USER-CENTRED INTELLIGENT SPATIAL ANALYSIS OF POINT DATA

Stan Openshaw and Tim Perrée

WORKING PAPER 95/16

SCHOOL OF GEOGRAPHY • UNIVERSITY OF LEEDS

Summary

This paper outlines the development of a new approach to spatial analysis in GIS environments. An attempt is made to develop an abstract but intuitively obvious spatial analysis process that is designed to be readily understandable to the typical non-statistical minded end-user of GIS. The precision and complexity of conventional spatial statistics and quantitative geography are replaced by a visual form of presentation that provides a simpler and more direct means of communicating and interpreting advanced spatial analysis technology. A prototype system is described, and assessed using synthetic data.

Introduction

This paper presents a new approach to spatial analysis that is designed to be relevant to the current data rich GIS era. The first GIS revolution involved the management rather than the analysis of geographic information. Its very success has led to a geographic information explosion and created the need for a second GIS revolution; one that is focused upon the analysis and use of the geographical information that now exists. This requires the development of new analytical tools that are sufficiently powerful to make good use of the information riches and thus permit users to further capitalize on their investments in GIS technologies (Openshaw 1994b, 1994c). A very important need is to create new and more efficient ways of performing exploratory spatial data analysis (EDA) tasks relevant to GIS environments. According to Haining (1990) EDA involves the description of spatial data, the identification of statistical properties, and the preliminary identification of data structure, ".. with the objective of encouraging hypothesis formulation from the data" (p4). Many users are probably more interested in packaging EDA as a function they can activate whenever they wish to explore spatial information for pattern or relationships. Unfortunately the spatial dimension presents many problems. Previous attempts to resolve these problems has produced spatial analysis methods that are too complex, perhaps too theoretically sophisticated, and usually too hard for many GIS users to cope with (Openshaw 1991). As a result there have been few routine and successful applications of spatial analysis in GIS by end-users working outside of research environments.

Spatial data is inherently complex. While some significant patterns can be found using existing methods, many more undoubtedly remain undiscovered. The challenge of uncovering new patterns in spatial data is important, and its importance is emphasised by the large number of spatial datasets that now exist. Openshaw (1993) talks about the criminal waste of geographical information by both non-analysis and poor analysis of those key data that need to be properly analysed, either because it is in the public good to do so or because there are overwhelming commercial or other imperatives for its full and proper analysis. However, in seeking to use geographical information it is also important to ensure safe and appropriate use; else another type of GIS crime might be invoked (Openshaw, 1994a).

At one time it was thought likely that GIS systems would be able to service the complete spectrum of user needs and that the principal obstacle in a spatial analysis context was the coupling of existing analysis technology to proprietary GIS systems; see for example, Anselin et al (1993). This now seems to be increasingly irrelevant as many of the end-user needs would appear to lie outside the areas for which existing, largely research focused, spatial analysis techniques exist. There is a strong argument that the safe and appropriate analysis of geographical information requires the development of new and easy to use tools based on new thinking, and new systems that are focused on the spatial analysis of geographic information in the context of the applied end-user. It is important to recognize that increasingly the end-users are no longer solely academic researchers or statistical experts, and they have a different set of needs. Neither spatial statistics nor quantitative geography has yet faced up to this particular challenge.

There are also some other major user related problems that need to be addressed. Specifically, many users are not used to data riches, there is a lack of appropriate methods to help them exploit this new situation, and there is confusion about what analysis technology is needed. The novelty of the emerging new opportunities means that there is no pre-existing set of clearly identified user needs and requirements that can be used to specify computer systems. As Openshaw (1995a) argues, '..in many ways the potential applied spatial analysis opportunities that may be perceived by researchers to exist still need to be defined, demonstrated, and then converted by a programme of awareness raising into new sets of user needs and tools'. Maybe there is a need to shake both users and developers out of a highly complacent state so that they can start to become more aggressively adventurous in their exploitation of the new opportunities that GIS has provided them with, and to persuade them of a need to move on from where the more conventional methods stop. Maybe the key issue is to determine what users would want from spatial analysis if they actually knew what was now possible. One way of approaching this is to develop prototype systems with which end-users can explore new possibilities for analysis.

There are five principal issues, from the end-user's point of view, that need to be faced:

1. The apparent complexity of existing spatial analysis technology that often instils a neurosis of fear in the user about their own statistical inadequacies. For example, excellent books such as Haining (1990), Upton and Fingleton (1985), Griffith (1988), Cressie (1991), Anselin (1988), and Ripley (1981) are unintelligible and will remain unreadable to most GIS end-users.

2. The inherent inability of existing methods to provide useful results in many important practical applications.

3. The difficulty that end-users often experience in understanding what the results mean because of the nature of the statistical language that is commonly used to report them.

4. It is very difficult to use results that are not understood as the basis for decisions or actions.

5. Finally, there is a need to communicate the meaning and significance of the spatial analysis results to others.

These end-user concerns are essentially to be able to understand and trust the results of spatial analysis so that they can be used in managerial and decision making processes. This is quite different from the traditional research oriented focus of the spatial analyst or the quantitative geographer. A Type I Error probability has probably little relevance to many decision and policy making contexts. The decision maker wants a black or white answer not a qualified one; for instance, either there is either a problem here or there is not. The fuzzy or qualified 'maybe' outcome does not help anyone as it implies uncertainty, suggests that the analyst does not know what is going on, and provides no information that end-users can cope with. This might be the correct outcome from an academic and research perspective but it is unhelpful to the policy or decision maker because it provides no basis for decisions. It is, therefore, no longer sufficient to 'sit on fences' but to try and meet the needs of the end-user more explicitly by developing appropriate technology that provides what they want in a form they can understand.

Some Generic Design Principles

In attempting to develop an end-user oriented approach to spatial analysis some important issues emerge that question much of the conventional thinking that underpins current approaches. The principal need is to develop a style of spatial analysis that the users of GIS can use, feel comfortable with, and believe in. Experience suggests that these aspects cannot easily be added as an afterthought to methods that were created by experts for experts. Instead, end-user friendliness and ease of comprehension need to be built into both the methods and the software from the beginning. It is also striking that virtually every aspect of current spatial analysis technology is either seeking too much precision or attempting to be too complex and sophisticated. The analysis process is regarded as a science, based on rigor and as much precision as possible. Rigor is clearly an important virtue but it does not necessarily require the highest possible degree of precision regardless of the nature of the application. Consider, for example, mapping; the cartographic origins of GIS are still so dominant that it is usually considered that mapping has to be a highly precise process. This is an admirable goal for a national mapping agency but is it relevant for mapping data in a spatial analysis context in which a subtle change of class intervals can completely change the results? Conventional cartography is quite often too precise given the uncertain nature of the data it represents. It provides a level of accuracy that is spurious and in practice unnecessary. Often the end-user merely wishes to see results that matter and, probably nothing else.

There are two important design issues that need to be resolved: (1) the <u>form</u> of visualisation being used to communicate information to the end-user; and (2) the <u>content</u> of the information being communicated. Conventional cartography of raw or semi-processed data from a GIS may look colourful but it is usually inappropriate as a basis for analysis and communication. The acute observer may, with luck, good eye sight, and a sufficiently simple pattern, discover something of interest; but in practice this seldom happens! In spatial epidemiology the last pattern to be discovered by these means was over 150 years ago. GIS has done virtually nothing to improve this situation. Analysing spatial data by simply mapping it is a most inefficient, and wholly

GIS inappropriate approach to the analysis of inherently complex spatial information. Colourful map wallpaper is for interior decorators not spatial analysts!

Deciding what to communicate is also fraught with traditions that can point in the wrong direction. It is hard to imagine a less appropriate approach to spatial analysis in a GIS context than that which is sometimes advocated by spatial statisticians and quantitative geographers. The statistical technology is too complex; it is highly elitist, being understandable only by a few; it is very partial in what it offers; it is highly fragile, with its validity often resting on naive and, sometimes, suspect assumptions about the nature of spatial data; and the quality of the results is almost totally dependent on the skills of the user. It ignores the nearly universal lack of a high level of statistical expertise amongst the end-users of GIS. The need is for user friendly systems designed for clarity of exposition rather than sophisticated systems of massive statistical complexity. Even when performed to a highly competent professional level, what end-user could use results produced by methods they may never be able to comprehend?

A new class of spatial analysis methods is needed. They should be exploratory rather than confirmatory because the latter serves to mislead. If inference is used it should merely be seen as a measure of performance that has no critical thresholds, other than perhaps the feeling that small probability values are better than large. It may also be necessary to allow for multiple testing, but without putting too much emphasis on the results or relying too heavily on testing null hypotheses that may be wholly inappropriate or wrongly specified for a GIS context. Computational intensive statistical methods particularly bootstrapping and Monte Carlo simulation can be used to try to ensure that only meaningful and valid results are passed to the user. Finally, it is important to avoid being too precise; most applications do not need it. There are real limits to what can be achieved by geographical analysis in a GIS context. The major requirements and characteristics of this new form of spatial analysis are outlined in Table 1.

Table 1 Requirements and Characteristics of this New Approach to Intelligent Spatial Analysis.

1.	The user should not need to know in advance precisely WHERE, WHEN, or WHAT to look for.
2.	There is a need to incorporate knowledge, where it exists and when it is relevant.
3.	It is unreasonable to expect users to possess many genuine *a priori* hypotheses that can properly tested.
4.	The tools have to be easy to use by non-experts.
5.	The technology needs to be inherently safe so that the results are believable, appropriate to the context in which they were produced, and provide a basis for subsequent decisions and actions.
6.	The results should be robust given the realities and nature of spatial information;
7.	The technology should be flexible and comprehensive.
8.	The analysis procedures should be capable of providing new insights and acting as an intelligent assistant working with, rather than against, the user in a creative partnership;
9.	The methods of analysis have to be understandable to the end-users even if that understanding is at the level of plain English description.
10.	The methods should deal with the principal needs for generic and application independent forms of spatial analysis.
11.	The methods should resolve complex spatial data into simpler information about significant patterns or anomalies within the data.
12.	The technology should be able to look within maps and be independent of study region boundaries.

System Design

Searching spatial databases for patterns and relationships is a common requirement of many spatial analysis procedures. This search process is often hard because of data complexities, data overload, multiple data domains, and lack of prior knowledge of what to look for and where to find it. In some systems the user controls the search in an interactive graphical environment; see for example, SPIDER, (Haslett et al, 1990), REGARD (Unwin, 1992, 1994), and ISP (Nagel, 1994). This allows a user to understand aspects of the spatial structure of the data but its use is restricted to expert users and fairly simple low dimensional data sets. The other extreme is the fully automated and 'black box' approach to spatial analysis of Openshaw's Geographical Analysis Machine (GAM), Openshaw et al (1987). In GAM there is no user interaction and the results are automatically generated. This approach has some attractions; particularly, its spatial comprehensiveness and independence of user skill level. However, there are also some problems; mainly related to the computationally intensive brute force search, its 'black box' nature, and the complexity of the underlying technology. Moreover, there is nothing to see other than a final and fixed set of results that may convey little or no understanding of the structure in the data or how the results were obtained. It would be helpful to develop a new approach that retained the best features of both visual interactive graphics and GAM but with none of the disadvantages.

This leads to the conjecture that the GIS end-user is more likely to trust, understand and use an intelligent spatial analysis tool that they can watch as it searches a GIS database for localized patterns in spatial data. The hope is that the end-user can, by observing an animation of the search process, develop a better understanding of what the results mean, and moreover is able to improve this understanding by watching the analysis process being repeated on library data sets that contain different types of known data patterns.

There are four components to this system:

1. An automatic exploratory mechanism that searches spatial data for map patterns under its own control;
2. A measure of search result performance that can be used to guide the search;
3. An interpretation module that statistically processes the results to highlight the important information, taking into account data uncertainty and multiple testing effects; and
4. A means of visualising the search process via animation, emphasising the significant patterns and allowing user interaction with them.

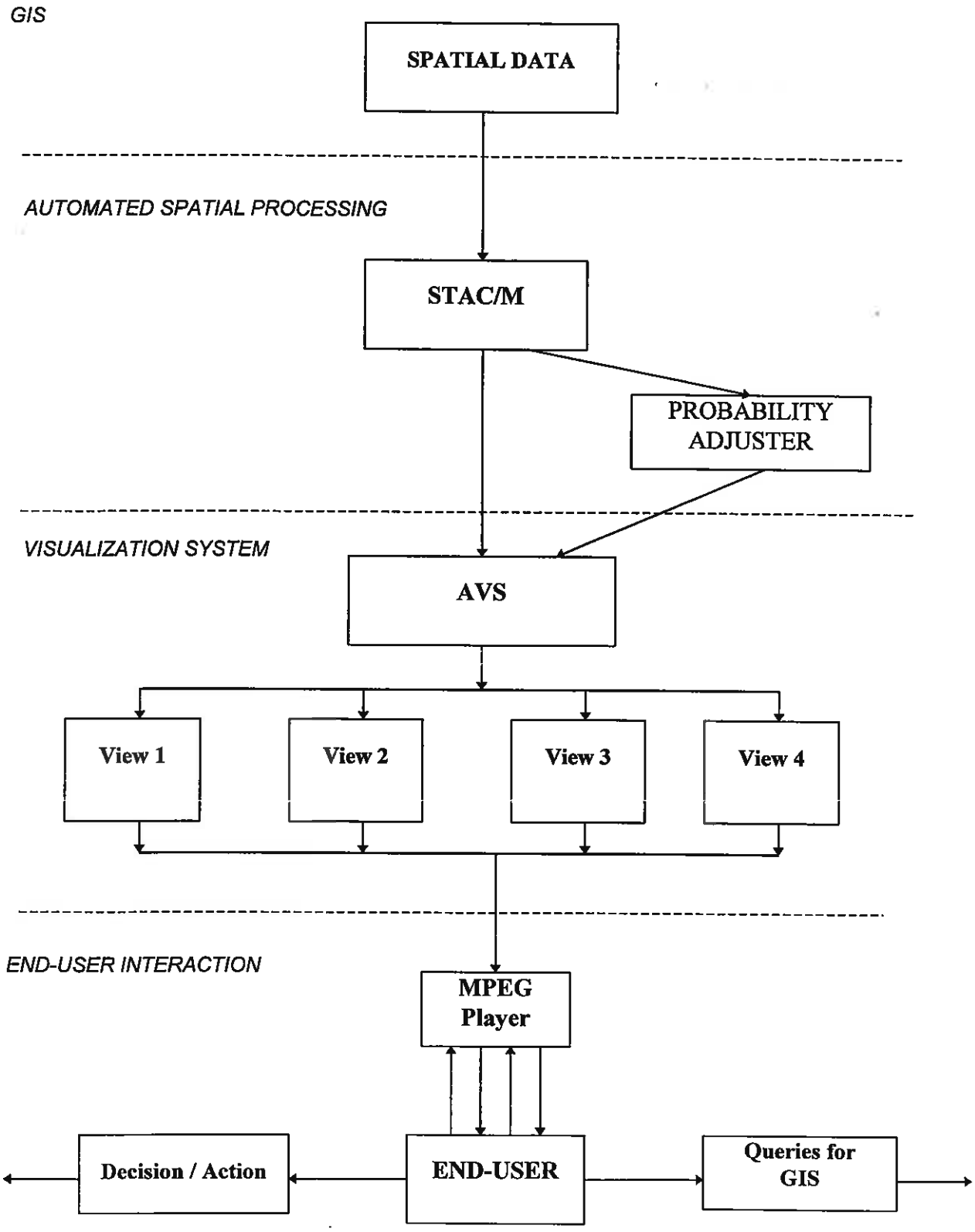
These components need to be complementary. The search process has to be efficient in its coverage of geographic space and has to provide results that help the user understand the nature of any spatial structure being highlighted. In the described system the visualisation takes place only after the analysis is complete. This separation of analysis from visualisation allows the use of highly complex analysis methods without being restricted to the need for instant results. This approach enables the analysis process to be performed on a remote machine, before the end-user interacts with the results of the analysis. The visualisation system is highly interactive, but uses results generated earlier. There is also no assumption that the user is an expert able to interact with a complex statistical analysis process or a visualisation package that may involve a learning curve at least as lengthy as a GIS system. The skill level of the end-user should not constrain the analysis process, but equally nor should the interactive visualisation of analysis results put limits on what the analysis process can do.

The system design assumptions are given in Table 2, and Figure 1 shows the overall system structure.

Table 2 Principal Design Assumptions relating to the system

1.	Separation of visualisation of the spatial analysis process from the spatial analysis process to allow for heterogeneous computing
2.	No real time interaction with the analysis process
3.	No user interaction with the animation creation process
4.	Computational methods used to ensure high quality and consistent results are produced
5.	Built-in self-checking mechanisms to identify inconsequential results
6.	Based on generic principles
7.	Capable of being upgraded in a modular fashion as the constituent technologies improve

Figure 1 System Structure



Although the final visualisation is a map of some kind, the map is not a traditional one, containing vast amounts of detail that impresses the bystander and oppresses the user. Detail that cannot be used is little more than map junk. Only the critical information should be presented and then in a highly simplified graphic format. If you ask the question, 'Where are the cancer clusters?', it should create a display of icons to broadly represent where they are located and their general distribution. There is no need for any further precision. The results are 'real' and significant if the underlying analysis technology is explicitly designed to ensure that this is so. Statistics that only experts understand, and then perhaps do not believe, are simply not relevant here. If the results are worth noting then they should be presented to the user in a form that can be clearly understood. It is an important function of the statistical processing that black and white decisions are made which are likely to be correct. The end-user does not want to know what the likelihood of being right or wrong is. Such knowledge is a distraction, it is seldom understood, and it only confuses. Of course this runs contrary to conventional academic opinion. It is not that detailed statistics are of no use, it is merely that the statistical processing system should be clever enough to take them into account before forming an opinion that the end user can simply understand, trust, and use or reject. Whether the resulting system can be trusted is a matter for objective evaluation, but dodging the issues by dumping a heap of largely incomprehensible statistics on the end-user, and then declaring that interpretation is their responsibility, helps nobody. So a simple iconic presentation of the results that matter is important in developing an end-user oriented approach to spatial analysis. It is designed to be geographically fuzzy rather than precise and to be abstract rather than detailed, because that is what the communication process seems to require.

Building A Smart Point Data Analysis System

The spatial search and analysis module

A useful spatial analysis method able to detect clusters in point referenced disease data is the Kth nearest neighbour method of Besag and Newell (1991). This method is a version of the Geographical Analysis Machine that uses a search based on observed cases. It was developed in the context of detecting what are termed 'data anomalies' in childhood cancer databases but is more generally applicable as a means of analysing point data for evidence of localized clustering. With this method a small positive integer K is chosen and is used to draw circles centred on each case of the disease in turn and passing through the location of its (K-1)th nearest neighbour. In each of the circles the number of individuals at risk is determined and if this number is unusually low (i.e. there are an unusually large number of cases per population at risk, so that resulting Poisson probability is sufficiently small) the circle is plotted on a map of the region. The plotted circles represent the anomalous areas, and a range of maps may be produced using different values of K. The results for any specific K value can be compared because the method takes into account the effect of variations in the underlying population distribution. This simple method suffers from a number of problems. In particular, the spatial search is incomplete as it only looks in regions around existing cases, which implies a highly localized cause and may well fail to adequately represent the locations and extent of any anomalies it finds; it ignores multiple testing with a tendency for a high false positive rate; and it cannot detect small clusters because of the loss of a degree of freedom, which probably only matters most in the analysis of very rare diseases. It is also a statistical method that most GIS end users may find hard to use or understand.

Openshaw (1994d, 1995b) describes the development of what is termed a Space Time Attribute Creature (STAC), an artificial creature that is created to search for patterns in GIS data bases under the control of a Genetic Algorithm (GA). This is used here to develop a much simplified version that implements the Besag-Newell Kth nearest neighbour method. The search is determined by three parameters; X and Y values for the circle centre, and a value for K. The search strategy no longer depends on the location of existing cases but can examine any location, to virtually any level of spatial precision, within the study area. This approach makes fewer arbitrary prior assumptions about the data, which may affect the results. The program that contains this GA implementation is called STAC/M, the Space-Time-Attribute Creature/Movie.

The STAC/M uses a basic genetic algorithm to drive the exploratory search. The three parameters are encoded as bit strings. The X,Y co-ordinates are each stored as 14 bit numbers and the value for K coded in 8 bits. This is sufficient precision for current purposes; giving the X,Y co-ordinates a 10m resolution and allowing values of K in the range from 0 to 256. The basic genetic algorithm operates as follows:

Step 1. Generate a population of PSIZE random X, Y, K bit strings
Step 2. Evaluate each location using a Poisson probability (P) to represent the fitness of any given X,Y,K triplet. The function 1-P is used.
Step 3. Create some children bit strings by applying genetic operators to the data. This occurs as follows: with probability of pc randomly select two parents according to fitness, apply a two point crossover of their bit strings, apply mutation (probability pm), and inversion (with probability pi). Only above average performing bit strings are subjected to crossover and there is an offspring restriction to avoid premature loss of diversity; see Goldberg (1989), Davis (1991) for details.
Step 4. Evaluate the performance of the M new bit strings.
Step 5. Randomly select M worst performing members from the original PSIZE strings and replace.
Step 6. Repeat Steps 3 to 5, a number of times; each constitutes one generation.

Here the GA parameter PSIZE is set at 64, pc is 0.95, pm is 0.01, and pi is 0.02.

GAs are a well established and highly effective optimisation procedures able to rapidly seek out the optimal values of very complex functions. Here the GA is being used to find X, Y, and K values that have the highest probability of being the location of data anomalies; in effect it is seeking to minimize the Type I error probability. Note that there is no explicit test of hypothesis as the Poisson probability is merely being used in a descriptive sense; as a measure of performance that happens to be normalized on the range 0.0 to 1.0. The intention is to use the search process generated in Steps 1 and Step 5 to visualize the spatial analysis function being performed. The version of the GA used here was developed to be gradual rather than rapid, ensuring that the developing search can be watched, and to be informative about the locations of possible multiple different anomalies. The GA searches for a global optimum result, but in the current context it would be more useful if it also located other significant but sub-optimal results. To this end, the GA crossover process was modified to include a spatial neighbourhood effect on the second parent selection process. The search for other sets of pattern can be achieved by removing the data that contributed to the first set of results and then re-running the STAC/M analysis on what is left.

The results for each GA generation are output to a series of ASCII files containing details of each bit string; the location, the search radius, the Poisson probability of the observed number of cases being due to chance; the number of cases within the search radius, and the corresponding population at risk. These files form the base data sets used in the subsequent visualisation of the spatial analysis process.

The statistical processing system

The STAC/M results only really constitute a data screening. The GA driven search process is highly effective in finding localized data anomalies, but some of the seemingly interesting results could well be due to small number effects or data unreliability. One way of handling this problem is to compute a more robust measure of the fitness function used in the search process. A simple and effective modification is to replace the simple Poisson probability calculation by a bootstrapped version. The objective is to avoid extreme results that might mislead the GA due to data uncertainty. The bootstrap operates as follows. For a particular X,Y,K combination retrieve the data that lies within the implicit circle focused on location X,Y that has K cases within it. Re-sample this data subset to generate 2,000 bootstrap samples of the same size, note that the sampling is done with replacement. Now select and use as a measure of fitness the median Poisson probability value. This avoids extremely small probabilities being generated due to small number effects which are highly variable.

Another problem is that of multiple testing. The GA selects the best result by testing multiple hypotheses and this might also produce misleading results; see Hochberg and Tamhane (1987). Geographers have long been guilty of ignoring the multiple testing problem; see for example, Getis and Ord (1992). Indeed, virtually every map generated that presents the results of statistical testing suffers from multiple testing problems; the larger the number of zones the greater the problem and the more misleading the extreme results may become. Once a result is declared significant then many will believe it is real and even attempt to explain it in terms of possible causes. Therefore it is important to correct the results for multiple testing and not mislead the end-user who may well be quite unaware of the problem.

There are a number of potentially suitable methods for correcting for multiple testing. Recently, Benjamini and Hochberg (1995) describe a modified sequential rejective method that is used here. The method works as follows:

Step 1 Assume the testing of N null hypotheses H_1, H_2.. H_N are based on the probability values P_1, P_2,.. P_N. They are sorted into ascending order.

Step 2 Let k be the largest i for which P_i is less than or equal to (i/N)alpha where alpha is the false discovery rate, which can be regarded as equivalent to the usual Type I error significance level.

Step 3 Reject all H_i i=1,2... k

The number of hypotheses being tested is the total number of unique X,Y,K values generated during the search process. The GA search is so effective that when clustered data are analysed the probabilities become extremely small and survive the adjustment procedure. This adjustment procedure can be improved by re-sampling so that the distribution characteristics as well as the extreme values of the distribution of results can be handled; see Westfall and Young (1993). The principal problem is the two to three orders of magnitude increase in compute times so this aspect is left for future development on a faster high performance computing platform.

The visualisation and animation system

Previous computer movies of mapped data have tended to focus on animation as the principal means of analysis. Dorling and Openshaw (1992) and Openshaw et al (1994) describing how to animate space-time data series that have been subject to two- and three-dimensional space-time smoothing. Here the objective is quite different. The purpose is to provide a meaningful visualisation of the spatial analysis process itself. The visualisation is designed to: (1) inform the end-user about the nature of the spatial analysis that has been performed, (2) to offer insight into the structure of the results by watching their emergence, and (3) to highlight areas of particular interest. These visualisations operate on both the raw and multiple testing adjusted data and are designed to offer additional information about the analysis process over and above that provided by the statistical aspects. For example, the geographical distribution and location of data abnormalities as they appear during the search process is itself of considerable interest. Both the trail of extreme results scattered over the map and those locations where the results seem to concentrate are offering very different insights. Some results will reflect the GA search process and how it interacts with the underlying data. Others may provide a meta-analysis of the results that transcends that offered by any of the statistics; for example, the locations of heaps of anomalies compared to much more spread out and even distribution is indicative of two very different types of spatial patterning. The challenge here is to develop the visualisation process so as to emphasize these potentially insightful and pattern reflective aspects so that an intuitive understanding can be developed. They should show the interactions between a spatial analysis tool and the data being analysed. It is argued that this concept can be generalized and extended into higher dimensional space. The aim here is to make visible not the statistical aspects of spatial analysis but the intuitive, the artistic, and the creative. There is no wish to test hypotheses or state results with a precise but probably misleading, and certainly difficult to comprehend, level of quantitative precision. Instead the aim is to provide visualisations of significant information about patterns and process that may be of interest, that would appear to be unusual in relation to reference benchmarks, and which encapsulate intuitive notions regarding the feel of the spatial data under analysis. The hypothesis is that the viewer may gain more understanding of the structure within the spatial data set by visualising the progress of the whole search process, than by only visualising the final set of results.

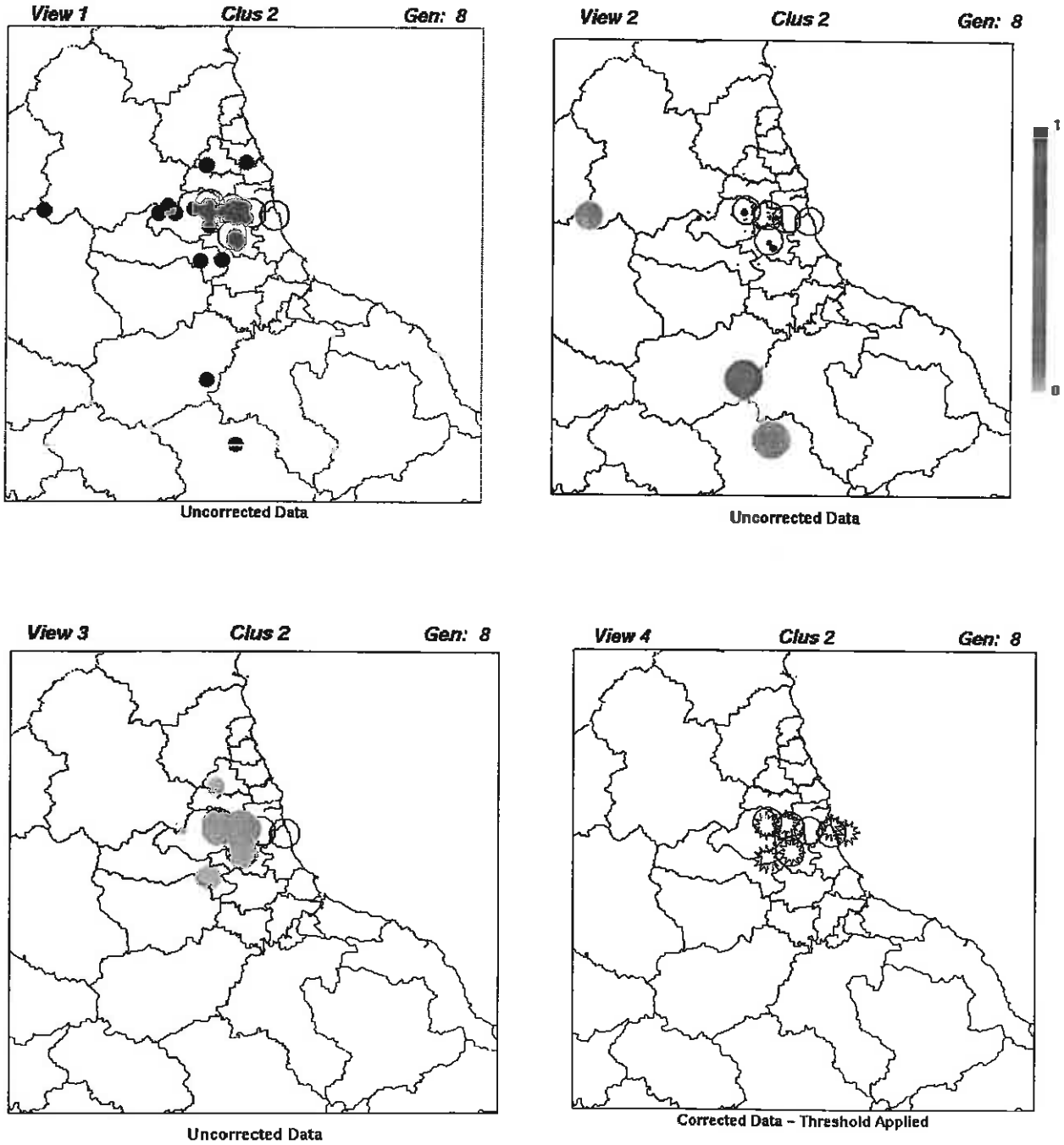
The UNIX-based Application Visualisation System (AVS) is used here because this state-of-the-art scientific visualisation software offers a flexible range of visualisation options, including an animation capability. AVS is one of the most commonly used visualisation software packages. It is designed to be easily extended through user-written modules, and many additional modules are freely available. AVS is also well suited to the rapid prototyping of different visualisation methods - the so-called 'plug, play and throw away!' approach to development. However, the need for an end-user orientation precludes letting the user have access to AVS. That would violate the design objective of ease of use. Instead the plan was to offer the user a series of prepared views of the spatial analysis process that emphasized different aspects, and use AVS purely as a means of generating the animations.

Perrée (1994) describes some early experimentation with AVS, using various visualisation approaches to present animated views which attempted to maximize understanding of the STAC/M results. The initial work

focused on using circles against a background map of the study area. The location of the circles corresponded to the location of the STAC/M X,Y values and the circle radius and colour represented attributes such as the size of the search area and the probability of the result. This initial visualisation method mapped the performance measures to the circle colour using a range of colours from blue to red. The red circles (i.e. the hottest) represented the best search locations and the blue circles (i.e. the coldest) represent the worst. The K values represents a distance value and this is mapped to the circle radius. The resulting visualisation gave the first views of the STAC/M data as a series of various coloured circles against a background map.

These initial studies revealed the need for some improvements in the visualisation of the STAC/M results. As a result of considerable experimentation, a number of visualisations have been developed, each of which emphasizes different aspects of the spatial analysis process. See Figure 2 for examples of the different views.

Figure 2 Four Standard Views
Showing different aspects of the same STAC/M search results.



View 1 The X,Y locations of STAC/M bit strings are mapped as white circles of uniform size. This view helps the end-user to focus on the STAC/M search process itself without the distraction of additional information on bit string performance and search radii.

View 2 The STAC/M bit strings are mapped as coloured circles. The performance of the bit strings is represented by the colour; red for the best and blue for the worst. The circle diameter represents the geographical search radius. This view focuses on where the search process is taking place.

View 3 The bit strings are mapped as red circles, with the circle size corresponding to performance. A non-linear mapping, performance raised to the power of 4, is used to emphasize the location of the best performers.

View 4 The best performing STAC/M bit strings are represented using a star icon. The stars represent the locations of the most significant results. The stars do not overlap each other, and each star may represent one or several high performance bit strings. This iconic representation emphasizes where the best performing STAC/M bit strings are located.

The AVS system is used to create these views and to save them as MPEG format movies that the user can play back later on a variety of different hardware platforms. The animation of the results is an important feature of this approach to intelligent spatial analysis; a sequence of still images on paper cannot convey the results in the way that an animated movie sequence can. The user's task is to watch and identify whether any of the results appear surprising or interesting. The availability of the same standard views for different data sets, including purely synthetic data with known results, allows users to train themselves and to compare real data analysis against library benchmarks. This is useful both to improve their analytical performance and to gain confidence that the technology works.

Results

This new method is evaluated using synthetic data containing known spatial structure. Four different datasets that simulate a rare disease with a frequency similar to that of childhood leukaemia are used. The data consist of a random component with a small systematic clustered part. The clustering process is represented by a two-dimensional gaussian distribution. Data 1 consists of a single big cluster focused at a single site. The method homes straight in on it. It is so obvious that all four views pick it up. The iconic display clearly and unambiguously identifies it. Data 2 is more difficult in that there are a number of much weaker clusters, located fairly close together. The question now is whether or not the methods find one or some or all of them. The task is harder but the iconic display identifies all the clusters. Data 3 is even more difficult in that the clustering is weaker and located in areas that might be expected to be even more difficult to find. Indeed, the iconic display fails to detect any clustering. However, all the other views are giving indications of fairly weak clustering in the correct areas of the map. Data 4 is purely random and the iconic display suggests nothing. The other views either show nothing at all or pick up weak clustering around the edges of the map where the search process has identified locations with very weak clustering due to edge effects. This illustrates the effectiveness of the genetic search process. These unintended data artefacts are detected as such. These results look very different and are much weaker than the results reported for data 3.

The results for the four different datasets are contained in the series of MPEG format movies that were created using the four standard views. These MPEG movies are publicly available via the Centre for Computational Geography's Home Page on the World Wide Web at http://www.geog.leeds.ac.uk/research/ccg.html. The results are also summarized in Table 3.

In datasets 1 to 3, there is a competition between the genetic algorithm's search for extreme results and the need to quickly find ever more extreme results to withstand the effects of the multiple testing adjustments. The cost of this correction is that the sensitivity of the method is reduced and the weak cluster in Data 3 is lost. However, this is probably for the best since in practice cluster detection causes considerable public anxiety so being conservative is a very useful attribute.

Table 3 Summary of results

	Data 1	Data 2	Data 3	Data 4
Number of evaluations	648	1510	1445	1378
Number of uncorrected p<0.01 at generation				
5	1	6	1	1
10	42	12	1	1
15	64	20	1	1
20	64	24	1	1
25	64	24	1	1
Number of corrected p<0.01 at generation				
5	0	4	0	0
10	41	4	0	0
15	64	9	0	0
20	64	8	0	0
25	64	15	0	0
Number of uncorrected p<0.05 at generation				
5	1	6	1	1
10	43	12	1	1
15	64	20	1	1
20	64	24	1	1
25	64	24	1	1
Number of corrected p<0.05 at generation				
5	0	5	0	0
10	42	11	0	0
15	64	19	0	0
20	64	23	0	0
25	64	23	0	0

Conclusions

The paper has outlined a new approach to spatial analysis that is regarded as being of some general utility to GIS. It has argued that current spatial analysis methods are often too complex, too assumption dependent, too precise, and too narrowly focused to be of much practical use to most of the potential end-user community. A new approach is advocated that seeks to combine the benefits of sophisticated analysis methods with the visual appeal of interactive map graphics to devize a simpler and more end-user friendly technology. This different style of approach seeks to combine intelligent automated methods of analysis, that use computational methods to ensure robustness, with a very simple visual presentation of the results based on an iconic interpretation of what they mean. The user is not expected to be, or to become, a statistician, a quantitative geographer or a computer scientist. Instead the technology has been designed to be almost instantly understandable in an intuitively meaningful way. This new approach is designed to support vague types of spatial analysis question; such as 'What should I know about what is happening?', and 'Whereabouts should I be looking?'. The system responds by offering equally abstract responses; such as 'It's over here' , 'It's here, there and everywhere', 'It's very rare'; and so on. The precise map details and data queries that contribute to the detected patterns and relationships, together with the underlying statistics and data, still exist. It's just that they have been abstracted out and removed from view.

The question of how believable the results might be still needs to be answered. The curious can always look more closely at the details that supported the conclusions being presented. However, the technology is designed to be intrinsically safe, so that the user could simply trust the results that are presented. This might seem at first sight to be a recipe for disaster because it is counter to current thinking about how to apply spatial statistical methods. But it can be no worse than not having analysis technology or having to rely on experts as the interface, most of whom are quite unable to communicate what the results mean to end-users, because they are not the end-users. It is far better to build intelligent software that can answer the questions that interest users and at the same time yield reliable results. The most the user should be asked is to specify their required level of certainty that the results are significant, so that only results above this threshold are displayed. This is purely a matter of deciding how risk-averse to be.

The immediate objective here is to specify, in a general way, a spatial analysis technology that operates in an abstract manner and produces understandable high quality results that users can cope with. This would consist of a readily understandable interface to some of the most complex spatial analysis technology available. A longer term task is to demonstrate that users can develop and rely on an intuitive, soft, qualitative, impressionistic understanding of spatial analysis that transcends the standard approaches. It is argued that the general paradigm described here provides a basis for developing a new generation of user friendly, intelligent and appropriate spatial analysis tools relevant to data rich GIS environments.

Appendix - Further Information Available on World Wide Web

The MPEG movies created in this study, and information on further developments and ongoing research, can be accessed from the Centre for Computational Geography's home page on the World Wide Web. The URL is:

http://www.geog.leeds.ac.uk/research/ccg.html

Acknowledgement

Some of this work is based on data provided with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown and Post Office.

Upton, G.J., Fingleton, B., 1985, *Spatial Statistics by Example. Vol 1: Point pattern and quantitative data*, Wiley, New York.
Westfall, P.H., Young, S.S., 1993, *Resampling based multiple testing*, Wiley , New York.

Views expressed in Working Papers
are those of the author(s) and not
necessarily those of The School of
Geography