

Only a few autocorrelation statistics (including I and c) have well defined distributional properties that make it possible to use them in conventional statistical tests. One of the major contributions of (Cliff and Ord (1973, 1981a) has been a thorough specification of the distributional properties of these and a few other statistics. Alternative methods of significance testing using permutation methods have been presented by Hubert et al. (1981), but the material in this volume concentrates on statistics that have known distributional properties so that they can be used in conventional tests of hypotheses.

1.3 Why Are Things Autocorrelated in Space?

Spatial autocorrelation statistics measure a basic property of geographic data-the extent of their interdependence with data at other locations. Many geographic data series may be interdependent because the data are affected by processes that connect different places, including spatial interaction and spatial diffusion processes; or by phenomena that extend over space to occupy regions rather than point locations.

Spatial interaction, which is the movement of goods, people, or information over space, means that events or circumstances at one place can affect conditions at other places if the places interact. Further, these movements or interactions among places usually vary with distance in systematic ways.[1] For example, prices and supplies in a set of spatially separated markets may be related if the markets are close enough to exchange commodities. In fact, prices at one location are unlikely to be independent of prices at other locations if they are near enough for supplies of a commodity to be moved between them and prices, along with other market-related data, are likely to be more similar between nearby markets than distant markets.[2] Spatial diffusion, which is the dispersion of phenomena from a set of origins, implies that the frequency or intensity of some phenomenon, such as an innovation in agricultural technology, may depend on distance from an origin. Locations that are near to one another are likely to be at similar distances from the origin and hence to experience similar frequencies for the phenomenon.

The fact that many geographic phenomena extend over space to occupy regions rather than single point locations means that sets of neighboring locations are affected by the same phenomena. The space or region that is occupied may not always be well defined. For example, legal systems are usually confined within well-defined boundaries and applied uniformly within those boundaries, but other phenomena such as "climates," "cultures," and "housing markets" occupy space in more ambiguous ways. Even so, they do exert similar influences on neighboring locations. Rainfall totals are likely to be autocorrelated mong nearby weather stations because they experience the same (or very nearly the same) weather events and prices for neighboring houses are likely to be autocorrelated because they are influenced by similar conditions of supply and demand. Hence a single event may affect several locations if the event involves an extended region. The passage of a single front may affect rainfall totals at several weather stations, for example, and the opening of a freeway interchange may affect the prices of housing in an entire neighborhood.

Interactions among places and the extension of many phenomena over space mean that events and circumstances at one location are unlikely to be independent of conditions at nearby locations. This interdependence among places lends pattern and structure to geographic data and autocorrelation statistics can be used to investigate hypotheses about how data for a particular variable are organized in space. Statistics such as I and c can be used to test the null hypothesis that the data values are independent of values at other locations and hypotheses about particular distance relationships or other spatial relationships may be formalized in the definition of the spatial weighting function that is used to calculate these statistics. This makes it possible to use the spatial autocorrelation statistics to conduct fairly extensive investigations into the organization of a single variable over space. Some fairly serious statistical problems are encountered, however, when other statistical methods are employed to investigate relations among sets of variables when the data are autocorrelated.

[1] Spatial interaction is the topic of another volume in this series (Haynes & Fotheringham, 1984)

[2] The logic of this example can be formalized as the spatial price equilibrium model. See Casetti (1972).

1.4 Spatial Autocorrelation and Statistical Problems

Spatial autocorrelation statistics are numerical measurements of some basic properties of geographic phenomena-the extent and nature of their relations with phenomena at other locations. As it happens, this basic property of geographic phenomena means that geographic data will often fail to fulfill a basic condition for most of the statistical methods that are used to analyze relations among variables-that of independence among the observations. Autocorrelation statistics that are designed to measure interdependence, such as Moran's I and Geary's c, are, of course, reliable when the data are not independent, but many other statistical tests are unreliable if the independence condition is not fulfilled. Autocorrelated data generate biased estimates of the standard errors used in most hypothesis tests and this causes the tests to be misleading if the data do not fulfill the independence condition. For example, Haggett, Cliff, and Frey (1977, pp. 330-334 and 374-377) show how autocorrelated observations lead to biased estimates of the standard error for tests of the difference between means based on Student's t distribution-a test that might be used to compare two sets of spatially distributed data.

Similar difficulties occur for other statistical methods and the greatest concern in geography has been with the effects of autocorrelation on regression models that are fitted to data for areas or regions (Cliff & Ord, 1981b; Haining, 1980). Independence of the errors from a regression model is a condition for valid hypothesis tests and, for regression models fitted to spatial data, this means that the residuals from the models should not be spatially autocorrelated. In fact, autocorrelation in the residuals would indicate that some source of variation has been omitted from the model or that the functional form of the model is not correct (Miron, 1984). Statistical tests of regression models are also likely to be mistaken if the errors are autocorrelated among the regions. Estimates of the standard errors of regression coefficients will be biased if the errors are autocorrelated and, where the autocorrelation is positive, these standard errors will be underestimated. This will, in turn, inflate the calculated values of test statistics for the significance of regression coefficients and may lead to the mistaken conclusion that variables are related when they are not.

Spatial autocorrelation has, therefore, a dual nature. Autocorrelation, or dependence among places, is a basic characteristic of most geographic processes and most spatial distributions. Autocorrelated data, on the other hand, make it difficult to investigate these same processes and distributions by using standard statistical methods. This duality has caused autocorrelation to be widely regarded as a statistical difficulty rather than a reflection of spatial processes. It is, in fact, both of these things, and the appropriate treatment of autocorrelation will depend on objectives of a particular research project.

In some cases, it may be appropriate to investigate relations variables in ways that abstract those relations from their spatial context. If it makes sense to abstract processes from their spatial contexts it will be appropriate to treat autocorrelation in the data as a statistical nuisance-a problem of the data that derives from the fact that they are embedded in a spatial sequence. The problem is then one of fitting a model to data that are affected by processes that are not accounted for by the model, and this may be done by transforming the data or by enlarging the statistical method so that reliable inferences are available from autocorrelated data. It is by no means a simple matter to enlarge the methods in this fashion and the enlarged methods will, in most cases, require a numerical specification of the autocorrelation in the data.

The objectives of geographic research are usually to understand phenomena *within* their spatial contexts and in this case the basic questions shift from questions about how to test nonspatial hypotheses with autocorrelated data to questions about how the data came to be autocorrelated and what this indicates about the relations among variables. It is then appropriate for a statistical investigation to involve hypotheses not merely about how variables are related, but how they are related over space. This usually requires statistical models to be enlarged to incorporate some representation of processes such as spatial interaction or regionalization. The enlargement of statistical models in this way may be difficult to accomplish and may not always resolve the statistical problems that are associated with the use of autocorrelated data.

The serious inferential problems that are associated with the interdependence of spatial data mean that spatial autocorrelation statistic play an especially crucial role in spatial analysis. First, these statistics can be used alone, as the sole means of testing hypotheses about organization and structure over space. Inferential problems associated with the independence condition do not arise in that case. They do arise when most other statistical methods are applied to spatial data, however, and in these cases spatial autocorrelation statistics are important for investigating and diagnosing problems with those models. Both applications of autocorrelation statistics are investigated in the remainder of this volume.

1.5 Further Reading

The two volumes by Cliff and Ord (1973, 1981a) are basic references for spatial autocorrelation. Their most important results establish the statistical distributions for autocorrelation statistics (1981a, Ch. 2). An introductory presentation of spatial autocorrelation is contained in Ebdon (1977) and the two-part article by Sokal and Oden (1978a, 1978b) provides a very readable presentation using examples from ecology and evolutionary biology. The volume by Upton and Fingleton (1985) includes extensive discussions of spatial autocorrelation as well as other statistical methods for spatial analysis. Hubert et al. (1981) integrate spatial autocorrelation statistics with a very broad inferential framework and the special statistical problems associated with autocorrelation in spatial data are discussed in Cliff and Ord (1981b) and Hammg (1980). There is considerable interest in spatial problems among statisticians and many of the developments in this area are discussed by Ripley (1981, 1984).

2 FRAMEWORKS FOR ANALYZING SPATIAL AUTOCORRELATION

Spatial autocorrelation statistics make it possible to measure interdependence in a spatial distribution and to use formal statistical methods to test hypotheses about spatial interdependence. These tests and measurements require that spatial or locational information be expressed in numerical terms and the utility of the statistical procedures depends on our success in translating locational information into a form that is suitable for statistical analysis.

2.1 The Logic of Statistical Testing

The process of using spatial autocorrelation statistics to test hypotheses about spatial interdependence follows the same general logic as most other statistical tests. First, a null hypothesis is specified. This null hypothesis usually states that the observed map of the values is produced by a process that assigns values to locations independently and at random. That is, every location is assumed to have the same chance of receiving any particular value; and the chance of receiving that value at any location is assumed to be independent of values at other locations. A map produced by such a process probably would not display an autocorrelated pattern. An autocorrelated pattern could occur purely by chance, but the chances are very small if the number of locations is reasonably large. The null hypothesis is, in fact, a simple *model* of the process that distributes data values over locations although it is a model that may be rejected on the basis of statistical testing.

The next logical step in testing is to derive some explicit consequences of the model by describing the chances for any level of autocorrelation under the null hypothesis. This can be done by deriving the probability distribution for the value of an autocorrelation statistic under the random and independent assignment of values to locations. The probability distributions for values of the commonly used autocorrelation statistics, including Moran's I, Geary's c, and the join-count statistics used for binary data, are asymptotically normal[3] under the null hypothesis of independence (Cliff & Ord, 1981a, pp. 34-65).

The mean and variance of the hypothesized distribution of the statistic are then obtained and compared with an observed value that is calculated for the data. If the observed value deviates from the mean by an amount that is large, compared to the variance, it is unlikely that the observed map was produced by random and independent assignments. The null hypothesis can then be rejected in favor of the conclusion that the data are interdependent over space. The normality of the distribution makes it possible to assign a probability to the occurrence of the observed value, under the null hypothesis, by using a set of statistical tables.

A random and independent assignment of values would produce an uninteresting map but a null hypothesis of random and independent assignments is more useful than it may appear at first. It may be helpful, before setting out to explain interdependence in a set of data, to establish that the data are, in fact, interdependent and not merely the result of a random and independent assignment of values to regions. Further, independence in regression residuals is a condition for making reliable inferences from regression models, as noted in the previous chapter. Complete independence is also an extreme situation that can serve as a benchmark for comparing alternative models of interdependence. The autocorrelation statistics themselves offer the possibility of investigating alternative explanations for interdependence because the spatial weighting functions can be used to express alternative hypotheses about the relations among places or regions. The autocorrelation values calculated for alternative weighting schemes can be compared to each other by comparing each of them to the distribution of values expected under a random and independent assignment.

The testing process described above involves at least two kinds of "translations" between different "languages." First, abstract postulates about a process that assigns values to locations (randomness and independence) are translated into a formal mathematical statement (the probability distribution for a statistic). Second, an observed spatial pattern is translated into mathematical terms (a calculated value for the statistic). These translations may, like linguistic translations, lead to statements that are awkward, incomplete, inaccurate, or misleading. The difficulty of the translations, like that of linguistic translations, depends on both the

[3] The term *asymptotically normal* means that the distributions for the statistics approach normal distributions as the sample size increases. If samples are very small, however, the distributions may not be normal and the test statistics may be misleading. See the section on Sample Sizes.

character of the languages and the message of the particular statement. There are special difficulties in translating spatial patterns into languages that are appropriate for statistical testing, but the severity of the problems often depends on the particular statement or model that is involved.

The problems involved in using autocorrelation statistics to compare a particular spatial pattern with the null hypothesis of random and independent assignments are not trivial and, since tests of this kind are emphasized in this volume, some of those problems are discussed in the remainder of this chapter. More serious problems may occur when an explicit model of an autocorrelated process is constructed as an alternative to the null hypothesis of randomness and independence. This null hypothesis may be too general for some investigations or, if it can be rejected, it may be useful to construct an alternative model that incorporates interdependence among locations and estimate parameters that measure the interdependence numerically. Some of the difficulties that are associated with formalizing more complex hypotheses and estimating parameters don't affect the simple tests for spatial autocorrelation. A few of these problems are discussed as well because they overlap with the problems for autocorrelation statistics and because analyses of autocorrelation often form a part of wider studies where these problems occur.

2.2 Types of Geographic Data

Geographic data, like other data, may consist of continuous valued variables in which the observations are real numbers; rank orderings in which an ordering of observations is available but not on a continuous numerical scale; and categorical variables in which the information about each observation consists of its membership in one of a set of discrete categories (Wrigley, 1985). Different statistical methods are generally required for each of these types of data so it is not surprising that different spatial autocorrelation statistics are required as well. Moran's I and Geary's c are the most familiar statistics for continuous variables. Analyses of autocorrelation in categorical variables can be carried out by means of the join-count statistics to be discussed in Chapter 3 and methods for analyzing autocorrelation in rank orderings have been presented by Royalty, Astrachan, and Sokal (1975) and by Sen and Soot (1977).

Geographic data are also data that can be mapped and the type of mapping that is possible also affects the kinds of statistical methods that are available. The main distinction is between data that can be continuously mapped and data that can only be mapped discontinuously. The process of mapping data consists of assigning values to locations in a two-dimensional space. All the locations in a two-dimensional space can be represented by a pair of coordinates u and v that take on continuous values. A *continuous mapping* of data occurs when it is possible to assign a data value to every coordinate location in the space *and* it is also possible for that value to differ between every pair of locations no matter how close together they may be.

Continuously mapped data are usually categorical and point pattern maps of the locations of towns, farmhouses, trees, or other phenomena are the most familiar examples. Data are available, in this case, for every possible pair of coordinates but in the form of a categorical variable that indicates presence or absence for every location. Hence a listing of these data usually consists of a listing of coordinate locations where the phenomena are present. An extensive set of statistical methods is available for analyzing point patterns and they are discussed in another volume in this series (Boots & Getis, 1988). The types of hypotheses that are tested using these methods are often similar to the hypotheses that are investigated with autocorrelation statistics even though the statistical calculations are different. Continuous data can also be mapped continuously but this kind of mapping involves assigning a value to every possible pair of coordinate values so this kind of map is usually produced by an operation on another kind of data, such as interpolation or smoothing of discontinuously mapped values.

Autocorrelation statistics are used for data that are discontinuously mapped. A *discontinuous mapping* occurs when data values can be assigned to only a finite subset of the locations in a space *or* when the values cannot vary between some pairs of locations. The first kind of discontinuous mapping, usually known as point data, occurs when data are available for a set of points but not for the intervening space. Figure 2.1a, which is a map of air pollution monitoring stations in the South Coast Air Basin of California, illustrates this kind of discontinuous mapping. (The South Coast Air Basin includes Los Angeles and Orange counties and parts of San Bernardino and Riverside counties.) Air pollution occurs continuously over the Los Angeles region but data are available only at the locations of monitoring stations.

The second type of discontinuous mapping, areal data, occurs when data are reported as single values for each member of a set of subregions. Figure 2.1b shows the counties of the state of Oregon. Data on variables such as per-capita incomes, population densities, or disease rates are often available for subregions such as these. A data value can be assigned to every pair of coordinates in this case, but only for information that has been aggregated to the county level, and the values change only at the boundaries of the subregions. The data may be continuous, rank ordered, or categorical for either type of discontinuous mapping.

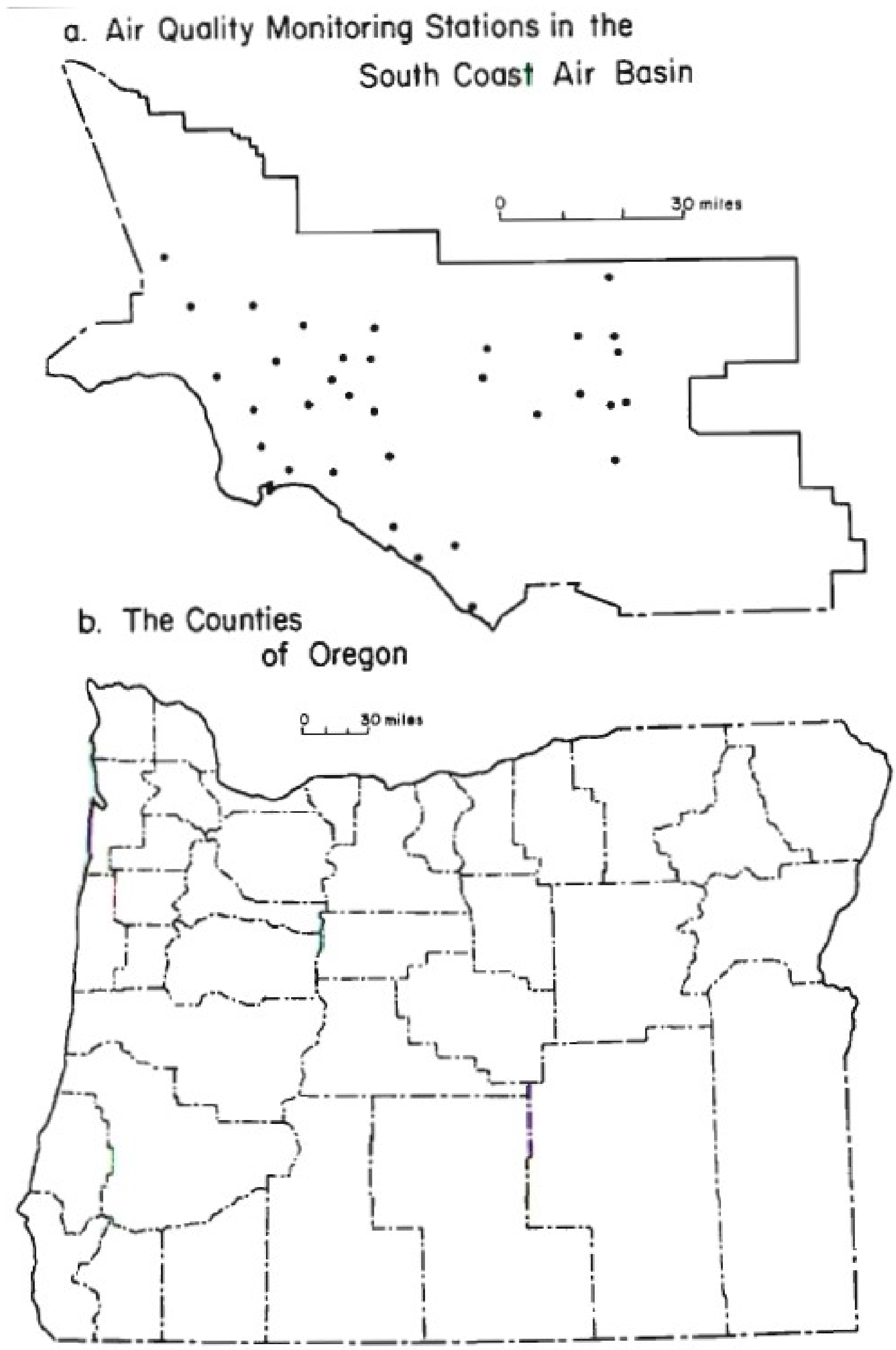


Figure 2.1 Discontinuous Mapping; Point and Areal Data

It is possible to make transformations between these discontinuous and continuous mappings. Values for air pollution at locations other than the monitoring stations can be inferred from data at those points and areal data can be transformed into a continuous surface of values by interpolation (Tobler, 1979). Areal data can also be transformed into a continuously mapped pattern of points (Getis, 1983, 1984). These transformations make it possible to use methods developed for continuously mapped data if they offer some advantage over autocorrelation statistics in a particular situation.

It is also possible to make transformations between the two types of discontinuous mappings. Areal data can be assigned to a set of points that correspond to the centers of subregions and the reverse operation can be performed by defining Thiessen polygons for a set of points. (The Thiessen polygon for each point is the region that contains all locations that are closer to that point than to any other.) The transformation is merely from one type of discontinuous mapping to another, however, and there is little to be gained since the problems of analyzing the two types of data with autocorrelation statistics are very similar.

2.3 Problems with Geographic Data

The data used in geography are usually nonexperimental and are often obtained from secondary sources such as census reports so the process of gathering the data is not controlled by the investigator. Any data from secondary sources may present problems because of inappropriate definitions of variables, doubts about reliability, and so forth. Geographic data are also arrayed over space, often in ways that the investigator might not have chosen, and the ways that the data are distributed in space may cause special problems.

IRREGULAR SPACING

It is often convenient to construct abstract models of spatial processes on the basis of a regular lattice of points (Figure 2.2a) or a regular grid of subregions (Figure 2.2b) because both these arrangements offer the advantage of *spatial neighborhood stationarity* (Tobler, 1979). That is, each point or region has the same number of neighbors and those neighbors have the same distance relations with a central point or region. Consequently, a set of rules for connecting the regions or points--a spatial weighting function--can define a set of distance relations that is uniform across the space when measured in a conventional metric such as miles or kilometers.

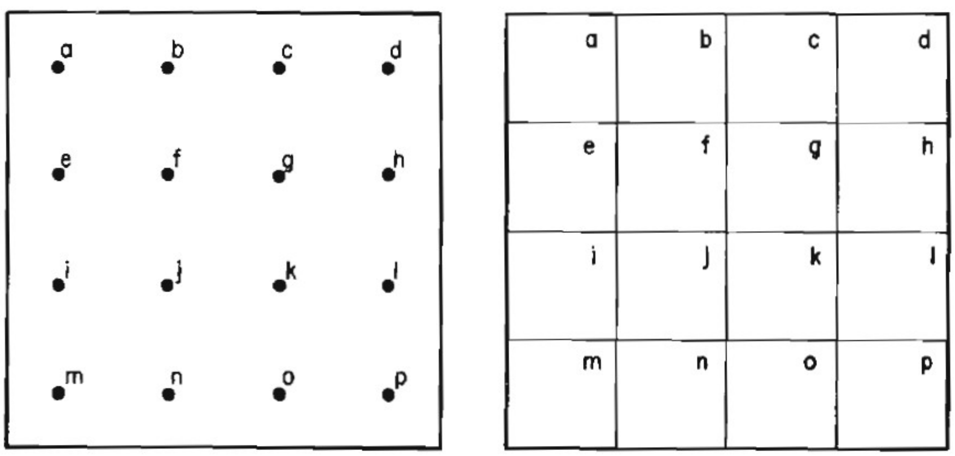


Figure 2.2 Points on a Regular Lattice and Regions on a Regular Grid

Each nonboundary region in a regular grid (or point on a regular lattice) has the same number of neighbors so long as we use a definition of “neighbor” that is the same for every location. The point labeled F (or the region labeled F) has four neighbors if a “neighbor” is the adjacent point in the same row or column. So does every other point or region that is not on a boundary. Further, all of these neighbors are at the same distance from the central point or, for regions, have the same distance relations with the central region. It may be desirable to include adjacent points or regions on the diagonals as neighbors along with those on the rows and columns, and this introduces some variation in the distances to neighbors, but their number and the set of distance relations remain uniform across all points or regions that are not on the boundaries.

A set of ”higher-order neighbors” can also be defined on a regular grid or lattice. For example, region E is a second-order neighbor of region G because it is a neighbor of the immediate neighbor of region G. Region E is also a third-order neighbor of region H and so on. Autocorrelation statistics for higher-order neighbors are the basis for constructing spatial correlograms, a topic that is discussed in Chapter 6.

Autocorrelation statistics can be calculated for regular grids or lattices and the extreme values of the statistics are associated with simple patterns on the regular grid or lattice. An example of a binary categorical variable is illustrated in Figure 2.3. Each region is assigned to one of two categories “black” or “white” and Figure 2.3a shows the extreme case of positive autocorrelation among first-order neighbors, where similar values are clustered as much as possible. The checkerboard pattern in Figure 2.3b shows an extreme case of negative autocorrelation among first-order neighbors. Values are still interdependent in this case but similar values are dispersed as much as possible.

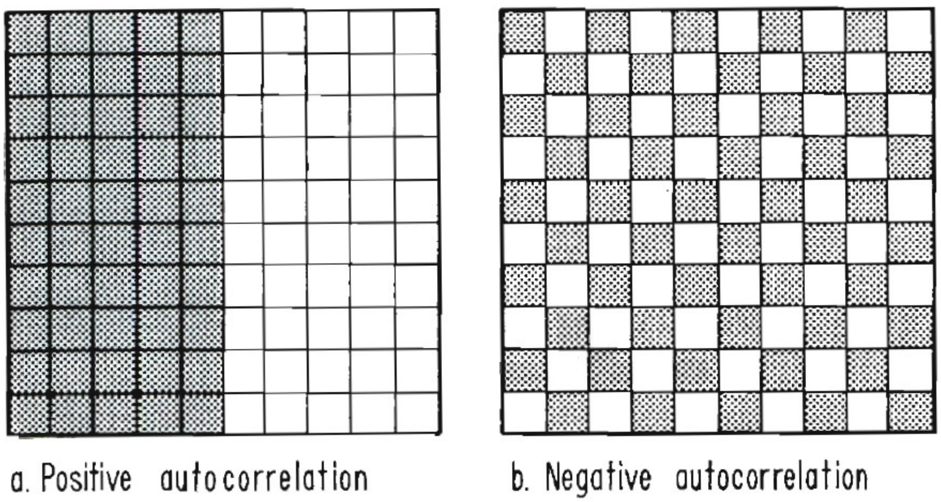


Figure 2.3 Patterns with Extreme Spatial Autocorrelation on a Regular Grid

Geographic data are rarely available on regular grids and lattices. There are some exceptions: Samples in ecology are often gathered on a regular grid (Jumars, Thistle, & Jones, 1977); remotely sensed data consist of a regular grid of pixels (or picture elements); and some census materials from both Japan and the United Kingdom are available on regular grids made up of one kilometer squares. In most cases, however, data are arrayed over space in an irregular way, more like the examples of counties and monitoring stations in Figure 2.1. The condition of spatial neighborhood stationarity is not fulfilled for data that are not regularly spaced and it becomes impossible to define spatial weighting functions so that each region or point has both the same number of neighbors and the same distance relations with its set of neighbors.

The sizes and shapes of the subregions used to report areal data are often highly variable. Counties in Oregon are generally larger in the southern and southeastern parts of the state and smaller in northwestern Oregon and some counties are nearly rectangular while others have very irregular shapes. The locations used to report point data may also be located in irregular ways. Some of the air monitoring stations in the Los Angeles region are close to their neighbors while others are relatively isolated. Some parts of the region have a high density of monitoring stations while other areas have few stations or none at all.

Irregular spacing of the data values makes it difficult to evaluate or compare autocorrelation statistics in terms of familiar distance metrics. It is necessary to define a spatial weighting function that yields a set of weights, w_{ij}, for every pair of areas (or points) in order to calculate an autocorrelation statistic but the autocorrelation statistic will not have an unambiguous relation with a distance metric such as miles or kilomters unless the w_{ij} have a regular relation with distance. Regular distance relations do hold for data on a regular lattice. If the neighboring points on a regular lattice are always separated by one kilometer, and we set $w_{ij} = 1$ for neighbors and $w_{ij} = 0$ otherwise, the corresponding autocorrelation statistic will be associated with events or circumstances one kilometer apart.

A weighting function for irregularly spaced data will, in contrast, assign varying numbers of neighbors to points or areas, or will assign neighbors at varying distances, or both. The simplest kind of spatial weighting function for Oregon counties would be one that defines adjacent counties as neighbors and assigns $w_{ij} = 1$ if the counties share a boundary and $w_{ij} = 0$ otherwise. This spatial weighting function would assign differing numbers of neighbors to the counties and some neighbors would share long boundaries while others might touch only at their corners. Further, the neighbors of some counties would form a compact region close to the central county (if the neighboring counties are small in area) while the neighbors of others would form extensive regions. Finally, extreme autocorrelation values generally will not be associated with simple patterns such as those illustrated in Figure 2.3 because the simplicity of those patterns depends on the regularity of the grid.

Autocorrelation statistics can be calculated and tested in these circumstances although the statistics will not have an unambiguous relation with distance metrics and it is important, in interpreting the results, to be aware of the characteristics of the spatial weighting function when it is applied to a particular map of data. The difficulties associated with irregular spacing are somewhat greater when investigation goes beyond tests based on autocorrelation statistics to the development of models to explain the interdependence that may be revealed by those statistics. Models of this kind might take the form of a spatial autoregression, a function that relates the value at some location to the values at neighboring locations. These models are relatively easy to construct for regular lattices. For example, the value of a variable X at any coordinate location u, v may be related to the four surrounding values by the equation

$$X_{u,v} = a + b_1 X_{u-1,v} + b_2 X_{u+1,v} + b_3 X_{u,v-1} + b_4 X_{u,v+1}$$

where a is a constant and the various b are regression coefficients. Variations in the b would indicate some directionality in the spatial process.

It is more difficult to define an autoregressive model for irregularly spaced data. Most definitions of neighbors will assign variable numbers of neighbors and, hence, variable numbers of terms to the equation. The neighbors would also be at variable distances and values of the regression coefficients, the b values, may not be independent of the distances between neighbors. Using the same coefficients for neighbors at variable distances would introduce a systematic bias into the estimates of $X_{u,v}$.

SPATIAL RESOLUTION

The sizes and shapes of sampling areas such as counties also limit the capacity of autocorrelation statistics, or other spatial statistics, to detect patterns in a set of areal data because these sizes and shapes limit the scale of the patterns that can be detected. Patterns that are manifest at small scales will not be evident when the information is aggregated into larger areas. The average *resolution* of a set of subregions is defined as the square root of the average of their areas (Tobler, 1984) and pattern elements less than twice this size cannot be detected by examining data for the subregions. For example, Tobler (1984) gives the average resolution of counties in the conterminous United States as 43 kilometers so pattern elements need to be at least 86 kilometers (or 54 miles) wide in order to be manifest in this array of data. This is an approximation since county sizes vary considerably from this average and the resolution also depends on the shapes of the areas relative to the orientation of the pattern. For example, a set of long narrow regions would offer better resolution for patterns oriented perpendicular to their long axes than for patterns oriented parallel to those axes.

The limits of resolution for any set of areas mean that a set of data with relatively coarse resolution, resulting from aggregation into large subregions, cannot be expected to show the effects of processes that occur at relatively fine scales. For example, central place theory implies a regular spatial periodicity in population distributions and, although this periodicity is evident in fine-scale or high-resolution data (Rayner & Golledge, 1972; Tobler, 1969), much of the pattern could not be detected in population data at the county level. In fact, any spatial patterns that depend on interactions over short distances, such as commuting or the personal contacts involved in disease contagion, will not be revealed by analyzing data for relatively large sampling areas.

SAMPLE SIZES

Small samples are often used in spatial analyses and hypothesis tests about autocorrelation statistics may be misleading if they are based on small samples. The distributions of the test statistics such as Moran's I are *asymptotically* normal, which means that their distributions approach normality as the sample size increases. Conversely, their distribution may not be normal for small samples and the use of the normal distribution could lead to mistaken inferences.

It is probably safe to assume normality for the distributions of the test statistics for samples of at least fifty, and possibly fewer, observations (Cliff & Ord, 1981a, p. 53; Sen, 1976). It is not possible to specify a minimum sample size for all situations because the distributions of the test statistics depend on the set of spatial weights and the distribution of the variable as well as the sample size (Cliff & Ord, 1981a, p. 54). Small samples are also more troublesome for categorical data, especially binary data, because the statistics for categorical data approach normality more slowly as sample sizes increase. Cliff and Ord (1981a, pp. 53-65) provide a series of methods for approximating the distributions of test statistics for small samples and, although their use may require additional calculations, they make it possible to test hypotheses even when the samples are very small.

The distribution of the test statistics is also likely to deviate from normality for a particular pattern of spatial weights: when a set of regions are all connected to a central region or node and unconnected or weakly connected with one another (Cliff & Ord, 1981a, p. 50). This is not strictly a problem of sample size but it is less likely to occur for large samples.

Aside from strictly statistical considerations, it should be obvious that small samples are less useful than large samples. A small sample of subregions will consist of either large subregions that offer poor spatial resolution or small subregions that cover only a small area in total. In either case, spatial patterns that can reveal something about spatial processes are less likely to develop over a small set of subregions.

BOUNDARIES

Empirical spatial analyses are always carried out in bounded regions and influences may intrude across the boundaries of a study area, especially if the boundaries are more or less arbitrary. There is no boundary problem if the processes that affect spatial patterns are confined within the boundaries and outside influences are effectively excluded. These conditions could be fulfilled for some situations, such as epidemics on an isolated island (Cliff, Haggett, Ord, & Versey, 1981), but study areas often must be defined by more permeable boundaries.

The autocorrelation statistics can be calculated, and tests can be performed against the null hypothesis of randomness and independence even in the presence of boundary effects because the statistics can be interpreted merely as measurements of pattern for a region. The boundary effects will, however, bias the parameter estimates for autoregressive models or other models whose parameters provide numerical measurements of the interdependence of locations. Even the process of comparing alternative spatial weighting functions by comparing the associated autocorrelation statistics may be affected. An arbitrary boundary eliminates some locations that might have been included in the analysis and this may affect the statistics for different weighting functions in different ways.

Griffith (1983) reviews a set of five alternative methods that have been used for dealing with boundary problems, including methods that define an outer perimeter or buffer zone consisting of locations near the boundary. He also proposes a set of statistical methods for estimating the parameters of autoregressive models in the presence of boundary effects.

The effects of boundaries are related to the sample size and the shape of the study area. Any two-dimensional study area will have a large proportion of the observations near its boundaries, and more subject to influences from beyond the boundaries, and this proportion will be greater for smaller samples. The effects of influences from beyond the boundary will also be greater for study areas that do not have a compact shape.

2.4 Spatial Weighting Functions

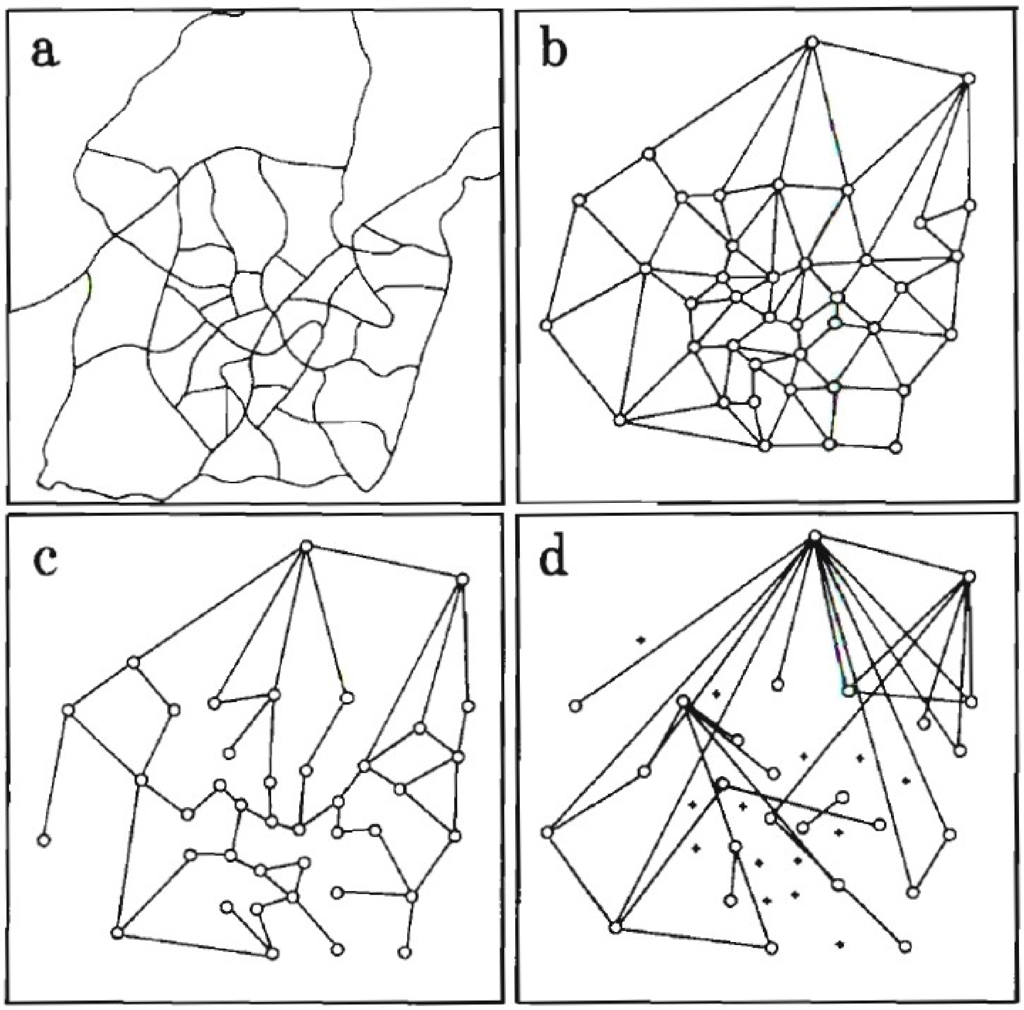
The selection of a spatial weighting function is the most important step in calculating a spatial autocorrelation statistic. A spatial weighting function is a set of rules that assign values or "weights" to every pair of locations in a study area and the value of an autocorrelation statistic will depend on these weights as well as the data for the locations. Spatial weighting functions are often defined to represent the arrangement of areas or points relative to one another in a conventional space (where locations can be identified by latitude and longitude) but it is more general to think of a spatial weighting function as a means of accommodating hypotheses about the relations among places. These hypotheses will often be simple variants of Tobier's (1970) "first law of geography: everything is related to everything else, but near things are more related than distant things." If "near" and "distant" refer to conventional distance metrics, such as miles or kilometers, the problem of defining a spatial weighting function is to represent nearness and distance as consistently as possible for a set of irregular regions or irregularly spaced points. Other hypotheses about the relations among places can also be represented by spatial weighting functions, however, and the flexibility in defining the weights makes spatial autocorrelation statistics a useful means of investigating alternative hypotheses about the relations among places.

Spatial neighborhood stationarity is not possible for irregular areas or irregularly spaced points, but a variety of spatial weighting functions are available to represent "nearness" or "distance" for these kinds of data. The simplest weighting function for areal data is a set of binary weights that have a value of one for areas that share a boundary and zero otherwise. This function assigns the same values to pairs of regions with very short boundaries and pairs with very long boundaries and a more detailed representation of proximity may be necessary. One approach is to scale the weights according to the length of the common boundaries, under the presumption that areas that share long boundaries are "closer" than regions that share short boundaries. For example, the value of w_{ij} may be the proportion of the total boundary of region i that is shared with region j. If it is desirable to have symmetric weights so that $w_{ij} = w_{ji}$ these weights can be redefined as $w_{ij} = (w_{ij} + w_{ji})/2$. A further improvement in the definition of proximity may be obtained by using distances between the centers of regions along with the lengths of the boundaries. Cliff and Ord (1981a, pp. 17-18) suggest that weights can be defined as $w_{ij} = d_{ij}^{-a} z_{ij}^{b}$ where d_{ij} is the distance between the centers of the regions, z_{ij} is the proportion of the boundary of i that is shared with j and a and b are parameters that are selected a priori. This scheme gives greater weights to areas whose centers are separated by shorter distances and that also share long boundaries. Any set of areas can also be represented by points located at their centers so that weighting functions defined for point data can also be applied to areal data.

Tobler (1975) lists a series of weighting functions for point data including binary functions where $w_{ij} = 1$ for all points within some fixed distance of point i; or $w_{ij} = 1$ for the some fixed number of points that are nearer to i than other points. The weights can also be scaled according to some function of distance between the points, usually an exponential function such as $w_{ij} = d_{ij}^{-b}$ so that the weights approach zero for large distances but do not become negative. Spatial weighting functions based on Gabrial graphs (Matula & Sokal, 1980) have been widely applied in biological studies of point data. Gabrial graphs are binary weighting functions where $w_{ij} = 1$ if no third point location lies within the circle whose diameter is the straight line connecting i and j.

These spatial weighting functions represent distance relations among places, subject to the limitations of irregular spacing, but tests for autocorrelation are not limited to patterns that are manifested solely over distance. The procedure of calculating an autocorrelation statistic is actually a special case of a set of very general procedures for comparing two sets of values (Hubert et al., 1981). Moran's I, for example, is a means of comparing the set of w_{ij} with the set of cross-products $(x_i - \bar{x})(x_j - \bar{x})$ to ascertain if patterns of variation in the cross-products are similar to patterns of variation in the weights. That comparison can be carried out for any set of w_{ij}, including values that have nothing to do the locations of i and j. When we calculate Moran's I for a set of w_{ij} that represent geographical proximities for a set of regions we are calculating a measure of the association between the proximities and the data at the locations but nothing in the logic of the procedure limits its application to sets of weights that represent only the proximities of the regions.

The generality of the procedures used in calculating spatial autocorrelation statistics means that their use can extend beyond tests for randomness in map patterns to comparisons of alternative hypotheses about the spatial processes that produce the patterns. Alternative hypotheses about the spatial processes can be formalized as different weighting functions and differences in the associated autocorrelation statistics can provide evidence for one hypothesis over another. Alternative weighting functions have been used, for example, to represent competing hypotheses about the spatial diffusion of epidemic diseases (Cliff, Haggett, Ord, Bassett, & Davies, 1975; Haggett, 1976). Adesina (1984) compares five alternative weighting functions in an effort to identify the details of the diffusion process for a cholera epidemic among subregions in the city of Ibadan (Figure 2.4). A model of local contagion, in which cholera would spread mainly between nearby areas, is represented by a binary weighting function where $w_{ij} = 1$ for regions that share a boundary. Alternative models include contagion patterns related to patterns of travel and personal contact, including a model based on the availability of road links and models based on the journey-to-work, the journey-to-school, and the journey-to-market. These hypotheses are also represented by binary weighting functions: The road network model has $w_{ij} = 1$ if a paved road connects the subregions and the other models have $w_{ij} = 1$ if a threshold value for the numbers of travelers between the zones is exceeded. The spatial weighting functions for three of these models are shown in Figure 2.4, where lines connect the centers of subregion i and j if $w_{ij} = 1$. Adesina's results indicate that the pattern of weights in the local contagion model has the strongest relation with the spread of the epidemic.



SOURCE: Adapted with permission from **Social Science and Medicine**, Vol. 18, H. O. Adesina, "Identification of the Cholera Diffusion Process in Ibadin, 1971," Figures 1a and 3. Copyright ©1984, Pergamon Journals Ltd.

Figure 2.4 Alternative Spatial Weighting Functions for an Analysis of a Cholera Epidemic in Ibadan

2.5 Further Reading

Problems and approaches in spatial analysis are discussed in the two-volume work by Haggett et al. (1977) and extensive reviews of spatial analysis are available in Bennett and Haining (1985) and Cliff and Ord (1975c). Tobler (1975, 1979) describes the special character of geographic information and Griffith (1980) discusses several of the special problems inherent in the statistical analysis of spatial data. Gatrell (1979) presents an extended range of possibilities for the interpretation of spatial weighting functions.

3 AUTOCORRELATION STATISTICS FOR CATEGORICAL DATA

Tests for spatial autocorrelation in categorical data are based on the proximities of members of the same or different categories. In most cases proximity is defined on the basis of simple contiguity and the statistics are frequencies or “join-counts” for the numbers of instances where like or unlike categories occupy adjoining locations. In order to carry out tests for autocorrelation it is necessary to choose a statistic; select an appropriate sampling assumption; and calculate an expectation and variance for the statistic under the null hypothesis of random and independent assignment of regions to categories. The discussion in this chapter is limited to binary classifications although the methods can be extended to systems with more than two categories (Cliff & Ord, 1981). The two categories in a binary classification are often referred to as “black” and “hite” and the possible types of joins are limited to black-black (BB), black-white (BW), and white-white (WW).

The use of these statistics should, as a general rule, be limited to data that are truly categorical. The join-count statistics are somewhat easier to calculate than the corresponding statistics for continuous data so there is some temptation to reduce continuous variables to a binary form by imposing some classification on the data. This is a risky practice that can lead to mistaken inferences. It is especially easy to overestimate autocorrelation if observations of a continuous variable are classified as “black” or “white” when they are above or below the mean for the variable.

3.1 Alternative Test Statistics

The test statistics for binary data are limited to frequency counts of either like (BB) or unlike (BW) joins. These values can be calculated for the data by assigning a binary variable x_i to each region with $x_i = 1$ if region i is “black” and $x_i = 0$ if region i is “white.” The observed values of the test statistics are then

$$BB = {}^1\!/\!_2 \Sigma\Sigma w_{ij} x_i x_j \tag{3.1}$$

$$BW = {}^1\!/\!_2 \Sigma\Sigma w_{ij} (x_i - x_j)^2 \tag{3.2}$$

where w_{ij} is the value assigned to region i and region j by the spatial weighting function and the double summation indicates summation over all pairs of regions. The constant, $^1/_2$, takes account of the double counting produced by the double summation (which is a summation over all pairs of regions). A WW statistic can be calculated by reversing the assignments of one and zero. These formulas correspond to simply counting the numbers of joins between like or unlike regions if w_{ij} takes on binary values that depend on the contiguity of the regions.

The two test statistics (actually three if WW is included) do not measure exactly the same things even though they are closely related and it is possible that they may not support the same conclusions for some sets of data. Is one of the tests preferable to the others on statistical grounds? Cliff and Ord (1975c) have examined this question by calculating the asymptotic relative efficiencies for the BB and BW statistics under a variety of circumstances. The asymptotic relative efficiency is related to the power of a test, which is the probability of rejecting the null hypothesis when it is false. They conclude that tests of the numbers of BW joins are preferable to tests of the number of BB joins. It may be useful to calculate all three statistics for some problems but the BW statistic is probably the most reliable.

3.2 Alternative Sampling Assumptions

The selection of a sampling assumption is part of the specification of the null hypothesis. Specifying a null hypothesis amounts to specifying a process that assigns values to regions and working out the moments (the mean and variance) of the distribution that would emerge if the specified process were repeated for a large number of trials. Different sampling assumptions are associated with slightly different rules for assigning values to regions and they lead to different variances for the expected join-counts.

Two sampling assumptions are available for binary data and they amount to sampling with replacement and sampling without replacement. In the case of sampling with replacement (sometimes called free sampling) the assignment of a value (black or white) to a region corresponds to sampling from a binomial distribution with a probability p for assigning one of the values to each region and a probability $1 - p$ for assigning the other. Repeated binomial sampling would not result in the study area having the same numbers of black and white regions for every trial. Sampling without replacement (or nonfree sampling) imposes fixed numbers of black and white regions on the study area. The spatial arrangement of black and white regions will vary in repeated trials, but the numbers of black and white regions always matches the number observed for the data. Sampling without replacement corresponds to sampling from a hypergeometric distribution.

The choice between these two sampling assumptions should be based on some knowledge about the underlying process that assigns values to regions. For example, sampling without replacement would be more appropriate if it is known that the number of black regions is limited to the observed value. Information about the underlying process often will not be sufficient to indicate a clear choice, however, and other factors should also be considered. The value of p, under sampling with replacement, must usually be estimated on the basis of the observed frequencies in the study area. Since this value is an estimate, the means and variances that are based on p will also have the status of estimates. The assumption of sampling without replacement, on the other hand, takes the observed numbers of black and white regions as given and is concerned only with their arrangement in space. Sampling without replacement will always produce a smaller variance for the null hypothesis than sampling with replacement, however, since fixing the numbers of regions in each category constrains the possible outcomes. Consequently, the hypothesis of spatial autocorrelation is more likely to be accepted under an assumption of sampling without replacement.

3.3 Means and Variances for the Join-Count Statistics

Tests are carried out by comparing observed values for the BB or BW statistics with the distributions of these values that would be expected under random and independent assignments. These distributions will differ under the differing assumptions of sampling with and without replacement.

SAMPLING WITH REPLACEMENT

The assumption of randomness means that every *region* has the probability p of assuming the value "black" and the probability $1 - p$ of assuming the value "white." The assumption of independence means that, for any pair of regions, the probability that both are "black" is p^2 and the probability that one is "black" and the other "white" is $p(1 - p)$. Substituting these values into equations 3.1 and 3.2 gives, for the expected means,

$$E(BB) = {}^1\!/_2\Sigma\Sigma w_{ij} p^2 \tag{3.3}$$

$$E(BW) = \Sigma\Sigma w_{ij} p(1 - p) \tag{3.4}$$

Expressions for the variances are more extended, especially when the w_{ij} are not binary weights and when the w_{ij} are not symmetric. It is useful to isolate two of the terms used in calculating the variances.

$$S_1 = {}^1\!/_2\Sigma\Sigma(w_{ij} + w_{ji})^2$$

$$S_2 = \Sigma(\Sigma w_{ij} + \Sigma w_{ji})^2$$

The first of these is a summation over the weights and, for binary and symetric weights, $(w_{ij} + w_{ij})^2$ is always equal to four so S_1 is simply four times the total number of joins in the entire study area. The second term, S_2, is obtained by counting the weights associated with each region (in both directions, w_{ji} as well as w_{ij}) and summing the squared values. For weights that are binary and symmetric this reduces to $S_2 = 4\Sigma[(\Sigma w_{ij})^2]$, which can be obtained by counting the total joins for each region, summing the squared totals and multiplying by four.

The variances of the two statistics are

$$\mathrm{Var}(BB) = {}^1\!/_4 p^2(1 - p)[S_1(1 - p) + S_2 p] \tag{3.5}$$

$$\mathrm{Var}(BW) = {}^1\!/_4\{4S_1 p(1 - p) + S_2 p(1 - p)[1 - 4p(1 - p)]\} \tag{3.6}$$

These fairly lengthy expressions are actually quite simple to calculate, especially when the weights are binary and symmetric.

Equations 3.3 and 3.5, or 3.4 and 3.6 are the means and variances of asymptotically normal distributions so a test for autocorrelation in moderately large samples can be performed by calculatin the standard normal deviate

$$z = \frac{BB - E(BB)}{[\mathrm{Var}(BB)]^{1/2}},$$

or the corresponding value for BW and looking up the calculated value of z in a table of the areas of the normal distribution.

SAMPLING WITHOUT REPLACEMENT

The numbers of "black" and "white" subregions are fixed under sampling without replacement so the assumption of random assignments means that the probability that any region is black is n_b/n where n_b is the number of "black" regions and n is the total number of regions. The probability that the members of any pair of regions are both "black" is $n_b(n_b - 1)/n(n - 1)$ and the probability that one member is "black" and the other "white" is $n_b(n - n_b)/n(n - 1)$. The expectations under randomness and independence are then

$$E(BB) = {}^1\!/\!_2\Sigma\Sigma w_{ij}[n_b(n_b - 1)/n(n - 1)] \tag{3.7}$$

$$E(BW) = \Sigma\Sigma w_{ij}[n_b(n - n_b)/n(n - 1)] \tag{3.8}$$

where the double summations indicate summation over pairs of regions. The variances are

$$\begin{aligned}\mathrm{Var}(BB) =& \frac{1}{4}\frac{S_1 n(n_b - 1)}{n(n-1)} + \frac{(S_2 - 2S_1)n_b(n_b - 1)(n_b - 2)}{n(n-1)(n-2)} \\ &+ \frac{[(\Sigma\Sigma w_{ij})^2 + S_1 - S_2]n_b(n_b - 1)(n_b - 2)(n_b - 3)}{n(n-1)(n-2)(n-3)} \\ &- [E(BB)]^2\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(BW) =& \frac{1}{4}\frac{2S_1 n_b n_w}{n(n-1)} + \frac{(S_2 - 2S_1)n_b n_w(n_b + n_w - 2)}{n(n-1)(n-2)} \\ &+ \frac{4[(\Sigma\Sigma w_{ij})^2 + S_1 - S_2]n_b(n_b - 1)n_w(n_w - 1)}{n(n-1)(n-2)(n-3)} \\ &- [E(BW)]^2\end{aligned}$$

3.4 An Example: Alcoholic Beverage Control in Georgia[4]

The map in Figure 3.1 shows the status of alcoholic beverage control in the state of Georgia, where each of the 159 counties of the state has the option of prohibiting the sale of alcoholic beverages. In total, 53 of the counties (colored "black" and commonly known as "dry" counties) prohibit the sale of distilled spirits everywhere within their boundaries. The remaining 106 counties permit the sale of distilled spirits everywhere or in certain cities in the county. The "wet" or "dry" decision made by each county presumably depends on local conditions of custom, religion, politics, or mythology. The conditions that lead a particular county to prohibit distilled spirits probably are not confined by county boundaries, and circumstances that lead to a county to prohibit the sale of alcohol are likely to be present in neighboring counties as well. The counties of Georgia are fairly small, averaging 960 square kilometers (370.5 square miles) so they offer a spatial resolution of roughly 31 kilometers (19.2 miles). This scale is probably fine enough for neighboring counties to share many of the same political or cultural conditions. Counties may also imitate their neighbors when choosing a policy for beverage control.

The map gives an impression of at least some clustering or positive spatial autocorrelation among "dry" counties in Georgia with sizable clusters of these counties in the northern and south-central parts of the state.

[4] I am grateful to William Berentsen for information on alcoholic beverage control in Georgia

Visual impressions of map patterns may not be reliable, however, and the join-count statistics can be used to perform a formal test for clustering, or positive spatial autocorrelation.

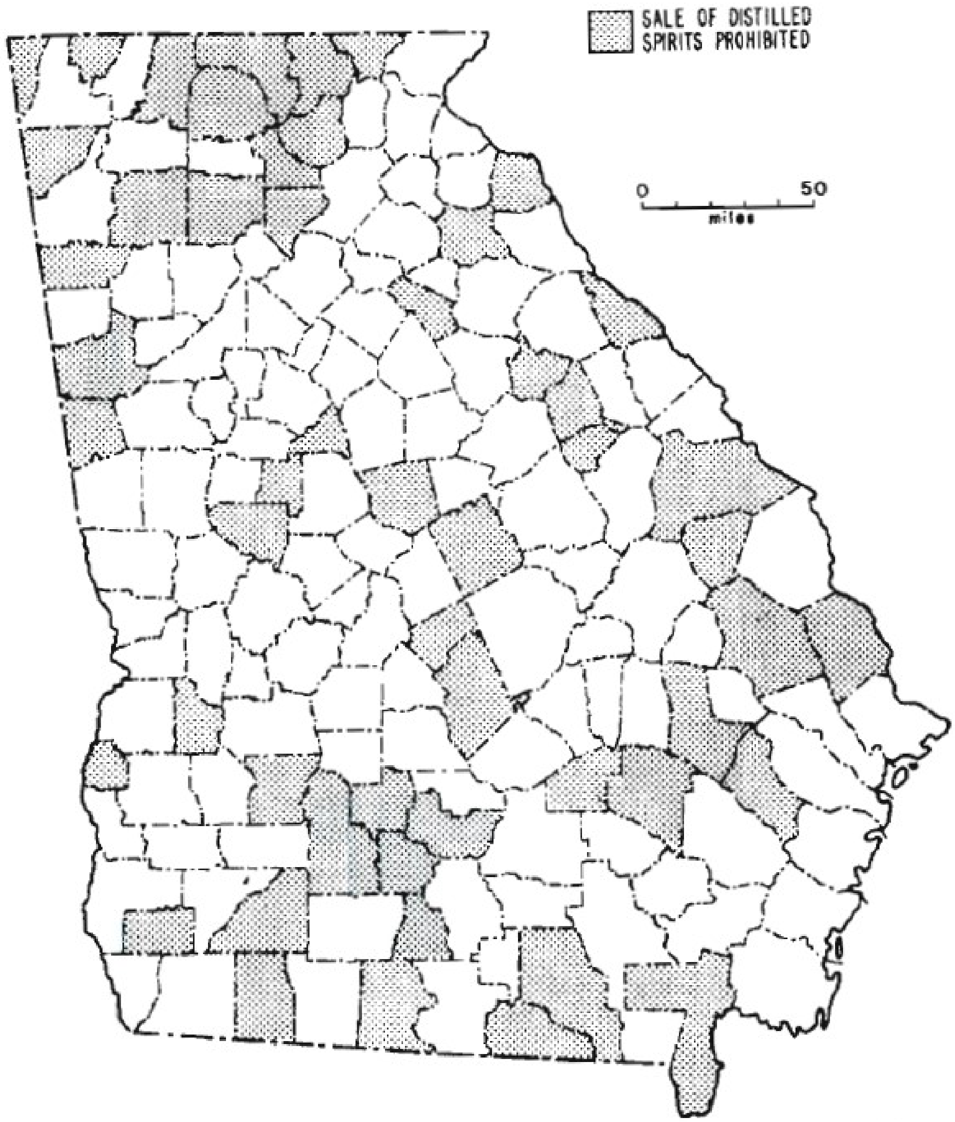


Figure 3.1 Alcoholic Beverage Control in Georgia Counties

Several matters have to be considered in carrying out this analysis. The reliability of the data, and any peculiarities that might affect inferences, should be examined in any statistical analysis. This includes any peculiarities in the spatial distribution of the data. One or more of the available tests must be selected, along with a sampling assumption, and a spatial weighting function must be defined.

The reporting of counties as “wet” or “dry” is probably reliable although there is some ambiguity in this classification since “wet” counties include some that allow sales everywhere and some that allow sales only in certain cities. The counties are small enough to offer reasonable spatial resolution and they are also fairly uniform in size and fairly compact in shape. The number of counties, 159, is reasonably large and sample size is important for the join-count statistics because their distributions converge to normality more slowly than those for continuous data.

Join-count statistics for binary data are at their most reliable when the two classes have equal numbers of members or, equivalently, when $p = 0.5$. The numbers are not equal in this case but simulation studies by Cliff and Ord (1981, p. 59) indicate better reliability for join-counts of the more frequent class. It may be useful, however, to calculate statistics for all three types of joins, coded as follows:

> *BB* joins are joins between two “dry” counties.
> *WW* joins are joins between two “wet” counties.
> *BW* joins are joins between a “wet” and a “dry” county.

Drastic disagreements among the statistics would indicate some kind of problem with the analysis.

An assumption of sampling without replacement, which constrains the number of “dry” counties to exactly 53, is probably more appropriate in this case even though it will produce a smaller variance and favor the hypothesis of positive spatial autocorrelation. It is difficult to justify the use of a single estimate, based on the contemporary values, for the probability of a county choosing a “wet” or “dry” policy because the decisions have been made at widely varying dates. The assumption of sampling without replacement sidesteps the question of how a particular county came to be “wet” or “dry” and focuses the analysis on the spatial arrangement of a fixed number “wet” and “dry” counties. The values used in calculating the expected values and variances are $\Sigma\Sigma w_{ij} = 792; S_1 = 1584;$ and $S_2 = 16968$.

The selection of a spatial weighting function is a critical step in evaluating spatial a w_{ij} for pairs of counties are $w_{ij} = 1$ if the counties share a boundary and $w_{ij} = 0$ if they do not. The calculation of the observed values of the autocorrelation statistic is reduced to merely counting the numbers of *BB* or *BW* joins for this weighting function. There are a total 396 joins among the 159 counties, including 40 *BB* (or “dry-dry”) joins, 177 *WW* joins, and 179 *BW* joins.

Results for all three join-count statistics are shown in Table 3.1 for sampling without replacement. Clustering of counties with similar beverage control laws would produce more *BB* and *WW* joins and fewer *BW* joins than a random and independent assignment of these values. The final column in Table 3.1 shows the probabilities that random and independent assignments of values would produce join-counts larger than the observed values for *BB* and *WW* joins, or smaller than the observed values for *BW* joins.

TABLE 3.1
Join-Count Statistics for Beverage Control in Georgia: Sampling Without Replacement

	Observed Value	*Expected Value*	*Variance*	*Standard Normal Deviate*	*Probability*
BW Joins	177	177.11	83.996	.0120	.5128
BB Joins	40	43.44	123.478	.3096	.3462
WW Joins	179	175.44	48.740	.5099	.6947

These results do not support a hypothesis of spatial autocorrelation or clustering among the counties but indicate that the spatial pattern of wet and dry counties is consistent with a random and independent assignment of values to the counties. The number of *BW* joins is very close to the mean for such assignments and there are actually fewer *BB* joins than expected under random and independent assignments. The number of *WW* joins does exceed the expected value but not by much random and independent assignments would produce more than the observed number of *WW* joins about 31% of the time. This doesn’t mean that beverage control policies were randomly selected in each county but it does indicate that the selection is independent of policies in adjoining counties.

4 AUTOCORRELATION STATISTICS FOR CONTINUOUS DATA

The process of testing for spatial autocorrelation in continuous data is similar to testing for autocorrelation in categorical data. Tests are performed against a null hypothesis of random and independent assignments; the distribution of a test statistic is established under these conditions, for a particular sampling assumption; and a calculated value is compared with the distribution expected under the null hypothesis.

4.1 Alternative Test Statistics

The test statistics that are usually used for continuous data are Moran's I;

$$I = \frac{n}{\Sigma\Sigma w_{ij}} \frac{\Sigma\Sigma w_{iji}(x_i - \bar{x})(x_j - \bar{x})}{\Sigma(x_i - x)^2}$$

and Geary's c;

$$c = \frac{n-1}{2\Sigma\Sigma w_{ij}} \frac{\Sigma\Sigma w_{ij}(x_i - x_j)^2}{\Sigma(x_i - \bar{x})^2}$$

where x_i is the value of a variable in region i, $\bar{x}$ is the mean for the variable and the w_{ij} are, as before, a set of weights. A much greater range of test statistics is available (Hubert et al., 1981) but both Moran's I and Geary's c have well-established distributions based on conventional sampling theory. Cliff and Ord (1975) find that the efficiency of Moran's I is generally a little better than that of Geary's c. The distributions of both statistics depend on sampling assumptions.

4.2 Alternative Sampling Assumptions

The distributions of Moran's I and Geary's c can be obtained under a sampling assumption of either normality or randomization. Normality means that the x_i are assumed to be independent events drawn from a normal distribution. The data are assumed to be one sample from this normal distribution and a distribution of I or c under the null hypothesis is established on the logical basis of repeated sampling from the normal distribution. Repeated sampling would not produce the same set of values, nor would repeated samples have exactly the same mean and variance as the data.

The numerical values of the data are assumed to be fixed under the assumption of randomization, but the association of values with locations is not. That is, the null hypothesis states that the observed set of values are randomly and independently distributed over the locations and a distribution for an autocorrelation statistic is based on the number of ways that the observed values could be assigned to locations. There are $n!$ equally likely assignments of data values to n locations under randomness and independence.[5] The randomization assumption restricts the possible outcomes to the values of the observed set of data (although any value could occur in any locality).

4.3 Means and Variances for Moran's I

Only the expectations for Moran's I are given here since that statistic is employed more frequently than Geary's c, and the equivalent values for care available elsewhere (Cliff & Ord, 1981a). The mean, or expected value of Moran's I, under either normality or randomization is

$$E(I) = \frac{-1}{n-1}$$

Notice that this value approaches zero for large samples.

[5] The $n!$ term indicates the factorial of n, which is

$$n(n-1)(n-2)\ldots(n-n+1)$$

The variance of Moran's I, under the assumption of normality, is

$$\mathrm{Var}(I) = \frac{n^2 S_1 - n S_2 + 3(\Sigma\Sigma w_{ij})^2}{(\Sigma\Sigma w_{ij})^2 (n^2 - 1)}$$

This expression is relatively simple because it is derived from the assumption of an underlying normal distribution. No underlying distribution is assumed under the randomization assumption and, since less is assumed, more must be calculated. The variance for randomization is based on the number of possible permutations of the n data values over the n locations,

$$\mathrm{Var}(I) = \frac{n[(n^2 - 3n + 3)S_1 - nS_2 + 3(\Sigma\Sigma w_{ij})^2]\left[\frac{1/n\Sigma(x_i - \bar{x})^4}{[1/n\Sigma(x_i - \bar{x})^2]^2}\right][S_1 - 2nS_1 + 6(\Sigma\Sigma w_{ij})^2]}{(n-1)(n-2)(n-3)(\Sigma\Sigma w_{ij})^2}$$

4.4 An Example: Respiratory Cancer in Louisiana

Maps of the incidence of particular diseases often reveal distinctive spatial patterns[6] and those patterns may provide clues to the etiology of the diseases (Mayer, 1983). Epidemics of contagious diseases are an obvious example because they often develop in systematic spatial patterns as the disease spreads from one or more origins. The patterns are systematic because transmission of an infective disease depends on interactions between infective and vulnerable individuals and the chances of sufficient contact depend very strongly on the locations and spatial behaviors of the individuals.

Spatial patterns of noncontagious diseases may also reflect distinctly geographic processes. Diseases are often associated with interactions between people and their environment, including contacts with toxic pollutants or with environments that harbor disease vectors. Consequently, the spatial pattern of disease may reflect the ways that people interact with their environment as well as the distribution of disease related elements of the environment. For example, people become infected with schistosomiasis through contact with an aquatic parasite, but the incidence of the disease may depend on human practices as well as the location of infected water bodies. People who practice irrigated agriculture are more vulnerable than others, and exposure to the disease may depend on gender roles if laundry is done in infected water bodies. Patterns of disease may also reflect cultural and social factors, such as dietary and smoking habits, which themselves have distinctive spatial patterns. Genetic predispositions are also a factor in some diseases, so the spatial pattern of disease may reflect the spatial distribution of particular ethnic populations.

Studies of the geography of disease may suggest some of the factors involved in the incidence of a disease or at least raise doubts about hypothesized etiologies that cannot explain the observed spatial patterns. The methods of spatial analysis may make valuable contributions to the study of disease although designing and interpreting such studies may depend on a knowledge of epidemiology as well as skill in spatial analysis.[7]

Figure 4.1 shows the spatial pattern of mortality rates from respiratory cancer (cancer of the trachea, bronchus, and lung) for white males among the parishes of Louisiana for the 1970-1979 period. (Parishes in Louisiana are the equivalent of counties elsewhere in the United States.) Mortality rates from lung cancer are highly variable within the United States (Blot & Fraurneni, 1982) and they are especially high in Louisiana. Riggen, Van Bruggen, Acquavella, Beaubeir, and Mason (1984) provide age-corrected mortality rates per 100,000 population for the states and counties of the United States for three time periods: 1950-1959, 1960-1969, and 1970-1979. These data indicate that death rates for white males were higher in Louisiana than in any other state for all three periods and death rates for nonwhites and for white females were also above national levels. Riggen et al. also calculate expectations of the numbers of deaths in each county (or parish) in the United States, based on the national rate. The numbers of respiratory cancer deaths among white males in the period 1970-1979 exceeded this expectation, at the 95% confidence level, in 26 of the 64 parishes of Louisiana.

[6] Some interesting maps of disease rates for the United States can be found in the atlas by Mason et al. (1975).

[7] Mayer (1983) reviews the methodological and epistemological problems of spatial analysis in medical geography and Greenberg (1983) provides an extensive study of the geography of cancer in the United States.

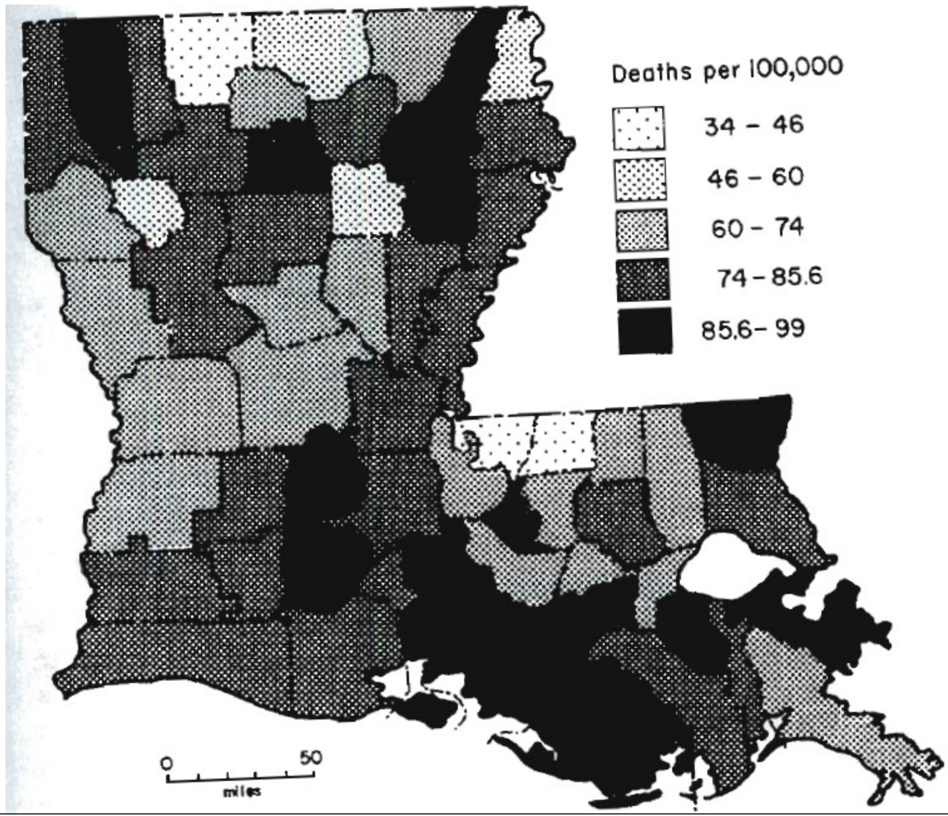


Figure 4.1 Rates for Respiratory Cancer in Louisiana Parishes, White Males, 1970-1979

The map in Figure 4.1 suggests that rates of respiratory cancer are not randomly distributed over the parishes of Louisiana. Parishes with high mortality rates appear to cluster together and mortality rates seem to be higher in South Louisiana than in the northern part of the state. That is, mortality rates from respiratory cancer seem to exhibit positive spatial autocorrelation among the parishes of Louisiana and this hypothesis can be tested using Moran's I.

The reliability and limitations of the cancer data should be considered before undertaking any analysis. Statistics on disease rates are often limited to mortality statistics, but the nature of this disease assures that mortality rates will be reasonably good indications of its prevalence. Respiratory cancer is not especially difficult to diagnose and when it is involved in a death it is very likely to be listed as the cause of death. This is not always true for some other diseases of possible geographic interest such as chronic bronchitis, emphysema, or hypertension. Respiratory cancer is also a fairly common cause of death so that reasonably reliable estimates of the rates can be based on numbers of deaths in comparatively small populations, such as the populations of Louisiana parishes. Even so, death rates for some parishes are based on fewer than twenty deaths over a decade.

The parishes of Louisiana are reasonably compact in shape and fairly uniform in size. They average 1,802 square kilometers (696 square miles), offering a spatial resolution of roughly 42.5 kilometers (26.3 miles). The simplest kind of spatial weighting function was chosen for this example. The value of w_{ij} is one when counties i and j share a boundary and zero otherwise.

Values of Moran's I for white male death rates among Louisiana parishes are shown in Table 4.1 along with the results of the test for spatial autocorrelation under the assumption of randomization. The rates exhibit strong spatial autocorrelation in all three decades although there is a slight decline in autocorrelation in the 1970s compared to the 1950s and 1960s. The variance under the randomization assumption is based on retaining the observed values of the rates and deriving the distribution of Moran's I that would result from permuting those values over the parishes in accordance with random and independent assignments. The probability values indicate that such a permutation would rarely produce a value of Moran's I as large as the observed value-less than 1% of the time for the 1950s and 1960s data and less than 3% of the time in the 1970s. The assumption of a normal distribution is associated with a variance of .00618857 so the same results would be obtained under that assumption.

TABLE 4.1
Autocorrelation Statistics for Rates of Respiratory Cancer in Louisiana: Variances Calculated Under Randomization

	Moran's I	*Expected Value*	*Variance*	*Standard Normal Deviate*	*Probability*
1950-1959	.4017	-.0159	.0062	5.1750	.0001
1960-1969	.1749		.0062	2.4308	.0075
1970-1979	.1358		.0061	1.9354	.0265

These results indicate that cancer mortality is not only higher in Louisiana but is also organized in a systematic way within the state. A more extensive analysis would be necessary to identify the spatial processes that are associated with this spatial pattern, but one worthwhile hypothesis is that spatial variations in cancer mortality are associated with spatial variations in smoking habits. Cigarette smoking is often associated with lung cancer and rates of cigarette smoking are higher in southern Louisiana than in northern Louisiana (Correa & Johnson, 1983). Spatial variations in lung cancer rates have been associated with spatial variations in smoking rates in other parts of the United States (Weinberg, Keller, & Redmond, 1982). Occupational hazards and exposure to environmental conditions (Greenberg, 1983, pp. 5-13).

Spatial analyses of cancer mortality rates that are much more thorough than the example presented here have been carried out by Glick (1979a, 1979b, 1982) and by Kennedy (in press). Glick (1979a) shows how alternative (nonspatial) models of the development of cancer can imply different spatial patterns in cancer rates. For example, exposure to ultraviolet radiation, which depends mainly on latitude, is probably involved in the development of two types of skin cancer (melanoma and nonmelanoma) but the ultraviolet radiation plays a more complex role in the development of the latter and this leads to different patterns of spatial, or latitudinal, variation in their rates. Glick also examines spatial patterns in respiratory cancer and identifies differences between the spatial patterns in respiratory cancer and identifies differences between the spatial patterns of male and female rates. (Glick, 1982).

Differences in spatial patterns between male and female rates may provide important clues to cancer risks that are linked to occupation because these differences may depend on differences in gender roles, especially gender-related differences in occupation and spatial behavior. Kennedy (in press) has carried out an analysis of spatial autocorrelation for respiratory cancer rates in 1,086 counties in eleven states in the Southeast and Gulf Coast. Female rates show a generally increasing trend from West to East buy only weak spatial autocorrelation. Male rates, on the other hand, exhibit strong spatial autocorrelation among nearby counties, a pattern that is consistent with long distance commuting to work sites where occupational hazards are localized. These differences in spatial organization may reflect differences in the etiology of cancer between men and women and, since differences in spatial pattern are difficult to attribute to biological differences between the sexes, they may reflect differences between gender roles, especially differences related to occupation.

5 AUTOCORRELATION AND REGRESSION MODELS

Spatial autocorrelation statistics were introduced in the preceding chapters as statistics that can be used own their own to investigate interdependence in spatial patterns. Other statistical models are also used to investigate spatial data and autocorrelation statistics can play an important role when they are used in conjunction with these other models. Many of these statistical models predict the values of some variable at a set of locations and this means that one result of applying the model is a map of predicted values for the variable. Autocorrelation statistics can be used to diagnose the shortcomings of the model by identifying systematic differences between the maps of predicted and actual values.

The difference between the actual value for an observation (or location) and the value predicted by a statistical model is known as an error or residual and statistical models are often evaluated on two general criteria: (1) the errors should be small and (2) the errors should be independent. Spatial autocorrelation has nothing to do with the first criterion but, when the second criterion is applied to spatial models, it means that the errors should not be autocorrelated over space. Spatial autocorrelation in the errors from a statistical model indicates that the observed values have some systematic spatial organization that is not accounted for by the model.

Regression models are widely used in geography and spatial autocorrelation statistics are especially important in fitting regression models to spatial data. Independence of the errors is a condition that a regression model must fulfill before it can be a reliable basis for testing hypotheses and tests for that condition are based on spatial autocorrelation statistics. The residuals from regression models should not be autocorrelated and spatial autocorrelation in regression residuals has two closely related implications (see Miron, 1984).

(1) Spatial autocorrelation in the residuals indicates that the model is incorrect or, at least, incomplete. A complete and correct model would explain all of the systematic spatial organization in the data and leave residuals that displayed no spatial organization at all. The model needs to be corrected in order to provide a suitable explanation of how the phenomena are organized in space.

(2) Spatial autocorrelation in the residuals indicates that the regression model fails to fulfill an important independence condition. Consequently, it will not be a reliable basis for making statistical inferences and any inferences based on the model are likely to be mistaken. The model needs to be corrected before it can be a reliable basis for testing hypotheses.

Spatial autocorrelation has sometimes been regarded as a statistical difficulty or nuisance peculiar to spatial data because of the second implication----that statistical problems and mistaken inferences are associated with autocorrelation m regression residuals. The problems do not arise primarily from unsuitable data, however, but from unsuitable models. The simplest kinds of regression models were not developed to accommodate the needs of spatial analysis and the simple regression models often must be enlarged in order to accommodate an analysis of spatial phenomena. These enlargements, and their relation to autocorrelation statistics, are discussed in this chapter.

5.1 The Linear Regression Model

The general linear regression model is very widely used to examine relations among variables and, in geography, the observations on these variables are frequently arrayed in a spatial sequence. For example, the cancer mortality rates for Louisiana parishes may be associated with other variables that could also be recorded for the parishes, such as the prevalence of cigarette smoking, or exposure to environmental and occupational hazards. A simple linear regression model, if it were fitted successfully for these data, would measure these relations in a numerical way, in the form of a linear equation that could be applied to the data for each parish:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} \ldots + b_k x_{ik} + e_i. \tag{5.1}$$

The terms in this equation are as follows.

y_i is the value of a dependent variable (such as a mortality rate) at location i (such as a parish).

$x_{i1}, x_{i2} \ldots x_{ik}$ are the values of k independent variables at the same location, such as measures of smoking or environmental hazards in the same parish.

b_0 is an estimate of a numerical constant.

b_1, b_2 and so on are estimates of a set of numerical coefficients that indicate how the variation of the dependent variable depends on variation in each of the independent variables.

e_i is an estimate of the error in predictingts y_i,or a "residual" for location i.

An estimate of the value of the independent variable for each location can be obtained by calculating the sum

$$b_0 + b_1 x_{i1} + b_2 x_{i2} \ldots + b_k x_{ik}$$

and the error term or residual, e_i is the difference between this estimate and the observed value y_i. These errors are used to calculate the mean squared error, which is an estimate of the variance of the errors. The mean squared error is used, directly or indirectly, in all of the hypothesis tests that might be performed using a regression model. The presence of spatial autocorrelation in these same errors indicates, however, that the estimate of the mean squared error will not be reliable and neither will any of the hypothesis tests that are based on that estimate.

Where the residuals are spatially autocorrelated the autocorrelation is likely to be positive and, in that case, the estimate of the mean squared error will be an underestimate of its true value. If this biased estimate is used to evaluate the model it will indicate that the model fits the data better than it actually does, and related statistics, such as the coefficient of determination, will be misleading in the same way. The mean squared error is also used to calculate standard errors for each of the regression coefficients (the b_k values) and these standard errors are then used to test the hypothesis that these regression coefficients are significantly different from zero (usually by means of t-tests or partial F-tests). A value of b_k that is significantly different from zero indicates a significant relation between the dependent variable and the variable x_k. The standard errors used in these significance tests will be underestimated whenever the mean squared error is underestimated, however, and the tests are likely to be misleading. In fact, they are likely to lead to the erroneous conclusion that variables are related when they are not.

Tests for spatial autocorrelation in regression residuals diagnose the failure to fulfill a very general condition for independence in the errors of the regression model- a condition that must be fulfilled for any regression model but that is especially likely to be violated when data are arrayed in a spatial sequence or a time series. A regression model does not provide a valid basis for statistical inferences if this condition is not fulfilled but the "problem" is not the presence of autocorrelation in the residuals but the absence of an explanation for autocorrelation in the model. The "solution" is to develop a model that does account for autocorrelation along with the effects of independent variables. Autocorrelation in the residuals usually stems from some spatial process or some peculiarity of spatial data that is not accounted for by the model, and it may be possible to develop a satisfactory model by accounting for a spatial process that is associated with autocorrelation. This may involve adding variables to the model, perhaps variables that account for interactions among locations, or revising the functional form of the model. In other cases, it will be impossible to resolve the residual autocorrelation by enlarging the explanatory variables, but satisfactory models can still be estimated by including measurements of residual autocorrelation within the model in the form of an autoregressive error term. These approaches are discussed in the last part of this chapter.

5.2 Regression Models for Spatial Data

The form of the general linear regression model can be enlarged to accommodate spatial analyses although it may be difficult to estimate the coefficients of the spatial model. The regression model has a logical basis in sampling theory and, under this logic, the observed set of y_i are presumed to be just one of many possible samples. The regression coefficients, the b_{ik}, and the residuals, the e_i, would differ for different samples and they are estimates of underlying values β_k and ϵ_i These underlying values characterize the population that underlies the presumed sampling process. It is useful to rewrite the regression model, using Greek letters to indicate the underlying population values, and using the notation of matrix algebra,

$$Y = X\beta + \epsilon \tag{5.2}$$

The left-hand term in this equation is a column vector of the values of y_i at each location; X is an n by k matrix of values for the independent variables; β is a k-element row vector of parameters; an ϵ is an n-element column vector of errors. This matrix notation is equivalent to a set of linear equations like equation 5.1 with one equation for each location.

The model is not complete until the distribution of the errors, ϵ has been specified. The errors have a probability distribution for each location, under the logic of sampling theory, even though only one member of this distribution is observed in a set of data. In the simplest kinds of regression models the errors are assumed (1) to have the same variance for every observation (or location) and (2) to be uncorrelated between pairs of observations (or locations). This condition can be written in matrix form as

$$\epsilon\epsilon' = \sigma^2 I \tag{5.3}$$

where σ^2 is the variance of the errors for each location, that is the variance that would emerge from repeated sampling. The term ϵ' is the transpose of ϵ (obtained by transforming the column vector ϵ into a row vector) and the product $\epsilon\epsilon'$ is the n by n matrix of expected variances and covariances of the errors for each observation (or location). The term I on the right-hand side is an n by n identity matrix, which has ones on the principal diagonal (where row and column both identify the same location) and zeros elsewhere. The whole specification states merely that the variance of the errors is uniform over all locations and that the covariances (and hence the correlations) of the errors are zero for all pairs of locations.

The model in equations 5.2 and 5.3 constitutes a null hypothesis for the values of y_i although estimates of the values of β must be obtained. A part of this null hypothesis, equation 5.3, states that the value of the errors are randomly and independently distributed over the locations--a hypothesis that is very similar to the null hypothesis used in testing for autocorrelation in Chapter 4. If the hypothesis represented in equation 5.3 is, in fact, true, a set of ordinary least squares estimates, b, for the β can be obtained by solving the following set of linear equations,

$$b = [X'X]^{-1}X'Y. \tag{5.4}$$

The matrix $[X'X]^{-1}$ is the inverse of the product matrix $[X'X]$. The estimates of the regression coefficients, b, can be used to calculate a set of estimates for each of the values y_i of the dependent variable along with a set of residuals, e_i. This set of estimated errors makes it possible to calculate standard errors for each regression coefficient, b_j, and test hypotheses about the significance of those regression coefficients. These standard errors will not be reliable, however, and the statistical tests will not be valid, unless the condition in equation 5.3 is fulfilled and a test for spatial autocorrelation in the observed residuals is the necessary test for this condition.

Regression models can still be defined even when the errors are autocorrelated but it is necessary to enlarge the simple specification in equations 5.2 and 5.3 to include a numerical measurement of the residual autocorrelation. Spatial autocorrelation in the errors means that the error at each location depends on the errors at other locations and if a spatial weighting function can be defined for the locations, this dependence can be written as an autoregressive function:

$$\epsilon_i = \Sigma\rho w_{ij}\epsilon_j + \upsilon_i \tag{5.5}$$

where w_{ij} are spatial weights, ρ is a parameter, and υ_i is another error term. The spatial weights for an autoregressive function are specified in the same way as the spatial weights for an autocorrelation statistic but it is necessary for the weights centered on each location to sum to one. Otherwise the parameter ρ implies a spatial trend in the value of e_i. That is, $\Sigma w_{ij} = 1$ for every i where the summation is over all j. This formulation divides the error term into two components, an *autoregressive* term that is the autocorrelated portion of the error; and υ_i, a residual that should be identically and independently distributed (not autocorrelated).

The same formulation is written in matrix terms as

$$\epsilon = \rho W\epsilon + \upsilon \tag{5.6}$$

where W is an n by n matrix of spatial weights. Substituting this for ϵ in equation 5.2 gives an enlarged model which allows for autocorrelation in the errors:

$$Y = X\beta + \rho W\epsilon + \upsilon \tag{5.7}$$

$$\upsilon\upsilon' = \sigma^2 I. \tag{5.8}$$

This model, which includes an autoregressive error term, is a much more general model for spatial analysis than the simple model in equations 5.2 and 5.3. It includes a parameter ρ that measures spatial association in the residuals and, if an estimate for ρ can be obtained, the condition in equation 5.8 should be fulfilled since the autocorrelation is accounted for by ρ[8]. This model cannot be fitted by simply solving a set of equations such as equation 5.4, however, and the major problem in using the model is to obtain a suitable estimate for ρ.

It is not always necessary to estimate ρ because it may be possible to formulate a model that accounts for autocorrelation by enlarging the independent variables or changing the functional form of the model. If regression residuals are autocorrelated the problem might be corrected in some cases by defining one or more independent variables that account for the spatial variation of the dependent variable. For example, one or more independent variables might be defined to account for interactions among locations. Spatial autocorrelation will also result if the functional form of the model is inappropriate. For example, residuals from a linear model are likely to be autocorrelated if the true relation is nonlinear. Revising the functional form or adding missing variables may be sufficient to account for the residual autocorrelation and, in that case, the value of ρ will be zero. The model in equations 5.7 and 5.8 then reduces to the simpler model in equations 5.2 and 5.3 and the ordinary least squares estimates in equation 5.4 will be suitable.

If the independence condition cannot be satisfied by a model with additional variables or a different functional form then autocorrelation can be measured by the parameter ρ and reliable inferences can be obtained from the extended form with autoregressive errors (equations 5.7 and 5.8). More complex methods are necessary to estimate the parameters of this model (Hepple, 1976; Ord, 1975).

5.3 Tests for Spatial Autocorrelation in Regression Residuals

Tests for autocorrelation in regression residuals follow the same logic as the tests already discussed but the expected values and variances for regression residuals are not the same as for other data. The independence condition that regress ion models must fulfill (equation 5.3 or equation 5.8) is defined for an unobserved matrix of sampling variances and covariances but tests of the condition are based on the observed residuals from a regression model. The appropriate null hypothesis states that the underlying population values are not autocorrelated, but even if this is true the observed residuals may be autocorrelated if the independent variables are autocorrelated.

The calculated residuals are, in matrix notation,

$$e = Y - X\beta \tag{5.9}$$

If β is a set of ordinary least squares estimates substitution from equation 5.4 and multiplication gives the matrix of variances and covariances of the calculated residuals as

$$ee' = \sigma^2[I - X(X'X)^{-1}X] \tag{5.10}$$

Since the values of ee' are not independent of X, the values of the independent variables, it is necessary for the means and variances used in testing for autocorrelation in the residuals to include measures of the autocorrelation of the independent variables. This also makes it difficult to construct a test based on a randomization assumption since there is no obvious way of randomizing the values of the residuals over the locations while maintaining a fixed level of autocorrelation in the independent variables. Consequently, the tests for autocorrelation in regression residuals are based on an assumption of normalization.

[8]It is possible that the model might still have residual autocorrelation even if an estimate of ρ is included. Unfortunately, no test for autocorrelation in the residuals from a spatial autoregression model is yet available (Cliff & Ord, 1981a, p. 240).

The expected value and variance for Moran's I are more complex expressions for regression residuals than for original data because they depend on the autocorrelation in the original data. The expected value under the null hypothesis of independence, is

$$E(I) = \frac{-(1 + I_{1x})}{n - k - 1} \tag{5.11}$$

for a regression model with k independent variables and n observations (Cliff & Ord, 1972; see also Cliff & Ord, 1981a, pp. 200-203). The term I_{1x} accounts for the effect of the independent variables on the expectation of autocorrelation in the residuals and, for the case of just one independent variable, this is just the calculated value of Moran's I for that variable (Cliff & Ord, 1972). For several independent variables a slightly more complex calculation is necessary:

$$I_{1x} = \frac{n\Sigma\Sigma w_{ij}[x_i'(X'X)^{-1}x_j]}{\Sigma\Sigma w_{ij}} \tag{5.12}$$

Notice that this is very similar to an ordinary I value for one variable except that the matrix term within the square brackets replaces the usual cross-product terms.

The equation for the variance of Moran's I for regression residuals is rather lengthy and it is helpful to define the matrix term used above as $d_{ij} = [x_i'(X'X)^{-1}x_j]$. It is also helpful to define $w_{i.}$ and $w_{.j}$ as the sums of the weights centered on region i or region j. That is, $w_{i.} = \Sigma w_{ij}$ where the summation is over regions other than region i, and $w_{.j} = \Sigma w_{ij}$ where the summation is over regions other than j. The variance is then

$$\begin{aligned} \mathrm{Var}(I) = & \frac{n}{(n-k)[\Sigma\Sigma w_{ij}]^2} \frac{n^2 S_1 - nS_2 + 3[\Sigma\Sigma w_{ij}]^2}{n^2} \\ & + (1/n)\Sigma\Sigma(w_{i.} + w_{.j})(w_{i.} + w_{.j})d_{ij} + 2[\Sigma\Sigma w_{ij} d_{ij}]^2 \\ & - [\Sigma\Sigma\Sigma(w_{ik} + w_{ki})(w_{jk} + w_{kj})d_{ij} + \Sigma\Sigma(w_{ij} + w_{ji})^2 d_{ii}] \\ & + (1/n)\Sigma\Sigma\Sigma(w_{ij} + w_{ji})(w_{ik} + w_{ki})(d_{ii}d_{jk} - d_{ij}d_{ik}) \\ & - \frac{1}{(n-k)^2} \end{aligned} \tag{5.13}$$

These values are useful for testing for autocorrelation in the residuals from ordinary least squares estimates of β and since those residuals depend on the independent variables (equation 5.9), randomization tests are not available and the formula for the variance is rather long. Residuals from some alternative estimators that do not have these problems are discussed and evaluated by Bartels and Hordijk (1977), Brandsma and Ketellapper (1979a). These estimators may not, however, possess all the desirable properties of the ordinary least squares estimators.

5.4 Fitting Spatial Regression Models

A test for spatial autocorrelation should be applied to the residuals whenever a regression model is fitted to spatial data and the autocorrelation tests can be very helpful in developing a useful and reliable model. Spatial autocorrelation in the residuals indicates that a particular model cannot support reliable inferences but the spatial pattern of residuals from this model may be very helpful in identifying the reasons for autocorrelation and it may be possible to define a suitable model by enlarging or correcting the model in some way. Autocorrelation in the residuals may be the result of the omission of a variable, especially a variable that is associated with interactions between places. Alternatively, spatial autocorrelation may result from an incorrect functional form. A map of the residuals is often very useful in diagnosing the problem.

The general autoregressive formulation (equations 5.7 and 5.8) offers a great deal of flexibility in defining a useful model because autocorrelation may be accounted for by the independent variables or measured by the parameter ρ, Variables that are added to a regression model in order to account for autocorrelation can be interpreted in the same way as other independent variables and their addition may be sufficient to account for the autocorrelation in the residuals. The parameter ρ would be reduced to zero in that case and the model could be fitted using ordinary least squares estimates.

Revising the model by adding variables or changing the functional form often will not be sufficient to account for all of the autocorrelation in the residuals but a suitable model can still be fitted by obtaining an estimate of the autoregressive parameter, ρ, in equation 5.7. The estimate of ρ will be only a numerical measurement of the residual autocorrelation and will not have the same possibilities for interpretation as an independent variable-it will indicate how much autocorrelation here is without indicating why things are autocorrelated-but the model will satisfy the independence condition in equation 5.8 and will provide reliable inferences.

The full autoregressive model cannot be fitted by using the relatively simple ordinary least squares estimates, however, and the necessary maximum likelihood estimators (Bivand, 1984; Hepple, 1976; Ord, 1975) are not as simple to calculate as the least squares estimates and are not included in standard packages of computer programs. The following sections discuss the possibilities for ordinary least squares estimates as well as models with autoregressive error terms.

ENLARGING THE EXPLANATION

The possibilities for enlarging the independent variables to account for the spatial processes that lie behind autocorrelated residuals can be illustrated by the process of fitting a simple model of a spatial labor market. One of the very simplest models of a regional labor market is based on a division of the local economy into an "export" sector and a "local" sector. Employment in the export sector depends on extraregional demand for the region's products and employment in the local sector depends on employment in the export sector. It is often difficult to categorize employment as "export" or "local" but a useful empirical model might be developed by estimating the relation between manufacturing employment and total employment. Suitable employment data are likely to be available for counties in the United States and, at the scale of counties most manufacturing is likely to be oriented to markets outside the county. A preliminary version of the model might be

$$E_i = \beta_0 + \beta_1 M_i + e_i \tag{5.14}$$

where E_i is total employment in county i, measured as the number of residents of county i who have jobs, and M_i is manufacturing employment in the same county, measured as the number of manufacturing jobs available in county i. The parameter β_1 is a ratio between manufacturing employment in the county and the total employment among the residents of the county. This model has been fitted to 1970 data for the 92 counties of Indiana (Odland, 1976) and the results are shown (as Model I) in Tables 5.1 and 5.2.

TABLE 5.1
Parameter Estimates for Four Models of Employment in Indiana Counties

Model I	$Y_i = 3332.9869$	+	$2.3665M_i$			
Model II	$Y_i = 1185.4134$	+	$2.3782M_i$	+	$.0501A_i$	
Model III	$Y_i = 1105.3234$	+	$2.3796M_i$	+	$49.1540A_i$	
Model IV	$Y_i = 1070.8286$	+	$2.3971M_i$	+	$.1552A_i$	
			(.0558)		(.0516)	$R^2 = .9546$

SOURCE: Odland(1976).

TABLE 5.2
Tests for Spatial Autocorrelation in Four Models of Employment in Indiana Counties

	Model I	*Model II*	*Model III*	*Model IV*
Calculated I	.112	.089	.128	.005
Expected Value	.011	.013	.013	.013
Standard Deviation	.047	.046	.046	.046
Standard Normal Deviate	2.151**	1.653*	2.484**	.172

SOURCE: Odland(1976)
*Significant at the .05 level; **Significant at the .01 level.

The regression model that corresponds to equation 5.14 has residuals that are strongly autocorrelated (Table 5.2) so that version of the model should be rejected as a basis for making inferences. It is not difficult to see why the residuals might be autocorrelated, however, and it may be possible to modify the model by accounting for the autocorrelation. The residual autocorrelation indicates that the labor markets in neighboring counties are interdependent and it seems likely that commuting between the counties is an important factor in this interdependence. The multiplier effects of manufacturing employment could spread between counties if commuters earn income outside their home county but spend most of their income in their home county, where their spending supports nonmanufacturing employment. Thus the total employment in county i probably depends on the manufacturing employment within commuting distance of county i and not simply on the manufacturing employment within the county.

The effect of commuting to employment outside the home county will depend very strongly on the spatial resolution of the set of counties used in the analysis. An important proportion of the work force can be expected to commute across county boundaries only if the counties are relatively small compared to the lengths of work trips. The 92 counties of Indiana are indeed fairly small, compared to reasonable commuting distances. They average about 1,000 square kilometers (390.5 square miles), offering an approximate spatial resolution of 31.7 kilometers (19.7 miles). Notice that the approximate spatial resolution is also an approximation of the distance between the centers of the counties. In fact, the census figures for 1970 indicate that almost 17% of the work force in Indiana held jobs outside their county of residence.

The effect of manufacturing employment in nearby counties can be included in the model if a variable can be defined to measure the availability of manufacturing employment outside the county,

$$E_i = \beta_0 + \beta_1 M_i + \beta_2 A_i + e_i$$

where A_i measures the availability of manufacturing employment to commuters from county i.

Four different models have been fitted to data for employment in Indiana counties in 1970 (Odland, 1976). The first (Model I) is the nonspatial model inequation 5.14. The other three incorporate competing definitions of A_i. Model II is based on a definition of A_i as the total manufacturing employment in counties adjacent to county i, where an "adjacent" county is one that shares a boundary with county i. The actual availability of employment to commuters may vary if the sets of adjacent counties vary in size. Variations in commuting distance to employment in adjacent counties are incorporated in Model III, in a crude way, by defining A_i as the ratio of total employment in adjacent counties to the total area of county i and the counties adjacent to county i. Models II and III are attempts to modify the model by incorporating information about the spatial distribution of employment opportunities outside the home county. Model IV incorporates a direct measure of spatial behavior, using census data on the numbers of persons in each county who are employed outside their county of residence. The value of A_i in that model is the total manufacturing employment in counties adjacent to county i multiplied by the percentage of workers who reside in county i but are employed elsewhere.

The results of fitting the four alternative models by ordinary least squares are shown in Tables 5.1 and 5.2. The tests for residual autocorrelation shown in Table 5.2 are based on a spatial weighting function in which $w_{ij} = 1$ if counties i and j share a boundary and $w_{ij} = 0$ otherwise. The residuals are significantly autocorrelated for Models I, II, and III but the residuals from Model IV are not autocorrelated. This indicates that the parameters for Model IV, as shown in Table 5.1, can be used as a reliable basis for making statistical inferences. The model also indicates that commuting across county boundaries plays an important role in the economies of Indiana counties and that the total employment in a county depends on the manufacturing employment in surrounding counties as well as local manufacturing employment. The omission of this variable from Model I was apparently the reason for autocorrelation in the residuals of Model I, and the reason why Model I would not be a reliable basis for testing hypotheses. Standard errors for the coefficients of the successful model (Model IV) are shown in parentheses in Table 5.1, along with a correlation coefficient for that model. The equivalent values for the other models are known to be incorrect and are not shown.

5.5 Fitting an Autoregressive Error Term

Efforts to define regression models that fulfill the independence condition by enlarging their explanatory content to include spatial processes will not always succeed. In fact, the example of Indiana employment may be a fortunate exception rather than a typical case. Even in that example, two of the three efforts to enlarge the models failed to account for autocorrelation (Models II and III). In fact the only successful model, Model IV, incorporated an explicit measure of spatial behavior, commuting, rather than a measure of the spatial distribution of employment. It is likely, in fact, that limitations on the data and uncertainties about spatial processes will mean that spatial interdependence will have to be treated as part of the unexplained error in many regression models.

Models that incorporate spatial interdependence in the form of an autoregressive error can support reliable inferences although these models relegate spatial interdependence to an error term (as indicated in equation 5.7) so they do not offer much basis for explaining interdependence in terms of some process, such as commuting. They also require the use of likelihood methods in order to estimate the parameters Hepple, 1976; Ord, 1975; see also Cliff & Ord, 1981a, chs. 6 and 9) and the application of these methods requires substantially more effort than the use of the familiar least squares methods for estimation.

The likelihood methods are necessary because dependencies in spatial data may extend in all directions. This is in contrast to time-series data where later observations can depend on earlier observations but not vice versa. This logical feature of time sequences simplifies the construction of statistical models for a time series since a value at any time, say time t, can depend on values before time t, but those earlier values are independent of values at time t. Values in a spatial sequence are, in general, mutually interdependent and although the value at one location, say at coordinates i and j, may depend on nearby values, those nearby values themselves depend on the values at location i, j. This complicates the sampling theory for spatial models and means that ordinary least squares, which is the simplest and most familiar method for estimating parameters, cannot be used for models of spatial autoregression.[9]

The importance of taking the extra trouble to estimate a model with an autoregressive error term is demonstrated by comparisons between models with autoregressive errors and models that have been fitted to the same data without accounting for spatial interdependence. For example, Hepple (1976) analyzes data on used car prices among the states of the United States. A nonspatial model for these data seems to indicate that interstate variation in the price of used cars depends on interstate variation in new car prices. A model that accounts for interdependence among nearby states indicates, however, that used car prices are autocorrelated between states but that there is no relation with variations in new car prices. The apparent significance of the nonspatial model is apparently a consequence of spatial autocorrelation.

The use of models with autoregressive errors is also illustrated by an extensive study of variations in cardiovascular mortality rates among British towns by Pocock, Cook, and Shaper (1982; see also Cook & Pocock, 1983). They find significant associations between cardiovascular mortality and several other variables, including water hardness and climatic variables. They also find that errors from an ordinary least squares model are autocorrelated. It is difficult to postulate a mechanism that would account for the spatial interdependencies in mortality rates but the model can at least be corrected for the effect of spatial autocorrelation. A corrected version of the model indicates the same basic relations between mortality and the independent variables although their effects on mortality are less pronounced than the uncorrected model would indicate.

5.6 Applications with Other Statistical Models

The relations between spatial interdependence and the regression model have been emphasized in this chapter because that model is widely applied and concern about the effects of autocorrelated errors is widespread (Miron, 1984). Spatial autocorrelation statistics can also have important applications in conjunction with other statistical models, however. Independence assumptions of some kind are involved in most statistical models and autocorrelation statistics can be used to check for spatial interdependence. Cliff and Ord (1975a) have examined the effects of autocorrelated data on tests using the t distribution and the effects of autocorrelated frequency counts on χ^2 statistics are discussed by Cliff, Martin, and Ord (1975).

[9] The estimation problem may actually be simpler for a time series of spatial data, where there are observations for each location for a series of time periods. The presence of a temporal ordering provides for a one-way dependence in that case since the value at some location at time t may depend on the values for neighboring locations at an earlier time period, but the reverse is not true. Simultaneous dependence and the associated problems are still present, however, if the value at a location also depends on contemporary values at neighboring locations.

Spatial autocorrelation statistics may also be used to diagnose systematic differences between the spatial patterns predicted by some model and the observed spatial patterns, provided that the model produces some kind of predicted map. Models that produce predicted maps are fairly common in geography. For example, the Hägerstrand-type diffusion models produce a series of predicted maps and singly constrained spatial interaction models produce predicted maps of arrivals or departures. Spatial autocorrelation in the differences between these predicted maps and the actual patterns indicates some inadequacy in the current version of the model, just as spatial autocorrelation in regression residuals indicates an inadequacy in a regression model.

5.7 Further Reading

The autoregressive formulation that has been used for the errors from regression model is a very general spatial model that may be applied to original data as well as regression residuals. In fact, equation 5.7 becomes the autoregressive model

$$Y = \rho WY + \upsilon$$

if there are no independent variables in the regression model. This spatial autoregressive model measures the dependence of values at each location on values at neighboring locations. Fitting a value for the autoregressive parameter ρ and testing the hypothesis $\rho = 0$ is very similar to the process of testing for autocorrelation using the methods of Chapters 3 and 4, but the estimation problems of autoregressive models make this approach more difficult to implement. These estimation problems stem primarily from the simultaneous dependence of values in a spatial sequence.

Useful discussions of autoregressive models for spatial data are found in Upton and Fingleton (1985), Haining (1979, 1980), Bennett (1979, ch. 7), and Cliff and Ord (1981 a, ch. 6). The seminal article on sampling theory for spatial autoregressive models is Whittle (1954).

6 AUTOCORRELATION AT DIFFERENT SCALES

The applications of autocorrelation statistics discussed so far have been limited to autocorrelations among neighboring locations. These statistics, which are sometimes called "first-order spatial autocorrelations" because they are calculated for immediate or first-order neighbors, measure the relations between neighboring elements in a spatial pattern, but the important relations in a spatial pattern are not necessarily limited to relations among immediate neighbors. Autocorrelation statistics can also be calculated to measure relations among more distant locations by defining the spatial weighting function so that the statistic is calculated for pairs of locations separated by greater distances or spatial lags. A series of autocorrelation statistics, called a *spatial correlogram,* can be obtained if the weighting function is redefined in a systematic way to include sets of locations that are more and more remote. A spatial correlogram shows spatial autocorrelation as a function of spatial lags and allows autocorrelation at different spatial lags to be analyzed and compared.

6.1 Scale Variation in Respiratory Cancer Rates

The calculation of a spatial correlogram, and some of the possible applications, can be illustrated by reviving the example of respiratory cancer rates that was introduced in Chapter 4. The results presented earlier (Table 4.1) were first-order autocorrelations, calculated on the basis of a spatial weighting function that assigned parishes as first-order neighbors if they shared a boundary, and the results indicated a strong association between rates in neighboring parishes. This indicates something about the relations in the spatial pattern of cancer rates at a particular scale, that of adjacent parishes, or, given the spatial resolution of Louisiana parishes, roughly 43 kilometers. It provides no information about relations in the pattern at other scales, although the behavior of spatial autocorrelations at different scales may reveal important information about the spatial pattern and perhaps about the processes responsible for the pattern. For example, spatial autocorrelations are often greatest for small scales, or neighboring locations, and diminish as the scale of analysis increases. This is especially likely where the autocorrelations depend on some kind of interaction behavior such as commuting. The scale where the autocorrelations diminish to zero is an important feature of the pattern, since it indicates how distant pairs of locations must be in order for events at the locations to be independent. This scale is a characteristic of a spatial pattern but it would also be characteristic of any spatial process that is responsible for the interdependence of the values in the pattern.

Autocorrelation statistics can be calculated for different scales by redefining the spatial weighting function in a systematic way so that more and more remote locations are used in calculating the statistic. In the case of the Louisiana data the spatial weighting function for the first-order spatial autocorrelation has already been defined for the calculations in Chapter 4: The weight w_{ij} is one if parishes i and j share a boundary and these neighboring parishes are called "first-order neighbors." A weighting function that defines a larger scale of analysis can be defined on the basis of "second-order neighbors." Parishes i and j are second-order neighbors, separated by second-order spatial lags, if they are not first-order neighbors but if they do have at least one neighbor in common. Equivalently, parishes i and j are second-order neighbors if it is necessary to cross only one intervening parish (or two parish boundaries) in order to travel between them. Third-order and higher-order neighbors are defined in a similar way. Two parishes are third-order neighbors if it is necessary to cross two intervening parishes to travel between them and so forth. This definition of higher-order neighbors is widely used in analyzing areal data and an algorithm to identify the sets of higher-order neighbors using matrix algebra is given by Haggett et al. (1977, pp. 319-320). Other schemes can be used to define higher-order neighbors, but the essential feature of such schemes is that they define successively more remote sets of neighbors.

Spatial correlograms for respiratory cancer rates in Louisiana parishes for three time periods are shown in Figure 6.1. These correlograms depict autocorrelation as a function of spatial lag in graphic form and the height of the bars indicates the value of autocorrelation at each spatial lag. Autocorrelations are shown only out to the fourth-order neighbors since the definition of higher neighbors is artificially restricted by the boundaries of Louisiana. Spatial autocorrelation drops off rapidly after the first order in each case and, in fact, only the first-order autocorrelations are significant. This indicates that rates are independent among more distant pairs of parishes and suggests that any explanations for the spatial association in the rates would involve processes that operate at a localized scale.

6.2 Autocorrelation, Distance, and Direction

The correlograms in Figure 6.1 show autocorrelation as a function of spatial lags where spatial lags are defined on the basis of adjacency among the parishes. The relation of the autocorrelations to distances in miles or kilometers is not precise because there are no precise relations between these spatial lags and distance measures. Since the respiratory cancer rates are available only for a set of regions of irregular size and shape, the distance relations between pairs of first-order or second-order neighbors are not uniform for all pairs and a rough estimate of the distances involved at any order is the best that can be obtained.

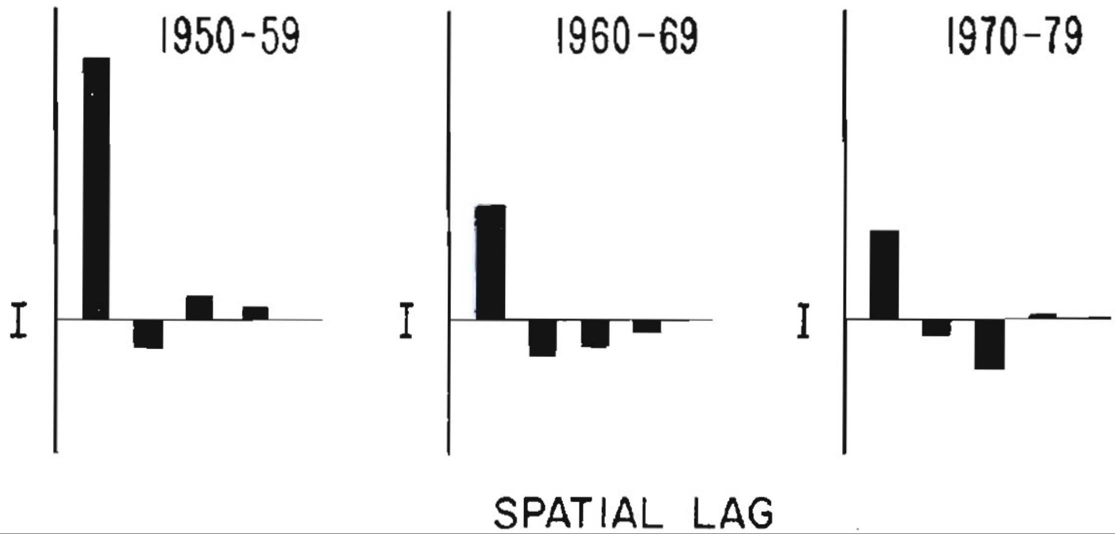


Figure 6.1 Spatial Correlograms for Respiratory Cancer Rates among White Males in Louisiana Parishes for Three Time Periods

Spatial autocorrelations can be estimated as unambiguous functions of distance where data are available on a regular grid or lattice, as in Figure 2.1. Particular spatial lags correspond to particular distances if data are regularly spaced in these ways and measures of autocorrelation can be calculated for particular distance separations, in miles or kilometers. Further, there is also a repetitive set of directional relations among the locations on a regular grid or lattice so that autocorrelations can be calculated for spatial lags that correspond to particular directions as well as particular distances. This makes it possible to calculate a two-dimensional correlogram that reveals directional regularities in the structure of spatial dependence as well as distance regularities.

Regularly spaced data are not common in geography but it is sometimes possible to aggregate irregularly spaced data onto a regular grid in order to calculate a two-dimensional correlogram. Sibert (1975) has calculated the two-dimensional autocorrelation structure of urban land values for Detroit and his results show a decline in autocorrelations over distance as well as a directional bias associated with the orientation of the street network. Gatrell (1979b) analyzed Christaller's (1966) original data for the populations of central places in southern Germany by calculating two-dimensional autocorrelation functions. His results indicate spatial dependence at considerable distances and include negative autocorrelations at short lags that match the spacing of the lowest-order central places. Two-dimensional autocorrelation functions in which autocorrelations have a precise relation to distance lags can also be used as the basis for calculating the two-dimensional spectrum of a spatial pattern, a method that makes it possible to measure the important frequencies in a two-dimensional pattern (see Rayner & Golledge, 1972).

6.3 Applications of Correlograms

Spatial correlograms provide useful information about the scale that is typical of variation in a spatial pattern but they may also be used to examine other kinds of hypotheses about spatial patterns. Differences in correlograms indicate that the associated spatial patterns are different and Sokal and Oden (1978a, pp. 215-221) have demonstrated that distinctive spatial patterns can yield different and recognizable correlograms.

For example, a regular gradient of values across a region is associated with significant positive autocorrelations at the smallest lags, declining to zero at intermediate values. The values in the correlogram then become significant but negative at high-order lags because the distant values on the gradient are related, but have negative correlations. A series of patches of homogeneous values, on the other hand, could have similar positive autocorrelations at small lags but the autocorrelations would level off at zero rather than becoming negative at high-order lags. The diagnosis of spatial pattern on the basis of correlograms must be informal and tentative, however. Formal statistical tests are available to test whether a particular correlogram is compatible with a random and independent pattern of values but no *formal* methods are available to test whether the correlogram is diagnostic of a particular spatial pattern, such as a gradient. Further, the values in a correlogram are likely to depend not only on the pattern of values in a region but also on the overall shape of the region and the locations of values within that shape, especially for the autocorrelations at higher-order lags. For example, the highest-order lags for a rectangular region are confined to lags between pairs of locations at opposite corners of the region. Consequently, the higher-order autocorrelations may reflect directional patterns (diagonal to the region's boundaries) that don't affect the calculations for the low-order lags.

Correlograms can provide very useful insights into the processes that are responsible for spatial patterns. For example, Sokal and Mennozi (1982) have analyzed information on the frequencies of blood groups for contemporary populations at 58 localities in Europe in order to examine the hypothesis of demic diffusion of early farmers into Europe from the Middle East. Demic diffusion is a process where a new population enters a region and intermarries with a resident population and, under this hypothesis, agriculture was introduced into Europe by waves of migrants who also mixed genetically with local populations. The alternatives are cultural diffusion, in which the innovation passed between neighboring populations who remained sedentary, and displacement, where invading agriculturalists either killed the local population or forced them into other regions. Neither of the alternative hypotheses involves genetic mixing on a large scale and neither implies any particular spatial pattern for the genetic characteristics of a population of descendants. Demic diffusion, on the other hand, implies that genetic characteristics are likely to become differentiated along gradients that are aligned with the routes of migration. Sokal and Menozi find that correlograms characteristic of such gradients prevail in the spatial patterns of the European blood group data, especially when the weighting matrix reflects the likely routes of migration of early agriculturalists. This confirms the earlier results of Menozzi et al. (1978) who used different methods to analyze spatial patterns in the same data.

The use of spatial correlograms to investigate spatial processes is also typified by analyses of the spatial spread of measles epidemics in Cornwall Cliff et al. (1975), who calculate correlograms for the numbers of reported measles cases in each of 27 subregions in Cornwall. Correlograms were calculated for each week in a four-year period but the important results are found for the weeks that cover two measles epidemics. Positive spatial autocorrelation is found for low-order lags during the epidemic periods, indicating that the disease spread between neighboring subregions. Positive autocorrelations at high-order lags are also characteristic of one of the epidemics, however. These high-order lags are mainly lags between urban subregions, while the lower-order lags are generally between rural subregions or between an urban and a rural subregion. The significance of both types of lags indicates that the disease probably spread hierarchically as well as spatially, traveling between distant urban centers at about the same rate as it spread to nearby rural areas. A disease epidemic is a process that develops over time as well as in space and a number of valuable analyses can be carried out by examining autocorrelations in space *and* time. Space-time autocorrelations are the subject of the next chapter.

7 AUTOCORRELATION IN SPACE AND TIME

Spatial patterns can change and develop over time and some of the most important research in geography is concerned with the evolution of spatial patterns. An analysis of the development of a spatial pattern will involve a time series of spatial data and may require statistical methods that go beyond the scope of this volume (see Bennett, 1979; Bennett & Haining, 1985), but many of the methods that are useful in the analysis of spatial-temporal data are close relatives of the autocorrelation statistics introduced in earlier chapters. These relatively straightforward methods can often be used to analyze some important hypotheses about the development of a spatial pattern.

A time series of spatial data can be thought of as a sequence of maps that show how a spatial pattern in a particular area changes over time. In some cases a spatial pattern may be invariant over long periods or it may show changes that are merely the result of random variation. Many spatial patterns are, however, the result of spatial-temporal processes in which spatial patterns change in some systematic way and the pattern at any time is related to earlier patterns. The objective of a statistical analysis will be to test hypotheses about the relations between a map pattern and those patterns that precede it. That is, spatial-temporal processes are processes that transform a map for an earlier time into the map for a later time and statistical models can be used to investigate hypotheses about these processes.

7.1 Space-Time Processes

A great variety of space-time processes is possible since almost any explanation for a spatial pattern can be associated with a process of change or development over time. For example, price competition among localized retailers may lead to distinctive patterns of price adjustments in time and space (Haining, 1983) or changes in shopping behavior may result in a shift in the spatial pattern of retailing (Bennett & Haining, 1985), a shift that may take time to develop. Some fairly general methods for analyzing space-time data are available (Bennett, 1979; Hooper & Hewings, 1981) but the most revealing analyses are often provided by statistical models that focus on the distinctive properties of a hypothesized space-time process to develop a set of hypotheses about the associated spatial and temporal patterns.

Epidemic diffusion processes are a class of space-time processes that are important in geographical research and that have distinctive properties that can provide a focus for statistical testing. Space-time methods are introduced here by concentrating on statistical analyses for this particular type of process. An epidemic diffusion process need not involve the spread of a disease but is a general type of process in which a phenomenon spreads through the conversion of individuals who remain at fixed locations--as opposed to relocation diffusion in which a phenomenon spreads through the movement of converted individuals (Brown, 1981, pp. 27-28). The spatial diffusion of innovations, as described in Hägerstrand's (1967) models, is probably the type of epidemic diffusion that is best known in geography. Hägerstrand's models describe the spread of an innovation through a spatially dispersed population. A pattern of locations where the innovation has been adopted develops as localized individuals are converted from "nonadopters" to "adopters." In the simplest of Hägerstrand's models, this conversion depends only on exposure to information and information is available from earlier adopters. The probability that an individual in a particular locality would adopt an innovation depends on his or her location with respect to persons who have already adopted the innovation because the availability of the information depends on proximity to those earlier adopters.

This type of diffusion process, which hinges on interactions between adopters and nonadopters, causes the spatial pattern to evolve in a distinctive fashion in which the pattern of adopters at any time is related to the pattern at preceding times. The locations of the earliest adopters may be random in space, but those locations exert an influence on the locations of later adoptions, which tend to cluster near the earlier adopters. As the innovation spreads, these clusters become larger and information becomes more widely available. Most of potential converts eventually become adopters, so the pattern of adopters finally approximates the distribution of population and the close relation between the locations of earlier and later adopters weakens as the innovation becomes universal.

The spatial-temporal process of innovation diffusion described by Hägerstrand's models is one of a general class of epidemic diffusion processes that also includes the spread of contagious diseases (Bailey, 1975; Cliff et al., 1981). These processes have two essential elements. First, they progress as individuals undergo a change, such as the conversion from nonadopter to adopter or, in the case of a disease epidemic, from vulnerable to infected. Second, the chances for an individual to undergo this change are conditioned by interactions with individuals who have already changed. The intensity of those interactions will usually depend, of course, on the relative locations of existing and potential converts.

These processes may be complicated by other factors. Indeed, the spread of infectious diseases is usually complicated by several elements that may not be present in the diffusion of innovations. The "converts" in a disease process may be infectious for only a limited period before they recover (or die) and cease to infect others. Further, patterns of immunity acquired during one disease outbreak will complicate the environment for the diffusion of the next epidemic by leaving a group of individuals who are immune to the disease. The processes may also be complicated by the presence of other factors that affect the conversion of individuals. For example, Odland and Barff (1982) show that the deterioration of urban housing spreads over time and space as an epidemic, in which housing is more likely to deteriorate if it is near to existing deteriorated units, but the development of a spatial pattern of deteriorated housing is complicated by another factor because the chances of deterioration also depend on the age of the housing.

7.2 Statistical Analyses for Space-Time Patterns

Despite the important differences between particular diffusion processes, they all share some common features in their development, particularly the clustering of "converted" individuals at some stages of development and these features can provide the basis for testing hypotheses about the processes. These hypotheses can center on distinctive patterns that develop over time, over space, and over space and time jointly, and spatial-temporal data can be used to examine hypotheses about all three types of patterns.

TEMPORAL PATTERNS

First, diffusion processes are associated with distinctive temporal patterns of events even if the locations of the events are ignored. That is, events do not merely accumulate over time-they accumulate in particular patterns over time. These patterns depend on one of the essential features of an epidemic diffusion process--its progress depends on interactions between converted individuals and potential converts. The simplest kinds of diffusion processes involve only one kind of event, a conversion from potential convert to convert, or from "vulnerable" to "infected." The number of these events depends on the frequency of contact between the two groups and, if the locations of the individuals are ignored, the frequency of contact depends on the relative numbers of individuals in each group.[10] The number of events, or conversions, will be small in the early stages of the process because few converts are present to influence others. The number of conversions will increase at an increasing rate as more and more individuals are converted and will reach a peak at an intermediate stage (Figure 7.la). The number of events will then decline as the number of potential converts in the population is exhausted. This behavior describes the "epidemic curve" for a contagious disease (Bailey, 1975, pp. 31-57) and some kind of epidemic curve over time is characteristic of contagious diffusion processes. In the case of a simple process where converts never leave the converted state (such as the Hägerstrand models) the cumulative number of converts is described by an S-shaped or logistic function of time (Figure 7.1b). An epidemic curve or logistic function in the development of a process over time is consistent with the operation of a contagious diffusion process in which conversions or adoptions depend on interaction with earlier converts. An epidemic curve in the temporal pattern of conversions provides evidence in favor of epidemic diffusion but it does not eliminate alternative explanations. It is possible that other processes could lead to the same temporal patterns of events. For example, one year of data on cases of hay fever would show a pattern very similar to an epidemic curve, with the peak centered on the major pollen season, even though the number of hay fever cases does not depend in any way on interactions among the victims of hay fever.

[10] A formula for an epidemic curve can be obtained from a mathematical model of the diffusion process (see Bailey, 1975, pp. 31-80), but nonspatial models of epidemics depend on an assumption of "homogeneous mixing," which states that the probability of a contact sufficient to transmit the infection is uniform for all pairs of individuals in the population. This assumption is not tenable for a population of individuals who are localized in space and, in that case, the shape of the epidemic curve would not be strictly independent of the pattern of locations. Spatial models, which incorporate contact probabilities that do depend on location, involve some mathematical complication. See Bailey (1975) or Mollison (1977).

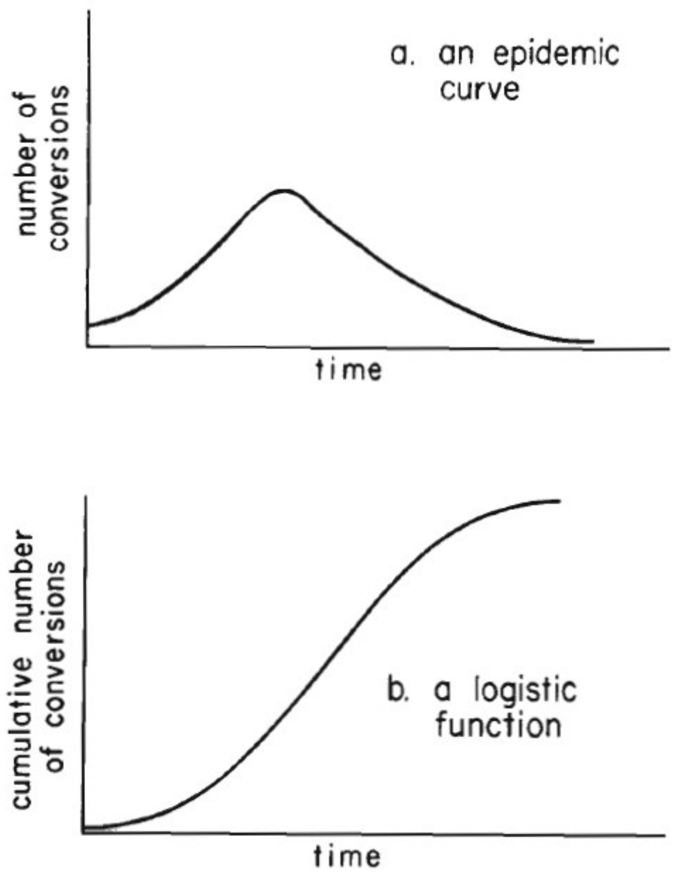


Figure 7.1 Temporal Behavior of an Epidemic Diffusion Process

SPATIAL PATTERNS

Distinctive spatial patterns are also characteristic of contagious diffusion processes if the interactions between converts and potential converts are affected by distance. The locations of the earliest converts may be unpredictable, but the pattern of converts is likely to assume a clustered distribution at intermediate stages of the process, with the clusters forming in the vicinity of the earliest adopters. (The conditions that must be fulfilled for recognizable spatial *patterns* to emerge from a spatial diffusion *process* have been derived by Mollison [1977].) The pattern of converts may eventually approach the pattern of population distribution at later stages if a substantial portion of the population is converted. The time series of maps of the contagious diffusion process develops in a particular way, therefore, with clustering or strong positive autocorrelation characteristic of the intermediate stages of pattern development.

The tests for spatial autocorrelation described in earlier chapters can be applied to each map in a time series of maps of a diffusion process. If the development of the pattern depends on interactions, and these interactions depend on distance, then the patterns of existing converts should be autocorrelated for at least some stages of the time series. Further, the strength of the autocorrelation should increase from the early to the intermediate stages of the diffusion process and may diminish toward the end of the process. These patterns are consistent with the operation of the hypothesized process but, like the hypothesized temporal patterns, they are not exclusive to that process. The same patterns could be produced by other spatial processes so, although the tests may provide evidence in favor of the hypothesis of contagious diffusion, they do not eliminate all other possibilities.

SPATIAL-TEMPORAL PATTERNS

The pattern of an epidemic curve over time, or the pattern of clustering in space, are aspects of a contagious diffusion process that are manifest when data are examined from a solely temporal or solely spatial point of view. Contagious diffusion processes also produce distinct patterns when data are examined in *both* space and time because an epidemic diffusion process is one in which the timing of events is not independent of the locations of the events, provided that the events are related by some process that depends on distance. That is, the timing of events in an epidemic diffusion process will be spatially autocorrelated so long as the contagion is affected by distance.

The timing of events in a contagious diffusion process will be spatially autocorrelated because pairs of events that are related by some kind of contagious diffusion will tend to occur at about the same time and these related events will also tend to occur close together in space. A spacetime pattern consists of a set of events (or conversions) that take place at particular locations and particular times and the events that make up such a pattern can be arrayed as pairs of events. There will be $n(n-1)/2$ pairs in a pattern that includes n events. The distance and time between the members of the pair can be recorded for each pair and the two values, a time interval and a distance interval, will not be independent if the events are related by a contagious diffusion.

The pairs of events can be arrayed in a contingency table such as the one in Figure 7.2 if the information on distance and timing is used to classify the pairs of events as "close" or "distant" in space and close or distant in time. The classification of events as near or distant in space has already been discussed at some length-it is merely the problem of assigning spatial weighting functions that occurs for any test of spatial autocorrelation. A similar problem has to be resolved in a temporal context to designate pairs of events as near or distant in time. A contagious diffusion process will produce a pattern of events in which those events that occur close together in space will also tend to occur close together in time. That is, the number of events that are close together in *both* time and space in an epidemic diffusion process would exceed the expectation for that number under a null hypothesis of independence in the location and timing of the events. An expectation for the number of pairs that would be close together in both time and space under the null hypothesis of independence can be calculated from the marginal totals of the contingency table. The expected number of events that are close in both space and time if spacing and timing are independent is the product of the number of pairs that are "close in space" times the number of pairs that are "close in time," divided by the total number of pairs.

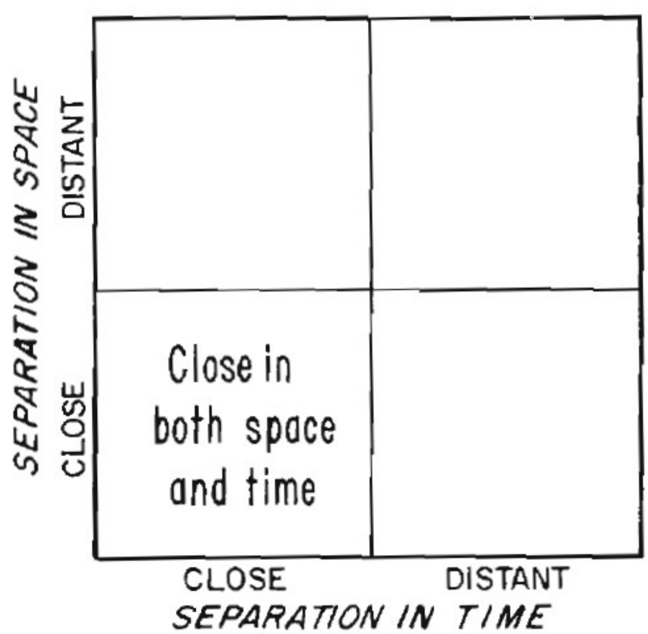


Figure 7.2 Classifiction of Pairs of Events in a Space-Time Process

The pattern of space-time autocorrelation, in which pairs of events that occur close together in time also occur close together in space, is also known as a pattern of *space-time interaction* or *space-time clustering.* Like the purely spatial or purely temporal patterns associated with epidemic diffusion, the pattern of space-time interaction is consistent with epidemic diffusion and not exclusive to such a process, although there are usually few alternative explanations for a pattern of space-time interactions.

7.3 Tests for Space-Time Autocorrelation

Tests for space-time autocorrelation were introduced by Knox (1964), who studied space-time clustering in cases of childhood leukemia, using a classification of time and distance similar to Figure 7.2. The tests have been refined by others (Mantel, 1967; Vere-Jones, 1978) and the available tests for space-time autocorrelation have been reviewed by Glick (1979). These tests have been placed in the framework of generalized cross-product statistics by Hubert et al. (1981; see also Cliff & Ord, 1981a, pp. 22-33, 1981c; Upton & Fingleton, 1985, pp. 204-208). A general space-time autocorrelation statistic can be written, in the form of the generalized cross-product statistics discussed in the second section of Chapter 1, as

$$\Gamma_{st} = \Sigma\Sigma w_{ij} y_{ij}$$

where w_{ij} is a measure of spatial proximity for the pair of events i and j, and y_{ij} is a measure of their temporal proximity. The values of w_{ij} are assigned by a spatial weighting function and those of y_{ij} by an equivalent function for temporal separation.

The space-time autocorrelation statistic Γ_{ij} is equivalent to autocorrelation statistics discussed in earlier chapters for certain choices of the function that assigns values of y_{ij}. The statistic Γ_{ij} amounts to one of the join-count statistics for categorical data if regions are categorized on the basis of the date when the process reaches the region. That is, the regions can be categorized as “black” or “white” if they are “early” or “late” in experiencing conversions. Moran’s I could be used to evaluate spacetime autocorrelation if a continuous value is used to record the timing of an event in each region. The value of y_{ij} is then the cross-product term for when the conversions occur in regions i and j.

The distributional properties discussed in earlier chapters can be used to test for space-time autocorrelation if one of the established autocorrelation statistics is used to measure space-time autocorrelation. The distributions of other test statistics must be established. The approach of Knox (1964), who classified events as near or distant in space and time (Figure 7.2), is especially useful. The statistic is merely a count of the number of pairs of events that are nearby in both space and time. (This count is calculated as a generalized cross-product by assigning $w_{ij} = 1$ for nearby pairs of locations and $w_{ij} = 0$ for distant pairs, while values of y_{ij} are one and zero for pairs of events that are close or distant in time.) The value of this count has a Poisson distribution and the distribution theory that is necessary for statistical testing has been provided by Cliff and Ord (1981a, pp. 52-53, 1981c).

REFERENCES

Adesina, H. O. (1984). Identification of the cholera diffusion process in Ibadan, 1971. *Social Science and Medicine, 18,* 429-440.

Agterburg, F. P. (1970). Autocorrelation functions in geology. In D. F. Merriam (Ed.), *Geostatistics* (pp. 113-141). New York: Plenum.

Anselin, L. (1982). A note on small sample properties of estimators in a first-order spatial autoregressive model. *Environment and Planning A, 14,* 1023-1030.

Bachtel, D. C. (1984). *The Georgia county guide (4th* ed.). University of Georgia, Cooperative Extension Service, College of Agriculture.

Bailey, N.T.J . (1975). *The mathematical theory of infectious diseases.* London: Charles Griffen.

Bannister, G. (1975). Population change in southern Ontario. *Annals of the Association of American Geographers, 65,* 177-188.

Bartels, C.P.A. (1979). Operational statistical methods for analyzing spatial data. In C.P.A. Bartels & R. H. Ketellapper (Eds.), *Exploratory and explanatory analyses of spatial data* (pp. 5-50). Boston: Martinus Nijhoff.

Bartels, C.P.A., & Hordijk, L. (1977). On the power of the generalized Moran contiguity coefficient in testing for spatial autocorrelation among regression disturbances. *Regional Science and Urban Economics,* 7, 83-101.

Bennett, R. J. (1979). *Spatial time series.* London: Pion.

Bennett, R. J., & Haining, R. P. (1985). Spatial structure and spatial interaction: Modelling approaches to the statistical analysis of geographical data. *Journal of the Royal Statistical Society, 148,* 1-36.

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

Blot, W. J., & Fraumeni, J. F. (1982). Changing patterns of lung cancer in the United States. *American Journal of Epidemiology, 115,* 664-673.

Bodson, P., & Peeters, D. (1975). Estimation of the coefficients of a linear regression in the presence of spatial autocorrelation: An application to a Belgian labour-demand function. *Environment and Planning, 7*, 445-472.

Boots, B., & Getis, A. (1988). *Point pattern analysis.* Newbury Park, CA: Sage.

Brandsma, A. S., & Ketellapper, R. H. (1979a}. Further evidence on alternative procedures for testing of spatial autocorrelation amongst regression disturbances. In C.P.A. Bartels & R. H. Ketellapper (Eds.), *Exploratory and explanatory analyses of spatial data* (pp. 113-136). Boston: Martinus Nijhoff.

Brandsma, A. S., & Ketellapper, R. H. (1979b). A biparametric approach to spatial autocorrelation. *Environment and Planning A, 11*, 51-58.

Brown, L.A. (1981). *Innovation diffusion: A new perspective.* New York: Methuen.

Burridge, P. (1981). Testing for a common factor in a spatial autoregressive model. *Environment and Planning A, 13,* 795-800.

Casetti, E. (1972). Competitive spatial equilibrium by mathematical programming: An alternative frame of reference. *Geographical Analysis, 4,* 368-372.

Christaller, W. (1966). *Central places in southern Germany* (C. W. Baskin, Trans.). Englewood Cliffs, NJ: Prentice-Hall.

Cliff, A. D., Martin, R. L., & Ord, J. K. (1975). A test for spatial autocorrelation in choropleth maps based upon a modified χ^2 statistic. *Transactions and Papers of the lnstitute of British Geographers, 65,* 102-129.

Cliff, A. D., & Ord, J. K. (1969). The problem of spatial autocorrelation. In A. J. Scott (Ed.), *Studies in regional science* (pp. 25-55). London: Pion.

Cliff, A. D., & Ord, J. K. (1972). Testing for spatial autocorrelation among regression residuals. *Geographical Analysis, 4,* 267-284.

Cliff, A. D., & Ord, J. K. (1973). *Spatial autocorrelation.* London: Pion.

Cliff, A. D., & Ord, J. K. (1975a). The comparison of means when samples consist of spatially autocorrelated observations. *Environment and Planning,* 7, 725-734.

Cliff, A. D., & Ord, J. K. (1975b). The choice of a test for spatial autocorrelation. In J. C. Davis & M. J. McCullagh (Eds.), *Display and analysis of spatial data* (pp. 54-77). New York: John Wiley.

Cliff, A. D., & Ord, J. K. (1975c). Model building and the analysis of spatial pattern in human geography. *Journal of the Royal Statistical Society, Series B, 37,* 297-348.

Cliff, A. D., & Ord, J. K. (1977). Large-sample distributions of statistics used in testing for spatial correlation: A comment. *Geographical Analysis, 9,* 297-299.

Cliff, A. D., & Ord, J. K. (1981a). *Spatial processes: Models and applications.* London: Pion.

Cliff, A. D., & Ord, J. K. (1981b). The effects of spatial autocorrelation on geographical modelling. In R. L. Craig & M. L. Labovitz (Eds.), *Future trends in geomathematics* (pp. 108-137). London: Pion.

Cliff, A. D., & Ord, J. K. (198lc). Spatial and temporal analysis: Autocorrelation in space and time. In N. Wrigley & R. J. Bennett (Eds.), *Quantitative geography: A British view* (pp. 86-91). London: Routledge & Kegan Paul.

Cliff, A. D., Haggett, P., Ord, J. K., Bassett, K., & Davies, R. B. (1975). *Elements of spatial structure: A quantitative approach.* London: Cambridge University Press.

Cliff, A. D., Haggett, P., Ord, J. K., & Versey, G. R. (1981). *Spatial diffusion: An historical geography of epidemics on an island community.* London: Cambridge University Press.

Cook, D. G., & Pocock, S. J. (1983). Multiple regression in geographical mortality studies, with allowance for spatially correlated errors. *Biometrics, 39,* 361-371.

Costanzo, C. M. (1983). Statistical inference in geography: Modern approaches spell better times ahead. *Professional Geographer, 35,* 158-164.

Curry, L. (1983). Elements of spatial statistical systems analysis. *Professional Geographer, 35,* 149-157.

Dacey, M. F. (1968). A review of measures of contiguity in two and K-color maps. In B.J.L. Berry & D. F. Marble (Eds.), *Spatial analysis* (pp. 479-495). Englewood Cliffs, NJ: Prentice-Hall.

DeJong, D., Sprenger, C., & Van Ween, F. (1984). On extreme values of Moran's I and Geary's C. *Geographical Analysis, 16,* 17-24.

Dorien, P. (1981). Estimating linear models with spatially distributed data. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 359-388). San Francisco: Jossey-Bass.

Ebdon, D. (1977). *Statistics in geography: A practical approach.* Oxford: Basil Blackwell.

Gabriel, K. R., & Sokal, R.R. (1969). A new statistical approach to geographic variation analysis. *Systematic Zoology, 18,* 259-270.

Gatrell, A. C. (1977). Complexity and redundancy in binary maps. *Geographical Analysis, 9,* 29-41.

Gatrell A. C. (I 979a). Autocorrelation in spaces. *Environment and Planning A, 11,* 507-516.

Gatrell, A. C. (1979b). The autocorrelation structure of central places in southern Germany. In N. Wrigley (Ed.), *Statistical applications in the spatial sciences* (pp. 111-126). London: Pion.

Gatrell, A. C. (1983). *Distance and space: A geographical perspective.* Oxford: Clarendon.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *Incorporated Statistician, 5,* 115-145.

Getis, A. (1983). Second-order analysis of point patterns: The case of Chicago as a multi-center urban region. *Professional Geographer, 35,* 73-80.

Getis, A. (1984). Interaction modeling using second-order analysis. *Environment and Planning A, /6,* 173-184.

Glick, B. J. (1979a). The spatial autocorrelation of cancer mortality. *Social Science and Medicine, 13D,* 123-130.

Glick, B. J. (1979b). Distance relationships in theoretical models of carcinogenesis. *Social Science and Medicine, 13D,* 253-256.

Glick, B. J. (1979c). Tests for space-time clustering used in cancer research. *Geographical Analysis, 11,* 202-208.

Glick, B. J. (1982). The spatial organization of cancer mortality. *Annals of the Association of American Geographers, 72,* 471-481.

Greenberg, M. R. (1983). *Urbanization and cancer mortality: The United States experience, 1950-75 .* New York: Oxford University Press.

Greenberg, M. R., McKay, F., & White, P. (1980). A time-series comparison of cancer mortality rates in the New Jersey-New York-Philadelphia metropolitan region and the remainder of the United States, 1950-69. *American Journal of Epidemiology, 111,* 166-174.

Griffith, D. A. (1980). Towards a theory of spatial statistics. *Geographical Analysis, 12,* 325-329.

Griffith, D. A. (1983). The boundary value problem in spatial statistical analysis. *Journal of Regional Science, 23,* 377-387.

Griffith, D. A. (1984a). Theory of spatial statistics. In G. L. Gaile &C. J. Willmott (Eds.), *Spatial statistics and models* (pp. 3-15). Dordrecht: D. Reidel.

Griffith, D. A. (1984b). Reexamining the question "are locations unique?" *Progress in Human Geography, 8,* 82-95.

Hägerstrand, T. (1967). *Innovation diffusion as a spatial process.* Chicago: University of Chicago Press.

Haggett , P. (1976). Hybridizing alternative models of an epidemic diffusion process. *Economic Geography, 52,* 136-146.

Haggett, P., Cliff, A. D., & Frey, A. E. (1977). *Locational analysis in human geography* (2nd ed.). London: Edward Arnold.

Haining, R. (1979). Statistical tests and process generators for random field models. *Geographical Analysis, 11,* 45-64.

Haining, R. (1980). Spatial autocorrelation problems. In D. T. Herbert & R. J. Johnson (Eds.), *Geography and the urban environment, progress in research and applications* (Vol. 3, pp. 1-44). New York: John Wiley.

Haining, R. (1981). Spatial and temporal analysis: Spatial modelling. In N. Wrigley & R. J. Bennett (Eds.), *Quantitative geography: A British view* (pp. 86-91). London: Routledge & Kegan Paul.

Haining, R. (1983). Modeling intraurban price competition: An example of gasoline pricing. *Journal of Regional Science, 23,* 517-528.

Haynes, K. J., & Fotheringham, A. S. (1984). *Gravity and spatial interaction models.* Newbury Park, CA: Sage.

Hepple, L. W. (1976). A maximum-likelihood model for econometric estimation with spatial series. In I. Masser (Ed.), *Theory and practice in regional science* (pp. 90-104). London: Pion.

Hepple, L. W. (1979). Bayesian analysis of the linear model with spatial dependence. In C.P.A. Bartels & R. H. Ketellapper (Eds.), *Exploratory and explanatory statistical analysis of spatial data* (pp. 179-199). Boston: Martinus Nijhoff.

Hodder, I., & Orton, C. (1976). *Spatial analysis in archaeology.* London: Cambridge University Press.

Hooper, P. M., & Hewings, G.J.D. (1981). Some properties of space-time processes. *Geographical Analysis, 13,* 202-223.

Hordijk, L. (1974). Spatial correlation in the disturbances of a linear interregional model. *Regional and Urban Economics, 4,* 117-140.

Hubert, L. J. (1978). Non-parameteric tests for patterns in geographic variation: Possible generalizations. *Geographical Analysis, 10,* 86-88.

Hubert, L. J., & Golledge, R. G. (1982a). Measuring association between spatially defined variables: Tjostheim's index and some extensions. *Geographical Analysis, 14,* 273-278.

Hubert, L. J., & Golledge, R. G. (1982b). Comparing rectangular data matrices. *Environment and Planning A, 14,* 1087-1095.

Hubert, L. J., Golledge, R. G., & Costanzo, C. M. (1981). Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis, 13,* 224-233.

Jumars, P.A. (1978). Spatial autocorrelation with RUM: Vertical and horizontal structure of a bathyal benthic community. *Deep-Sea Research, 25,* 589-604.

Jumars, P. A., Thistle, D., & Jones, M. L. (1977). Detecting two-dimensional spatial structure in biological data. *Oecologia, 28,* 109-123.

Kennedy, S. (in press). A geographic regression model for medical statistics. *Social Science and Medicine.*

Kennedy, S., & Tabler, W. (1983). Geographic interpolation. *Geographical Analysis, 15,* 151-156.

Knox, E. G. (1964). The detection of space-time interactions. *Applied Statistics, 13,* 25-29.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research, 27,* 209-220.

Martin, R. L. (1974). On autocorrelation, bias and the use of first spatial differences in regression analysis. *Area, 6,* 185-194.

Martin, R. L., & Oeppen, R. L. (1975). The identification of regional forecasting models using space time correlation functions. *Transactions and Papers, Institute of British Geographers, 64*, 95-ll8.

Mason, T. J., McKay, F. W., Hoover, R., Blot, W. J., & Fraumeni, J. F. (1975). *Atlas of cancer mortality for U.S. counties, 1950-1969.* Washington, DC: U.S. Public Health Service National Institutes of Health.

Mayer, J.D. (1983). The role of spatial analysis and geographic data in the detection of disease clustering. *Social Science and Medicine, 17*, 1213-1221.

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

Mead, R. (1974). A test for spatial pattern at several scales using data from a grid of contiguous quadrants. *Biometrics, 30,* 295-307.

Menozzi, P., Piazza, A., & Cavelli-Sforza, L. (1978). Synthetic maps of human gene frequencies in Europeans. *Science, 201,* 786-792.

Miron, J. (1984). Spatial autocorrelation in regression analysis: A beginner's guide. In G. L. Gaile & C. J. Willmott (Eds.), *Spatial statistics and models* (pp. 201-222). Dordrecht: D. Reidel.

Mollison, D. (1977). Spatial contact models for ecological and epidemic spread. *Journal of the Royal Statistical Society, Series B, 39,* 283-326. ·

Moran, P.A.P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B, 37,* 243-251.

Morrill, R. L. (1983). The nature, unity and value of geography. *Professional Geographer, 35,* 1-9.

Oden, N. L. (1984). Assessing the significance of a spatial correlogram. *Geographical Analysis, 16,* 1-16.

Odland, J. (1976). *The estimation of employment multipliers in a spatial context.* Proceedings of the Business and Economics Section of the American Statistical Institute.

Odland, J., & Barff, R. (1982). A statistical model for the development of spatial patterns: Applications to the spread of housing deterioration. *Geographical Analysis, 14,* 326-339.

Ord, J. K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association, 70,* 5-50.

Ord, J. K. (1980). Tests of significance using non-normal data. *Geographical Analysis, 12,* 387-392.

Pocock, S. J., Cook, D. G., & Shaper, A.G. (1982). Analysing geographic variation in cardiovascular mortality: Methods and results. *Journal of the Royal Statistical Society, Series A, 145,* 313-341.

Rayner, J. N., & Golledge, R. G. (1972). Spectral analysis of settlement patterns in diverse physical and economic environments. *Environment and Planning A, 4,* 347-371.

Rayner, J. N., & Golledge, R. G. (1973). The spectrum of U.S. Route 40 reexammed. *Geographical Analysis, 4,* 338-350.

Riggan, w. B., Van Bruggen, J., Acquavella, J. F., Beaubeir, J., & Mason, T. J. (1984). *U.S. cancer mortality rates and trends 1950-79.* Washington, DC: Government Ponting Office.

Ripley, B. D. (1981). *Spatial statistics.* New York: John Wiley.

Ripley, B. D. (1984). Spatial statistics: Developments 1980-3. *International Statistical Review, 52,* 141-150.

Royalty, H. H., Astrachan, F., & Sokal, R. R. (1975). Tests for patterns in geographic variation. *Geographical Analysis, 8,* 175-184.

Sakai, A. K., & Oden, N. L. (1983). Spatial pattern of sex expression in silver maple (Acer Saccherinum L.): Morisita's index and spatial autocorrelation. *American Naturalist, 122,* 489-508.

Sen, A. K. (1976). Large sample-size distributions of statistics used in testing for spatial correlation. *Geographical Analysis, 9,* 175-184.

Sen, A. K. (1977). Large sample-size distributions of statistics used in testing for spatial correlation: A reply. *Geographical Analysis, 8,* 300.

Sen, A. K., & Soot, S. (1977). Rank tests for spatial correlation. *Environment and Planning A, 9,* 897-903.

Sibert, J. (1975). *Spatial autocorrelation and the optimal prediction of assessed values* (Geographical publication no. 14). Ann Arbor: University of Michigan.

Sokal, R. R. (1979a). Testing the statistical significance of geographically varying patterns. *Systematic Zoology, 28,* 227-232.

Sokal, R. R. (1979b). Ecological parameters inferred from spatial correlograms. In G. P. Patil & M. L. Rosenzweig (Eds.), *Contemporary ecology and related econometrics* (pp. 167-179). Fairland, MD: International Cooperative Publishing House.

Sokal, R. R. (1983). Analyzing character variation in geographic space. In J. Felsenstien (Ed.), *Numerical taxonomy* (pp. 384403). New York: Springer-Verlag.

Sokal, R. R., & Friedlander, J. (1981). Spatial autocorrelation analysis of biological variation on Bougainville island. In M. H. Crawford & J. H. Mielke (Eds.), *Current developments in anthropological genetics: Population structure and ecology* (pp. 205-227). New York: Plenum.

Sokal, R.R., & Menozzi, P. (1982). Spatial autocorrelations of HLA frequencies support demic diffusion of early farmers. *American Naturalist, 119,* 1-17.

Sokal, R. R., & Oden, N. L. (1978a). Spatial autocorrelation in biology 1: Methodology. *Biological Journal of the Linnean Society, 10,* 199-228.

Sokal, R.R., &Oden, N. L. (1978b). Spatial autocorrelation in biology 2: Some biological applications of evolutionary and ecological interest. *Biological Journal of the Linnean Society, 10,* 229-249.

Sokal, R. R., & Wartenberg, D. E. (1981). Space and population structure. In D. A. Griffith & R. D. MacKinnon (Eds.), *Dynamic spatial models* (pp. 186-213). Alpen van den Rijn: Sijthoff & Noordhoff.

Sokal, R.R., & Wartenberg, D. E. (1983). A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics, 105,* 219-238.

Student. (1914). The elimination of spurious correlations due to position in time or space. *Biometrika, 10,* 179-180.

Tobler, W.R. (1969a). Geographical filters and their inverses. *Geographical Analysis, 1,* 234-253.

Tobler, W. R. (1969b). The spectrum of U.S. 40. *Papers of the Regional Science Association, 23,* 45-52.

Tobler, W.R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, 46,* 234-240.

Tobler, W. R. (1975). Linear operators applied to areal data. In J. C. Davis & M. J. McCullagh (Eds.), *Display and analysis of spatial data* (pp. 14-38). London: John Wiley.

Tobler, W.R. (I979a). Cellular geography. In S. Gale & G. Olsson (Eds.), *Philosophy in geography* (pp. 379-386). Dordrecht: D. Reidel.

Tobler, W. R. (I979b). Smooth pycnophlyactic interpolation for geographic regions. *Journal of the American Statistical Association, 74,* 519-536.

Tobler, W.R. (1984). Applications of image processing techniques to map processing. In K. Brassel (Ed.), *Proceedings of the international symposium on spatial data handling* (pp. 140-144). Zurich.

Tobler, W.R., & Kennedy, S. (1984). *Smooth multidimensional interpolation.* Unpublished manuscript, University of California, Santa Barbara, Department of Geography.

Unwin, D. J., & Hepple, L. W. (1974). The statistical analysis of spatial series. *Statistician, 23,* 211-227.

Upton, G.J.G., & Fingleton, B. (1985). *Spatial data analysis by example.* New York: John Wiley.

Vere-Jones, D. (1978). Space-time correlations for micro earthquakes: A pilot study. *Advances in Applied Probability, 10,* 73-87.

Weinberg, G. B., Keller, L. H., & Redmond, C. K. (1982). The relationship between geographic distribution of lung cancer incidence and cigarette smoking in Allegheny county, Pennsylvania. *American Journal of Epidemiology, 115,* 40-58.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika, 41,* 434-449.

Wrigley, N. (1985). *Categorical data analysis for geographers and environmental scientists.* London: Longman.

ABOUT THE AUTHOR

JOHN ODLAND is Associate Professor of Geography at Indiana University. He earned his Ph.D. at the Ohio State University and his major research interests are in economic geography, urban geography, and migration. Much of his research has been concerned with mathematical and statistical modeling, including models of the development of patterns in space and time.

SCIENTIFIC GEOGRAPHY SERIES

This series presents the contributions of scientific geography in small books or modules. Introductory modules are designed to reduce learning barriers; successive volumes gradually increase in complexity, preparing the reader for contemporary developments in this exciting field. Emphasizing practical utility and real -world examples, this series of modules is intended for use as classroom texts and as reference books for researchers and professionals.

Volume 1
CENTRAL PLACE THEORY *by Leslie J King*

Volume 2
GRAVITY AND SPATIAL INTERACTION MODELS *by Kingsley E. Haynes* & *A. Stewart Fotheringham*

Volume 3
INDUSTRIAL LOCATION *by Michael J Webber*

Volume 4
REGIONAL POPULATION PROJECTION MODELS *by Andrei Rogers*

Volume 5
SPATIAL TRANSPORTATION MOD ELING *by Christian Werner*

Volume 6
REGIONAL INPUT-OUTPUT ANALYSIS *by Geoffrey J. D. Hewings*

Volume 7
HUMAN MIGRATION *by W.A.V. Clark*

Volume 8
POINT PATTERN ANALYSIS *by Barry N. Boots* & *Arthur Getis*

Volume 9
SPATIAL AUTOCORRELATION *by John Odland*

Volume 10
SPATIAL DIFFUSION *by Richard Morrill. Gary L. Gaile* & *Grant Ian Thrall*

SAGE PUBLICATIONS
The Publishers of Professional Social Science
Newbury Park Beverly Hills London New Delhi