INT. J. GEOGRAPHICAL INFORMATION SCIENCE, 2001, VOL. 15, NO. 3, 239–254

Research Article

Providing spatial statistical data analysis functionality for the GIS user: the SAGE project

STEPHEN WISE, ROBERT HAINING and JINGSHENG MA

Sheffield Centre for Geographical Information and Spatial Analysis, Department of Geography, University of Sheffield, Sheffield S10 2TN, UK

(*Received 19 October 1999; accepted 2 September 2000*)

Abstract. Geographical Information Systems (GIS) are being used in a growing number of application areas. As a consequence there have been frequent calls to expand the range of spatial analysis tools available to users of GIS but a reluctance on the part of GIS software vendors to include such tools in standard software packages. An alternative approach is to link extra tools to GIS packages which raises a series of issues, such as, What sort of tools should be included? How should the linkage be done? To what extent can the functionality of the GIS be used? This paper draws on the results of a project in which software for statistical spatial data analysis (SSDA) was linked to ARC/INFO to produce a software system called SAGE. The statistical tools implemented included those which were felt to be useful to the general GIS user (as opposed to the specialist spatial statistician or econometrician), and they were linked to ARC/INFO using a client server architecture. The GIS was used within the context of SSDA for map drawing, spatial queries and operations on the topology of the spatial data, although it was found that the map drawing facilities of ARC/INFO were not well suited to the needs of this application. One of the conclusions of the project was that many of the techniques of exploratory spatial data analysis, such as providing graphical data summaries and linking these to cartographic views of the data could be easily integrated into existing GIS packages, providing a useful addition to their functionality for many GIS users. Many of the other SSDA facilities are probably still best provided in specialist software, but there is a need for a robust and standardised means for such software to extract information about the topology of spatial data from within GIS packages.

1. Introduction: the case for developing SAGE for GIS users

A feature of the GIS research agendas of both the UK Regional Research Laboratories and the American National Centre for Geographic Information and Analysis (Masser 1988, NCGIA 1989, Openshaw 1990) was the perceived need for GIS users to have access to a greater range of facilities for undertaking spatial analysis. Spatial analysis techniques have be defined as those 'whose results are dependent on the locations of the objects or events being analysed' (Goodchild *et al.* 1992). For example, computing the arithmetic mean of a set of values located across an area would not be a spatial analysis technique (since any re-arrangement of the

International Journal of Geographical Information Science
ISSN 1365-8816 print/ISSN 1362-3087 online © 2001 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/13658810010010655

values on the map would leave the arithmetic mean unchanged) but fitting a trend surface to the set of values would be because the order of the surface and the parameter estimates would be affected by where values were located.

Wise and Haining (1991) identified three types of spatial analysis which might be of interest to those working with GIS: map-based analysis, spatial modelling and statistical spatial data analysis (SSDA). SSDA, which is the subject of this paper, may be described as the analysis of empirical spatial data using statistical methods. There is a considerable degree of concensus as to the types of techniques that ought to be made available to the GIS community to facilitate SSDA (Bailey *et al.* 1994, Haining 1996, Anselin 1996). Two types of SSDA can be identified, although there is some overlap between them. Exploratory spatial data analysis (ESDA) is concerned with detecting spatial patterns in data, identifying unusual or interesting spatial features of the data (such as spatial outliers), formulating hypotheses which are based on or which are about the geography of the data and validating spatial models (Haining *et al.* 1998). Confirmatory spatial data analysis (CSDA) is concerned with model building, which normally involves the estimation of parameters (and their errors) and usually includes hypothesis testing as part of the process of model specification (Haining *et al.* 2000).

Current GIS contain only limited support for GIS-relevant SSDA (Goodchild 1987, Burrough 1990) which is not surprising since early developments in GIS were driven by a need for mapping, map-based analysis, and facilities management. However a number of factors have led to a growing interest in facilities for more sophisticated analyses of spatial data. First, there is a growing number of large spatial databases, including topographic, environmental and socio-economic data. It is possible to integrate datasets within GIS and this capability has significant practical implications for academic research in many areas where processes are multi-variate across social, economic and environmental explanatory variables. The analysis of fine grained geocoded multivariate datasets has the potential to make a uniquely important contribution to disentangling associations and assessing the importance of local conditions in fields like spatial epidemiology or environmental criminology.

Second, many organisations in both the public and private sector are under increasing pressure arising from increasing 'demand' and spiralling costs to increase efficiency by targetting resources, managing and auditing more rigorously. Geographical targetting and geographically based auditing are elements in a resource allocation policy given greater impetus by government pressures for public agencies to work together ('joined up government'). In many cases spatial analysis can be a useful tool in supporting the strategic and operational needs for organizations. In the field of health services provision, people's health and use of the health service often correlate with patterns in the social and economic circumstances of the population. Obtaining a reliable picture of these patterns, and of areas with unusually raised incidence can lead to a more targetted use of resources, but requires more than simply mapping the basic incidence rates, because these can be affected by other factors such as the age and sex structure of the area and the size of the population at risk. Similar examples can be cited in the areas of criminology, housing and education. There is a potential market here of GIS users for whom certain aspects of SSDA functionality would be useful.

Since GIS were not defined with SSDA in mind, it can be argued that it is preferable to build specialist SSDA software, and import the necessary spatial data

from the GIS (Haslett *et al.* 1990, Unwin *et al.* 1996, Dykes 1996, Brunsdon 1998). However it is then not possible to use those features of GIS which can help support SSDA, such as the ability to draw maps and the handling of information on the geometrical and topological properties of the spatial objects (Wise and Haining 1991). In addition there is the inconvenience of transferring files and the danger of creating multiple versions of the same dataset within the different packages. An alternative veiw is that SSDA facilities are likely to be most widely and effectively used if they can be linked to the GIS packages used to store and manipulate these spatial databases. One way of doing this is to use the GIS as the main software platform, using its customization facilities to supply the additional functionality (Ding and Fotheringham 1992, Batty and Yichun 1994, Kehris 1990a, 1990b). Another approach is to link the GIS with another package (Cook *et al.* 1996, Anselin and Bao 1997).

This paper reports on the main findings of a project, one of whose aims was to explore whether it was possible to develop an SSDA package linked to a GIS, which could provide general purpose facilities for both ESDA and CSDA and take advantage of the features of the GIS. To this end, a software system called SAGE (Spatial Analysis in a GIS Environment) was designed and implemented which included a direct link to the ARC/INFO GIS. The functionality of SAGE has been described elsewhere (Wise *et al.* 1997, Haining *et al.* 2000) and complete documentation for the package is available online (Ma *et al.* 1997). The purpose of the present paper is to focus particularly on the role which standard GIS functionality plays in SAGE, and drawing on our experience of SAGE and other similar packages, to dicuss some of the issues relating to the inclusion of SSDA into GIS.

The paper next describes the criteria which governed the selection of the SSDA functionality to be included in SAGE, and the most appropriate means of incorporating ARC/INFO. The features of the system are illustrated using an example and the paper concludes with a discussion of some of the implications of this work for future developments in the provision of SSDA facilities for the GIS community.

2. Design criteria for SAGE

SAGE was designed for the analysis of what Cressie (1991) calles 'lattice data' in which the spatial units are fixed, and interest lies in the variability of the attributes across the units. Areas are the commonest example of suich spatial units, but many of the techniques of area-based analysis can also be applied to the analysis of data for fixed points. Such data can be analysed by SAGE, by attaching a notional area to each point (normally by the generation of a Dirichlet tesselation around the points).

The intention was to provide a wide range of both ESDA and CSDA methods, allowing the analyst to proceed all the way from the initial exploration of a set of data, through to model specification, calibration and validation. ESDA is often used as a means of suggesting hypotheses about data which can then be tested more formally using CSDA methods. However, the division is not always so clear cut and the two approaches are best seen as complementary. For instance, although regression modelling uses formal tests of significance for model specification, visual, exploratory methods also have an important role to play in model specification, and can be used to check assumptions and examine the results of the analysis.

EDA methods (Tukey 1977) are characterised by an emphasis on visual and statistically robust means of exploring data, and this same emphasis has been carried over into ESDA. What is particularly important in the case of spatial data however,

is the connection between the attribute values and the location of the areal units in geographical space. If a boxplot reveals a distributional outlier, then one of the first questions the analyst is likely to ask is 'where is that case on the map?' This link is provided as standard functionality in most modern GIS software, but the connection is normally between the selection of one or more records from the database, and the selection of areas on a map. Haslett *et al.* (1990) were among the first to demonstrate that by making a much richer set of links, between various forms of statistical graph and a map, a wide range of analyses become very easy to perform. This linked windows or brushing facility is now provided by most software packages which have been developed for ESDA (MacDougall 1992, Dykes 1996, Brunsdon and Charlton 1996, Cook *et al.* 1996, Unwin *et al.* 1996) and was regarded as an important element in the functionality of SAGE.

Area-based data have two characteristics which led to a number of other design criteria for SAGE. Firstly, many spatial phenomena are characterised by a degree of spatial autocorrelation, particularly positive spatial autocorrelation which is the tendency for nearby locations to have similar values of a given attribute. This means that samples taken from neighbouring locations cannot be regarded as independent. This invalidates one of the central assumptions of (parametric and nonparametric) classical statistical methods, and may render the results of significance tests and the estimation of confidence intervals unreliable. A variety of techniques and models have been developed to deal with this situation (Haining 1990) and one of the aims of writing SAGE was to make some of these available in an easily accessible fashion. Autocorrelation effects are also of interest in their own right, including the identification of first order effects (What trends exist in attribute values across space?) and second order effects (To what extent are attribute values in neighbouring areas correlated?). This type of analysis requires information about the spatial arrangement of the areal units which is normally handled as a matrix (often called the connectivity or **W** matrix) representing the nature of the links between all pairs of areas, e.g. whether the areas are contiguous or not (Haining 1990). To construct this matrix requires information about the topology of the areal units, information which can often be extracted from a GIS, although as Goodchild (1987) points out, GIS are not generally designed to handle the information in matrix form. One of the aims was that SAGE would not only be able to handle the **W** matrix, but would allow the analyst to modify it in order to model different assumptions about the nature of the links between areas. A simple example is the situation where a single administrative zone is divided by a river, thus producing areas in the GIS data which are apparently not connected. One way of dealing with this is to modify the **W** matrix to reflect the fact that the areas are neighbours.

The second important feature is that the majority of area-based data is derived by aggregating values for individual items (people, households, houses etc.). The areas are modifiable in that the aggregation could be done at any one of a number of different scales, and at any given scale in numerous different (but equally plausible) ways. It has long been known that different aggregations of the same data can lead to different analytical results (Kendall 1939, Openshaw 1984, Openshaw and Rao 1995). In practice, areas for which data are available often have a real significance in the sense that they represent divisions of responsibility for an organisation—examples are health and police areas. However for any analysis which is trying to investigate patterns and processes in the underlying variables which have been aggregated, there may be several reasons why it is useful to be able to re-aggregate

the data into new zones. First, it is common in area-based analyses to convert absolute count data into rates, using the population of the areas as the demoninator. However, the realiability of such rates will vary since it is a function of the size of the populations (Clayton and Kaldor 1987, Kennedy 1989) and one solution to this is to aggregate the areas into larger units with approximately equal populations. Second, rates calculated for areas with small populations will be particularly sensitive to inaccuracies in the data, such as errors in the basic count variable, or errors in assigning individual events to areas. This sensitivity can be reduced by aggregating the original zones into a series of larger ones. Third, with a large number of small areas, broad trends may be lost in local detail, and so aggregation is one way of identifying broad scale trends in the data.

A number of other design criteria were adopted because of the particular nature of the project. Because of the level of resources made available, it was important to use existing software wherever possible, rather than writing code from scratch. Since funding was from a UK research council, it was important that any resulting software be freely available to academics in the UK (and as far as possible, elsewhere). ARC/INFO was chosen as the GIS, because of its widespread availability in the GIS community, and because it is available to UK academics at a heavily discounted price due to a centralized purchase (Wise 1990).

3. The architecture of SAGE

In this section we consider how much of the functionality needed to support the specific programme of SSDA could be provided by the GIS itself. We only consider ARC/INFO as this was the GIS used in the work. However because ARC/INFO has a particularly rich set of functionality (one of the reasons it was selected for widespread use in the UK academic community (Wise 1990)), this does not restrict the generality of the discussion although clearly other systems may well have particular strength which would have lead to a slightly different division of labour between the GIS and the other pieces of software. We will return to the general issue of how best GIS might support spatial analysis in the concluding section.

The operations needed to support the functionality outlined in the design section can be classified as shown in table 1. The table presents the functionality at

Table 1. Operations needed to support ESDA functionality.

	High level	Fundamental level
Data management		Manage Spatial data
		Manage Attribute data
		Create **W** matrix
Data manipulation	Calculation of rates	Calculations on attributes
	Statistical tests	Classification and regionalisation algorithms
	Classification and regionalisation	Spatial element selection
	Selection of subsets of data	Database queries
	Editing **W** matrix	Polygon dissolve
Data display	Map drawing	Graphical display
	Tabular data display	Window management
	Statistical graphs	
	Linked windows	

two levels—as seen by the user (the High Level), and broken down into the underlying technical operations needed to support the high level functionality (the Fundamental Level).

Many of the high level operations cannot currently be performed by ARC/INFO, but in many cases it does possess the underlying technical capability to support them. This can be seen in the functions listed under data display for instance. ARC/INFO has comprehensive map drawing capabilities, some ability to display attribute data in tabular and graphical form, but no linked windows facility. However, using AML and the graphical drawing primitives of Arcplot, it is possible to implement these features entirely within ARC/INFO. For example, Batty and Yichun (1994) implemented a model of urban land use in ARC/INFO in which two different views of the model, a map and a graph, are drawn within the same Arcplot window. Using ARC/INFO's macro language (AML) the two views were linked so that when elements were selected from one view, the same elements were highlighted in the other view. Similar experiments were carried out during the writing of SAGE, but it was found that this was not an appropriate way to build a general-purpose linked windows facility. All the graphics must be contained in the same window, which is inflexible and cumbersome and the use of AML, an interpreted command language, makes the response of the system very slow.

To write SAGE it was necessary to provide some of the functionality outside ARC/INFO and to link this with ARC/INFO itself. Given the need for a rapid and responsive system, the majority of the graphical capabilities were provided outside ARC/INFO, simply using ARC/INFO for drawing the maps. ARC/INFO would also be used to perform the selection of spatial subsets of data (e.g. all areas within a defined polygon), again since it already has comprehensive capabilities in this area. Many of the numerical elements of SAGE are computationally intensive (especially the classification/regionalisation elements) and would perform more efficiently if written in a language such as C++. Those parts of SAGE which related to the handling of topological data are split between those which are central to ARC/INFO and which it would be foolish to re-write (generating the **W** matrix from the polygon data and dissolving a set of polygons as a result of reclassification) and those which it would be hard to do in ARC/INFO (editing the **W** matrix).

In summary, the major functions for which ARC/INFO was suitable within SAGE were map drawing and querying, generating the basic topological data and the polygon dissolve operation. All the other functions of SAGE would be provided using other software. The next question to be considered was how best to link ARC/INFO with the other elements of SAGE. Other workers have used a number of approaches for this task-loose coupling via the transfer of files (Anselin *et al.* 1993), close coupling in which the GIS calls routines written in other languages (Ding and Fotheringham 1992) and client-server computing (Cook *et al.* 1996). Since the intention was that SAGE should operate rapidly and responsively, especially when the graphical, exploratory tools were being used, methods such as loose coupling were discarded because they would be too slow. Close coupling would be quick, but it was felt that the capabilities which could be provided might then be constrained by what could be achieved from within ARC/INFO.

The client server architecture (Umar 1993) provides flexibility in the way that various software components may be linked, and for this reason this approach was chosen for SAGE (Haining *et al.* 1996). A component is considered a client if it requests the services of other components to complete a certain task, or as a server

if it provides services for clients. The communications between clients and servers are handled efficiently through a set of well-defined Application Program Interfaces (APIs) utilising, for example, remote procedure calls (RPCs) (Simon 1996, pp. 65–8).

As described above, SAGE has two major software components—ARC/INFO and a purpose written module for providing all the other functionality. These could have been implemented so that both could function as either client or server which is the approach used by Cook *et al.* (1996) to link ArcView and Xgobi. However, this results in a system in which the user must know when to use each component as the client and it was decided that it would be simpler if the purpose written SSDA module was the client, calling ARC/INFO as a server for mapping and dataset management.

This approach has a number of advantages. Firstly, it allows the client and the server to be implemented independently, communicating with each other only through pre-defined APIs. Secondly, it allows existing and tested code appropriate for their implementation to be re-used wherever possible without being constrained by each other. Therefore, the implementation workload could be reduced and the reliability of the system could be expected to be high. Thirdly, because the client and the server communicate through a pre-defined API, it offers the possibility of using a different GIS to replace ARC/INFO, thus making the SAGE client potentially portable. Fourthly, since the client-server model can be used for distributed computing (using RPCs for example) the SAGE server and the SAGE client could be run on different platforms on the network. This would be useful for analysing spatial data held remotely. All these advantages were exploited during the implementation of SAGE where networked SUN workstations running X-Windows were used.

The full architecture of the system is shown in figure 1. It can be seen from figure 1 that the main link between the SAGE client and ARC/INFO is via a linking

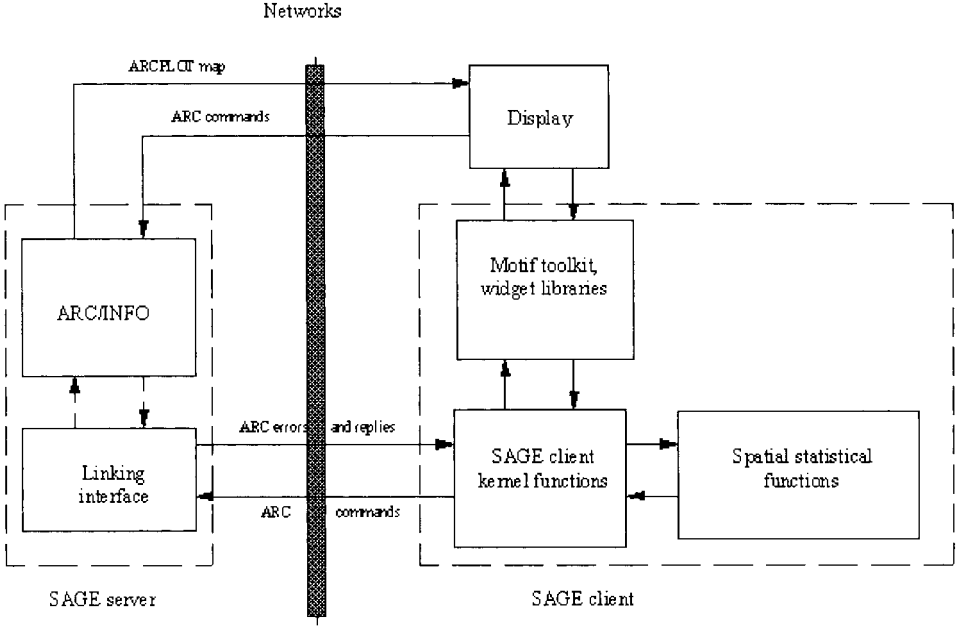

Figure 1. The architecture of SAGE.

interface which translates requests from the client into a series of ARC/INFO commands (some of which are actually purpose-written in AML) and converts the responses from ARC/INFO into a form suitable for the client.

4. An example SAGE session

A comprehensive description of the facilities provided in SAGE is available elsewhere (Ma *et al.* 1997, Haining *et al.* 2000). This section will attempt to give an overview of how the system operates by describing how a simple SAGE session might work. The example has been chosen to illustrate the sort of facilities which might be of interest to the general GIS user, and so the focus is on graphical, exploratory methods rather than the statistical modelling techniques which are also provided in SAGE.

The example is taken from work undertaken on behalf of the Trent Region of the UK National Health Service Executive, examining trends in the pattern of ill health in the region. The only health variable which will be considered here is the proportion of people in each area who have a limiting long term illness (LLTI) i.e. one which they consider limits their ability to work. It is well known that material deprivation is strongly linked to people's health, and so in an initial consideration of LLTI the questions which might be posed would include:

- What is the pattern of LLTI variation in the area?
- Is there a relationship between LLTI and deprivation?
- Is there any evidence of spatial cluster of high rates of LLTI?

Figure 2 shows a screen shot of a SAGE session. The window in the top left (labelled SAGE) is the user interface to the SAGE client. The map, which has been drawn by ArcPlot, shows the rates of LLTI in the region's 871 wards. Figure 3 shows the location of the region which roughly corresponds with the Eastern half of the English Midlands. In order to draw the choropleth map in SAGE (figure 2) several steps are needed. First the LLTI values must be classified, to identify which class each ward will fall in. This is done using the classification option on the main SAGE window, and the resulting classification is saved as a new variable (labelled LLTI-5) which is displayed in the main SAGE window (figure 2). In this case the variable was grouped into five classes, with equal numbers of wards in each class (a quintile classification), and so it was necessary to create a palette of five shades of grey, ranging from dark to light, in order to produce the final map.

What is of particular interest is the distribution of high rates of LLTI which can be identified in several different ways in SAGE. Figure 2 shows a boxplot of the LLTI rates with the points above the upper quartile selected graphically. This causes the corresponding areas on the map to be highlighted, in this case using a cross-hatch shading. There seem to be two main areas of high LLTI rates—along the coast north of Skegness, and in the main urban areas (Leicester, Nottingham, Sheffield, Rotherham). In order to check that these high values are not simply the result of wards with very small populations, a histogram of the population in each ward is also shown in figure 2. When the high rates are selected in the boxplot window, the same areas are highlighted in the histogram window showing that some high values do occur in the least populated wards. Rates based on small populations are not robust so the regionalization module of SAGE was used to merge some of the smaller wards together. This module is fully described in Wise *et al.* (1997) and Haining *et al.* (1998). It allows areas to be combined together to produce new

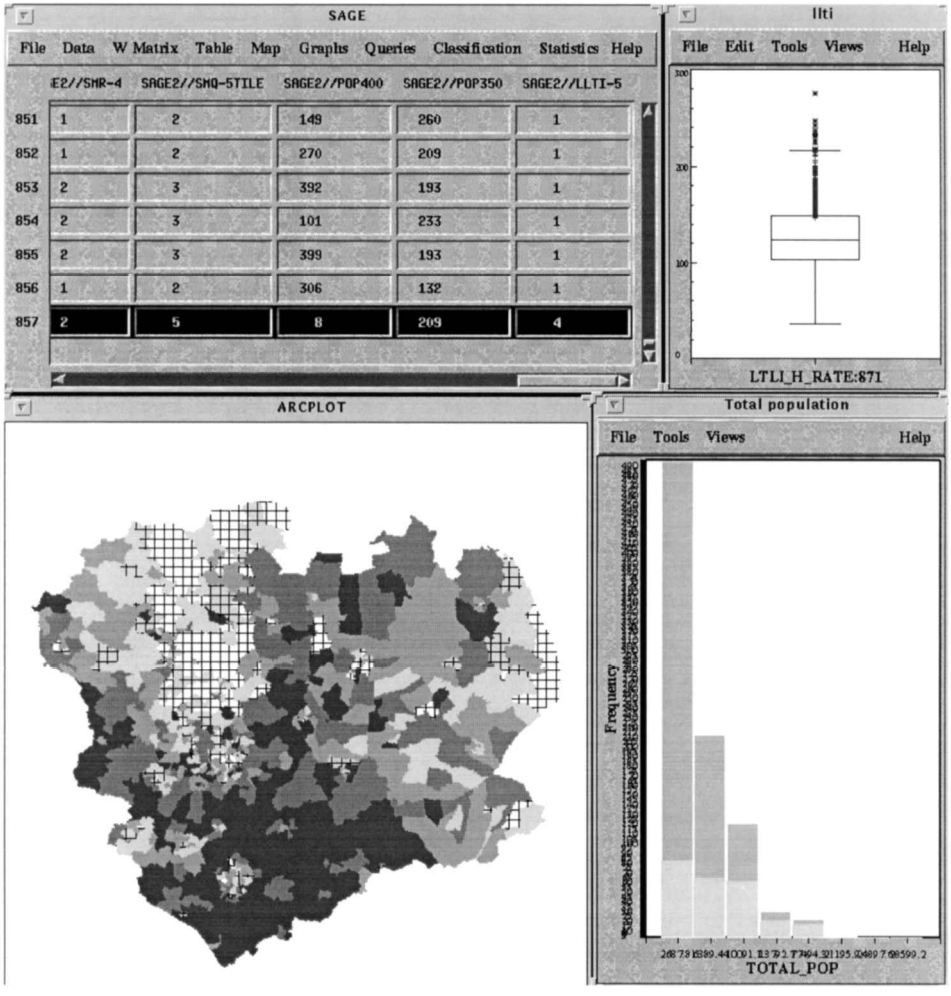

Figure 2. Screenshot of a SAGE session. The main user interface is via the menu options on the tabular view of the data. The map shows rates of Long Term Limiting Illness for wards in the Trent region on the NHS executive and the same rates are shown in the boxplot. The histogram shows total population in each ward. For an explanation of the highlighted features, see the text.

regions, which satisfy one or more of three criteria: homogeneity in terms of one or more variables, equality of the total value of one variable and compactness of shape. The original 871 wards, with population values between 837 and 30450, were combined to produce 500 regions using the criteria of population equality, and homogeneity in values of the Townsend deprivation index. Although the minimum population in the new regions only rises to 1136, none of the high LLTI rates is now found in a zone with a small population.

In order to look at the relationship between LLTI and deprivation a scatter plot is drawn using these two variables as shown in figure 4. The scatterplot suggests that deprivation is an important factor in the distribution of ill health in this area. The graph also shows three areas where the rate of LLTI is considerably higher than

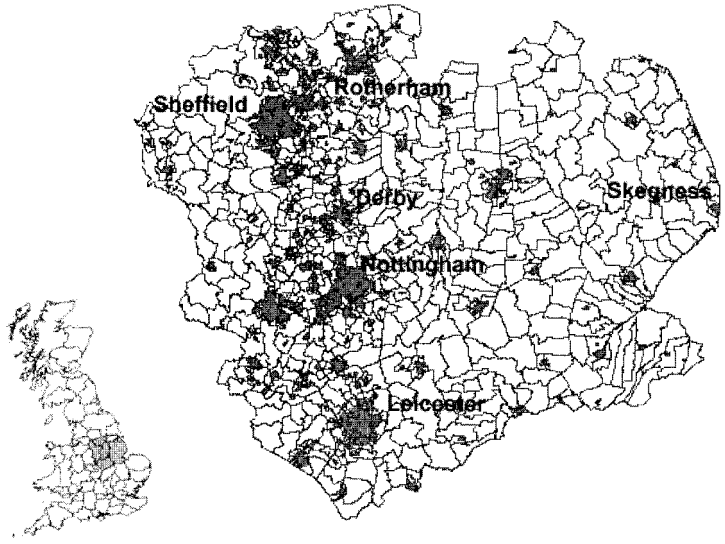

Figure 3. Location of the Trent region of the NHS. The larger map shows the ward boundaries, urban areas and places named in the text.

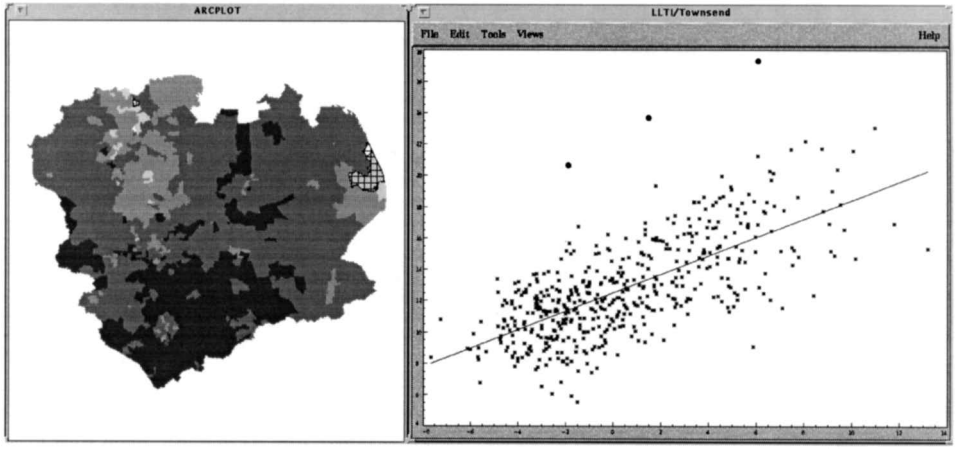

Figure 4. Relationship between Long Term Limiting Illness (LLTI) and Material Deprivation (measured using the Townsend Index) is shown on the scatterplot. The map shows LLTI rates, with the six outliers from the scatterplot identified by cross-hatching.

expected given the levels of deprivation. These are outliers from the regression fit (i.e. the standardized residuals are more than three standard deviations above the regression line) and to see where they are located, they have been selected in the scatter plot window. The map shows that one is located on the outskirts of Rotherham in the north, the other two on the coast.

The analysis so far has suggested that wards with high rates of LLTI tend to occur in urban and coastal areas but this has been based on considering each ward independently. By considering each ward in relation to its neighbours, it is possible to identify clusters of wards with high LLTI values, and this has been done in figure 5.

The Getis-Ord statistic (Getis and Ord 1992) has been calculated for the LLTI

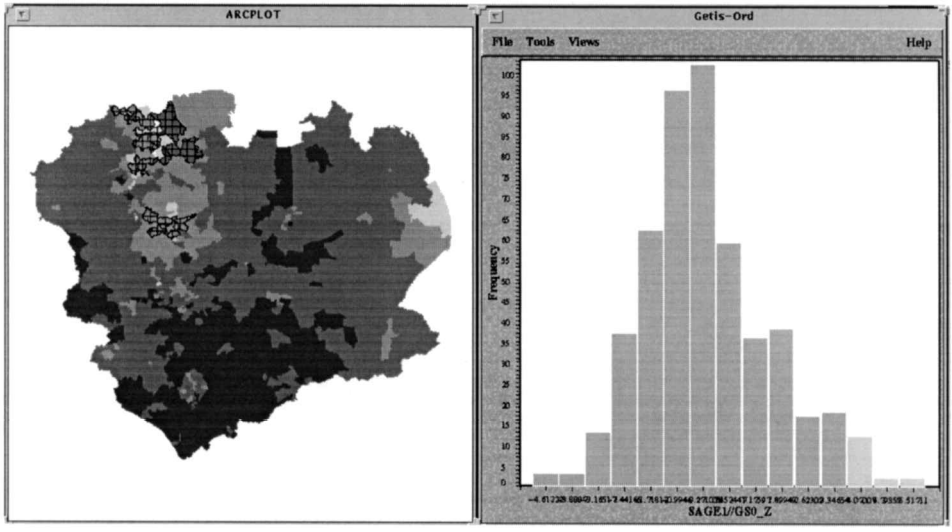

Figure 5. The detection of clusters using SAGE. The histogram shows values of the Getis-Ord statistic, a measure of local autocorrelation. Areas with high positive values have been selected, and are shown on the map by cross-hatching.

values and the histogram of the resulting values is shown on the right of figure 5. A high positive value indicates a ward with a high LLTI value with neighbours which also have high LLTI values, although individually they need not necessarily fall in the upper tail of the LLTI histogram. Selecting the top three categories from the histogram reveals the location of these clusters on the map. None of these are located on the coast. There is a large cluster around the Sheffield/Rotherham conurbation particularly in those areas which house workers in the steel industry. The other major cluster is further south around Derby. Although this lies in one of the areas previously identified as having generally high rates (figure 2) it is not clear why this particular area should appear to have a cluster of high values. Indeed, this may simply be an artefact of the area boundaries, something which could be explored by repeating the analysis using a different set of regions.

This is a fairly brief analysis of this set of data but even so the example illustrates how the use of relatively straightforward numerical and graphical techniques within SAGE can reveal useful information about spatial data.

5. Discussion: operational benefits and other approaches

Section 1 presented the case for building SAGE in terms of GIS user needs. In this section, we discuss the extent to which the SSDA facilities in SAGE benefited from being linked to a GIS. Within SAGE, ARC/INFO performs data management tasks, provides topological information and displays maps. The topological data is needed for the calculation of spatial statistics, such as the Getis-Ord statistic, and in the regionalization process. For both of these, the contiguity information about the area boundaries is extracted by SAGE from ARC/INFO. The regionalisation algorithm assigns each of the original areas a code which identifies which new region it will belong to, and this information is passed back to ARC/INFO so that it can perform a polygon dissolve operation to create a polygon coverage for the new regions.

With spatial data analysis the analyst should be able to manipulate the spatial framework and there is a strong case for making use of the functionality of existing GIS to do this. There is a large difference between the ability to import or calculate the topological information for a set of area boundaries, as is done by systems like cdv (Dykes 1996) or LiveMap (Brunsdon 1998), and the ability to alter those boundaries and update the topology and this seems one area of ESDA where a link to GIS is of great benefit.

The cartographic facilities of ARC/INFO work well in the case of selecting spatial subsets of the data and for implementing the brushing technique. However, it would be useful to make this facility more dynamic by being able to change the selection of areas interactively, with the related graphical plots changing simultaneously (dynamic brushing). However, given the need to communicate each change between the server and client in the current SAGE design, this would be far too slow using the current architecture. For this to work at an appropriate speed, the cartographic and graph drawing facilities would have to be provided in the same piece of software.

There are other respects in which the cartographic facilities of ARC/INFO are less suitable for a system like SAGE. ARC/INFO's Arcplot module was originally written with the production of paper maps in mind, not interactive visualisation. For example, it is difficult to draw a greyscale choropleth map of a variable unlike in systems such as cdv (Dykes 1996), MANET (Unwin *et al.* 1996) and Descartes (Andrienko and Andrienko 1999).

In order to derive some general conclusions about the possible role for GIS in supporting spatial analysis, it is necessary to consider the extent to which our experience in building SAGE might have been different if we had used a different GIS package. Our view is that the majority of GIS packages provide a very similar mix of functionality to that provided in ARC/INFO—none appear to possess a wider range of standard statistical graphs, or numerical statistical facilities. Some, such as MapInfo, support linked windows, but this appears to be limited to a single link between the map and a tabular display of attribute values. However some other packages might have been more suited to providing some of the functionality provided by ARC/INFO. In the case of map drawing this is almost certainly the case. More recent desktop mapping packages, such as ArcView and MapInfo for example have been designed with interactive map viewing in mind, and provide methods which allow data to be viewed very quickly on screen.

The one area in which ARC/INFO is perhaps strongest is in the accessibility of the topological information. Many of the current generation of vector GIS packages are based on the topological link and node data structure (Peucker and Chrisman 1975) which stores the identity of the two polygons bordering each link. This information can be used to construct two of the three **W** matrices which SAGE uses—that based on simple contiguity, and that based on the length of the border between areas.

In the case of ARC/INFO the data structure can be held internally, or it can be stored in a relational table in the INFO database for both line and area coverages making it directly accessible to the user. ARC/INFO is unusual in allowing such easy access to this information, and there are at least two reasons why most other software systems do not. Firstly, since the topological data is fundamental to the integrity of the GIS database, it is not desirable to store it in such a way that it can potentially be corrupted by the user. Secondly, there should be no need for the user to have direct access to this low level data—all of the operations which require it

should be provided as high level functions. However, as SAGE has shown, certain types of spatial analysis which are not currently supported by GIS also require access to this topological data. Given the reluctance of vendors to incorporate SSDA techniques into standard GIS (Maguire 1995) this makes it important that the mechanisms provided to link GIS to other software include some means of accessing the topological data.

The current trend in the GIS industry is towards the greater use of inter-operability as a means of building software solutions, especially using an object oriented approach (Graham 1999), and the SAGE project and the work of others (Cook *et al.* 1996) has shown that this case can be used as a means of providing SSDA functionality to GIS users. However, in the current offerings from vendors, the topological information is still seen as something to be handled internally, rather than as information which might be transferred between objects. The OpenGIS consortium has produced a specification of what it calls an Essential Model and it is to be hoped that this process will lead to the development of a standard way of providing information on the spatial relationships between spatial objects. However this process is at an early stage, with the only proposal which has reached any sort of fruition being the one which states how systems will deal with simple, geometrical forms such as lines and whole polygons (Buehler and McKee 1998, OpenGIS Consortium 2000).

There is a strong case for incorporating some of the techniques from systems such as SAGE directly into mainstream GIS software, and in this contect we identify two particular types of functionality: the first is the linked windows or brushing facility. This is a common feature of the majority of systems written to explore different methods of providing ESDA functionality (Andrienko and Andrienko 1999, Haslett *et al.* 1990, Cook *et al.* 1996, Dykes 1996, Unwin *et al.* 1996, Anselin and Bao 1997, Brunsdon 1998) and has been found to support a wide range of analytical operations. The second, which is related to the first, is the provision of graphical and other techniques which support spatial, as opposed to non-spatial, exploratory analysis. Features such as lagged boxplots and Moran plots are simple, intuitive means of exploring spatial data and regionalization allows the user to experiment with the effects of altering the spatial framework (Haining 1990). None require a detailed knowledge of spatial statistics in order to use them, but they do provide useful insights into the spatial variations which may exist in a set of data and there would seem to be a strong case for adding them directly into mainstream GIS software.

6. Conclusions

There are a number of conclusions which are of potential interest to both the GIS community and GIS software vendors. It should be clear from this work, and the work of other software developers, that there are a range of techniques which will be of interest to a wide range of GIS users, who already have GIS databases of area-based data and wish to do more with that than simply draw maps or make routine queries. In particular, the provision of relatively simple graphical displays, boxplots, scatterplots and choropleth maps, in a graphical environment in which the windows are dynamically linked has been shown by many researchers to provide a very powerful tool for undertaking many forms of ESDA. What is more, experience suggests that such techniques are intuitively simple to use, and so will be accessible to a wide range of users. There is a clear potential here for GIS software vendors to

add such functionality into standard GIS packages and this would be a major benefit to many users.

The more esoteric, specialist statistical tools of SDA are likely to be of interest only to a minority of users, although we believe that this is a significant minority. As indicated in the introduction, the range of science-based and policy-based research issues that call for rigorous SSDA is growing in both the public and private sectors. In the past, vendors have been reluctant to invest in developing such methods as part of their existing products (Maguire 1995), and this may continue to be the case. The examples of SAGE and other systems (Cook *et al.* 1996) have demonstrated that this type of functionality can be linked to GIS software, so it is not strictly necessary for GIS vendors to provide it within the GIS. What is needed to facilitate this linking process is a means of passing the topological information between the two applications. The current offerings of vendors in the area of software objects which can be used to build special-purpose GIS do not appear to support this, viewing the topological information as something to be encapsulated within the spatial objects. Work in the OpenGIS consortium does appear to have recognised the importance of information about the relationships between spatial features as important data in its own right, and it is to be hoped that this leads to better access to this type of information.

In summary, as the range of GIS users grow, the case for incorporating certain types of SSDA techniques directly into GIS also grows. Whilst the list of appropriate techniques will undoubtedly evolve, the linkage is also technically appropriate since GIS have many features that facilitate the implementation of SSDA techniques.

Acknowledgments

Our thanks to two anonymous referees whose comments did much to improve the quality of this paper. The authors wish to acknowledge receipt of ESRC research grant R000234470 'Developing spatial statistical software for the analysis of area based health data linked to a GIS'. The ward boundaries shown in the figures are from the EDLINE Digitised Boundary Data, and are Crown and EDLINE copyright (ESRC and JISC purchase). The urban areas in figure 3 are from the Bartholomew Map Data Great Britain 1:200000 database, Copyright Bartholomew Limited. (CHEST Bartholomew agreement). The population data are derived from the 1991 Census, Crown Copyright (ESRC purchase). The health data were provided by the Trent Region of the NHS Executive.

References

Andrienko, G. L., and Andrienko, N. V., 1999, Interactive maps for visual data exploration. *International Journal of Geographical Information Science*, **13**(4), 355–374.

Anselin, L., 1996, The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In *Spatial Analytical Perspectives on GIS*, edited by M. Fischer, H. Scholten and D. Unwin (London: Taylor and Francis), pp. 111–125.

Anselin, L., and Bao, S., 1997, Exploratory spatial data analysis linking SpaceStat and Arc View. In *Recent Developments in Spatial Analysis: Spatial statistics, behavioural modelling and neuro-computing*, edited by M. Fischer and A. Getis (Berlin: Springer-Verlag), pp. 35–59.

Anselin, L., Dodson, R. F., and Hodak, S., 1993, Linking GIS and Spatial Data Analysis in Practice. *Geographical Systems*, **1** (1), 3–23.

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

Batty, M., and Yichun, X., 1994, Urban Analysis in a GIS Environment: Population Density Modelling using ARC/INFO. In *Spatial Analysis and GIS*, edited by S. Fotheringham and P. Rogerson (London: Taylor and Francis), pp. 189–220.

Brunsdon, C., 1998, Exploratory spatial data analysis and local indicators of spatial association with XLISP-STAT. *Journal of the Royal Statistical Society Series D: The Statistician*, **47** (3), pp. 471–484.

Brunsdon, C., and Charlton, M. E., 1996, Developing an exploratory spatial analysis system in XLisp-Stat. In *Innovation in GIS 3*, edited by D. Parker (London: Taylor and Francis), pp. 135–145.

Buehler, K., and McKee, L., 1998, *The OpenGIS guide*. 3rd edition. [Online document] http://www.opengis.org/techno/guide/guide/Guide980629.pdf. [Visited 5th May 1999]

Burrough, P. A., 1990, Methods of spatial analysis in GIS. *International Journal of Geographical Information Systems*, **4** (3), pp. 221–223.

Clayton, D., and Kaldor, J., 1987, Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, pp. 671–681.

Cook, D., Maguire, J. J., Symanzik, J., and Cressie, N., 1996, Dynamics graphics in a GIS: exploring and analyzing multivariate spatial data using linked software. *Computational Statistics*, **11**, pp. 467–480.

Cressie, N. A. C., 1991, *Statistics for Spatial Analysis* (New York: John Wiley and Sons).

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

Dykes, J., 1996, Dynamic maps for spatial science: a unified approach to cartographic visualization. In *Innovation in GIS 3*, edited by D. Pankew (London: Taylor and Francis), pp. 177–187.

Fischer, M., Scholten, H., and Unwin, D., 1996, *Spatial Analytical Perspectives on GIS* (London: Taylor and Francis).

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

Goodchild, M. J., Haining, R. P., Wise, S. M., and 12 others, 1992, Integrating GIS and spatial data analysis: problems and possibilities. *International Journal of Geographical Information Systems*, **6**, 407–423.

Goodchild, M. G., 1987, A spatial analytical perspective on geographical information systems. *International Journal of Geographical Information Systems*, **1**, 327–334.

Graham, L., 1999, NT-based GIS rises to the occasion. *GeoEurope*, **8**, 34–39.

Haining, R., 1994, Designing spatial data analysis modules for geographical information systems. In *Spatial Analysis in GIS*, edited by A. S. Fotheringham and P. Rogerson (London: Taylor and Francis), pp. 45–63.

Haining, R., 1996, Designing a health needs GIS with spatial analysis capability. In *Spatial Analytical Perspectives on GIS*, edited by M. Fischer, H. Scholten and D. Unwin (London: Taylor and Francis), pp. 53–65.

Haining, R. P., 1990, *Spatial Data Analysis in the Social and Environmental Sciences* (Cambridge: Cambridge University Press).

Haining, R. P., Ma, J., and Wise, S. M., 1996, The design of a software system for interactive spatial statistical analysis linked to a GIS. *Computational Statistics*, **11**, 449–466.

Haining, R. P., Wise, S. M., and Ma, J., 1998, Exploratory spatial data analysis in a geographic information system environment. *The Statistician*, **47**, 457–469.

Haining, R. P., Wise, S. M., and Ma, J., 2000, Designing and implementing software for spatial statistical analysis in a GIS environment. *J. Geographical Systems*, **2**(3), 257–286.

Haslett, J., Willis, G., and Unwin, A. R., 1990, SPIDER—an interactive statistical tool for the analysis of spatially distributed data. *International Journal of Geographical Information Systems*, **4**, 285–296.

Kehris, E., 1990a, Spatial Autocorrelation Statistics in ARC/INFO. North West Regional Research Lab Report 16.

Kehris, E., 1990b, A Geographical Modelling Environment Built Around ARC/INFO. North West Regional Research Lab Report 13.

Kendall, M. G., 1939, The geographical distribution of crop productivity. *Journal of the Royal Statistical Society*, **102**, 21–48.

Kennedy, S., 1989, The small number problem and the accuracy of spatial databases. In *The accuracy of spatial databases*, edited by M. F. Goodchild and S. Gopal (London: Taylor and Francis), pp. 187–196.

Ma, J., Haining, R. P., and Wise, S. M., 1997, SAGE users guide [online document] http://www.shef.ac.uk/~scgisa

MacDougall, E. B., 1992, Exploratory analysis, dynamic statistical visualisation and geographic information systems. *Cartography and Geographic Information Systems*, **19**, 237–246.

Maguire, D. J., 1995, Implementing spatial analysis and GIS applications for business and service planning. In *GIS for Business and Service Planning*, edited by P. Longley and G. Clarke (Cambridge: Geoinformation International), pp. 171–191.

Masser, I., 1988, The Regional Research Laboratory Initiative: a progress report. *International Journal of Geographical Information Systems*, **2**, 11–22.

NCGIA, 1989, The research plan of the National Centre for Geographic Information and Analysis. *International Journal of Geographic Information Systems*, **3**, 117–136.

O'Kelly, M., 1994, Spatial analysis and GIS. In *Spatial Analysis in GIS*, edited by A. S. Fotheringham and P. Rogerson (London: Taylor and Francis), pp. 65–79.

OpenGIS Consortium, 2000, *OpenGIS Consortium Inc.—Frequently Asked Questions (FAQs)*. http://www.opengis.org/FAQs.htm#q9 (Updated 18 April 2000).

Openshaw, S., 1984, *The modifiable areal unit problem*. Concepts and Techniques in Modern Geography 38 (Norwich: GeoAbstracts).

Openshaw, S., 1990, A spatial analysis research strategy for the regional research laboratory initiative. Regional Research Laboratory Initiative Discussion Paper Number 3, Department of Town and Regional Planning, University of Sheffield.

Openshaw, S., and Rao, L., 1995, Algorithms for re-engineering 1991 census geography. *Environment and Planning A*, **27**, 425–446.

Peuker, T. K., and Chrisman, N., 1975, Cartographic Data Structures. *American Cartographer*, **2**, 55–69.

Simon, E., 1996, *Distributed Information Systems—from clinet-server to distributed multimedia* (London: McGraw Hill).

Tukey, J. W., 1977, *Exploratory Data Analysis* (Reading, Mass.: Addison Wesley).

Umar, A., 1993, *Disturbed Computing and Client-Server Systems* (New York: Prentice Hall).

Unwin, A., Hawkins, G., Hofman, H., and Siegl, B., 1996, Interactive graphics for data sets with missing values—MANET. *Journal of Computational and Geographical Statistics*, **5**, 113–122.

Wise, S. M., 1990, Evaluating GIS software for use in Higher Education. *Mapping Awareness*, **4**, 41–43.

Wise, S., and Haining, R., 1991, The role of spatial analysis in Geographical Information Systems. *Proceedings of the 3rd National Association for Geographic Information Conference*, 3.24.1–3.24.8.

Wise, S. M., Haining, R. P., and Ma, J., 1997, Regionalisation tools for the exploratory spatial analysis of health data. In *Recent Developments in Spatial Analysis—Spatial Statistics, Behavioural Modelling and Neurocomputing*, edited by M. Fischer and A. Getis (Berlin: Springer).