17

Interactive techniques and exploratory spatial data analysis

L ANSELIN

This chapter reviews the ideas behind interactive and exploratory spatial data analysis and their relation to GIS. Three important aspects are considered. First, an overview is presented of the principles behind interactive spatial data analysis, based on insights from the use of dynamic graphics in statistics and their extension to spatial data. This is followed by a review of spatialised exploratory data analysis (EDA) techniques, that is, ways in which a spatial representation can be given to standard EDA tools by associating them with particular locations or spatial subsets of the data. The third aspect covers the main ideas behind true exploratory spatial data analysis, emphasising the concern with visualising spatial distributions and local patterns of spatial autocorrelation. The geostatistical perspective is considered, typically taken in the physical sciences, as well as the lattice perspective, more familiar in the social sciences. The chapter closes with a brief discussion of implementation issues and future directions.

1 INTRODUCTION

Recent developments in computing hardware and GIS software have made it possible to interact directly with large spatial databases and to obtain almost instantaneous results for a wide range of GIS operations. The sophistication in storage, retrieval, and display provided by the rapidly evolving GIS technology has created a demand for new tools to carry out spatial analysis in general and spatial statistical analysis in particular (see, among others, Anselin and Getis 1992; Bailey 1994; Goodchild 1987; Goodchild et al 1992; Openshaw 1991). This demand grew out of an early awareness that the implementation of 'traditional' spatial analysis techniques was insufficient to address the challenges faced in a GIS environment (Goodchild and Longley, Chapter 40). The latter is often characterised by vast numbers of observations (hundreds to several thousands) and 'dirty' data, and some go so far as to completely reject 'traditional' spatial analysis that is based on statistical inference (Openshaw and Alvanides, Chapter 18; Fischer, Chapter 19; Openshaw 1990, 1991). While this rather extreme viewpoint is not shared by many, it is widely recognised that many of the geographical analysis techniques of the 1960s fail to take advantage of the visualisation and data manipulation capabilities embodied in modern GIS. Specifically, most spatial statistical techniques, such as tests for spatial autocorrelation and spatial regression models, are primarily *static* in nature, allowing only limited interaction between the data, the models, and the analyst. In contrast, *dynamic* or *interactive* approaches to data analysis stress the user interaction with the data in a graphical environment, allowing direct manipulation in the form of instantaneous selection, deletion, rotation, and other transformations of data points to aid in the exploration of structure and the discovery of patterns (Buja et al 1996; Cleveland 1993; Cleveland and McGill 1988).

The importance of EDA to enhance the spatial analytical capabilities of GIS has become widely

recognised (Anselin 1994; Anselin and Getis 1992; Bailey and Gatrell 1995; Fotheringham and Charlton 1994). The EDA paradigm for statistical analysis is based on a desire to let the data speak for themselves and to impose as little prior structure upon them as possible. Instead, the emphasis is on creative data displays and the use of simple indicators to elicit patterns and suggest hypotheses in an inductive manner, while avoiding potentially misleading impressions given by 'outliers' or 'atypical' observations (Good 1983; Tukey 1977). Since spatial data analysis is often characterised as being 'data rich but theory poor' (Openshaw 1991), it would seem to form an ideal area for the application of EDA. However, this is not a straightforward exercise, since the special nature of spatial data, such as the prevalence of spatial autocorrelation, may invalidate the interpretation of methods that are based on an assumption of independence, which is the rule in mainstream EDA (Anselin 1990; Anselin and Getis 1992). Hence, the need has arisen to develop specialised methods of exploratory spatial data analysis (ESDA) that take the special nature of spatial data explicitly into account (for recent reviews, see Anselin 1994; Anselin and Bao 1997; Bailey and Gatrell 1995; Cook et al 1996; Cressie 1993; Majure and Cressie 1997).

This chapter reviews the ideas behind interactive and ESDA and their relation to GIS. Many of the ESDA techniques have been developed quite recently and this remains an area of very active research. Therefore, the emphasis will be on general principles, rather than on specific techniques. The latter will only be used to illustrate the overall framework and no attempt is made to cover a comprehensive set of methods. The bulk of the chapter considers three important aspects of the integration of ESDA and interactive methods with GIS. First, an overview is presented of the principles behind interactive spatial data analysis, based on insights from the use of dynamic graphics in statistics and their extension to spatial data. This is followed by a review of spatialised EDA techniques, that is, ways in which a spatial representation can be given to standard EDA tools by associating them with particular locations or spatial subsets of the data. The third aspect covers the main ideas behind true exploratory spatial data analysis, emphasising the concern with visualising spatial distributions and local patterns of spatial autocorrelation (Getis, Chapter 16). The chapter closes with a brief discussion of implementation issues and future directions.

2 PRINCIPLES OF INTERACTIVE SPATIAL DATA ANALYSIS

The principles behind interactive spatial data analysis can be traced back to the work on dynamic graphics for data analysis in general, originated by the statistician John Tukey and a number of research groups at AT&T Bell Laboratories. An excellent review of the origins of these ideas is given in the collection of papers edited by Cleveland and McGill (1988), and early discussions of specific methods are contained in the papers by, among others, Becker et al (1987), Becker and Cleveland (1987), and Stuetzle (1987). More recent reviews of methods for the dynamic analysis of high-dimensional multivariate data and other aspects of interactive statistical graphics can be found in papers by, among others, Becker et al (1996), Buja et al (1991, 1996), Cleveland (1993), and Cook et al (1995).

Dynamic graphical methods started as enhancements to the familiar static displays of data (e.g. histograms, bar charts, pie charts, scatterplots), by allowing direct manipulation by the user that results in 'immediate' change in a graph (see Elshaw Thrall and Thrall, Chapter 23, for some examples). This had become possible by the availability of workstations with sufficient computational power to generate the statistical graphs without delays and to allow interaction with the data by means of an input device (light pen or mouse). The overall motivation was to involve the human factor more directly in the exploration of data (i.e. exploiting the inherent capabilities of the brain to detect patterns and structure), and thereby gain richer insights than possible with the traditional rigid and static display. This was achieved by allowing the user to delete data points, highlight (brush) subsections of the data, establish links between the same data points in different graphs, and rotate, cut through, and project higher-dimensional data. Furthermore, the user and not a preset statistical procedure determined which actions to perform. Interactive statistical procedures become particularly effective when datasets are large (many observations) and high-dimensional (many variables), situations where characterisation of the data by a few numbers becomes increasingly unrealistic (for an early assessment see, for example, Andrews et al 1988: 75). While dynamic graphics for statistics were originally mostly experimental and confined to research environments, they have quickly become pervasive features of the EDA capability in modern commercial statistical software packages.

An important aspect of dynamic graphics is the representation of data by means of multiple and simultaneously available 'views', such as a table, a list of labels, a bar chart, pie chart, histogram, stem and leaf plot, box plot, or scatterplot. These views are shown in different windows on a computer screen. They are linked in the sense that when a location in any one of the windows (e.g. a bar on a bar chart or a set of points on a scatterplot) is selected by means of a pointing device (brushing), the corresponding locations in the other windows are highlighted as well (see Becker et al 1987). While geographical locations have always played an important role in dynamic graphics (see the many examples of Cleveland and McGill 1988), it is only recently that the 'map' was introduced explicitly as an additional view of the data, for example by Haslett et al (1990, 1991), MacDougall (1991), and Monmonier (1989).

The most comprehensive set of tools to date that implement dynamic graphics for exploring spatial data is contained in the Regard (formerly Spider) software of Haslett, Unwin and associates, which runs on a Macintosh platform (see also Bradley and Haslett 1992; Haslett and Power 1995; Unwin 1994). Regard, and its successor Manet (Unwin et al 1996) allow for the visualisation of the distribution and associations between data for any subset of locations selected on a map display. Similarly, for any subset of data highlighted in a non-spatial view, such as a category in a histogram, the corresponding locations are highlighted on the map. This is illustrated in Figure 1, where attention focuses on suggesting promising multivariate relations pertaining to electoral change in the new German Bundesländer (formerly East Germany). Six types of dynamically linked views of the data are included, consisting of a map with highlighted constituencies, a bar chart,

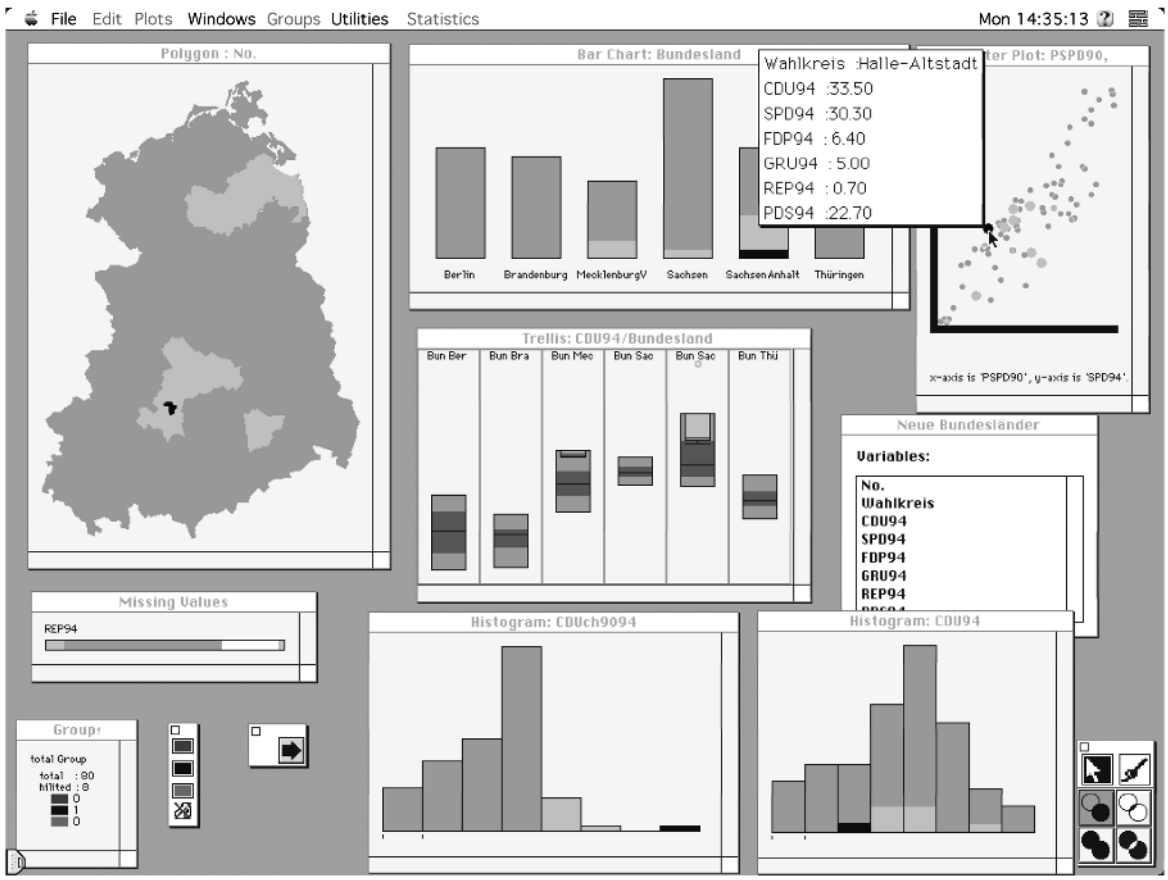

Fig 1. Interactive dynamic graphics for exploring spatial data with Manet.

conditional (trellis) plot, histogram, scatterplot and missing value chart, as well as lists with variable names and values observed at a specific location. (For details on the Manet approach, see Unwin et al 1996 and *http://www1.Math.Uni-Augsburg.de/~theus/Manet/ManetEx.html.)* While highly dynamic in its statistical graphics, the Spider–Regard–Manet approach is still somewhat limited in terms of the spatial aspects of the data, in the sense that it is based on a fixed map and does not take advantage of GIS functionality, such as specialised data models to facilitate spatial queries and overlays (see also Hazelhoff and Gunnink 1992).

Several ideas from the methodology of dynamic statistical graphics are reflected in the design of current GIS and mapping software. For example, the ArcView GIS (ESRI 1995b) is organised around several linked 'views' of the data (a map, a table, and several types of charts). These allow a limited degree of dynamic interaction in the sense that a selection made in any of the views (spatial selection of features on a map, records in a table) is immediately reflected in all other views. While Version 2.1 is rather limited in terms of its built-in statistical (exploratory) analysis capabilities, enhancements to make ArcView into a tool for interactive ESDA have been developed by linking it to other software modules. For example, at the Statistics Laboratory of Iowa State University, a 2-directional link was established between the XGobi dynamic graphics software of Buja et al (1991, 1996) and ArcView (Cook et al 1996; Majure et al 1996a, 1996b; Symanzik et al 1994, 1995, 1996; *http://www.gis.iastate.edu/XGobi-AV2/X Gobi-AV2.html*). Similarly, the SpaceStat software for spatial data analysis of Anselin (1992, 1995a) was linked with ArcView in a Microsoft Windows environment (Anselin and Bao 1996, 1997; *http://www.rri.wvu.edu/utilities.htm*). In many respects, these and similar efforts achieve a functionality close to that of Regard, although not as seamless and considerably slower in execution. For example, in Figure 2, ArcView scripts were used to construct a histogram for the median values of housing in West Virginia counties, linked to a map (a view in ArcView). Using a selection tool to click on a given bar (interval) in the histogram, the relevant counties in the map are highlighted (for further details on the dataset and the procedures, see Anselin and Bao 1996, 1997). In contrast to Regard, the linked frameworks allow the exploitation of the full functionality of the GIS to search for other variables that may display similar patterns, using queries and spatial overlays (for example, see Cook et al 1996).

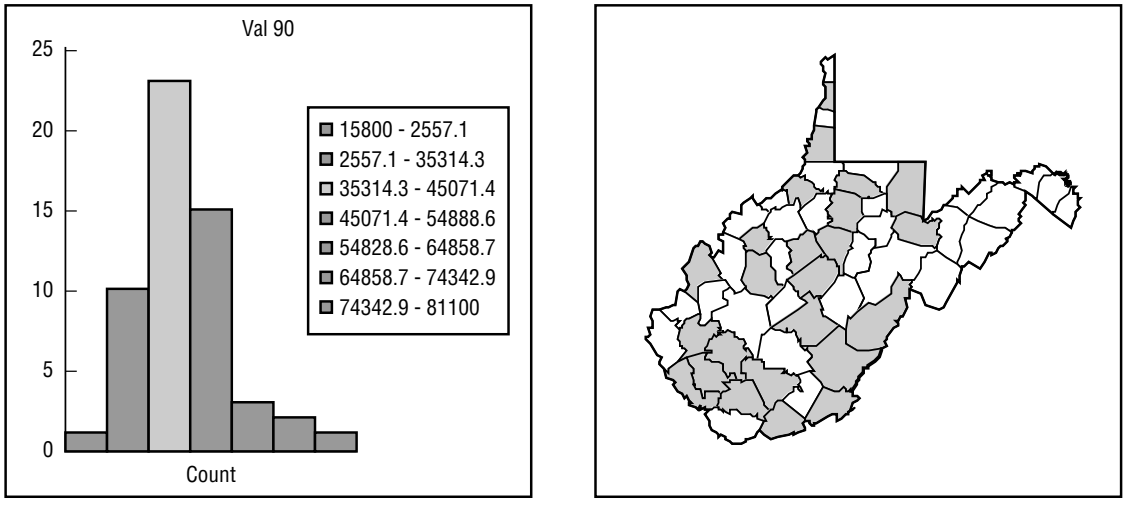

Fig 2. Linked histogram and map in ArcView–SpaceStat.

3 SPATIALISED EXPLORATORY DATA ANALYSIS

Whilst a widely available commercial implementation of interactive and dynamic spatial data analysis integrated with a GIS does not exist at the time of writing, the use of EDA with GIS has become fairly common. For example, in the 'archaeologist's workbench' of Farley et al (1990) and Williams et al (1990), standard EDA tools such as box plots and scatterplots were applied to geographical data, by exporting information from a GIS to a statistical package (a 1-directional link). However, the latter is not ESDA in the sense used by Cressie (1993) and Anselin (1994), but rather non-spatial EDA applied to spatial data (see also Anselin and Getis 1992).

Spatialised EDA (Anselin 1994) is one step closer to true ESDA in the sense that location is combined with a graphic description of the data in the form of a bar chart, pie chart, or various icons. The most familiar example of this may be the positioning of Chernoff faces at geographical locations on a map, such as coordinates of cities or centroids of states, as illustrated by Fotheringham and Charlton (1994) and Haining (1990: 226) (but for a critical assessment see Haslett 1992). The facility to add bar charts and pie charts to areal units on a map is by now a familiar feature in many commercial GIS and mapping packages.

A more meaningful combination of location and data description is obtained when summaries of spatial distributions are visualised for different subsets in the data, providing initial insight into spatial heterogeneity (i.e. different for spatial subsets in the data, such as a north–south differential) or suggesting a spatial trend (a systematic variation of a variable with location, such as an east–west trend). For example, Haining (1990: 224) organises box plots for standardised mortality rates by distance band away from the centre of the city, revealing a clear spatial trend. Similarly, spatialised EDA techniques may be used to carry out a form of exploratory spatial analysis of variance, in which the interest centres on differences in central tendency (mean, median) of the distribution of a variable between spatial subsets (or spatial regimes) in the data. In Figure 3 this is illustrated for the West Virginia data. Two box plots refer respectively to counties at the outer rim and inner counties (generated by applying a spatial selection operation in a GIS). A comparison of the two graphs suggests a systematically higher value for counties at the rim, although a few counties in either group do not fit the pattern. In an interactive data analysis, this could easily be addressed by sequentially removing or adding counties to one or the other subset, providing the groundwork for a spatial analysis of variance (for other examples see Anselin et al 1993). However, it is well recognised that potential spatial autocorrelation among these observations could invalidate the interpretation of any analysis of variance or regression analysis. Therefore, techniques only qualify as true ESDA when this is addressed explicitly.

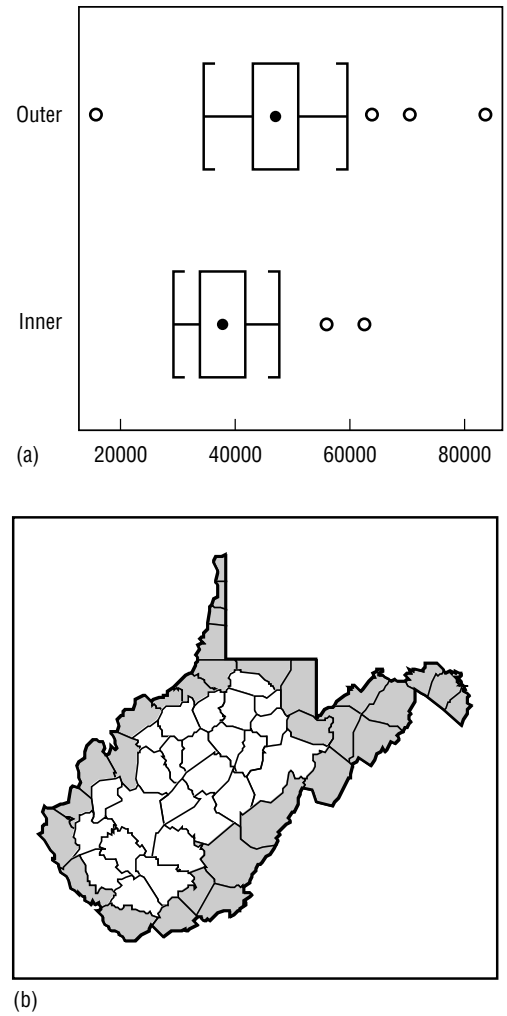

Fig 3. Exploratory spatial analysis of variance.

4 EXPLORATORY SPATIAL DATA ANALYSIS

ESDA can be broadly defined as the collection of techniques to describe and visualise spatial distributions, identify atypical locations (spatial outliers), discover patterns of spatial association (spatial clusters), and suggest different spatial regimes and other forms of spatial instability or spatial non-stationarity (Anselin 1994; see also Beard and Buttenfield, Chapter 15). Central to ESDA is the concept of spatial autocorrelation, that is, the phenomenon where locational similarity (observations in spatial proximity) is matched by value similarity (correlation).

Spatial autocorrelation has been conceptualised from two main perspectives, one prevalent in the physical sciences, the other in the social sciences. Following Cressie's (1993) classification, the so-called *geostatistical* perspective considers spatial observations to be a sample of points from an underlying continuous spatial distribution (surface). This is modelled by means of a variogram, which expresses the strength of association between pairs of locations as a continuous function of the distance separating them (for comprehensive reviews see Cressie 1993 and Isaaks and Srivastava 1989). By contrast, in the so-called lattice perspective, spatial locations are discrete points or areal units, and spatial data are conceptualised as a single realisation of a spatial stochastic process, similar to the approach taken in the analysis of time series. Essential in the analysis of lattice data is the concept of a *spatial weights matrix*, which expresses the spatial arrangement (topology, contiguity) of the data and which forms the starting point for any statistical test or model (for extensive reviews see Cliff and Ord 1981; Cressie 1993; Haining 1990; Upton and Fingleton 1985).

Juxtaposed on the distinction between the geostatistical and lattice perspective is that between global and local indicators of spatial association. Global indicators, such as the familiar Moran's I and Geary's c spatial autocorrelation statistics, summarise the overall pattern of dependence in the data into a single indicator (see Getis, Chapter 16). A major practical drawback for GIS analysis is that these global indicators are based on a strong assumption of spatial stationarity, which, among others, requires a constant mean (no spatial drift) and constant variance (no outliers) across space. This is not very meaningful or may even be highly misleading in analyses of spatial association for hundreds or thousands of spatial units that characterise current GIS applications. The main contribution of ESDA with respect to GIS lies therefore in visualising local patterns of spatial association, indicating local non-stationarity and discovering islands of spatial heterogeneity (Anselin 1994; Cressie 1993). In the remainder of this section, first some techniques are considered to visualise spatial distributions, with a particular focus on identifying outliers and atypical observations. These techniques are more specialised than the methods for visualisation for GIS discussed by Kraak (Chapter 11). This is followed by a short review of ESDA techniques to visualise and assess spatial autocorrelation, for both geostatistical and lattice perspectives.

4.1 Visualising spatial distributions

Many of the spatialised EDA techniques described above can be successfully applied to gain insight into the distribution of data across locations in a GIS. These methods can also be integrated in a dynamic interactive framework in a fairly straightforward way, for example as in the Manet software. A more explicit focus on identifying spatial outliers is offered by the so-called *box map*, the extension of a familiar quantile choropleth map (a standard feature in most GIS and mapping software) with highlighted upper and lower outliers, defined as observations outside the 'fences' in a box plot (Cleveland 1993). A box map can easily be implemented in many current GIS and mapping packages (e.g. Anselin and Bao 1997). By comparing box maps for different variables using overlay operations in a GIS, an initial look at potential multivariate associations can be obtained (e.g. see Talen 1997). Other approaches to identify outliers in spatial data can be envisaged as well, for example by constructing spatial queries for those locations whose values exceed some criterion of 'extremeness'. Such devices can be readily implemented in most currently available commercial GIS.

A more rigorous approach, geared towards the geostatistical perspective, consists of the estimation of a spatial cumulative distribution function (SCDF), that is, a continuous density function for all observations in a given region. This is implemented in the ArcView–XGobi linked framework mentioned earlier. The linkage allows users to highlight regions of the data on a map in ArcView and to find an SCDF plot in XGobi, to brush areas on the map to find the corresponding subset in the SCDF, and to brush quantiles of the estimated SCDF and find the matching locations on the map. For example, in Figure 4 (from Majure et al 1996a), the two SCDF functions for forest health indicators in the graph on the left-hand side correspond to the two large sub-regions of New England states in the map on the

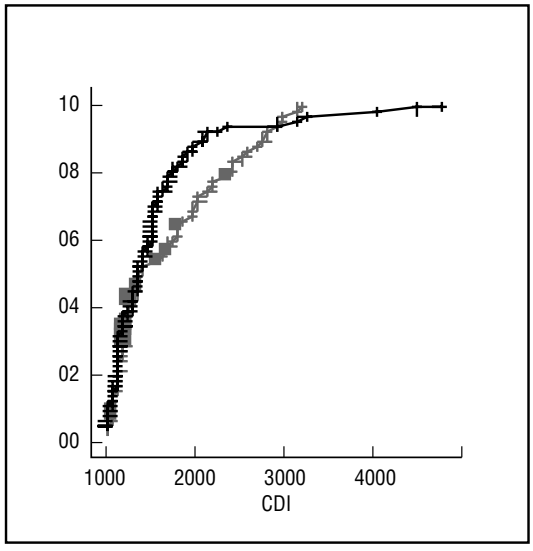

Fig 4. Spatial cumulative distribution function (SCDF) in ArcView–Xgobi.
Source: Majure et al 1996a

right. An advantage of this form of linkage is that the GIS can be used to overlay other data onto the sample points, in order to suggest potential multivariate associations. Clearly, an approach such as SCDF could be integrated into a more comprehensive Manet-type dynamic interactive ESDA framework, although this has not been implemented to date.

4.2 Visualising spatial autocorrelation: the geostatistical perspective

The main focus of ESDA in geostatistics is on identifying 'unusual' and highly influential (pairs of) locations in order to obtain more robust estimates of the variogram. Such locations are referred to as spatial outliers, or pockets of local non-stationarity, and they require closer scrutiny before proceeding with geostatistical modelling or spatial prediction (Kriging). The basic tools are outlined by Cressie (1993) and include the *variogram cloud*, the *variogram box plot* and the *spatial lag scatterplot*. A variogram cloud is a scatterplot of squared differences (or of square root absolute differences: see Cressie 1993) between all pairs of observations, sorted by distance band. An implementation of this device in an interactive dynamic graphics framework consisting of ArcView and XGobi is illustrated in Figure 5 (from Majure et al 1996). By brushing points in the cloud plot, lines are drawn between pairs of observations on the map, suggesting potential regions that are spatial outliers. A similar, but more encompassing approach is implemented in the Regard software, where the variogram cloud is included as one of the linked views of the data to facilitate a search for local pockets of spatial

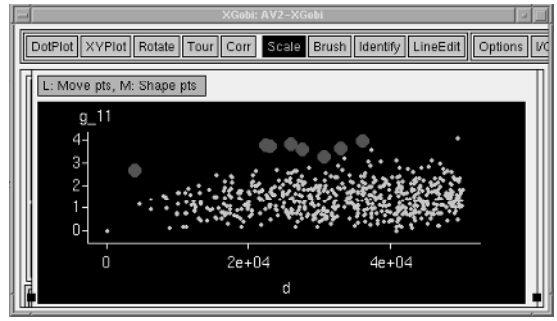

Fig 5. Brushed variogram cloud plot in ArcView–Xgobi.
Source: Majure et al 1996a

non-stationarity (Bradley and Haslett 1992; Haslett 1992; Haslett et al 1991; Haslett and Power 1995). The spatial lag scatterplot (also referred to as a lagged scatterplot) and the variogram box plot provide two different summary views of the information in the cloud plot. The spatial lag scatterplot focuses on the observation pairs that belong to a given distance class, that is, a subsection of the variogram cloud between two distances. The value observed at each point is plotted against the value observed at the 'lagged' point (a point separated from it by a distance belonging to the given distance band). The spatial lag scatterplot identifies potential influential locations as points that are far-removed from the 45 degree line (Majure and Cressie 1997). The variogram box plot consists of a box plot for each distance band in the variogram cloud, as in the left-hand side of Figure 6, illustrating the spatial dependence in the West Virginia housing values. For several distance bands, outliers may be identified as points outside the fences of the box plot. These outliers can be associated with the pairs of locations to which they correspond, as in the right-hand side of Figure 6, typically obtained in an interactive manner (and in a way similar to the procedure illustrated in Figure 5).

Extensions of both types of plots are possible in many ways, for example by using robustified measures of squared difference, by focusing on different directions (anisotropy), or by including multiple variables (for extensive examples see Majure and Cressie 1997). ESDA techniques based on the geostatistical perspective can be found in many academic as well as a number of commercial geostatistics software packages (e.g. S+SpatialStats, MathSoft 1996a), although the linkage to GIS is still limited or non-existent at the time of writing.

4.3 Visualising spatial autocorrelation: the lattice perspective

Central in the lattice perspective to spatial autocorrelation is the concept of a *spatial weights matrix* and associated spatially lagged variable or *spatial lag*. The non-zero elements of the spatial weights matrix indicate for each location which other locations potentially interact with it (the so-called spatial neighbours). Furthermore, the value of the non-zero elements is related to the relative strength of this interaction (for technical details see Cliff and Ord 1981; Haining 1990; Upton and Fingleton 1985). A spatial lag is constructed as a weighted average (using the weights in the spatial weights matrix) of the values observed for the neighbours of a given location (see Anselin 1988).

The matching of the value observed at a location with its spatial lag for a given spatial weights matrix provides useful insight into the local pattern of spatial association in the data. More precisely, when a high degree of positive spatial autocorrelation is present, the observed value at a location and its spatial lag will tend to be similar. Spatial outliers will tend to be characterised by very different values for the location and its spatial lag, either much higher or much lower in the location compared to the average for its neighbours. The association between a variable and its spatial lag can be visualised by means of so-called *spatial lag pies* and *spatial lag bar charts* (Anselin 1994;

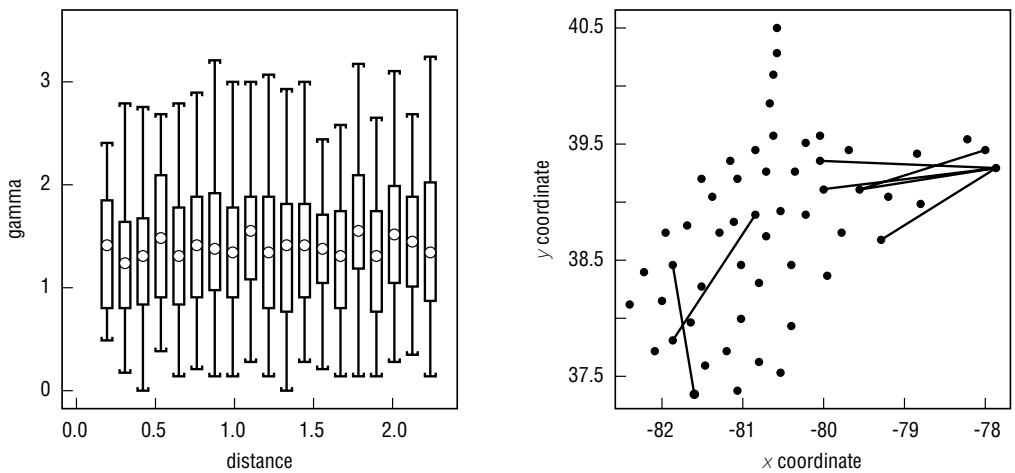

Fig 6. Variogram box plot with outlier pairs identified by location.

Anselin et al 1993; Anselin and Bao 1997). Both of these are made up of visual devices (size of the pie or length of the bar) that indicate the relative value of the spatial lag compared to the value at a location, as illustrated in Figure 7. Other visualisation schemes are possible as well, for example based on the difference, absolute difference, squared difference, or ratio between the value observed at a location and its spatial lag. These devices can be implemented in most GIS and mapping software in a straightforward way. In addition to the usual zooming and querying facilities available in an interactive GIS, the use of spatial lag pies or spatial lag bar charts could be made dynamic by allowing an interactive definition of the spatial weights matrix. It is envisaged that systems implementing these ideas will be available in the near future.

A more formal approach towards visualising spatial association can be based on the concept of a *Moran scatterplot* and associated *scatter map* (Anselin 1994, 1995b, 1997). It follows from the interpretation of the Moran's *I* statistic for spatial autocorrelation as a regression coefficient in a bivariate spatial lag scatterplot. More precisely, in a scatterplot with the spatial lag on the vertical axis and the value at each location on the horizontal axis, Moran's *I* corresponds to the slope of the regression line through the points. When the variables are expressed in standardised form (i.e. with mean zero and standard deviation equal to one), this allows for an assessment of both global spatial association (the slope of the line) as well as local spatial association (local trends in the scatterplot). The latter is obtained by the decomposition of the scatterplot into four quadrants, each corresponding to a different type of spatial association: positive association between high values in the upper right and between low values in the lower left quadrants; negative association between high values surrounded by low values in the lower right and the reverse in the upper left quadrant. An illustration of this decomposition for the West Virginia data is given in Figure 8. The spatial locations that correspond to

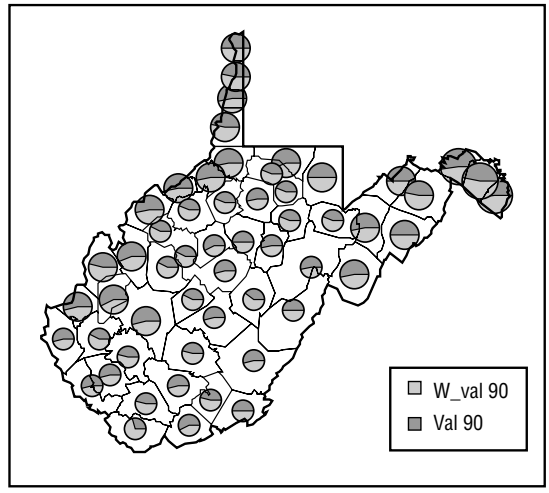

Fig 7. Spatial lag pie chart in ArcView–SpaceStat.

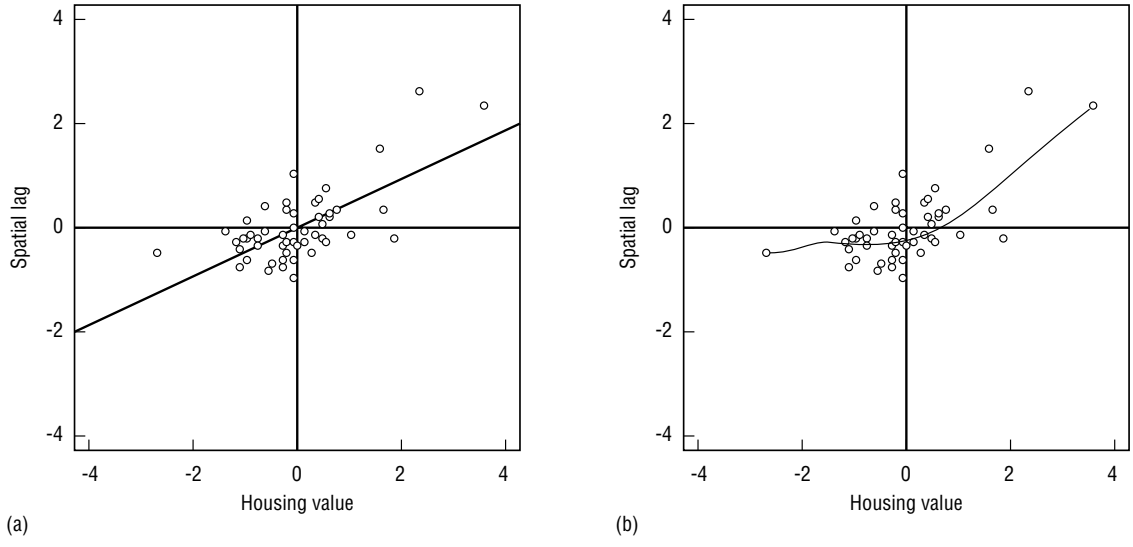

Fig 8. Moran scatterplot with linear and loess smoother.

the points in the scatterplot can be found in a linked map, where each quadrant is represented by a different shade or colour, as in Figure 9. By interactively identifying particular points in the graph (e.g. extreme values), the corresponding location can be shown on the map. This is a straightforward extension of the notion of brushing scatterplots to assess local spatial association.

Two additional interpretations of the Moran scatterplot are useful in an interactive ESDA setting. One is to identify outliers or high leverage points that unduly influence the slope of the regression line (i.e. the measure of global spatial association). Such outliers can be found by means of standard regression diagnostics and are easily identified on a map in a linked framework. They can also be related to the significance of local indicators of spatial association (LISA) statistics (Getis, Chapter 16; Anselin 1995b; Getis and Ord 1992; Ord and Getis 1995). In conjunction with a map of significant LISA statistics, the Moran scatterplot provides the basis for a substantive interpretation of spatial clusters or spatial outliers (further details are given by Anselin 1995b, 1996). A second interpretation is to consider the extent to which a non-linear smoother (such as a loess smoother; Cleveland 1979) approximates the linear fit in the scatterplot. Strong non-linear patterns may indicate different spatial regimes or other forms of local spatial non-stationarity. For example, on the right-hand side of Figure 8, the loess function suggests two distinct slopes in the graph, one considerably steeper than the other. The Moran scatterplot and associated map (Figure 9) can easily be implemented in a dynamic graphics setting, for example using the ArcView–SpaceStat linked framework.

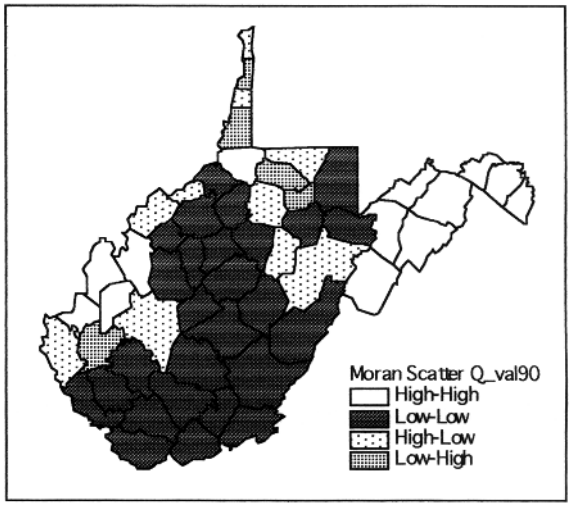

Fig 9. Moran scatter map.

5 IMPLEMENTATION AND FUTURE DIRECTIONS

To date a fully interactive ESDA functionality is not yet part of commercial GIS. However, several partial implementations exist, where a spatial statistical 'module' is added to an existing GIS (a point also made by Aspinall, Chapter 69; Boots, Chapter 36; Fischer, Chapter 19; and Getis, Chapter 16). Early discussions of these approaches were primarily conceptual, and a number of different taxonomies for integration have been advanced, primarily focusing on the nature of the linkage – closely coupled versus loosely coupled – and the types of statistical function that should be included (e.g. Anselin and Getis 1992; Goodchild et al 1992). Building on the general framework outlined by Anselin and Getis (1992), a schematic overview of the interaction between different analytical functions of a GIS is given in Figure 10 (based on Anselin 1998; see Getis, Chapter 16; and Goodchild and Longley, Chapter 40, for related conceptual schema). Following the usual classification of GIS functionality into four broad groups (input, storage, analysis, and output), the analysis function can be further subdivided into selection, manipulation, exploration and confirmation. Anselin et al (1993) considered the first two of these to form a 'GIS module' while the latter two formed a 'data analysis module' to emphasise the practical division of labour between typical commercial GIS software and the specialised (add-on) software needed to carry out spatial data analysis. However, this distinction is becoming increasingly irrelevant, since many statistical software packages now have some form of mapping (or even GIS) functionality, and a growing number of (spatial) statistical functions are included in GIS software. More important than classifying these functions as belonging to one or other module is to stress their interaction and the types of information that must be exchanged between them, as illustrated by the linkages in Figure 10. While many other taxonomies are possible, the main point of the classification in Figure 10 is that selection and manipulation (shown on the left) are present in virtually all advanced systems and have become known as 'spatial analysis' in the commercial world (e.g. ESRI 1995c: Lesson 8). By contrast, the spatial

data analysis functions (shown on the right) are essentially absent in commercial systems.

The essence of any integration as in Figure 10 is that spatial information (such as location, topology, and distance) must be transferred from the GIS to the statistical module and location-specific results of the statistical analysis must be moved back to the GIS for mapping. Apart from the self-contained approach taken in Spider-Regard-Manet, most implementations to date of ESDA functionality in a GIS are extensions of existing systems by means of macro-language scripts. This typically hides the linked nature of the analysis routines from the user. Recent examples are extensions of ARC/INFO with non-spatial EDA tools, such as scatterplots (e.g. Batty and Xie 1994), and routines for the computation of global and local indicators of spatial association (e.g. Ding and Fotheringham 1992; Bao et al 1995). An alternative is a closely-coupled linkage between two software packages that allow remote procedure calls (in Unix) or dynamic data exchange (in a Microsoft Windows environment). This approach is taken in the only commercial implementation that exists to date of an integrated data analysis and GIS environment, the S+Gislink between the S-Plus statistical software and the ARC/INFO GIS (MathSoft 1996b). On Unix workstations a bi-directional link is established that allows data to be passed back and forth in their native format. In addition, the linkage allows users to call S-Plus statistical functions from within ARC/INFO. A similar approach is taken in the ArcView–XGobi integration at the Statistics Laboratory of Iowa State University. A much looser

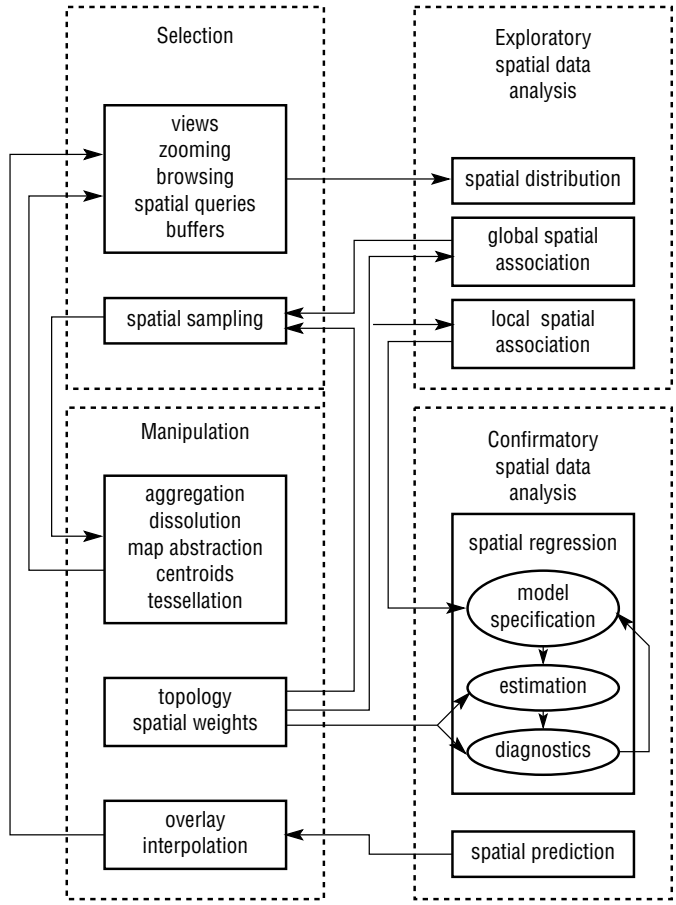

Fig 10. Spatial analysis in GIS.

coupling is implemented in the SpaceStat–ArcView linkage. Both of these efforts focus explicitly on ESDA, while the S-Plus–ARC/INFO linkage pertains primarily to traditional non-spatial EDA.

Several promising research directions are being pursued in the quest to develop more powerful tools for spatial analysis in GIS in general, and interactive spatial data analysis in particular. Highly relevant ongoing efforts include the use of the Internet to facilitate interactive mapping and visual data exploration (e.g. the Iris framework of Andrienko and Andrienko 1996; and see Batty, Chapter 21), the extension of data mining techniques to spatial data (e.g. Ng and Han 1994), and the use of massive parallel computing for the estimation of local indicators of spatial association (e.g. Armstrong and Marciano 1995). The extent of commercial and academic research activity devoted to methodological and computational facets will likely lead to a much-enhanced ESDA functionality in the GIS of the near future. This is an area of rapid change, and it is hoped that the general principles outlined in this chapter may provide a basis for the interpretation and assessment of future developments.

References

Andrews D F, Fowlkes E B, Tukey P A 1988 Some approaches to interactive statistical graphics. In Cleveland W S, McGill M E (eds) *Dynamic graphics for statistics*. Pacific Grove, Wadsworth: 73–90

Andrienko N, Andrienko G 1996 *IRIS, a knowledge-based system for visual data exploration.* See also *http://allanon.gmd.de/and/java/iris/Iris.html*

Anselin L 1988 *Spatial econometrics: methods and models.* Dordrecht, Kluwer

Anselin L 1990a What is special about spatial data? Alternative perspectives on spatial data analysis. In Griffith D A (ed.) *Spatial statistics, past, present, and future*. Ann Arbor, Institute of Mathematical Geography: 63–77

Anselin L 1992 *SpaceStat: a program for the analysis of spatial data*. Santa Barbara, NCGIA, University of California

Anselin L 1994a Exploratory spatial data analysis and geographic information systems. In Painho M (ed.) *New tools for spatial analysis*. Luxembourg, Eurostat: 45–54

Anselin L 1995a *SpaceStat version 1.80 user's guide*. Morgantown, Regional Research Institute, West Virginia University

Anselin L 1995b Local indicators of spatial association – LISA. *Geographical Analysis* 27: 93–115

Anselin L 1997 The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In Fischer M, Scholten H, Unwin D (eds) *Spatial analytical perspectives on GIS in environmental and socio-economic sciences*. London, Taylor and Francis: 111–25

Anselin L 1998 GIS research infrastructure for spatial analysis of real estate markets. *Journal of Housing Research* 8

Anselin L, Bao S 1996 *SpaceStat.apr user's guide.* Morgantown, Regional Research Institute, West Virginia University

Anselin L, Bao S 1997 Exploratory spatial data analysis: linking SpaceStat and ArcView. In Fischer M, Getis A (eds) *Recent developments in spatial analysis – spatial statistics, behavioural modelling and neurocomputing*. Berlin, Springer

Anselin L, Dodson R, Hudak S 1993 Linking GIS and spatial data analysis in practice. *Geographical Systems* 1: 3–23

Anselin L, Getis A 1992 Spatial statistical analysis and geographic information systems. *Annals of Regional Science* 26: 19–33

Armstrong M P, Marciano R 1995 Massively parallel processing of spatial statistics. *International Journal of Geographical Information Systems* 9: 169–89

Bailey T C 1994 A review of statistical spatial analysis in geographical information systems. In Fotheringham A S, Rogerson P (eds) *Spatial analysis and GIS*. London, Taylor and Francis: 13–44

Bailey T C, Gatrell A C 1995 *Interactive spatial data analysis*. Harlow, Longman/New York, John Wiley & Sons Inc.

Bao S, Henry M, Barkley D, Brooks K 1995 RAS: a regional analysis system integrated with ARC/INFO. *Computers, Environment, and Urban Systems* 18: 37–56

Batty M, Xie Y 1994a Modelling inside GIS: part I. Model structures, exploratory spatial data analysis and aggregation. *International Journal of Geographical Information Systems* 8: 291–307

Becker R A, Cleveland W S 1987 Brushing scatterplots. *Technometrics* 29: 127–42

Becker R A, Cleveland W S, Shyu M-J 1996 The visual design and control of Trellis display. *Journal of Computational and Graphical Statistics* 5: 123–55

Becker R A, Cleveland W S, Wilks A R 1987 Dynamic graphics for data analysis. *Statistical Science* 2: 355–95

Bradley R, Haslett J 1992 High interaction diagnostics for geostatistical models of spatially referenced data. *The Statistician* 41: 371–80

Buja A, Cook D, Swayne D F 1996 Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics* 5: 78–99

Buja A, McDonald J A, Michalak J, Stuetzle W 1991 Interactive data visualisation using focusing and linking. In Nielson G M, Rosenblum L (eds) *Proceedings of Visualisation 91*. Los Alamitos, IEEE Computer Society Press: 155–62

Cleveland W S 1979 Robust locally weighted regression and smoothing scatter plots. *Journal of the American Statistical Association* 74: 829–36

Cleveland W S 1993 *Visualizing data*. Summit, Hobart Press

Cleveland W S, McGill M E (eds) 1988 *Dynamic graphics for statistics*. Pacific Grove, Wadsworth

Cliff A, Ord J K 1981b *Spatial processes: models and applications*. London, Pion

Cook D, Buja A, Cabrera J, Hurley C 1995 Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics* 4: 155–72

Cook D, Majure J, Symanzik J, Cressie N 1996 Dynamic graphics in a GIS: a platform for analysing and exploring multivariate spatial data. *Computational Statistics* 11: 467–80

Cressie N A C 1993 *Statistics for spatial data*, revised edition. New York, John Wiley & Sons Inc.

Ding Y, Fotheringham A S 1992 The integration of spatial analysis and GIS. *Computers, Environment, and Urban Systems* 16: 3–19

ESRI 1995b *ArcView 2.1, the geographic information system for everyone*. Redlands, ESRI

ESRI 1995c *Understanding GIS, the ARC/INFO method*. Redlands, ESRI Inc.

Farley J A, Limp W F, Lockhart J 1990 The archaeologist's workbench: integrating GIS, remote sensing, EDA and database management. In Allen K, Green F, Zubrow E (eds) *Interpreting space: GIS and archaeology*. London, Taylor and Francis: 141–64

Fotheringham A S, Charlton M 1994 GIS and exploratory spatial data analysis: an overview of some research issues. *Geographical Systems* 1: 315–27

Getis A, Ord J K 1992 The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24: 189–206

Good I J 1983 The philosophy of exploratory data analysis. *Philosophy of Science* 50: 283–95

Goodchild M F 1987 A spatial analytical perspective on geographical information systems. *International Journal of Geographical Information Systems* 1: 327–34

Goodchild M F, Haining R P, Wise S et al 1992 Integrating GIS and spatial analysis – problems and possibilities. *International Journal of Geographical Information Systems* 6: 407–23

Haining R P 1990 *Spatial data analysis in the social and environmental sciences*. Cambridge (UK), Cambridge University Press

Haslett J 1992 Spatial data analysis – challenges. *The Statistician* 41: 271–84

Haslett J, Bradley R, Craig P, Unwin A, Wills G 1991 Dynamic graphics for exploring spatial data with applications to locating global and local anomalies. *The American Statistician* 45: 234–42

Haslett J, Power G M 1995 Interactive computer graphics for a more open exploration of stream sediment geochemical data. *Computers and Geosciences* 21: 77–87

Haslett J, Wills G, Unwin A 1990 SPIDER – an interactive statistical tool for the analysis of spatially distributed data. *International Journal of Geographical Information Systems* 4: 285–96

Hazelhoff L, Gunnink J L 1992 Linking tools for exploratory analysis of spatial data with GIS. *EGIS 92, Proceedings Third European Conference on Geographical Information Systems*. Utrecht, EGIS Foundation: 204–13

Isaaks E H, Srivastava R M 1989 *An introduction to applied geostatistics*. Oxford, Oxford University Press

MacDougall E B 1991 A prototype interface for exploratory analysis of geographic data. *Proceedings, Eleventh Annual ESRI User Conference* Vol. 2. Redlands, ESRI Inc.: 547–53

Majure J, Cook D, Cressie N, Kaiser M, Lahiri S, Symanzik J 1996a Spatial CDF estimation and visualisation with applications to forest health monitoring. *Computing Science and Statistics* 27: 93–101

Majure J, Cressie N 1997 Dynamic graphics for exploring spatial dependence in multivariate spatial data. *Geographical Systems*

Majure J, Cressie N, Cook D, Symanzik J 1996b GIS, spatial statistical graphics, and forest health. *Proceedings, Third International Conference/Workshop on Integrating GIS and Environmental Modeling, Santa Fe, 21–26 January*. Santa Barbara, NCGIA.

MathSoft 1996a *S+SpatialStats user's manual, version 1.0.* Seattle, MathSoft, Inc.

MathSoft 1996b *S+Gislink*. Seattle, MathSoft, Inc.

Monmonier M 1989 Geographic brushing: enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis* 21: 81–4

Ng R, Han J 1994 *Efficient and effective clustering methods for spatial data mining*. Technical Report 94–13. Vancouver, University of British Columbia, Department of Computer Science

Openshaw S 1990 Spatial analysis and geographical information systems: a review of progress and possibilities. In Scholten H, Stillwell J (eds) *Geographical information systems for urban and regional planning*. Dordrecht, Kluwer: 153–63

Openshaw S 1991c Developing appropriate spatial analysis methods for GIS. In Maguire D, Goodchild M F, Rhind D (eds) *Geographical information systems: principles and applications*. Harlow, Longman/New York, John Wiley & Sons Inc. Vol. 1: 389–402

Ord J K, Getis A 1995 Local spatial autocorrelation statistics: distributional issues and applications. *Geographical Analysis* 27: 286–306

Stuetzle W 1987 Plot windows. *Journal of the American Statistical Association* 82: 466–75

Symanzik J, Majure J, Cook D 1996 Dynamic graphics in a GIS; a bidirectional link between ArcView 2.0 and XGobi. *Computing Science and Statistics* 27: 299–303

Symanzik J, Majure J, Cook D, Cressie N 1994 Dynamic graphics in a GIS: a link between ARC/INFO and XGobi. *Computing Science and Statistics* 26: 431–35

Symanzik J, Megretskaia I, Majure J, Cook, D 1997 Implementation issues of variogram cloud plots and spatially lagged scatterplots in the linked ArcView 2.1 and XGobi environment. *Computing Science and Statistics* 28

Talen E 1997 Visualizing fairness: equity maps for planners. *Journal of the American Planning Association*

Tukey J W 1977 *Exploratory data analysis*. Reading (USA), Addison-Wesley

Unwin A 1994 REGARDing geographic data. In Dirschedl P, Osterman R (eds) *Computational statistics.* Heidelberg, Physica: 345–54

Unwin A, Hawkins G, Hofman H, Siegl B 1996 Interactive graphics for data sets with missing values – MANET. *Journal of Computational and Graphical Statistics* 5: 113–22

Upton G J, Fingleton B 1985 *Spatial data analysis by example*. New York, John Wiley & Sons Inc.

Williams I, Limp W, Briuer F 1990 Using geographic information systems and exploratory data analysis for archeological site classification and analysis. In Allen K, Green F, Zubrow E (eds) *Interpreting space: GIS and archaeology*. London, Taylor and Francis: 239–73