AN INTEGRATED SPATIAL STATISTICS PACKAGE FOR MAP DATA ANALYSIS

Barry J. Glick and Stephen A. Hirsch
PAR Technology Corporation
P.O. Box 2950
Landover Hills, Maryland 20784

ABSTRACT

In the analysis of geocoded statistical data, common practice has been to treat the data in isolation from its locational or spatial characteristics. This results in a potentially critical loss of the spatial information that is contained in the mapped representation of the statistical data but not in the application of aspatial statistical techniques such as cross-sectional regression. One explanation for the domination of aspatial methods is the absence of an integrated software package to carry out explicitly spatial data analysis on geocoded databases. In this paper we describe the functional elements of such a package, based on our development of a typology of general-purpose statistical map data analysis needs.

The design of the integrated spatial statistics package includes a module for transforming digital coordinate data (e.g. polygonal boundary files) into concise mathematical representations of relative location for use in the analytic procedures. We distinguish between several categories of locationally-referenced data including point patterns, point-sampled data, grid-cell data, irregular polygon-based data and network and flow data. Within these classes we consider a hierarchy of spatial analytic objectives: quantitative pattern description and randomness testing, univariate modeling, spatial interpolation, space series analysis, multivariate spatial analysis, spatial forecasting and control, and dynamic and space/time analysis. Examples of the application of the package to typical spatial statistical problems are discussed, with emphasis on the hierarchial nature of the analytic process.

INTRODUCTION

The growing acceptance of the concept of a geographic information system or a spatial data handling system reflects the realization of the uniqueness of spatial data. This uniqueness has motivated the search for input techniques, data structures, and query languages appropriate to spatial data processing. These specialized resources are integrated within a geographic information system to provide the capability to enter, store, retrieve, manipulate, and display spatial data.

An additional function included in most detailed definitions of geographic information systems is spatial data analysis. In contrast to spatial data manipulation or massaging, spatial data analysis is usually defined as the process of deriving useful information from the data

as opposed to realligning or transforming the data. In addition, the goal of analysis is the interpretation of the derived results whereas the goal of manipulation is the attainment of data in a form more suitable for subsequent storage, display, or analysis; that is, data manipulation in itself does not normally produce information that is useful in problem solving.

The analytic components of existing systems contain diverse capabilities. In part this pluralism reflects the variety of applications and problem areas that geographic information systems address and the wide variety of analytic tools that may be appropriate for specific applications. Nevertheless, a small set of generic analytic tools are common to many geographic information systems - these include simple measurements such as the length of line segments and area of regions, polygon overlay, point-in-polygon determination, and perhaps some form of spatial interpolation and contouring. This existence of a defacto set of generic spatial data analysis functions suggests the possibility of developing a scheme for classifying analytic functions and identifying, for a particular application, those that may be appropriate to include in a particular geographic information system.

A category of spatial data analysis that has been largely ignored by designers of geographic information system is statistical spatial analysis. When statistical techniques are included at all in a system (perhaps through linkage to an existing package of statistical methods), they are nearly always standard, aspatial techniques. The use of these techniques to analyze spatial data always results in a potentially significant loss of information and sometimes results in misuse of methods due to the violation of assumptions normally not met with spatial data. To paraphrase Tobler's analogy (Tobler, 1980), the use of aspatial techniques - e.g. regression analysis - on spatial data is akin to arranging monthly time-series data according to the alphabetical order of the months and then analyzing it. The availability, in a geographic information system, of precise digital descriptions of location along with attributes of those locations, provides an ideal opportunity for making use of truly spatial analytic tools.

The purpose of this paper is to present a preliminary functional description for a general purpose statistical spatial data analysis element that could be included, in part or in whole, in a geographic information system. This functional design is based upon an attempt to identify the various classes of spatial data analysis. Following the functional description, a brief example of the application of some of the tools discussed is presented.

A TYPOLOGY OF SPATIAL ANALYSIS FUNCTIONS

There are many alternate approaches that can be taken to identify and classify statistical spatial analysis functions. For effectiveness and simplicity of presentation, an approach based on a classification of statistical

and spatial data types will be used. Three types of spatial data are portrayed on maps: point, line (or network), and area. In addition, three major levels of measurement for statistical data exist: nominal (presence/non-presence or binary), ordinal or rank, and interval/ratio. The combinations of spatial data type and level create the need for alternative statistical map types. For example, a dot map may be used to graphically depict the spatial pattern of nominal-level point-located phenomena and a graduated point symbol map used for ratio-level point-located or point-sampled phenomena.

The cross-classification of spatial data by level of measurement and spatial dimensionality yields nine major categories. For the purposes of this paper, rank-ordered data is not considered as generating substantially distinctive spatial analytic requirements and is therefore considered together with the ratio-level data category (although some specialized spatial analytic tools have been developed for rank-ordered data: see, for example, Glick 1982).

In addition to the definition of classes of spatial data to be used in statistical spatial analysis, a hierarchical categorization of generic types of analyses is possible. For example, at an initial problem solving stage, information can be extracted individually from single-variable statistical maps. We refer to this as univariate spatial analysis. In the bivariate or more general multi-variate class of analysis, the relationship between two or more statistical maps is investigated. Finally, map data may be used to generate new maps; for example, maps of future conditions or maps covering areas where empirical data is not available. For lack of a better lable, we refer to this category of analysis as spatial forecasting.

Univariate Analysis: Nominal Data

Three fundamental objectives of univariate statistical spatial analysis are to provide: 1) concise descriptions of important spatial aspects of the mapped data; 2) inferential tests related to the descriptive measures; and 3) modeling of the spatial variation in the data. As shall be discussed in more detail, these three goals can be viewed as steps in an iterative process for deriving quantitative information from mapped data.

Point Patterns - The analysis of nominal-level point data (i.e. point pattern analysis) is relatively well-developed methodologically - probably due to its application in a wide variety of disaplines including plant ecology, geography, archeology, and geology. A wide range of descriptive measures are available, some of which provide useful descriptive information but do not interface well with follow-on statistical analysis procedures. For example, the well-known centrographic measures such as the bivariate median and standard distance (Bachi, 1963) are useful in some applications but lack a strong connection to inferential testing and model-building.

Among the more popular and useful measures are distance-based measures (e.g. nearest-neighbor distances) and quadrat or cell-count measures. Many refinements of the original work in these areas have been and continue to be made. Good references that describe these measures and their relative strengths and weaknesses are available (e.g. Rogers 1974; Ripley 1981). Polygon techniques, in which the spatial point pattern is transformed into a polygon structure (usually based on a Thiessen or Delaunay polygon approach) can also be used to produce descriptive measures of a point pattern (Boots, 1974).

The distance-based, quadrat count, and polygon-based measures can be used in hypothesis-testing due to their connection with theoretical statistical models of spatial variation. For example, the simplest and earliest distance based measure, the mean distance between nearest neighbors, has an expected value of one-half the square root of the density parameter for the region under study. This expectation is based on an underlying two-dimensional Poisson process - a process that is commonly used as the baseline independent random mechanism for placing points in the plane.

This connection of measures of spatial point patterns to theoretical random processes of point allocation immediately suggests the most fundamental of inferential tests on spatial patterns - randomness tests. The purpose of a randomness test for a spatial pattern is to evaluate the possibility that the points are spatially distributed such that the location of other points has absolutely no bearing on the location of a particular point. Thus, if the null hypothesis of randomness cannot be rejected, the conclusion is that the pattern could have been generated by the planar Poisson process.

The distance-based and quadrat methods can also be used to test the liklihood that an observed spatial point pattern could have arisen from a particular non-random generating process. Models of both clustering and more-regular-than-random processes exist (see Getis and Boots 1978) and inferential tests have been employed to assess the liklihood that an observed pattern could have been generated by these processes.

If a generating model can be identified for a point pattern it is possible to estimate the parameters of the model. The estimated model can then be used as a concise description of the spatial characteristics of the observed phenomenon, as a source of interpretable information about the possible mechanism that has produced the pattern, as a basis for forecasting.

Line Patterns and Networks

The statistical spatial analysis of nominal one-dimensional or linear data is much less developed than for point or area data. In this case we have a linear pattern in space and the goal is to attain information on the locational aspects of the pattern. Independent random generating processes, analogous to the Poisson point process, have

been suggested for generating a random pattern of straight lines in a plane (e.g. Bartlett 1975). Applications for this approach include the testing of the randomness of line transects in spatial sampling situations.

Many descriptive measures exist for assessing the geometric and topological attributes of networks. These include measures of circuitry, connectedness, redundancy, and hierarchical structure of the network. In addition, we can measure the complexity of line segments and determine bifurcation or branching measures, etc. When a network is represented (with attendant loss of information) as a connectivity matrix of modes, the measures and methods of graphtheory provide a powerful analytic methodology (Christofides 1975). The application of statistical inference and modeling in network aralysis has been undertaken by geomorphologists and hydrologists interested in the statistics of the topological properties of stream networks (Haggett and Chorley, 1969).

Area/Polygon Pattern

Nominal area data can be represented as simple polygon outline maps. One set of descriptive measures related to this class of spatial data are those used to describe the shape and form of areal features. Shape measures are numerous and range from highly simplified one-parameter indices to complex representations containing information on boundary complexity, compactness, and directional bias (Bookstein 1978; Mandelbrot 1977).

Analysis may also be carried out on the polygon net or tesselation of the plane taken as a whole. Typical descriptive measures are the mean number of polygon sides, mean polygon perimeter, mean length of a side, and mean area of polygons. For Thiessen polygons, expectations for these measures based on using a Poisson point process to generate the polygons, have been derived. Another approach to creating a random polygon net is by partitioning an area by a system of random lines (Haggett et al 1977).

Univariate Analysis: Ratio-Level Data

In nominal-level spatial data analysis, the objects of the analysis are the digital coordinates that define the location of the entity. The ratio-level case takes this data into consideration but also uses the corresponding set of numerical attribute or descriptive data relating to the locational entities.

Ratio-level point data, in many real-world applications, are point samples of spatially continuous phenomena. In these cases it is customary to use the sample points to estimate a continuous surface, as can be represented by isoplethic mapping techniques. In our scheme, methods for carrying this transformation out are considered to be in the spatial forcasting/interpolation category of analyses.

In the special case of points that fall on vertices of a regular lattice such as a square a triangular grid,

methods generalized directly from time-series analysis can be used. Thus we have two-dimensional correlograms or spatial autocorrelation functions, two-dimensional periodograms, and two dimensional spectral density functions. These descriptive measures of spatial pattern provide robust and precise information on key characteristics of the spatial data - for example, clustering of similar values, spatial periodicity, major directional biases or non-isotropy and scale effects. Two-dimensional spectral density and autocorrelation measures are also very well suited to inferential testing and model-building (Besag 1974; Ripley, 1981).

Through the use of schemes to define neighboring or connecting data units (i.e. a weighting matrix), the spatial spectral density and autocorrelation methods can be extended to the more general case of irregular data points and irregular polygons (Cliff and Ord, 1973). Thus, for both point and area-based data spatial spectral density and autocorrelation analysis provide tools that support description, inferential testing, and model-building.

There are, of course, many other useful analytic techniques appropriate to ratio-level univariate spatial data analysis. For example, one productive tool appropriate for both regular grid cell data and some irregular polygon data is the analysis of spatial scale variance (Moellering and Tobler 1972). This straight-forward application of nested analysis of variance partitions variance into several user-defined hierarchial levels of grid-cells or irregular polygons.

Table 1 summarizes the categorization of univariate statistical spatial analysis techniques according to data type and measurement level. For linear data, two major classes of methods are network/flow analysis and one-dimensional spatial autocorrelation/spectral density analysis. The various methods for describing and modeling flow patterns in space generally focus on flows through a topological network whereas one-dimensional spatial autocorrelation/spectral density analysis considers the distance relationships along a single transect or straight line.

Table 1: Typology of Univariate Statistical Spatial Analysis Techniques

Data Type	Measurement Level	
	Nominal	Ratio
Point:	nearest-neighbor analysis quadrat/cell counts centrography measures polygon techniques spectral analysis	spatial auto-correlation analysis spatial spectral density analysis
Line:	line pattern analysis network measures graphic theoretical measures	network flow analysis one-dimensional spatial autocorrelation anaylsis

Area:	shape analysis	spatial auto-
	polygon net analysis	correlation analysis
		spatial spectral
		density analysis
		scale decomposition

BIVARIATE ANALYSIS

In bivariate statistical spatial data analysis, the overriding goal is to obtain information on the relationship between two statistical data maps. Two aspects of the study of pairwise relationships, often performed sequentially, are the analysis of the degree of similarity between maps and the analysis or modeling of transformations required to convert an input map into an output map.

The methods available to be used for bivariate spatial analysis largely make use of the descriptive measures discussed in the previous section. Many of the methods are analogous to the standard aspatial statistical methods of correlation and regression analysis. For example, Tobler (1979) has introduced a technique to transform a given input nominal-level point map into a specified output map where each point in the input map corresponds to a point in the output map. This technique, known as bidimensional regression yields information on the spatial pattern of distortion needed to implement the transformation. Similarity of pairs of networks has also been assessed using a variety of correlation-like techniques (e.g. Cummings et al, 1973).

For ratio-level data, the techniques of spatial autocorrelation and spectral density analysis provide a powerful methodology for bivariate analysis. The nature of the relationship between two maps can be investigated using spatial cross-correlation and cross-spectral analyses (Rayner, 1971; Cliff et al, 1975; Bennett, 1979). These methods not only provide information on the overall strength of similarity between two maps but also yield strength-of-similarity estimates at different spatial scales or distance lags. This added insight is significant because two data maps may be closely related at one spatial scale (e.g. regional) but unrelated at others (e.g. local). In addition, it is also possible that spatially-lagged effects exist in the relationship between two maps; that is, the value of the dependent variable at a given location may be related to the value of the independent variable at neighboring locations.

Spatial cross-correlation and cross-spectral density analyses provide tools for a flexible methodology for transforming a known input map into a known output map. This approach, known as transfer function modeling, takes advantage of the information contained in the cross-correlation and cross-spectral functions in order to build a transformation model that includes lagged effects as well as autoregressive effects (Bennett 1979).

INTEGRATED METHODOLOGY: AN EXAMPLE

The methods discussed in previous sections of this paper are not unrelated - in a typical application, they tend to be used in a logical sequence; for example, from description to inferential testing to model building and to forecasting. For the class of ratio-level area-based spatial data, typically displayed using choropleth mapping, a methodological flow outline suitable for implementation in an intelligent automated system can be developed.

The example presented here is based on spatial autocorrelation analysis. The objective is to learn as much as possible about a mapped pattern and its relationship to a second mapped pattern. The basic approach taken is to decompose spatial variation by scale and statistical property. The decomposition is sequential with the residuals from any given step providing the input for a subsequent step until all spatial order or patterning has been modeled and the residuals are indistinguishable from a realization of a Poisson process (Table 2).

Table 2: Flow Outline - Univariate Analysis

1. Choose study area, variable of interest.

2. Preprocessing: For map study area, obtain contiguity matrix, polygon centroids, proportion of shared boundary.

3. Define weighted connectivity matrix (possibly interactively).

4. Obtain spatial autocorrelation function - if trend is suspected continue; otherwise skip to step 6.

5. Model spatial trend and remove (i.e. obtain residuals) (trend - surface models, spatial differencing).

6. Calculate autocorrelation function.

7. Identify and specify neighborhood effects model (e.g. spatial autoregressive, moving average model).

8. Estimate neighborhood effects model and obtain residuals.

9. Calculate autocorrelation function for residuals, test for randomness.

10. If residuals random, stop; otherwise go back to steps 5 and/or 7.

For bivariate analysis, the transfer function approach-based on spatial cross-correlation analysis - is outlined in Table 3. Note that the bivariate method relies on results of the univariate method for the input map. The result of the bivariate method is an estimate transfer function model that describes the input-output relationships of the two data maps.

Table 3: Flow Outline: Bivariate Analysis

1. Select input and output maps

2. Process input map through univariate analysis procedure.

3. Use the autoregressive/moving average model estimated for the input series (in step 8 of univariate processing) to transform output series.

4. Access the residuals of both the input and output series.

5. Calculate spatial cross-correlation function for the residual data.

6. Calculate the spatial impulse response function between the input and output maps.

7. Identify the order of the transfer function model.

8. Estimate the identified transfer function model.

Empirical Example

To illustrate the flavor of a univariate statistical analysis using these techniques, an empirical example using stream-bed elevation data is presented. Because this data is linear, the spatial series can be considered one-dimensional. This makes this illustration simpler than would be the case with two-dimensional data. For example, the autocorrelation functions and spectral functions are themselves one-dimensional. Figures 1 through 7 present the results of the analysis.

CONCLUSION

Spatial data requires specialized analytic tools. Geographic information systems and general-purpose statistical mapping packages would benefit by incorporation of a linkage to a generic spatial analysis capability. It is possible to associate specific spatial data types with particular techniques of statistical spatial analysis. This provides input for developing functional specifications for an analytic component appropriate to a particular spatial data handling or statistical mapping application.

In developing this capability, attention needs to be paid to utility routines to prepare the data for the appropriate methods. For statistical spatial analysis, information on contiguity of polygons, nearest neighbors for points, and connectivity for networks are examples of information requirements. Appropriate data structures can assist in making the derivation or retrieval of this information efficient.

REFERENCES

Bachi, R. 1963, Standard Distance Measures and Related Methods for Spatial Analysis, Regional Science Association, Papers and Proceedings, Vol. 10, pp. 83-132

Bennett, R.J. 1979, Spatial Time Series, Pion, London

Besag, J. 1974, Spatial Interaction and the Statistical Analysis of Lattice Systems, Journal of the Royal Statistical Society, Vol. B36, pp. 192-236.

Bookstein, F.L. 1978, The Measurement of Biological Shape and Shape Change, Lecture Notes in Biomathematics, Vol. 24.

Boots, B.N. 1974, Delaunay Triangles: An Alternative Approach to Point Pattern Analysis, Proceedings of the Association of American Geographers, Vol. 6, pp. 26-29.

Christofides, N. 1975, Graph Theory: An Algorithmic Approach, Academic Press, NY.

Cliff, A.D., Haggett, P., Ord, J.K., Bassett, K., and Davies, R.B., 1975, Elements of Spatial Structure, Cambridge Univeristy Press, London.

Cliff, A.D. and Ord, J.K. 1973, Spatial Autocorrelation Pion, London

Cummings, L.P., Manly, B.J., and Weinand, H.C. 1974, Measuring Association in Link-Node Problems, Geoforum Vol. 33, pp. 43-51.

Getis, A. and Boots, B. 1978, Models of Spatial Processes Cambridge University Press, London.

Glick, B.J. 1982, A Spatial Rank-Order Correlation Measure, Geographical Analysis, Vol. 14, pp. 177-181.

Haggett, P. and Chorley, R. J. 1969, Network Analysis in Geography, Edward Arnold, London.

Haggett, P. Cliff, A.D., and Frey, A. 1977, Locational Analysis in Human Geography, Edward Arnold, London.

Mandelbrot, B.B. 1977, Fractals: Form, Chance, and Dimension, W.H. Freeman, San Francisco.

Moellering, H. and Tobler, W. R. 1972, Geographical Variances, Geographical Analysis, Vol. 4, pp. 34-50.

Rayner, J.N. 1971, An Introduction to Spectral Analysis, Pion, London.

Ripley, B.D. 1981, Spatial Statistics, Wiley, New York.

Tobler, W.R. 1979, Bidimensional Regression

Tobler, W.R. 1980, Statistical Cartography: What Is It? Proceedings, AutoCarto IV, pp. 77-81.