INT. J. GEOGRAPHICAL INFORMATION SYSTEMS, 1992, VOL. 6, NO. 5, 407–423

# Integrating GIS and spatial data analysis: problems and possibilities‡

MICHAEL GOODCHILD

National Center for Geographic Information and Analysis,
University of California, Santa Barbara, U.S.A.

ROBERT HAINING, STEPHEN WISE and 12 others†

Department of Geography, University of Sheffield,
Sheffield, England, U.K.

(The workshop was sponsored by the Economic and Social Research Council as part of its Regional Research Laboratory (RRL) initiative. Receipt of ESRC grant A 504 265004 is gratefully acknowledged).

**Abstract.** This article is an agreed summary of a workshop held in Sheffield between 18–20 March 1991. The focus here is on three of the themes of the workshop: the mutual benefits of closer links between geographical information systems (GIS) and the methods of spatial data analysis (SDA); the specific areas of SDA that should be linked with GIS; how the linkage should be made in practice. Directions for future research are also reviewed. The emphasis throughout is on statistical SDA and principally from the perspective of human rather than physical geography.

## 1. Introduction

Between 18–20 March 1991, a workshop was held at the University of Sheffield with the aim of discussing the issues associated in putting data analysis, and in particular spatial data analysis (SDA), techniques into GIS to the mutual benefit of both fields. (For an earlier statement on the need for research in this area see Openshaw 1990). The workshop brought together academics from three areas: those with specific interests in GIS (both theory and applications), those with interests in applied spatial analysis and those with interests in the theory and methods of spatial data analysis. A list of the participants is given at the end of the article. They prepared brief discussion papers, each of which was used to introduce one area of the agenda, the remaining time being allocated to discussion. The collection of these discussion papers is available as a Regional Research Laboratory (RRL) working paper (Haining and Wise 1991).

The detailed objectives of the workshop were:

1. To identify the types of spatial analysis tools that would be of particular value to the GIS user community. This objective includes identification of the techniques that are currently ready in a theoretical or conceptual sense to become routinely available. (This was thought to be a particularly important issue in the relatively new area of spatial statistics).

† See list of participants.

‡ Parts of Section 2 are reprinted with permission from *GIS/LIS '91, Vol. 1*, copyright 1991, by the American Society for Photogrammetry and Remote Sensing: "Spatial Analysis and GIS: Problems and Prospects," Michael F. Goodchild, pp. 40–48.

2. To discuss the problems associated with developing appropriate data analysis packages that can be interfaced with a GIS. This objective includes:

   (*a*) Identification of the techniques that are already available and interfaced with a GIS.

   (*b*) Identification of the technical and computational problems of developing new software and how these tools might be integrated with a GIS.

The next section draws heavily on the contributions by Goodchild and Openshaw who overviewed the case for closer links between SDA and GIS. The incorporation of SDA techniques offers the opportunity for GIS to realise its potential as a general tool for handling spatial data. For SDA the links with GIS will enable efficient data retrieval and encourage greater visualization for both exploratory analysis and model checking. This trend towards visualization is consistent with advances that are being encouraged in other areas of modern science.

The workshop focused on spatial analysis techniques used by social, rather than environmental scientists, and particularly on statistical SDA. The third section summarizes discussions on data preparation (Flowerdew, Green, Martin), data assessment and sampling (Arbia, Griffith), exploratory and explanatory data analysis (Brunsdon, Diggle, Griffith, Haining, Openshaw) and econometric modelling (Anselin, Hepple). The fourth section summarises discussions on how best to link SDA and GIS (Bossard, Brunsdon, Green, Wise). In the final section we review opinions and conclusions regarding directions for further research.

## 2. GIS and spatial data analysis: the case for closer linkage

There seems to be widespread agreement in the GIS community on two simple propositions: that as a technology, GIS has the potential to support many different types of analysis and that this potential has not yet been realized. This theme is recurrent in the collection of papers on GIS edited by Worrall (1990) and reflected in Openshaw's oft-quoted comment:

> Such systems are basically concerned with describing the Earth's surface rather than analysing it. Or if you prefer, traditional 19th-century geography reinvented and clothed in 20th-century digital technology. (Openshaw 1987, p. 431)

The potential to support analysis is reflected in many discussions of GIS:

> Geographic information systems evolved as a means of assembling and analysing diverse spatial data. (Star and Estes 1990, p. 14)
>
> ...the Geographic Information System...is as significant to spatial analysis as the inventions of the microscope and telescope were to science, the computer to economics, and the printing press to information dissemination. (Department of the Environment 1987, p. 8)

But from a strictly pragmatic viewpoint, the reality of GIS today might be summed up in the following:

> A database containing a discrete representation of geographical reality in the form of static, two-dimensional geometrical objects and associated attributes, with a functionality largely limited to primitive geometrical operations to create new objects or to compute relationships between objects, and to simple query and summary descriptions.

GIS clearly needs stronger analysis and modelling capabilities if it is to meet its potential as a general-purpose tool for handling spatial data.

Spatial data analysis (SDA) is a set of techniques devised to support a spatial perspective on data. To distinguish it from other forms of analysis, it might be defined as a set of techniques whose results are dependent on the locations of the objects or events being analysed, requiring access to both the locations and the attributes of objects (Goodchild 1987). Its techniques range from simple descriptive measures of patterns of events to complex statistical tests of whether a set of events could have been generated by specific, well-defined, processes (Ripley 1981, Getis and Boots 1978). Note that SDA as defined here does not include techniques that use only the attributes of objects.

The case for the development of methods for SDA rests on the argument that explanation, understanding and insight can come from both seeing and examining data in their spatial context. There seem to be at least four separate arguments for this spatial perspective. First, space can provide a simple and useful indexing scheme. An archaeologist, for example, might record the locations of artifacts as they are unearthed at a dig simply in order to index them for later access. Geographical coordinates provide a kind of hashing code, on the assumption that two artifacts are unlikely to be unearthed at exactly the same location. A map provides a simple means of displaying the index, and retrieving information from the database given the human eye's extraordinary power to digest two-dimensional information. At the very least spatial indexing is a useful tool for handling large amounts of data.

Second, the spatial perspective allows easy access to information on the relative locations of objects and events, and proximity can indeed suggest insight. The Snow map showing the clustering of cholera victims in London during an outbreak of the disease in 1854 led directly to explanation, in the form of drinking water from a polluted well, and to an effective remedy for the outbreak (Gilbert 1958). Although the explanation was immediately apparent from the map, it would have been virtually impossible to arrive at the same insight in any other way.

Third, a spatial perspective allows events of different types to be linked, in a process formalized in GIS as overlay. The fact that an event occurs in proximity to other events or objects can be very suggestive. For example, any environmental abnormality in the vicinity of a cluster of cancer cases is immediately suspect, because an individual is clearly more vulnerable to the local environment than to distant ones.

Finally, the distance between events or objects is often an important factor in interactions between them. In the physical sciences, distance appears as an explicit variable, as in the inverse square laws of gravitation or electromagnetics. In the social sciences it is more likely to be a surrogate for information (we are more likely to know about nearby places than about distant ones) or for contact (we know more people locally) or time spent travelling (we prefer local shops to distant ones, all other things being equal) (Gatrell 1983).

SDA provides a set of objective techniques to replace and augment subjective intuition. Unfortunately, while the spatial perspective can be very powerful as a source of insight, it can also be highly misleading. Ancient cultures found endless images in the patterns in the night sky. More recently, there are many examples in the literature of false inferences drawn from apparent spatial patterns that later turn out to be no different from the outcomes of random processes. Haggett and Chorley (1969), for example, found that the average number of edges in the Brazilian administrative boundary network is very close to 6. They concluded that this supported the contentions of Christaller's Central Place Theory. But the bubbles in a polystyrene coffee cup also have close to six edges on average, and analysis later showed that this is a necessary outcome of a theorem of Euler applying to any boundary network (Getis and Boots 1978). It is only through such objective analysis that it can be determined whether, for example, visual cancer clusters are in fact abnormalities, rather than random events.

GIS needs SDA if it is to reach the potential implied by many of its definers and proponents. SDA needs GIS, to take advantage of the capabilities in GIS for data input, editing, display and mapping, if it is to become readily accessible to a broad user community.

### 2.1. *Progress to date*

If the arguments for linking GIS and SDA are so strong, why has so little progress been made to date? First, developments in the GIS industry largely reflect the demands of the GIS marketplace. This has been dominated for the past decade by applications in resource management, infrastructure and facilities management, and land information. In these areas, GIS tends to be used more for simple record-keeping and query than for analysis. Although a small minority of companies has stressed analysis in their development and marketing, SDA has had little impact on the GIS mainstream. SDA is also of greater interest in academic and scientific applications of GIS than in local government or the private sector, where GIS budgets and expenditures tend to have been much higher.

Second, despite its promises, SDA remains a comparatively obscure field. There are few books or reviews, and there is no easy way of organizing or codifying SDA. Notable exceptions include the now classic collection edited by Berry and Marble (1968), Unwin (1981), Upton and Fingleton (1985) and the recent book by Haining (1990). There are no widely accessible courses in SDA, and there is some concern that the introduction of GIS into many university programmes may in fact have diverted resources from existing courses in SDA (Heywood 1990). In the long term, linkage with GIS may lead to greater awareness of SDA and greater availability of courses and texts, but in the short term there is a distinct shortage of knowledge, experience and training on the SDA side.

Third, many of the techniques for analysing spatial data were developed in the 1960s and 1970s when GIS was still in its infancy, and cartography a technology of pen and paper. Early efforts to implement these methods in a computational environment had to rely on source code programming, notably in FORTRAN (for examples, see Baxter 1976, MacDougall 1976). The 1970s saw the emergence of integrated statistical packages such as SAS and SPSS, and although they provided no explicit support for coordinates or spatial objects except for mapping and display, in practice they were the most readily available basis for implementing methods of spatial analysis. In addition, the statistical techniques on which these packages are based are in many cases explicitly non-spatial, making assumptions about the lack of spatial dependence which fly directly in the face of the spatial perspective. Spatial autocorrelation has often been treated as a problem to be removed (Odland 1988), rather than as an inescapable property of almost all spatial data.

It is only in the past few years, with the development of GIS, that a realignment of SDA with other explicitly spatial technologies such as cartography, GIS and remote sensing has finally begun. But much statistical SDA remains strongly linked to the aspatial environment of packages such as MINITAB and SAS (Griffith 1988).

Fourth, many techniques of SDA are complex and difficult, requiring a very different approach from the intuitive, synthetic view often promoted for GIS. A glance through the pages of a journal such as *Geographical Analysis* is sufficient to strike despair into the hearts of most GIS enthusiasts. As an academic specialty with little immediate connection with the world of practical application, SDA might be accused of emphazing mathematical sophistication at the expense of practicality. Simple, intuitive techniques for exploring data in a spatial context have often been ignored in favour of elegant formal modelling. This chain of thought suggests that the key to integrating GIS and SDA may lie in an emphasis, at least initially, on the more intuitive, exploratory techniques in the SDA toolkit.

### 2.2. *The role of data models*

A GIS database captures real geographical variation in the form of a finite number of discrete, digital objects. Because geographical variation is fundamentally continuous and complex, this process of capturing reality must involve abstraction, generalization or approximation. The rules by which the objects and their relationships are defined is termed a data model (Tsichritzis and Lochovsky 1977). The variety of data models used in GIS is one of its complications and at the same time one of its strengths.

Reviews of GIS data models have been published by Peuquet (1984) and Goodchild (1992). Data models take two broad forms, depending on whether reality is perceived as an empty space populated by objects, or as a set of layers or fields, each defining the spatial variation of one variable. In very broad terms, the former view is more relevant to analysis and modelling in the social sciences, where discrete entities are conceived as interacting over space, and the latter is more relevant to the environmental and physical sciences; but exceptions abound.

Objects are normally modelled as points, lines or areas, after appropriate generalization of form. Fields are modelled in GIS in at least six ways:

— a raster of cells, each defining the average value of the field within the cell (e.g., a remote sensing scene);
— a raster of regularly-spaced point samples (e.g., a digital elevation model);
— a set of non-overlapping, space-exhausting polygons, each defining a class (e.g., a map of soil or vegetation cover);
— a set of irregularly-spaced point samples (e.g., a weather map);
— a set of digitized isolines (e.g., a contour map);
— a set of non-overlapping, space-exhausting triangles, each assumed to approximate elevations within the triangle with a simple plane (the triangulated irregular network (TIN) model).

Data models not only define how geographical variation is represented, but also determine the set of processes and analyses that can be undertaken. For example, it would be appropriate to use a set of point samples representing a field of atmospheric temperature to interpolate a contour map or create an oblique view, using the attributes of each point to determine the elevation of the interpolated surface. But it would be meaningless to perform the same operation on a set of point objects representing cities with attributes of population. Despite this, the two models may be stored identically in the GIS and the user may be unaware that a potentially meaningless operation is being performed.

Data models provide a logical and useful way of organizing the functionality of a GIS. They may also provide a framework for discussion of methods of SDA, since these are in principle extensions of basic GIS functionality. However, many methods of SDA treat the issue of data modelling as a matter of implementation, rather than as an intrinsic property of the method of analysis. For example, suppose that a hydrological analysis requires the determination of ground slope as an important factor in soil erosion. Slope is a well-defined property of any continuously differentiable mathematical surface. But it is not well-defined everywhere on the real landscape, which is characterized by frequent breaks of slope, and it is not defined independently of the data model in any discrete representation of the land surface. In the TIN model, for example, it is well-defined and constant within triangles but indeterminate on their edges, and curvature is everywhere zero or indeterminate. In a contour model of the same real surface, slope must be inferred by some additional, as yet undefined process of interpolation.

In summary, an additional problem facing any effort to integrate GIS with SDA is that data modelling must be explicit in any use of GIS, but is often left undefined in SDA. Use of GIS forces the analyst or modeller to confront the issue of discretization directly.

## 3. Techniques of spatial data analysis for GIS

There are two principal classes of statistical SDA techniques that are of potential importance to GIS, those that use locational data only and those that use a combination of locational and attribute data. The first set of methods is concerned with the analysis of spatial distributions. These methods include methods of point pattern analysis for object-based data sets where the objects are points or areas that are represented as points. The analysis sometimes concerns the distribution of the points in a space which is assumed to be continuous and homogeneous (as in nearest neighbour analysis). In other cases the point distribution may be analysed with respect to another set of points, or lines, or an inhomogeneous surface. For example, the point incidence of infected individuals (associated with the outbreak of a disease) might be analysed with respect to the point distribution of the total susceptible population, to indicate whether the occurrence is clustered or not.

The second set of methods analyses the spatial variation in attribute values. These can be subdivided into methods that are appropriate for spatially-continuous or at least quasi-continuous data (field data sets) and methods that are appropriate for object-based data sets where the objects may either be points or areas (that may or may not be contiguous). All these methods can be further classified in terms of the level of measurement associated with the recorded attribute properties (i.e., nominal, ordinal, interval or ratio). In the case of object data sets, attribute values associated with areas often refer to the collection of point events that fall within the boundaries of each area (e.g., census data where attribute values refer to properties of the set of households or individuals falling within each enumeration district). In the case of field data sets the attribute values may refer to sample points on the surface or to the component areas of some partition of that surface. In the latter case attribute values may refer to aggregate properties of each area or average levels of some variable, for example.

Within each of these categories it is useful to distinguish a second level of classification; that between descriptive methods which are used to summarize data properties and confirmatory methods which are associated with the process of data modelling and concerned with systematic analysis of data and hypothesis testing based on specified assumptions.

While the importance of this secondary classification is recognized, implicit within it is a distinction between methods of statistical data analysis, that would be of value to the GIS community as it currently defines its needs and are largely of a descriptive type, and those techniques, largely confirmatory, that would sit well within the structure of GIS (in terms of how data are handled and visualized) and so, if integrated into GIS, would contribute to their evolution.

### 3.1. *Spatial analysis methods and current needs*

There are four areas where statistical developments might strengthen current GIS practice: data rectification, data assessment, data sampling and initial data exploration.

A common problem in GIS occurs when the user wishes to compare variables which are defined for the same study area but for a different and incompatible set of zones (such as postcodes, enumeration districts, wards). Assigning sensible attribute values to the intersection zones can be done, for categorical and intensive data (e.g., proportion or ratio data), by simply allowing such zones to inherit the values from the original zones. This is not sensible for extensive data (totals). For these data the pycophylactic method (Tobler 1979) or the areal weighting method (Goodchild and Lam 1980) could be used. Flowerdew and Green (1990) are developing a method based on the *E-M* algorithm where the values for the subzones are regarded as missing data and estimated using information from ancillary variables.

The *E-M* algorithm is an iterative procedure consisting of an *E*-step, which computes the conditional expectation of the missing data given the model and the observed data, and an *M*-step, which fits the model by maximum likelihood to the complete data set, treating the values from the *E*-step as real observations. If, for example, $y_{s,t}$ is the (unknown) value for variable $Y$ for subzone $(s, t)$ which is the intersection of zones $s$ and $t$ from the two partitions, then the problem is formulated as one of inferring the values for the variable $Y$ for the subzones $(y_{s,t})$ from the original data for variable $Y$ and any ancillary information and then totalling the estimates of $y_{st}$ to give interpolated estimates of $y_t$. For example, suppose $y_{st}$ is Poisson distributed (so $Y$ is a count variable) with parameter $\mu_{st}$. Let

$$\mu_{st} = \mu(\boldsymbol{\beta}, \mathbf{x}_t, A_{st}) \tag{1}$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters, $\mathbf{x}_t$ a vector of ancillary information on the target zone $t$ and $A_{st}$ is the area of the intersection zone $(s, t)$. At the *E*-step:

$$\hat{y}_{st} = E(Y_{st}|\hat{\mu}_{st}, y_s) = \hat{\mu}_{st}\, y_s \Big/ \sum_k \hat{\mu}_{sk} \tag{2}$$

where $\hat{}$ denotes an estimated value. At the *M*-step a maximum likelihood estimate of $\boldsymbol{\beta}$ is obtained using $\{\hat{y}_{st}\}$ as independent Poisson data.

Results with the method have been encouraging. The method has been implemented in ARC/INFO at Lancaster University by building a link to a statistical package (GLIM) which performs the analysis.

In some spatial data sets, values on one or more variables may be missing for some zones or sites (Griffith *et al.* 1989). A number of relatively simple, cartographical methods is available for estimating missing values which depend largely on geometrical relationships between sites. For example, methods of distance weighting estimate missing values as weighted averages of neighbouring observed values.

Recently, however, there has been interest in the application of methods based on the statistical attributes of univariate data sets to provide maximum likelihood estimates of these missing values (Martin 1984). Such methods usually depend on careful modelling of the mean and covariance properties of the data. If $y_m$ is a missing value, Martin (1984) shows that the maximum likelihood estimator of $y_m$ in the case of correlated normal data is given by

$$\mathbf{A}_m\hat{\boldsymbol{\theta}}+\hat{\mathbf{V}}_{m0}(\hat{\mathbf{V}}_{00})^{-1}(\mathbf{y}_0-\mathbf{A}_0\hat{\boldsymbol{\theta}}) \tag{3}$$

where $\mathbf{A}_m$ is a row vector of location co-ordinates for the missing value, $\mathbf{y}_0$ the vector of observed values, $\sigma^2\hat{\mathbf{V}}_{ij}$ is the estimated covariance matrix for the subvectors $\mathbf{y}_i$ and $\mathbf{y}_j$, and $\hat{\boldsymbol{\theta}}$ is a vector of estimated parameters associated with the mean of the surface.

In overviewing statistical developments in this area Martin (Haining and Wise 1991) commented: 'the case for using efficiently all possible information appears clear, but in fact the use of more complicated and computer intensive methods that may require human intervention may need to be closely argued from a detailed knowledge of typical variation in the application of interest. It is very unlikely that any automated method would perform well for a wide variety of applications'. The case for using predictors based on averages and other simple statistical methods remains a strong one, particularly for data on irregular areal partitions. Krug and Martin (1990) review methods while Bennett *et al.* (1984) provide an earlier review in a more specifically geographical context.

Currently much attention is being paid to the issue of error in spatial databases. (Goodchild and Gopal 1989). Error analysis is concerned with the study of the nature of error (its sources, properties, methods for detection and measurement) and of error effects, including the way errors may propagate through different types of operations (such as overlay). In a spatial database errors arise in measuring both location and attribute properties (source errors) and are also associated with those computerised processes responsible for storing, retrieving and manipulating spatial data. It is important to be as specific as possible about the nature of error and how it might vary across a spatial database. Griffith and Arbia noted at least three areas where current statistical methods may be of use to the GIS community: methods for characterizing sampling error and its spatial variation; methods for the detection of outliers in databases (including 'spatial' outliers which are outliers that are extreme with respect to the local spatial distribution of values, e.g., Cressie and Read 1989, Cressie 1984); and methods for estimating the effects of error propagation (e.g., Heuvelink *et al.* 1989, Haining *et al.* 1992).

GIS databases are typically very large. Methods of data sampling may have an important role to play in the extraction of information from such databases. Arbia described a sequential GIS-based procedure (called DUST—Dependent Areal Units Sequential Technique) which can be used to identify areas to be included in a sample so as to maximize the information content of the sample. It is possible to economize on the number of spatial units sampled by exploiting the dependency structure in the set of data. Spatial dependency implies a certain degree of redundancy in the additional information that is provided by the near neighbours of an already sampled area; the essential idea is thus to ensure that sample points are chosen that are separated to a degree that depends directly on the strength of spatial dependency in the data. The first sample area is drawn randomly and then DUST proceeds by assigning to each area $j$ a probability of selection proportional to $(1-\beta_{1j})$ where $\beta_{1j}$ is a measure of the spatial correlation between the first sample point and the $j^{\text{th}}$ area. At the third step each area is assigned a probability of being drawn which is proportional to $(1-\beta_{1j})(1-\beta_{2j})$. There has been preliminary analysis of the method involving sampling within an ARC/INFO database (Arbia 1991).

There was broad agreement at the workshop that general data exploratory methods would be of great value within GIS, especially in those situations where the data are of poor quality and there is a lack of genuine prior hypotheses. As Openshaw observed, GIS users will 'want to explore data without knowing either "where" to look or "what" to look for or "when" to look'. To some, including Openshaw and Brunsdon, this calls for the development of methods of spatial analysis that are special and different, customized for the applications environment within which most GIS users find themselves working. 'There is a need to develop new tools from basic principles rather than adapting already available methods into a spatial framework' (Brunsdon). To others, including Haining, there is much to be learned from using methods of exploratory data analysis developed within mainstream statistics for handling data in the 'uncomfortable sciences' Haining (1990, p. 4) where theory development is limited, controlled experimentation unavailable and data of uncertain quality. Exploratory methods have also been developed by statisticians specifically for describing spatial data where the exploratory questions include the search for data characteristics such as trends, patterns and spatial outliers. Methods for field data include those developed by Cressie (1984) and for object data, nearest neighbour and K-function plots (Ripley 1978). Methods of point pattern analysis which allow for spatial heterogeneity in the underlying susceptible population are of particular interest in the context of testing for disease clusters (Diggle *et al.* 1990). Where there is common ground between the two groups, however, is on the need for 'robust' methods that allow GIS users to simplify, spot patterns and identify relationships in the data (table 1).

An underlying problem in the description of spatial data is the sensitivity of findings to the areal framework or discretization. As noted above, working in a GIS environment brings this issue firmly into the foreground. In the case of object data this includes the size of the study area and the ability to compare results for different spatial subsets of the data areas; for field data and some object data sets that have been aggregated into areal units it also includes the areal partition, that is, the scale and configuration of the spatial units into which the field has been (artificially) discretised.

Table 1. Core techniques for exploratory analysis of GIS data.

| Brunsdon | Haining | Openshaw |
|---|---|---|
| *Point data* | 'Standard' statistical exploratory methods (including Box plots, stem and leaf, rankits, etc.) | Pattern spotters and testers |
| Kernal estimation | | Relationship seekers |
| Kernel regression | | Data simplifiers |
| Semi-variogram 'cloud' plots | | Edge detectors |
| Non-continuous cartogram | Identifying trends (filter mapping, polishing, trend surface, graph plots against distance etc.) | auto spatial response modellers |
| | | fuzzy pattern analysis |
| *Field data* | | visualisation enhancers |
| Smoothing of choropleth maps to surfaces | | spatial video analysis |
| Choropleth mapping of smoothed data | Identifying pattern (autocorrelation tests, correlation functions graphical plots) | |
| Continuous area cartogram projections | | |
| Plots of correspondence area values and neighbours | Outlier and spatial outlier detection methods | |
| | Plots relating area values to area attributes (size, shape etc.) | |

For areal data the ability to merge spatial units is already available within GIS, but Diggle drew attention to the usefulness of a 'dragging windows' utility for exploratory analysis, particularly in large object-based spatial data sets (Haslett *et al.* 1990, 1991). Such a utility would allow the GIS user to examine spatial subsets through a window that, as it was moved (by the user) across the screen, would give an immediate response on the properties of the data within the window. Diggle cited in particular his work with others at Lancaster University, on the detection of spatial clustering in cancer cases, but such a utility was felt to be of general interest. Diggle suggested that this procedure might be extended to several windows simultaneously, each containing relevant statistical descriptions, and to a hierarchy of user-defined sub-regions that nested within one another and across which the user could move to obtain statistical descriptions. In addition to assessing sensitivity to the areal framework it would be useful to be able to assess the sensitivity of results to individual cases or sets of cases. This requires the facility to delete cases and recompute statistics. It will be through facilities such as this, together with computational methods displayed in the form of graphical and cartographical output, that effective exploration of spatial data can be undertaken.

### 3.2. *Spatial analysis, GIS and future developments*

Spatial data analysis usually goes beyond data sampling, data assessment and data description into areas of modelling. The arguments in this section are based on the view that GIS (as a software system) is an appropriate vehicle for spatial data modelling. In particular, a GIS may represent a natural focus around which specialist software modules might be built that would allow GIS to be used for the full range of SDA exploratory and explanatory procedures, thereby widening the potential utility of GIS. (How such integration might take place and the forms it might take is the subject of the next section).

An exploratory analysis of object data relating to cancer cases might focus on describing the pattern of cases. It might focus on testing a specific hypothesis about that pattern, such as the tendency for levels of incidence to be higher near some point source. In this case a model is fitted in which intensity of occurrence at location **x** is defined by a spatially-varying parameter $g(\mathbf{x})$. Let

$$g(\mathbf{x}) = \lambda_0 g_0(\mathbf{x}) f(\mathbf{x}; \boldsymbol{\theta}) \tag{4}$$

where $\lambda_0$ is a measure of overall intensity, $g_0(\mathbf{x})$, represents background variation and $f(\mathbf{x}-\mathbf{x}_h; \boldsymbol{\theta})$ models the raised incidence near a potential hazard as a function of distance from the location of the hazard $(\mathbf{x}_h)$ and a vector of parameters. With suitably chosen forms for $f(.)$, a non-significant set of values for $\boldsymbol{\theta}$ implies that there is no raised incidence near the potential hazard (Diggle *et al.* 1990).

An area of spatial data analysis that brings some of the issues associated with data modelling into sharp focus is that of spatial econometrics which generally deals with object data (Anselin 1988, Hepple 1979). As Hepple noted, spatial econometrics is an offshoot of econometrics bringing to the analysis of spatial data approaches and methods developed in econometrics for time series data. Like econometrics, spatial econometrics is strongly model-driven, that is, data are used to test between alternative theoretical explanations or to identify the data generating process (Gilbert 1986). Modelling in this area requires that a number of diagnostic tests, model specification and model selection procedures be available to help in choosing between models which may include spatial effects in the form of spatially-lagged variables or spatially-correlated errors. Hepple and Griffith both pointed to the need for good numerical routines for undertaking these estimation and test procedures, including the ability to provide alternative representations of the areal system (the 'W' or connectivity matrix which is used to represent spatial relationships between areas (Haining 1990)) and to provide informative visual output. There is a need for good customized packages (as time series econometricians have available through, for example, SHAZAM and GIVE). In view of the special needs of spatial analysis there are great benefits in creating links between a good regression package and a GIS. If this application field of GIS were to be developed, appropriate computational software to fit the spatial econometric models would also need to be developed. Such software should include procedures for non-linear optimization and the computation of eigenvalues of large asymmetric 'W' matrices (these being required for maximum likelihood estimation of the parameters of some spatial models) and procedures for numerical integration (these being required for Bayesian model selection). Anselin (1990) is in the process of developing a spatial econometric package (SPACESTAT) which is a software package for spatial data analysis which includes both exploratory analysis and regression analysis with a range of diagnostics. The software is written and compiled in GAUSS, and a test version and preliminary manual are available. Griffith has produced a similar set of macros in MINITAB which were demonstrated at the workshop.

The view was held that it would be of considerable benefit to the GIS community if a good regression package able to fit *spatial* regression models were to be coupled with a working GIS.

## 4. Coupling SDA and GIS: Practical issues

How might the integration of GIS and statistical SDA best be achieved? In general terms there are two approaches: to incorporate spatial analysis tools into GIS software; and to modify statistical software so that it can be used to work with spatial data. Discussion here will concentrate on the first approach, although many of the remarks are probably of equal validity for the second.

Adding SDA tools to GIS software raises two problems: how to introduce spatial analysis tools into GIS software, and how to introduce spatial analytic GIS into the GIS marketplace. The first is a technical computing question, and although it is not a trivial one, it should be possible to find solutions; several examples were given at the workshop. The second point highlights the fact that the GIS market is driven by the needs of data management rather than by those of data analysis. This means that the current demand for analytical tools in the GIS market is confined to academic and scientific applications. For the commercial sector this raises the issue that, unless demand grows, there will be little point in investing resources in solving technical problems which relate to facilities for which demand may not permit development costs to be covered and profits generated. Perhaps the best way to generate demand for better spatial tools in GIS software, or for better spatial data handling in statistical packages, is to develop prototype software systems which could demonstrate the potential applications of this increased functionality. The best means of developing prototypes brings us back to the first, technical, question.

Methods of providing spatial analysis tools within GIS can be classified as follows:

(*a*) Stand-alone spatial analysis software.
(*b*) Loose coupling of existing GIS software with statistical software (either standard packages or purpose written).

(*c*) Close coupling of GIS software with spatial statistical software.
(*d*) Full integration of spatial analysis with GIS.

These will each be discussed in turn.

### 4.1. *Stand-alone spatial analysis software*

In general, only a commercial company would have the resources to produce a comprehensive spatial analysis package. However, there are some situations where writing software from scratch may be a sensible option, for example, to perform one specialised piece of analysis, which is not easily related to anything offered by a current GIS. A good example of this approach is the Geographical Analysis Machine (GAM) developed by Openshaw *et al.* (1987).

Here there were at least two good reasons for writing such a piece of software from scratch. First, the GAM represents a completely different approach to the analysis of spatial data. It uses computer power to repeat the same calculation many times for different positions in space. In contrast, most current GIS software is geared towards providing interactive querying and display facilities with the possible intention of forming hypotheses about the data for later statistical testing. Second, the GAM needs rapid and repeated access to the geographical database in order to perform the statistical calculations repeatedly as the search mesh is moved across the geographical space. The efficiency of the database routines was therefore critical, and required the implementation of a special data structure, the KDB tree, in order to minimise search times (Openshaw *et al.* 1987). Commercial GIS software, on the other hand, has to be written with flexibility in mind, since it will need to deal with many different types of query. For any one query maximum efficiency may not be achieved.

In general, however, writing separate pieces of software for spatial analysis is not a good strategy because any spatial analysis package will need facilities for data input, data editing, database management and data display. These are the very areas where GIS is strong, and it seems foolish not to take advantage of this. The question then becomes one of how best to do this.

### 4.2. *Loose coupling*

This approach means using a GIS package for the things for which it is best suited and, when other facilities are required, taking the output from the GIS system as input into other software. There are two possible approaches here:

1. Loose coupling by the interchange of ASCII files between packages.
2. Loose coupling by packages using a common binary file format.

#### 4.2.1. *Loose coupling via ASCII files*

Here data are passed from the GIS into another package by editing the ASCII output file produced by the GIS. This is the approach described by Gatrell (1987) in which output from ODYSSEY is passed to GLIM.

This approach has been taken a step further by Waugh (1986) in designing the Geolink software. Geolink is an attempt to provide a tool which will allow the easy and flexible linking of different software packages by providing a programmable 'editor' for converting ASCII files from one format to another using a simple sequence of commands. The system also contains facilities for menu handling and has been used successfully at Edinburgh to produce complex end-user systems from a combination of standard software packages.

The great strength of this approach is the logic of tackling each task with a package which is best suited to that task. There is also no need to write extra software (apart from the writing of Geolink itself). One drawback is its dependence on the format of the output from the package remaining constant over different versions of the software. It may also require the use of several large pieces of software at the same time (e.g., ARC/INFO and ORACLE in the Edinburgh work) which has implications for memory requirements.

### 4.2.2. *Loose coupling via binary files*

The alternative approach is to ensure that different packages can all read and write data files in an agreed format. This is an approach which has become very popular in the world of commercial software for the IBM PC with the growth of packages which can work with the data formats used by industry standard software (e.g., Lotus 1-2-3 .WKS files and dBase .DBF files) and with generally-supported file transfer formats (e.g., .DIF files).

The IDRISI package, developed by Clark University, is the closest analogy in the GIS marketplace in that this is not a package at all but a set of separate modules which share a common data structure (Eastman 1990). This approach has been taken quite deliberately to allow other users to develop their own IDRISI modules. It works well because the raster data structure is inherently very simple. At the workshop Bossard proposed a 'user-friendly GIS' made up of spreadsheet, database, statistical and GIS packages, all linked together via the transfer of data in common file formats. Tobler (1990) has described how several types of SDA could be performed using spreadsheets.

Although this approach is restricted to a single hardware platform, the current convergence of the GIS market on IBM PCs and Unix workstations (often capable of reading PC disks) makes this less limiting. What is required here is for some standard file format either to be agreed between vendors, or to emerge by vendors exploiting the success of a market leader.

### 4.3. *Close coupling*

Close coupling covers a variety of methods which all involve modifying the operations of the GIS software itself in some way. A variety of mechanisms is already offered by many packages including: macro languages which, at the simplest level, allow complex sequences of commands to be packaged up into a single macro while more complex ones offer true programming language capabilities; 'hooks' to user routines written in standard languages such as FORTRAN or C; libraries of user-callable routines to access the low-level data structures within the GIS. These are potentially very powerful because they offer the developer access both to the standard user-interface facilities of the package (for example, macros can be written so that they appear to the user simply as extra commands), and to the underlying data structures used for handling the locational and attribute data.

There is some doubt as to whether the facilities currently offered are sufficiently powerful to provide the facilities needed for spatial analysis. A case in point is the '**W**' connectivity matrix based on the length of the common boundary between adjacent areas. Many vector systems store information about area boundaries explicitly—the question is whether the information is available to a user-written routine which can then construct the '**W**' matrix. (See Yeumin Ding and Fotheringham (1991) for constructing '**W**' matrices in ARC/INFO).

In general, members of the workshop felt that the need was for good methods of incorporating user-written routines into the GIS. This would require first, well-defined mechanisms for allowing user-written routines to be called from within the normal user interface of the packages, and second, well-defined interfaces to the data structures held by the GIS. In the case of the attribute data this may well already exist in the form of an SQL interface to a relational database. A challenge for the future is the development of an equivalent standard query language for spatial data, to allow users to access such data without needing to know about the particular data structures being used within the database. The range of spatial data structures listed earlier indicates that this is not a small task.

### 4.4. *Full integration*

The final approach which might be taken is to integrate techniques of spatial analysis fully within GIS software. This would have a number of advantages: software would be documented and supported by the vendors; the techniques would be available to all users of the package, and not just to those who happened to have written their own routines to interface to it. However, despite these undoubted advantages there are problems in following this route. It might result in a few techniques becoming ossified within GIS systems and might require vendors to make major changes which would not be necessary for other users. As commented earlier, it would be difficult to persuade vendors to adopt such an approach without pressure from the market for such facilities.

## 5. Conclusions and directions for future research

Any spatial analysis software will require facilities for the input, management and display of spatially-referenced data. All these facilities are currently available within existing GIS packages. This is an important element in the case for using GIS packages as a vehicle for SDA (e.g., Haining 1990). GIS, if they are to develop as a general-purpose tool for handling spatial data, need to incorporate functionality which goes beyond data management to data analysis. This is an important element in the case for putting SDA techniques into GIS (e.g., Openshaw 1990).

Both views for linking SDA and GIS are based on the potential GIS has for SDA (Goodchild 1987), but the distinction that is drawn above reflects assumptions about the types of users, the type and quality of available data, the range of suitable methods (including appropriate inference frameworks) and the objectives of analysis. The common ground lies in the existence of exploratory SDA techniques that are **both** a useful range of easy-to-apply techniques suitable for the non-specialist **and** an important first step in developing an SDA library linked to GIS for the user who is a specialist in data analysis. Simple exploratory tools can yield non-intuitive insights into data that are of importance in their own right and as important adjuncts to data modelling. Methods that facilitate the detection of trends, patterns and outliers as well as more general variable relationships (see, for example, Openshaw 1990, and Haining 1990, Chapter 6) fall within this category. The next step is to prioritize the sorts of techniques that should be made available.

Irrespective of the specific exploratory techniques that are available within GIS, the workshop identified two broad types of spatial operations that should be made available. The importance of these operations is that any set of prioritized techniques is likely to draw on them.

Operation 1. The ability to examine in an interactive environment user-defined sub-regions of the full data set. These sub-regions may be discrete, overlapping or hierarchically arranged (Haslett *et al.* 1990, 1991). It should be possible to obtain descriptive statistics for such user-defined sub-regions for comparative purposes.

Operation 2. The ability to construct alternative connectivity or '**W**' matrices based on geometrical properties of the system (such as inter point distance, shared common boundaries of the area partition) or based on interaction data. The need to test hypotheses in terms of alternative 'orderings' of the spatial system or to examine the sensitivity of findings (model fits) to different 'orderings' is a quite basic issue in all areas of spatial analysis.

Over the longer term there is a need to incorporate both exploratory and confirmatory methods of SDA into GIS. The view was held in the workshop that because exploratory/descriptive methods would be of general use to specialist and non-specialist alike (in all areas of spatial data analysis) then these methods should be available in software packages that are closely coupled to a GIS or integrated into GIS packages themselves. On the other hand because explanatory/confirmatory methods are of interest only to the specialist interested in data analysis *per se* (rather than data management), they should be available in software packages that are loosely coupled to a GIS. Such marriages would be of benefit to both the GIS and SDA communities.

## Workshop convener

R. Haining, Department of Geography, University of Sheffield, Sheffield, England, U.K.

## List of workshop participants

G. Arbia, Department of National Accountancy, University of Rome.

L. Anselin, Department of Geography, University of California, Santa Barbara, U.S.A. (Discussion document tabled).

E. Bossard, Department of Urban and Regional Planning, San Jose State University, U.S.A.

C. Brunsdon, Centre for Urban and Regional Development Studies, University of Newcastle-upon-Tyne, Newcastle upon Tyne, England U.K.

P. Diggle, Department of Mathematics, Lancaster University, Lancaster, England, U.K.

R. Flowerdew, North West Regional Research Laboratory, and Department of Geography, Lancaster University, Lancaster, England, U.K.

M. Goodchild, Department of Geography, University of California, Santa Barbara, U.S.A.

M. Green, Centre of Applied Statistics, Lancaster University, Lancaster, England, U.K.

D. Griffith, Department of Geography, Syracuse University, Syracuse, New York, U.S.A.

R. Haining, Department of Geography, University of Sheffield, Sheffield, England, U.K.

L. Hepple, Department of Geography, University of Bristol, Bristol, England, U.K.

T. Krug, Department of Probability and Statistics, University of Sheffield, Sheffield, England, U.K.

R. Martin, Department of Probability and Statistics, University of Sheffield, Sheffield, England, U.K.

S. Openshaw, North East Regional Research Laboratory and Department of Geography, University of Newcastle upon Tyne, Newcastle upon Tyne, England, U.K.

S. Wise, Department of Geography, University of Sheffield, Sheffield, England, U.K.

## References

Anselin, L., 1988, *Spatial Econometrics: Methods and Models* (Dordrecht: Kluwer).

Arbia, G., 1991, GIS—based sampling design procedures. In *Proceedings EGIS '91 Second European Conference on Geographical Information Systems, Brussels, 2–5 April 1991*, **1**, edited by J. Harts, H. Offens and H. Scholten (Utrecht: EGIS Foundation), pp. 27–35.

Baxter, R. S., 1976, *Computer and Statistical Techniques for Planners* (London: Methuen).

Bennett, R. J., Haining, R. P., and Griffith, D. A., 1984, The problem of missing data on spatial surfaces. *Annals of the Association of American Geographers*, **74**, 138–56.

Berry, B. J. L., and Marble, D. F., 1968, *Spatial Analysis: A Reader in Statistical Geography* (Englewood Cliffs, NJ: Prentice-Hall).

Cressie, N., 1984, Towards resistant geostatistics. In *Geostatistics for Natural Resources Characterization*, edited by G. Verly, M. David, A. G. Journel and A. Marechal (Dordrecht: Reidel), pp. 21–44.

Cressie, N., and Read, T. R. C., 1989, Spatial data analysis of regional counts. *Biometrical Journal*, **6**, 699–719.

Department of the Environment, 1987, Handling Geographic Information: Report to the Secretary of State for the Environment of the Committee of Enquiry into the Handling of Geographic Information (London: HMSO).

Diggle, P. J., Gatrell, A. C., and Lovett, A. A., 1990, Modelling the prevalence of cancer of the larynx in part of Lancashire: a new methodology for spatial epidemiology. In *Spatial Epidemiology*, edited by R. W. Thomas (London: Pion), pp. 35–47.

Eastman, J. R., 1990, *IDRISI: A Grid Based Geographic Analysis System* (Worcester, Mass: Clark University).

Flowerdew, R., and Green, M., 1990, Inference Between Incompatible Zonal Systems using the EM Algorithm. North West Regional Research Laboratory, Research Report No. 6 (Lancaster: NWRRL).

Gatrell, A. C., 1983, *Distance and Space: A Geographical Perspective* (New York: Oxford University Press).

Gatrell, A. C., 1987, On Putting Some Statistical Analysis into Geographical Information Systems: with Special Reference to Problems of Map Comparison and Map Overlay. Research Report number 5 (University of Lancaster: Northern Regional Research Laboratory).

Getis, A., and Boots, B., 1978, *Models of Spatial Processes* (Cambridge: Cambridge University Press).

Gilbert, E. W., 1958, Pioneering maps of health and disease in England. *Geographical Journal*, **124**, 172–183.

Gilbert, C. L., 1986, Professor Hendry's econometric methodology. *Oxford Bulletin of Economics and Statistics*, **48**, 283–307.

Goodchild, M. F., 1987, Towards an enumeration and classification of GIS functions. *Proceedings, IGIS 87: the Research Agenda*, edited by R. T. Aangeenbrug and Y. M. Schiffman (Washington, DC: NASA), **II**, 67–77.

Goodchild, M. F., 1992, Geographical data modelling. *Computers and Geosciences* (in press).

Goodchild, M. F., and Gopal, S., 1989, *Accuracy of Spatial Databases* (London: Taylor & Francis).

Goodchild, M. F., and Lam, N. S.-N., 1980, Areal interpolation: a variant of the traditional spatial problem. *Geo-Processing*, **1**, 297–312.

Griffith, D. A., 1988, Estimating spatial autoregressive model parameters with commercial statistical packages. *Geographical Analysis*, **20**, 176–186.

Griffith, D. A., Bennett, R. J., and Haining, R. P., 1989, Statistical analysis of spatial data in the presence of missing observations: a methodological guide and an application to urban census data. *Environment and Planning A*, **21**, 1511–23.

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

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

Haining, R. P., and Arbia, G., 1992, Error propagation through map operations, *Technometrics*, (in press).

HAINING, R. P., and WISE, S. M., 1991, GIS and Spatial Data Analysis: Report on the Sheffield Workshop. Regional Research Laboratory Initiative Discussion Paper No. 11 (Sheffield: Department of Town and Regional Planning, University of Sheffield).

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

HASLETT, J., BRADLEY, R., CRAIG, P. S., WILLS, G., and UNWIN, A. R., 1991, Dynamic graphics for exploring spatial data, with application to locating global and local anomalies. *American Statistician*, **45**, 234–242.

HEPPLE, L., 1979, Bayesian analysis of the linear model with spatial dependence. In *Exploratory and Explanatory Statistical Analysis of Spatial Data*, edited by C. P. A. Bartels and R. H. Ketellapper (Leiden: Nijhoff), pp. 179–99.

HEUVELINK, G. B. M., BURROUGH, P. A., and STEIN, A., 1989, Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems*, **3**, 303–322.

HEYWOOD, I., 1990, Geographic information systems in the social sciences. *Environment and Planning A*, **22**, 849–854.

KRUG, T., and MARTIN, R. J., 1990, Efficient Methods for Coping with Missing Information in Remotely Sensed Data. Research Report no. 371/90, Department of Probability and Statistics, University of Sheffield.

MACDOUGALL, E. B., 1976, *Computer Programming for Spatial Problems* (New York: Wiley).

MARTIN, R. J., 1984, Exact maximum likelihood for incomplete data from a correlated Gaussian process. *Communications in Statistics: Theory and Methods*, **13**, 1275–88.

ODLAND, J., 1988, *Spatial Autocorrelation* (Newbury Park, CA: Sage).

OPENSHAW, S., 1987, Guest editorial: an automated geographical analysis system. *Environment and Planning A*, **19**, 431–436.

OPENSHAW, S., 1990, *A Spatial Analysis Research Strategy for the Regional Research Laboratory Initiative*. RRL Discussion Paper 3 (Sheffield: Department of Town and Regional Planning, University of Sheffield).

OPENSHAW, S., CHARLTON, M., WYMER, C., and CRAFT, A. W., 1987, A Mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, **1**, 335–358.

PEUQUET, D. J., 1984, A conceptual framework and comparison of spatial data models. *Cartographica*, **21**, 66–113.

RIPLEY, B., 1978, The analysis of geographical maps. In *Exploratory and Explanatory Statistical Analysis of Spatial Data*, edited by C. P. A. Bartels and R. H. Ketellapper (Leiden: Nijhoff), pp. 53–73.

RIPLEY, B. D., 1981, *Spatial Statistics* (New York: Wiley).

STAR, J. L., ESTES, J. E., 1990, *Geographic Information Systems: An Introduction* (Englewood Cliffs, NJ: Prentice-Hall).

TOBLER, W. R., 1979, Smooth pycophylactic interpolation for geographical regions. *Journal, American Statistical Association*, **74**, 251–7.

TOBLER, W. R., 1990, *The Resel Based GIS* (Unpublished MS), Department of Geography, University of California at Santa Barbara.

TSICHRITZIS, T. C., and LOCHOVSKY, F. H., 1977, *Data Base Management Systems* (New York: Academic Press).

UNWIN, D. J., 1981, *Introductory Spatial Analysis* (London: Methuen).

UPTON, G. J. G., and FINGLETON, B., 1985, *Spatial Data Analysis by Example* (New York: Wiley).

WAUGH, T. C., 1986, The Geolink system, interfacing large systems. In *Proceedings of Auto Carto London*, edited by M. J. Blakemore (London: Royal Institute of Chartered Surveyors), **1**, 76–85.

WORRALL, L., 1990, *Geographic Information Systems: Developments and Applications* (London: Belhaven Press).

YEUMIN DING, and FOTHERINGHAM, A. S., 1991, *The Integration of Spatial Analysis and GIS* (Unpublished MS) National Center for Geographic Information and Analysis, University of New York at Buffalo.