

# DEVELOPING APPROPRIATE SPATIAL ANALYSIS METHODS FOR GIS

S OPENSHAW

*The geographical information revolution demands a new style of spatial analysis that is GIS appropriate and GIS proof. The existing spatial analytical toolbox is largely inadequate, consequently there is an urgent need to create more relevant methods and also to educate users not to expect the impossible when analysing geographical data. The real challenge is the need to develop new, largely automated, spatial data exploratory techniques that can cope with the nature of both the geographical data created by GIS and the skill base of typical GIS users. It also needs to emphasize the creative, hypothesis generating, and artistic aspects of geographical analysis, avoid being too dependent on a blinkered and inadequate inferential statistical mentality, and recognize the limitations of working within a geographical domain. A number of useful and applicable techniques are described and optimism is expressed about the opportunities that abound for making good use of spatial analysis within GIS environments.*

---

## ON THE NATURE OF SPATIAL ANALYSIS

---

A good GIS will today probably contain over 1000 commands (or their equivalent) but few, typically none, will be concerned with what might correctly be termed spatial analysis rather than data manipulation. This distinction is critical since spatial data handling procedures such as buffering, overlay, and query are not 'real' analysis operations except in a data descriptive or cartographic sense. It is useful, therefore, to define first what is meant by spatial analysis and then briefly outline the available technology.

The origins of spatial analysis lie in the development of quantitative and statistical geography in the 1950s. Spatial analysis was originally based on the application of the available statistical methods to spatial data (Berry and Marble 1968). Later, it was extended to include mathematical model building and operational research methods (Taylor 1977; Wilson and Bennett 1985). Hagerstrand (1973:69) provided an adequate

definition of spatial analysis when he wrote 'to no small degree the recent quantitative analysis in geography represents a study in depth of the patterns of points, lines, areas, and surfaces depicted on maps of some sort or defined by co-ordinates in two- or three-dimensional space'. Most other definitions are similar, for example, Johnston, Gregory and Smith (1986: 446) define spatial analysis as 'quantitative (mainly statistical) procedures and techniques applied in locational analytic work'. Unwin (1981) presents spatial analysis as concerned with the arrangements on maps of four types of data portrayed there: points, lines, areas, and surfaces. The techniques allow both description of the arrangements on individual maps and the comparison of two or more maps so that relationships might be identified. A variety of statistical and geographical analysis procedures have been developed to serve these objectives (Goodchild 1988).

Clearly, spatial analysis is extremely relevant to GIS and the gradual absorption of spatial analysis

tools into GIS systems is inevitable. Spatial analysis offers a toolbox that can in principle be applied to all the standard types of geographical information and be performed in one-dimensional space, more commonly in two-, occasionally in three-, and rarely in four-dimensional space. It is important as a means of increasing the functionality of GIS by providing a link between the essentially cartographic domain in which the origins of GIS lie and key areas of applied quantitative, statistical and mathematical analysis, and modelling, of interest to many users of GIS. However, in seeking to meet these objectives the recommended technology has to be capable of coping with the peculiarly complex nature of the spatial data. Several years ago Ripley (1984) talked about the need for a revolution; this need still exists.

This chapter is a critique of existing spatial analysis techniques and their potential for use in GIS. The considerable differences between methods appropriate for environmental and socio-economic applications mean that, apart from some general remarks, it is not possible to consider them both here. Instead, the discussion concentrates on socio-economic applications. Burrough (1991 in this volume) and Bonham-Carter (1991 in this volume) make some relevant comments about environmental applications and useful reviews include Davis (1986), Nielson and Bouma (1985) and Oliver, Webster and Gerrard (1989a, 1989b).

---

### **A REVIEW OF THE EXISTING SPATIAL ANALYSIS TOOLBOX**

---

The newcomer to spatial analysis may well require a standard textbook from which to work. This is especially important because of the current lack of spatial analysis procedures in GIS systems and also in statistical packages. A good survey of a wide range of statistical procedures is provided in Upton and Fingleton (1985). Diggle (1983) and Ripley (1981) also provide a useful digest. Simpler introductions are given in Taylor (1977), Unwin (1981) and Wilson (1974) considers modelling applications. In addition, a plethora of quantitative geography and statistical geography textbooks outline most of the standard methods. Unfortunately, there is as yet no globally useful text designed to inform the GIS user specifically about

the complete range of spatial analysis methods that might be considered relevant and appropriate to GIS.

It is useful, therefore, to provide a brief summary of the range of available spatial analysis tools by identifying classes of methods appropriate for different geographical data types (Table 25.1). Note that the four basic geographical data types shown in Table 25.1 can often be mapped on to each other. For example, point data can be aggregated to areas, areas can be represented by a point reference, lines can be aggregated to areas, and data for areas converted into a surface and surface values estimated for both points and areas (Gatrell 1991 in this volume). Likewise, levels of measurement can be changed by recoding operations. It should be noted, however, that all spatial data operations involving aggregation and generalization are usually irreversible. This is because they result in the loss of original information and the possible addition of unwanted noise and, sometimes, pattern to the data. It is important that information is held in its most disaggregated form and that it is analysed at that level.

**Table 25.1** A simple typology of some spatial analysis methods

<b>Type of geographical data</b>	<b>Methods of analysis</b>
Point	Nearest neighbour Quadrat methods
Line	Network analysis and graph theoretic methods Fractal dimension Edge detection
Area	Shape measures Spatial autocorrelation Spatial regression Regionalization Spatial interaction Location-allocation modelling
Surface	Image processing Bayesian mapping

A common starting point in the analysis sequence is the map generated by GIS. This usually results in the user conjuring up a whole series of questions that involve spatial analysis. Do the map

patterns mean anything? Are they 'real' or are they likely to be a chance occurrence? What might be 'causing' a particular pattern? Can the patterns be modelled, predicted, and forecasted? Can the map patterns be manipulated using planning tools?

These questions reduce to two key types of spatial analytical activity: (1) spatial pattern description and (2) spatial pattern relationships. These can involve univariate as well as multivariate analysis. It is often statistical in nature but not exclusively so, with mathematical modelling and other forms of *ad hoc* geographical analysis procedures being of interest.

### Spatial pattern descriptors

In spatial pattern description, various numerical and statistical descriptions can be obtained to summarize the display. For point data various nearest neighbour and geostatistical methods can be used to summarize patterns; for example, centroid and standard distance for point patterns relating to selected attributes and mean distance to *k*th nearest neighbours. For area data, the measures of spatial autocorrelation are often employed (Cliff and Ord 1981). Various multivariate statistical methods can be used to summarize complex multi-layered (i.e. multivariable) map data sets. Line data types are generally more difficult to analyse, although various measures such as orientation and intersection frequencies might be useful. Network data can be described using various graph theoretic measures. Finally, surface data are often described by being fitted to various mathematical functions to yield different degrees of pattern; for example, different orders of polynomial trend surface. Sometimes the map patterns relating to data cannot easily be shown in cartographic form; for instance, flow data relating to a complete origin or destination table, although there are exceptions even here.

A few other areas which may cause problems concern the non-ideal nature of spatial data distributions, the lack of linearity of relationships, and the usual problems of interpreting spatially aggregate information (e.g. ecological inference error and the Modifiable Areal Unit Problem (MAUP); see Openshaw 1984). In some areas, specially adapted methods exist which can cope better with the special needs of spatial analysis. For example, the incorporation of a contiguity

constraint into cluster analysis, so that regions consist of spatially contiguous areas, is often a useful improvement to a standard cluster analysis procedure.

### Spatial pattern relationships

An interest in spatial pattern description soon leads to more sophisticated questions about spatial pattern relationships; indeed, pattern description is seldom an end in itself. For example, if a pattern exists what might be causing it? If there is a particular variable of interest which displays spatial patterning, then what are the principal spatial covariates? The standard approaches involve factor analysis and regression methods to analyse data for spatial associations. However, the use of standard statistical procedures necessitates ignoring the presence of spatial dependencies in the data. If regression is of interest, then it is appropriate to use an explicitly spatial regression model (e.g. see Anselin 1988; Anselin and Griffith 1988; Kennedy 1988).

### Problems with spatial pattern description and relationships

It is important to note that both spatial pattern description and relationship analysis methods can be applied in three markedly different contexts: (1) testing *a priori* hypotheses about patterns and relationships present in spatial data; (2) efficient spatial pattern and relationship description; and (3) analysis for purposes of decision support and spatial planning.

One problem with both description and relationship measurement is the need to generalize the results and, perhaps, compare findings in different study regions. This can be handled within an inferential framework. It is usually assumed that the user has a predefined *a priori* hypothesis that was not generated by examining the data on which it is to be tested. Often a general purpose null hypothesis is used; namely, that the map data have been generated by some kind of spatially random process; for example, the *k*th nearest neighbour distance is similar to what would be expected in spatially random data. If more detail exists then the hypothesis can be more explicit; for example, that

cancer incidence does not decline with distance from a nuclear installation or that one region has a higher value on some test statistic than another region. This process is fraught with difficulty. First, it is necessary to have knowledge of the sampling distributions of the test statistic under the null hypothesis; with spatial data standard approximations do not often apply. Second, *any* prior knowledge of the data invalidates the outcome; for instance, if the hypothesis was generated by looking at a map of the data then it could not be properly tested without access to a second, unseen, data set. This *post hoc* hypothesis testing problem is extremely important in a scientific context, yet it runs counter to the long prevailing style of exploratory data analysis used in geography. It also has major implications for spatial model building. One approach is first to run a model, then examine the residuals to define a missing variable, then re-build the model. That is fine provided no tests of significance are used to validate the model but then how is the user supposed to know whether the model is a good one?

Other problems with the use of inferential variants of spatial analysis methods concern: (1) the use of published critical test statistic values will almost certainly be inappropriate because of the spatially autocorrelated nature of geographical data; (2) the power of test statistics used in a spatial analysis context is not usually known; (3) it is not certain as to whether the sampling analogy is meaningful because many geographical data constitute the population and there is no notion of sampling; and (4) problems of multiple testing which often occur in exploratory studies or when mapping probabilities (each zone constitutes a separate hypothesis so critical significance levels need to be corrected downwards).

Some of the difficulties with inferential methods can be overcome. For instance, the use of Monte Carlo significance tests is a neat way of avoiding the need to make asymptotic assumptions about the distribution of test statistics (Besag and Clifford 1989). It is also a good way of dealing with spatially autocorrelated data. Other problems remain less well identified to hamper the unwary in an inferential context. For instance, the use of spatial autocorrelation statistics for ratio variables may invalidate standardized mortality rates. Another avenue which may be used to avoid these problems is to switch to Bayesian methods and

leave the frequentist domain altogether. In the long term this may be ideal but at present there are problems in computational tractability. Nevertheless, some GIS-relevant Bayesian mapping procedures exist; (see Clayton and Kaldor 1987, and Alexander, Ricketts and Williams 1989). They have the nice property of seeking to avoid some of the problems in mapping data by taking into account the spatially varying degrees of data reliability, albeit at the expense of low power and a high degree of arbitrariness.

Another way to apply spatial analysis methods is to move away from an inferential approach and regard description as the main purpose in spatial analysis. This involves searching out potentially interesting map patterns without necessarily being in any position to test any hypotheses relating to them. In general, the descriptive use of many statistical methods on spatial data is satisfactory provided no strong reliance is placed on significance testing to validate or test the results. However, this does have implications for comparative study and result generalization. It may also appear to degrade the utility of GIS and spatial analysis, but as is argued in greater detail later, this is quite reasonable given the nature of geographical information. It will be possible with time to develop better statistical procedures for use with spatial data, although this task is proving extremely difficult and in any case it is not necessary for most uses of GIS. There is an argument, therefore, to abandon the traditional geographical applications of statistical inference in favour of a more descriptive approach in which significance tests are used mainly as a results filtering mechanism.

The third view is to focus on the use of spatial analysis as a planning and decision support tool. The conventional concerns of science and statistical inference are now subverted by the need to make decisions based on the results of spatial analyses. Densham (1991 in this volume) provides further details of spatial decision support techniques.

One final aspect concerns the nature of the available spatial analytical procedures in GIS. Most of the methods described in Table 25.1 are not yet available within GIS. This is simultaneously a problem and an advantage. The lack of relevant methods is a problem and it is probably only possible to think of applying spatial analysis tools to problem areas where there is a pre-existing body of expertise in the technology and/or the problems

themselves are of sufficient importance to demand special attention. At the same time, the relative absence of unsuitable methods is also an advantage. Unfortunately, it is extremely easy to implement methods that are inherently unsafe and it would be a major error to include methods that exist in the literature, but which are not readily applicable or useful to the GIS era. It is important to try and avoid the worst problems of misuse likely to be generated soon by the availability of a comprehensive menu of spatial analysis tools, through which the user can, without regard to any of the underlying principles, run the same data through everything that exists. That would be fine if the methods were all appropriate to the task in hand.

---

### SOME BASIC PRINCIPLES

---

The real problem is, therefore, not with the definition of what spatial analysis means or with any associated philosophical limitations, but with identifying the nature of the technology needed to provide basic spatial analytical functionality relevant to GIS. For instance, how is it possible to detect patterns in two-dimensional space, or discover whether some arbitrary map coverages are related in some way, or analyse time-dynamic spatial information? Such questions are common to many different GIS-inspired analyses. Yet it should not be assumed that the existing spatial analysis toolbox that comes from either quantitative geography or spatial statistics is at all useful or appropriate to GIS. It has been argued elsewhere (Openshaw 1989a) that the spatial analytical methods which GIS needs are still mainly absent and await development.

Obvious key functions that are missing from current GIS are basic exploratory geographical analysis tools which can help the user to 'find' and 'describe' patterns and relationships that may exist in spatial databases. This strong emphasis on data exploration, rather than on confirmatory analysis and hypothesis testing, reflects the lack of applicable theory and prior hypotheses in most GIS applications.

The need for exploratory geographical analysis tools has also been stimulated by the vast explosion in geographically referenced information which is

creating many new opportunities for spatial analysis in areas where there is little previous research. In many instances, the purpose behind the analysis can only be exploratory and often results from the fact that suddenly some new spatial data exist and because they exist, they need to be analysed. For example, experience has shown that once disease databases are geographically referenced and thus available for spatial analysis, there is only a short time-lag before the data are analysed using the full battery of available methods. The driving force is no longer the academic sector, but a diverse range of applied users. Many of these have little interest in basic research and academic concerns and merely want to use GIS technology to answer pressing practical questions. There is a need, therefore, for simple usable spatial analytical methods that are relevant to GIS.

It is argued that the emerging mountain of real and potentially creatable geographically referenced data challenges the conventional manner by which spatial statistical analysis and modelling are currently performed in an academic context. The challenge can be viewed as involving the need for an automated and more exploratory *modus operandi* in situations which are data rich but theory poor. Some will argue that a mapping capability is all that is required. Indeed maps provide an excellent communications medium for presenting results in a form that most people think they can understand. However, maps provide a very poor form of analysis technology; the human brain is far too easily tricked and misled and the patterns that occur are often far too complex for easy visual recognition and interpretation (Monmonier 1977). There is a clear need for a quantitative exploratory style of spatial analysis that can complement the map-oriented nature of GIS. These analytical tools must be designed to meet such needs without becoming so rigidly statistical that they become: (1) purely tools for researchers; (2) so lacking in creativity that they offer little prospect of new insights; and (3) unable to answer any of the basic questions. Maybe GIS only needs simple technology as the intrinsic limitations of spatial analysis, combined with the absence of process knowledge, argue strongly in favour of a more relaxed, flexible, artistic and less statistical approach than may have been expected. Certainly the new technology has to be creative within limits defined by some statistical analysis process, but without being too constrained by an

over-reliance on inappropriate methods. This perception is fundamentally different from the emphasis in spatial analytical research of the last two decades, in which the goal seems to have been greater levels of statistical and theoretical sophistication rather than any strong concern for application. If statisticians and geographical methodologists wish to remain in touch with GIS then they need to develop understandable methods that can answer the questions that typical users are likely to ask.

There is no magic set of spatial statistical tools that can be incorporated into GIS in order to provide an adequate spatial analysis toolbox analogous to what exists for spatial data handling. Instead there is the prospect of a long hard struggle to develop suitable methods that can cope with the very hard problems that characterize this area. There are many practical difficulties that have to be faced. In particular, geographical data are inherently difficult to handle; current GIS seldom contain all the data structures and access paths needed for spatial analysis; spatial data suffer from endemic errors of various kinds and these errors propagate thereby contaminating 'clean' data sets and reducing further the quality of these data. There is probably not much that can be done to eliminate the causes of error and uncertainty in geographical information. What is more important is the development of methods of analysis that can handle data uncertainty rather than simply ignore it.

The research challenge is to build on these principles and assemble a new set of spatial analytical tools. This task is assisted by the recent rapid increases in affordable computer processing power. New computationally intensive numerical- and simulation-based styles of spatial analysis can be developed that may avoid some of the limitations and difficulties of traditional approaches. A computer-intensive route also brings with it a very different perspective that seems more relevant to the needs of GIS. It is possible to 'buy' solutions, which are probably good enough for most users even if they do not necessarily completely satisfy the intellect, merely by throwing computing power at the key problem areas. The improved availability of supercomputers and multiple-processor systems is fundamentally changing the amount of computer power that is available, by two or three orders of magnitude. It is now possible to think about new numerical-based approaches without worrying too

much about computer speed restrictions.

Development work on today's supercomputers can proceed in the secure knowledge that by the mid-1990s (maybe much earlier), similar levels of performance will characterize the popular workstations on which GIS will be run.

These computer hardware trends also make it feasible to start developing automated spatial analytical methods. This is important as a means of improving the efficiency of exploratory tools and for coping with the thousands of potentially interesting geographical data sets that now exist in the world and which have never been subjected to any form of spatial analysis. It is no longer feasible to think only in terms of hand-crafted manual analysis by experts with one or two years per data set! Nor is it sensible to ignore most of the data. It is not inconceivable that the information locked up in some of these unanalysed data, in the form of spatial patterns and relationships, and not yet identified deducible theories, could be of considerable public, commercial, and academic value. Unless spatial analysis methods relevant to GIS can be developed, many of these data will never be analysed. The real challenge is, therefore, to discover how to trawl through these geographical databases in search of interesting results with a high level of efficiency by devising new forms of spatial analytical methods, rather than castigating this objective as constituting the ultimate in unsound science.

---

### UNDERSTANDING THE LIMITATIONS OF SPATIAL ANALYSIS

---

In seeking to develop methods of spatial analysis relevant to GIS, it is important to start by being realistic about what spatial analysis can reasonably be expected to deliver. The lessons from the first quantitative revolution in geography suggest that it failed, partly because too many users held wholly unreasonable expectations about what might be obtainable from geographical analysis. In reality, even the most sophisticated spatial analysis procedure will probably not progress the user very far along the path of scientific understanding and in some ways the technology appears to be limited in what it can offer. Some users may find that geographical pattern analysis and description is not particularly useful in their search for process

knowledge and causal understanding. Yet in many instances, spatial analysis of the available geographical data is the only available option. The purpose of that analysis would typically be to develop insights and knowledge from any patterns and associations found in the data, which will either be useful in their own right or else provide a basis for further investigation at a later date using different, probably non-spatial and more micro-scale, methods. This function of 'pointing others in the right direction' is an important goal for spatial analysis.

Several problems conspire to restrict what spatial analysis can achieve. They include: (1) the lack of prior theory or hypotheses forcing the user to start by searching for the existence of patterns or relationships without knowing what to look for, or even whether there is anything to find; (2) the difficulties of operating in an exploratory context in which knowledge of the data greatly complicates the testing of *post hoc* hypotheses and models; (3) the available geographical data are usually only surrogates for other information which is missing and often unobtainable (e.g. the use of distance as a proxy for all manner of processes); (4) GIS may be data rich but in most applications few of the key process variables are present; (5) the ecological nature of most analysis is a limiting factor while data aggregation can change the nature of micro-level relationships and sometimes also create spurious new ones; (6) spatial data tend to be characterized by complexity; (7) there are endemic questions about data accuracy and quality; and (8) there are severe difficulties in coping with the time dimension. As a result, spatial analysis in a GIS context is unlikely to result in a greatly improved understanding of causality. This is not so much a deficiency of the methods rather than a recognition of the complexity and limitations under which GIS operate. It is also questionable whether this goal of process understanding based on knowledge of causality is in fact reasonable given the well-known problems of making causal inferences in the non-experimental sciences. At best, all that will be achieved will be some qualitative and descriptive story about how processes may work. A few GIS models may claim to offer process-relevant insights, but at the end of the day they will probably still be incapable of adequately representing the real causal mechanisms.

Spatial analysis in the short term can only be a

fairly primitive descriptive science, but maybe for many purposes that is sufficient. The ultimate difficulty is the inherently limited utility of geographical information in studies which attempt to understand process. An example from spatial epidemiology may help to clarify this key statement. Age/sex standardization of incidence rates is a common practice but neither age nor sex covariates are directly causal variables, they are proxies for other unmeasured and unknown process variables; for instance, age cannot cause cancer but other variables related to age might. In geographical analysis also, virtually all the available data are surrogates and proxies for other variables which are either missing from the database or incapable of measurement or not yet identified. It would be sheer folly to go beyond what these sorts of data can sustain. This is not to underestimate the immense potential utility of geographical information, or of the important insights about process that spatial analysis may provide, but merely recognition that there are limits to what can be achieved in a geographical context. Fortunately, many GIS users have only a basic shopping list and often only want to make simple statements about the presence or absence of patterns and relationships. For instance, it is still usually sufficient to identify spatial pattern as a departure from a spatially random process expectation, without having to define precisely what spatial process generating assumption would be appropriate. Despite this apparently simple requirement, current spatial analysis methods are nowhere near meeting these needs and are a long way from attaining their full potential. GIS is revolutionary technology which requires flexible fresh thinking unfettered by the past. In short, it requires a new way of thinking. Table 25.2 offers some guidance for spatial analysts.

---

#### DEVELOPING APPROPRIATE SPATIAL ANALYSIS FOR GIS

---

The problem is where to start. It is apparent that most developers of GIS have in the past seen little need to put much effort into spatial analysis. Also, there is no clear view of what spatial analytical functions are needed. There is very little merit in merely trying to include either a complete statistical package that cannot cope with the special nature of

**Table 25.2** Some basic guidelines for spatial analysis in GIS

1. Avoid highly formalized scientific designs.
2. Adopt an exploratory data analysis mentality.
3. Avoid being too statistically blinkered with an over-emphasis on inappropriate inference.
4. Stay within the limitations of geographical analysis.
5. Avoid any technique that either implicitly ignores or explicitly removes the effects of space.
6. Think carefully before using methods left over from the 1960s era of quantitative geography.
7. Avoid the use of asymptotic assumptions, use Monte Carlo simulation instead.
8. Remain aware of the possible effects that data problems can have on the results.

spatial data or even to code-up the complete spatial statistical technology according to Diggle (1983), Ripley (1981) or Upton and Fingleton (1985) for a GIS. This would be pointless because the chosen methods have to be appropriate for typical GIS environments, users, and analysis needs. For example, nearest neighbour methods which assume no positional uncertainty in point coverages are inapplicable, as are spatial regression models which can only function with no more than a few hundred zones and provide no automated predictor search mechanism. Interfacing complex statistical packages, such as GLIM, would be another largely irrelevant diversion of effort, as the methods cannot readily cope with the spatially dependent nature of the data they are meant to process.

One way forward is to define a small set of generic spatial analysis functions, which can be built in as standard GIS operations with their complexity hidden by the use of appropriate interfaces. Another would be to develop a more advanced set of analysis tools, which would seek to provide new analytical functions which are only possible within a GIS environment. One problem involves defining what spatial analysis functions and operations are sufficiently general and generic to justify their inclusion. A related issue concerns the nature of the spatial data handling operations that the spatial analysis methods may require to function and the need to ensure that GIS builders put the necessary hooks into their systems.

In addition, there is the practical necessity of ensuring that the methods can actually work in a

GIS environment. A basic set of design objectives for developers of new spatial analysis methods would have to include: an ability to handle large numbers of zones (say 10 000); the need to cope with the nature of geographical data including the presence of uncertainty and errors; a high degree of algorithmic portability; coverage of generic analysis needs; the prospect of an extension into the time domain; a high degree of automation; freedom from critical assumptions and an essentially exploratory *modus operandi*.

---

**A SHORT LIST OF GENERIC SPATIAL ANALYSIS FUNCTIONS**

---

To some extent the importance of spatial analysis to GIS is being recognized, although it is doubtful whether sufficient research resources are being devoted to building practical methods. The US National Center for Geographic Information and Analysis (NCGIA 1989) is clearly aware of the problem and the spatial analysis theme is implicit in several of its initiatives for the first three years. In the United Kingdom, the ESRC's Regional Research Laboratory (RRL) initiative has also identified spatial analysis as one of three major research objectives for the eight RRLs to investigate. Attempts have also been made to develop a research agenda (Openshaw 1990a) and six key areas have been identified (see Table 25.3). It remains to be seen whether any of these procedures migrates into the GIS systems of the 1990s.

**Table 25.3** Six key spatial analytical research topics

1. Response modelling for large data sets with mixed scales and measurement levels.
  2. Practical methods for cross area estimation.
  3. Zone design and spatial configuration engineering.
  4. Exploratory geographical analysis technology.
  5. Application of Bayesian methods.
  6. Application of artificial neural nets to spatial pattern detection.
-

Underlying all these topics is a concern for exploratory geographical data analysis, restricted to questions that the available geographical information might be capable of answering. This simplifies the technical task as it involves no more than pattern and relationship description. In an applied study, these activities would constitute only the first stage of a more extensive work programme, with the GIS role being limited to providing an indication as to where further research should be performed. However, this is still an extremely important role. Moreover, hypotheses obtained from the analysis of data for one region can be tested (with considerable power) in another. Pattern description and inductive styles of analysis may seem difficult but they can also be extremely creative and may result in new knowledge. Also, the results they produce can often become the basis for action without necessarily risking delaying a response until there is proof of causation (typically involving a 50- to 100-year delay); provided there can be some assurance that the patterns are real and not spurious. This raises another important technical point. The conventional 5 per cent statistical significance level may appear to be both too lax (if the results are likely to cause concern) and too stringent if descriptive power is being lost. It is very easy to confuse descriptive and suggestive results with validatory and confirmatory work. The latter is wholly concerned with Type I errors (finding pattern where none exists) while totally ignoring Type II errors (failing to find pattern when it exists). Quite often the latter may be more critical than the former in exploratory and descriptive GIS based spatial analyses.

The next stage in the argument in favour of exploratory geographical analysis is to define a set of basic generic functions; see Openshaw (1990b). A list is given in Table 25.4 and the key features are expanded here. This is viewed as important because the most difficult task is identifying what functions are required; the subsequent operationalization is often far more easily achieved. It is hoped that by identifying a list of generic methods, which reflect the previous discussion of the need for GIS-relevant spatial analysis technology, that their creation will be encouraged where they do not yet exist, and that they will be more widely used where they do exist.

The idea of a pattern spotter is simply an automated means of identifying evidence of geographical pattern in point data sets without any *a*

**Table 25.4** Basic generic spatial analysis procedures.

1. Pattern spotters and testers.
2. Relationship seekers and provers.
3. Data simplifiers.
4. Edge detectors.
5. Automatic spatial response modellers.
6. Fuzzy pattern analysis.
7. Visualization enhancers.
8. Spatial video analysis.

*priori* hypothesis to test. This problem may occur in a genuine exploratory situation or as a means of avoiding problems of *post hoc* model construction when prior knowledge of a specific geographical database renders suspect any hypothesis testing approach. One realization of this method is the geographical analysis machine discussed in the next section.

A relationship seeker is an attempt to develop a statistical procedure that mirrors the map overlay process. Relationships between a point data coverage and a set of  $M$  map overlays can be modelled using Poisson regression. An alternative, more geographical procedure involves a search among  $2^M - 1$  permutations of map coverages for evidence of spatial pattern being created by the interaction of the overlays with point data of interest. A prototype procedure, termed a Geographical Correlates Exploration Machine (GCEM) has been built (Openshaw, Cross and Charlton 1990). An interesting feature is its use of location as an additional level of surrogate variable. It will allow relationships which occur 'here' but 'not there' to be identified.

Data simplifiers are merely GIS-relevant versions of existing classification methods. They have existed for a few decades but are still absent from most GIS. Regionalization procedures (i.e. classification with contiguity constraints) provide an obvious means of simplifying very large and complex spatial databases to identify patterns. These procedures can deal with extremely large data sets and also handle flow data. Automated zone design procedures provide a solution to a whole range of spatial engineering problems (e.g. redistricting and customized zone design). One reaction to modifiable areal unit effects (Openshaw 1984) is to engineer spatial data aggregations to

possess required characteristics. Automatic zoning procedures can effectively achieve this objective providing approximately optimal solutions to various constrained and unconstrained problem formulations. GIS is removing all the historic restrictions on the types of zoning systems available for reporting spatial data and it is important that this new found 'freedom' is properly controlled and used.

Edge detection is a further area of spatial analysis relevant to GIS that is very undeveloped. Zones are usually stored in a vector-based GIS as a set of line segments with topological details. Why not develop spatial analytical tools for analysing the data in a GIS at this level instead of at the zonal scale? Pattern detection now becomes a problem in edge detection.

Spatial response modelling is also of considerable importance. Increasingly, GIS are being used to create multi-scale databases ranging from the micro to the macro. There is a need for automated response models to be developed in which the values of a dependent variable can be predicted by reference to whatever spatial predictor information might be available, and under circumstances where there is minimal prior knowledge of the functional forms that might be most appropriate. One response is a fully automated modelling system (Openshaw 1988). Another is to develop variants of the AID (Automated Interaction Detector: a survey method based on binary segmentation) technique called database modelling (see Openshaw 1989b). No doubt there are other possibilities, but it is important to remember the design objectives that GIS set; for instance, many rather than few possible predictors, no prior knowledge of model specification, non-linearity should be assumed, data errors would not be unusual, mixed measurement scales are not uncommon, and large data sets are standard rather than exceptions.

Fuzzy analysis procedures are clearly relevant to many areas of GIS because they provide a means of dealing with all types of data uncertainty. The question is basically how to incorporate fuzzy analysis into GIS. Currently there are few operational examples, although a fuzzy geodemographic targeting system has been proposed (Openshaw 1989b). The linkages with object-oriented programming should help stimulate academic interest in this area of GIS, but it could be

the next century before any practical methods emerge.

Visualization enhancers represent another approach to supplementing the communication power of the basic static map display (see Buttenfield and Mackenness 1991 in this volume). The increasing availability of geographical data with time coordinates revives interest in the use of computer movies as a basic but potentially extremely effective analysis procedure. A time-driven computer movie presentation would enliven an otherwise static display (Tobler 1970; Moellering 1973). The cost of basic computer movie technology based on video recording is now fairly low, although the use of sophisticated animation technology requires access to special video computer hardware. In a GIS application, maybe only simple procedures are needed (for instance, a succession of maps showing data in two or three dimensions). The utility of the visual images might be further enhanced by performing spatial analysis at each time slice. For example, the effects of space-time analysis might be best seen by displaying a movie of  $N$  different, but sequential in time, space-time analyses. An additional dimension could be provided by using sound to supply supportive information about the map patterns being viewed. However, this form of spatial analysis delegates the task of spotting patterns and being creative to the human observer. It also puts considerable emphasis on stimulating little understood human cognition systems by supplying selected visual and auditory information.

A final area for investigation concerns the possibility of analysing spatial data at the pixel level. The objective here is to redefine spatial analysis operations at the pixel scale, a form of representation which is common to all spatial data so that a common micro-analytical technology for spatial analysis might be developed. For this application, image processing technology is useful but not adequate and a new set of tools would have to be constructed. Besag (1986) gives one example of how it might be achieved by statistical means; another route could involve cellular automata (a form of computer modelling that creates simple holistic structures out of simple rules applied to microscopic level data). As processing power becomes less of a constraint, entirely different forms of spatial analysis might well emerge in GIS and this may be one of them.

### **SOME SIMPLE PROCEDURES FOR DETECTING PATTERNS AND RELATIONSHIPS**

Finally, a brief case study based on an extension of the author's Geographical Analysis Machine (GAM), for the analysis of point pattern data, might be helpful in illustrating the role of spatial analysis in GIS. The original GAM concept (Openshaw, Charlton and Wymer 1987; Openshaw *et al.* 1988) has been generalized to encompass a wide range of spatial pattern search methods. Openshaw (1989c, 1990c) describes a GAM ( $g, m, s$ ) procedure where the  $g$  parameter relates to the nature of the search geometry (i.e. circles, squares, equal population at risk areas,  $k$ th nearest neighbours),  $m$  refers to the choice of significance assessment procedure (i.e. descriptive, hypothesis testing, and significance estimation), and  $s$  is the type of search strategy employed (i.e. locationally comprehensive, case based, *a priori* site restricted, linear feature, etc.). Variants of this method can be run on a microcomputer and experience in running a GAM identifies many of the key technical issues that occur elsewhere in exploratory spatial analysis.

The original GAM involved the use of a circular pattern detector and a locationally comprehensive search based around intersection points on a lattice. The lattice mesh was set at 0.2 circle radii to ensure that the circles overlapped. This was considered important to allow for positional uncertainty in the point data while a grid search pattern was used to ensure that no locations were excluded. In the original GAM, a range of circle sizes was examined. A Poisson probability was used to screen the results so that only those circles with a relatively small probability of being due to chance were mapped. The map of circles was then used to identify subregions where there appeared to be evidence of departure from a spatially random distribution of points.

This approach can be criticized on the grounds that the significance threshold used to screen the circles needs to be corrected for multiple testing, the overlapping circles re-use some of the data, and no overall measure of Type I map error is computed. All these problems can be overcome by running a GAM on a supercomputer so that a Monte Carlo significance test procedure can be used. However, to some extent these criticisms constitute an irrelevant distraction because they

imply that the spatial analysis is concerned not only with pattern detection but also with validation. They also raise severe technical problems. In particular, the large number of implied hypotheses being tested (typically a few million) results in low power and only the most extreme results will tend to survive (Openshaw *et al.* 1989). Additionally, the results also depend on study region boundary and size; the larger the region of interest the greater are the dilution effects. Some of the lost power can be regained by switching from measures of whole map pattern (i.e. total counts of significance circles) to what is termed vicinity analysis. The number of circles can be reduced by defining a series of blobs (or superclusters), each consisting of a set of overlapping circles. The significance testing is based on the frequency with which blobs formed in data generated under a null hypothesis, have more extreme characteristics than the observed data blobs that they overlay (see Openshaw 1990c). There is, however, an argument that the original descriptive GAM is best and that maybe spatial analysis should stop after defining areas where it may be worth performing additional analysis using different data and more precise methods. Certainly fears that the original GAM was prone to large degrees of Type I error proved to be unfounded.

The question now is what variant of the GAM ( $g, m, s$ ) family can be immediately used in GIS. The need to use a supercomputer is primarily a result of using a locationally comprehensive search strategy. If a less comprehensive analysis is acceptable then the procedure outlined in Appendix A can be run on a microcomputer. This is based on Besag and Newell (1990) who questioned the need for a locationally comprehensive search and recommended a  $k$ th nearest neighbour circle method (see the alternatives in Appendix A) as a means of coping with rural-urban differences in population density. It is also computationally far easier to focus only on the observed cases. However, if positional uncertainty in the data is to be simulated then results from the Besag and Newell method and the original GAM would tend to converge. Finally, it is possible to adapt the Appendix A procedure to cope with a search along a buffered linear feature, for instance, an overhead wire. A similar alternative is the Cuzick and Edwards (1990) test.

A different type of spatial analysis problem occurs when it is necessary to test a hypothesis that

there is a raised incidence of, say, a disease near a set of *a priori* locations. Appendix B outlines a simple procedure for site-based hypothesis testing. To be scientifically valid, the sites have to be identified prior to any knowledge of the data and if there are multiple sites then an additional correction for multiple testing should be applied. The procedure in Appendix B is a variant of the Poisson maximum method (Stone 1988) which makes no assumption about which distance band is the critical one; it merely examines a wide range and corrects the result for multiple testing using a Monte Carlo method.

A third basic procedure is concerned with measuring the association between a point data coverage and  $M$  other coverages. One route is the best coverage permutation search used in GCEM. Another is a Poisson regression model with a focus on analysing the residuals from a well-fitting model for evidence of interesting spatial patterning. Another possibility requires that the dependent coverage is not a point coverage. A random sample of points is generated and an attempt made to identify a relationship using categorical analysis methods (namely, log linear modelling) or a database model of some kind. These types of analyses could be readily performed using standard packages.

---

## CONCLUSIONS

---

The task of developing appropriate methods of spatial analysis for GIS is extremely important at a time when there is large growth both in the availability of geographical information and in the numbers of users who are potentially interested in spatial analysis. The previous neglect of spatial analysis looks set to become a major impediment to the full exploitation of GIS. There is a danger that the growing imbalance between the availability of geographical data and the limited range of existing analytical technology may slow the growth of GIS and result in widespread failure to make full use of the available information. This discussion has attempted to provide a better understanding of the nature of those spatial analysis methods that seem to be most relevant to GIS. The key features that are considered important are an exploratory function and an emphasis on insight and

creativity. There are dangers in trying to be too statistically pure in situations where the appropriate methods have not been developed and which, in any case, can only really sustain low level description. Perhaps in its GIS guise, spatial analysis can only remain an art and will never aspire to being a science.

---

## APPENDIX A: A SIMPLIFIED GAM FOR A MICROCOMPUTER

---

### Basic Algorithm

*Step 1.* Create two point data coverages, one of cancer cases or some other rare data to which a Poisson assumption is applicable, the other a measure of the population at risk. The data could refer to points or to small zones which are point referenced. The two point data coverages should be merged so that there is one data set containing both the incidence data and the population at risk counts.

*Step 2.* For each point with at least one observed case (i), order all other points by distance from it. Apply a selection rule to determine the count of cases and population at risk within a critical distance of (i); see below for alternatives.

*Step 3.* Compute a Poisson probability of obtaining the observed number of cases given the population at risk under the null hypothesis.

*Step 4.* If 'significant' at the 5 per cent level then draw the 'circle'.

*Step 5.* Repeat Steps 2 to 4 for all observed cases

### Alternative Selection Procedures

Some alternatives could be: (a) use a fixed radius circle; (b) determine for each case the minimum circle radius sufficient to encompass at least  $K$  other cases; and (c) the minimum circle radius to yield a target population at risk or an equal expectation of cases.

### Other variations

The search geometry could be changed; the 'circles' could be replaced by 'squares' or 'segments of

circles' and so on. The effects of possible covariates could be removed by adjusting the Poisson probability calculation. The effects of positional uncertainty can be handled by repeating the procedure many times (say 99 or 999), each time wobbling (according to some error model) the data coordinates. An approximate measure of robustness can be obtained by computing for each case the amount by which the population at risk could be increased before the result became insignificant. Finally, the effects of multiple testing could be handled by Monte Carlo methods, but these would suffer from study region dependency. Nevertheless, it is important to be aware that if there are  $M$  observed cases and with a significance threshold of 5 per cent, then  $M*0.05$  significant results would be expected purely by chance. In this spatial context the test is no more than a screening process and, even if a smaller than expected number of significant results occurred in a 'strange' area, then there may well be something interesting occurring. Whether this 'something interesting' is an artefact of the data or is real would be a matter for subsequent investigation.

---

## APPENDIX B: A SITE BASED HYPOTHESIS TESTER

---

### Basic Algorithm

*Step 1.* Assemble a set of sites to be evaluated and create data as in Appendix A.

*Step 2.* Select a maximum radial search distance ( $d$ ).

*Step 3.* Order the data from an evaluation site by distance. For each different distance band out to distance ( $d$ ) compute a Poisson probability that the observed cumulative number of cases could have occurred by chance under the null hypothesis.

*Step 4.* Identify the distance band with the most extreme result.

*Step 5.* Repeat Steps 3 and 4 for 99 or 999 or 9999 spatial data distributions generated under the null hypothesis and compute the rank of the observed data result. Convert to a measure of probability as a test of the hypothesis.

### Observations

It is important that the list of sites to be evaluated is identified prior to obtaining any knowledge of the data. The list should not subsequently be added to in the light of the results. On the other hand, if the purpose is description and not significance testing then a variety of search methods might be employed to 'look' for maximally significant results. This is valid only if the probability of obtaining similarly extreme results in spatially random data was sufficiently small to make any interpretation interesting.

---

## REFERENCES

---

- Alexander F E, Ricketts T J, Williams J** (forthcoming) Methods of mapping small clusters of rare diseases with applications to geographical epidemiology. *Geographical Analysis*
- Anselin L** (1988) *Spatial Econometrics: methods and models*. Kluwer Academic Publishers, Dordrecht
- Anselin L, Griffith D** (1988) Do spatial effects really matter in regression analysis? *Papers of the Regional Science Association* **65**: 11–34
- Berry B J L, Marble D F** (eds.) (1968) *Spatial Analysis: A reader in statistical geography*. Prentice-Hall, Englewood Cliffs New Jersey
- Besag J E** (1986) On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B* **48**: 192–236
- Besag J E, Clifford P** (1989) Generalised Monte Carlo significance tests. *Biometrika* **76**: 633–42
- Besag J E, Newell J** (forthcoming) The detection of clusters in rare diseases. *Journal of the Royal Statistical Society B*
- Bonham-Carter G F** (1991) Integration of geoscientific data using GIS. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 171–84, Vol 2
- Burrough P A** (1991) Soil information systems. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 153–69, Vol 2
- Butenfield B P, Mackaness W A** (1991) Visualization. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 427–43, Vol 1
- Clayton D, Kaldor J** (1987) Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics* **43**: 671–81
- Cliff A D, Ord J K** (1981) *Spatial Process, Models, and Applications*. Pion, London

- Cuzick J, Edwards R** (1990) Tests for spatial clustering of events for inhomogeneous populations *Journal of the Royal Statistical Society Series B* **52**: 73–104
- Davis J C** (1986) *Statistics and Data Analysis in Geology*, 2nd edn. Wiley, New York
- Densham P J** (1991) Spatial decision support systems. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 403–12, Vol 1
- Diggle P H** (1983) *Statistical Analysis of Spatial Point Patterns*. Academic Press, London
- Gatrell A C** (1991) Concepts of space and geographical data. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 119–34, Vol 1
- Goodchild M F** (1988) A spatial analytical perspective on GIS. *International Journal of Geographical Information Systems* **1**: 327–34
- Hagerstrand T** (1973) The domain of human geography. In: Chorley R J (ed.) *Directions in Geography*. Methuen, London, pp. 67–87
- Kennedy S** (1988) A geographical regression model for medical statistics. *Social Science and Medicine* **26**: 119–29
- Johnston R J, Gregory D, Smith D M** (eds.) (1986) *The Dictionary of Human Geography*, 2nd edn. Blackwell, Oxford
- Moellering H** (1973) The automatic mapping of traffic crashes. *Surveying and Mapping* **23**: 467–77
- Monmonier M S** (1977) Maps, distortion and meaning. *Association of American Geographers Resource Paper* 75–4. AAG, Washington
- NCGIA** (1989) The research plan of the NCGIA. *International Journal of Geographical Information Systems* **3**: 117–36
- Nielson D R, Bouma J** (1985) *Spatial Analysis of Soil Data*. PUDOC, Wageningen
- Oliver M, Webster R, Gerrard J** (1989a) Geostatistics in physical geography. Part 1. *Transactions of the Institute of British Geographers NS* **14**: 259–69
- Oliver M, Webster R, Gerrard J** (1989b) Geostatistics in physical geography. Part 2. *Transactions of the Institute of British Geographers NS* **14**: 270–86
- Openshaw S** (1984) The modifiable areal unit problem. *CATMOG* **38** Geo Abstracts, Norwich
- Openshaw S** (1988) Building an automated modelling system to explore a universe of spatial interaction models. *Geographical Analysis* **20**: 31–46
- Openshaw S** (1989a) Computer modelling in human geography. In: Macmillan W (ed.) *Remodelling Geography*. Blackwell, Oxford, pp. 70–88
- Openshaw S** (1989b) Making geodemographics more sophisticated. *Journal of the Market Research Society* **31**: 111–31
- Openshaw S** (1989c) Automating the search for cancer clusters. *The Professional Statistician* **8** (9): 7–8
- Openshaw S** (1990a) Towards a spatial analysis research strategy for the Regional Research Laboratory initiative. In: Masser J, Blakemore M J (eds.) *Geographical Information Management: methodology and applications*. Longman, London
- Openshaw S** (1990b) Spatial analysis and GIS: a review of progress and possibilities. In: Scholten H J, Stillwell J C H (eds.) *Geographic Information Systems for urban and regional planning*. Kluwer, Dordrecht, 156–63
- Openshaw S** (1990c) Automating the search for cancer clusters: a review of problems, progress, and opportunities. In Thomas R W (ed.) *Spatial Epidemiology. London Papers in Regional Science* **21**. Pion, London, pp. 48–78
- Openshaw S, Charlton M, Wymer C** (1987) A Mark 1 Geographical Analysis Machine for the automated analysis of point data. *International Journal of Geographical Information Systems* **1**: 335–43
- Openshaw S, Charlton M, Craft A W, Birch J M** (1988) An investigation of leukaemia clusters by use of a geographical analysis machine. *The Lancet* **1**: 272–73
- Openshaw S, Cross A E, Charlton M E** (1990) Building a prototype Geographical Correlates Exploration Machine. *International Journal of Geographical Information Systems* **3**: 297–312
- Openshaw S, Wilkie D, Binks K, Wakeford R, Gerrard M H, Croasdale M R** (1989) A method for detecting spatial clustering of disease. In: Crosbie W A, Gittus J H (eds.) *Medical Responses to the Effects of Ionising Radiation*. Elsevier Applied Science, London, pp. 295–308
- Ripley B D** (1981) *Spatial Statistics*. Wiley, New York
- Ripley B D** (1984) Present position and potential developments: some personal views. *Journal of the Royal Statistical Society A* **147**: 340–48
- Stone R A** (1988) Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Statistics in Medicine* **7**: 649–60
- Taylor P J** (1977) *Quantitative Methods in Geography*. Houghton Mifflin, Boston
- Tobler W R** (1970) A computer movie simulating urban growth in the Detroit Region. *Economic Geography* **46**: 234–40
- Unwin D J** (1981) *Introductory Spatial Analysis*. Methuen, London
- Upton G, Fingleton B** (1985) *Spatial Data Analysis by Example. Volume 1. Point Pattern and Quantitative Data*. Wiley, New York
- Wilson A G** (1974) *Urban and Regional Models in Geography and Planning*. Wiley, London
- Wilson A G, Bennett R J** (1985) *Mathematical Methods in Human Geography and Planning*. Wiley, London