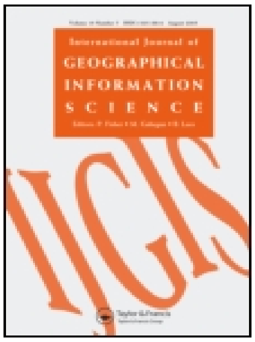


International Journal of Geographical Information System

ISSN: 0269-3798 (Print) (Online) Journal homepage: www.tandfonline.com/journals/tgis19

A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets

STAN OPENSHAW, MARTIN CHARLTON, COLIN WYMER & ALAN CRAFT

To cite this article: STAN OPENSHAW, MARTIN CHARLTON, COLIN WYMER & ALAN CRAFT (1987) A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets, International Journal of Geographical Information System, 1:4, 335-358, DOI: 10.1080/02693798708927821

To link to this article: https://doi.org/10.1080/02693798708927821

Published online: 06 Apr 2007.

INT. J. GEOGRAPHICAL INFORMATION SYSTEMS, 1987, VOL. 1, NO. 4, 335–358

A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets

STAN OPENSHAW and MARTIN CHARLTON

Centre for Urban and Regional Development Studies, Newcastle University,
Newcastle upon Tyne NE1 7RU, England

COLIN WYMER

Department of Town and Country Planning, Newcastle University,
Newcastle upon Tyne NE1 7RU, England

and ALAN CRAFT

Royal Victoria Infirmary, Newcastle upon Tyne, England

Abstract. This paper presents the first of a new generation of spatial analytical technology based on a fusion of statistical, GIS and computational thinking. It describes how to build what is termed a Geographical Analysis Machine (GAM), with high descriptive power. A GAM offers an imaginative new approach to the analysis of point pattern data based on a fully automated process whereby a point data set is explored for evidence of pattern without being unduly affected by predefined areal units or data error. No prior information or specification of particular location-specific hypotheses is required. If geographical data contain strong evidence of pattern in geographical space, then the GAM will find it. This technology is demonstrated by an analysis of data on cancer for northern England.

1. Introduction

The computerization of society and the development of Geographical Information Systems (GIS) are greatly increasing the supply of point-referenced data sets. The commodification of information is also emphasizing the importance of geographical analysis as value-adding technology (Openshaw and Goddard 1987). Yet there are few signs of much renewed methodological interest in the development of better spatial analytical techniques despite growing recognition of their importance. The Chorley Committee of Enquiry into opportunities for Geographic Information Handling in the United Kingdom comments that 'Existing statistical methodology which can be applied to spatial data is little developed: hence there is a need to develop this field'. (Department of the Environment 1987, p. 108). In particular, there is need for spatial statistical techniques that can both be utilized within a GIS framework and make best use of the new opportunities for exploratory analysis being created by GIS developments.

There are seemingly several reasons for this neglect of basic methodological research (Openshaw 1986). There is also a fundamental misunderstanding as to what spatial analytical techniques can and cannot reasonably be expected to achieve. It needs to be recognized that the task of detecting substantive processes from spatial patterns is always going to be difficult or uncertain and may often be impossible. The justification for spatial analysis returns to the more traditional objective analysis of map patterns itself. Indeed, it is precisely here where there is a demand for both better techniques and new opportunities. There is, of course, a well-established set of spatial statistical techniques that would claim to meet these needs (see, for example, Ripley 1981, Diggle 1983, Upton and Fingleton 1985). The problem is that, with the development of GIS technology and the vast growth in spatially-referenced data sets, the available tools are often no longer adequate for the tasks of data processing that they now face. There is also a tendency to limit the spatial analytical tool-kit to established statistical methods and thereby, implicitly, adopt a blinkered view of what is both possible and scientific. It should not be assumed that only statistical techniques operated in the traditional manual fashion are of interest; new methods, based on the linking of statistical spatial analysis to a GIS in an automated way, might well be possible and potentially far more useful as the basis for the exploratory analysis of point data sets.

Openshaw (1987) urged the development of fully-automated geographical analysis systems in order (1) to make best use of the information the data contain without being restricted by the dictates of state-of-the-art theories and methods, (2) to handle efficiently the growing number of important data sets of interest to many different subjects, (3) to allow the development of forms of analysis which can explicitly handle the special characteristics of spatial data without making untenable assumptions for purposes of statistical or mathematical tractability, (4) to better meet the needs of applied users who often want trustworthy answers to quite basic questions concerning spatial patterns and (5) to develop effective and unbiased search techniques capable of both generating and testing hypotheses related to spatial phenomena in an automated manner in an exploratory context under conditions of little and uncertain prior theoretical knowledge. There was also a desire to investigate new forms of spatial analysis which can exploit both the growing richness and size of spatial data sets and the great computational power of modern computers.

The objective then is clear enough. The aim is to develop a new automated approach to the analysis of point pattern data as a means of meeting current and future needs for geographical analysis whilst overcoming many of the statistical and scientific problems that currently exist. The immediate context for this work is the analysis of data on cancer but the general results are widely applicable to other types of point pattern data. In § 2 the thinking that lay behind the idea for a conceptual Geographical Analysis Machine is outlined. Its implementation on computer hardware is described in § 3 and its application to a point data set is discussed in § 4. Finally § 5 provides speculations on future developments.

2. Some outstanding technical problems

Point pattern analysis is complicated because it is not purely a statistical problem but also involves fundamental questions about the manner of conducting analyses and making inferences based on non-experimental research in a geographical domain. Additionally, there is some need to take into explicit account geography as an endogenous variable rather than one which is neutral and exogenous to the chosen method of analysis. These issues are discussed here with particular reference to epidemiological studies of cancer patterns because this is a good example of point data sets which are of considerable topical interest. However, the same set of problems also applies to other point data sets.

The basic question being addressed in many cancer studies is whether or not a point pattern shows signs of clustering in space and, perhaps, in time also. Typically, these studies have used various measures of incidence to identify areas with excess cancer rates and a few localized areas of excess incidence or 'clusters' have been identified (see, for example, Craft *et al.* 1985). There are problems with conventional studies because a number of technical issues undermine the validity of the results. Indeed, purely statistical methods are increasingly being viewed as unable to provide either accurate or unbiased scientific answers to the problem of detecting clusters (see, for example, Croasdale and White 1987). It is helpful, therefore, to enumerate the problems briefly because they provide design objectives that any new form of analysis has to meet.

The list of technical problems includes the following: (i) the results may well have been biased by the conscious or accidental selection of time periods, categories of disease, age-groups, study-regions and (if circular statistics are used) the radial distance; (ii) the hypothesis being tested on spatial data is often invalidated by prior knowledge of the data; (iii) there are often problems associated with determining the significance of results; (iv) the presence and potential impact of errors of both measurement and spatial representation in point data sets are ignored and (v) in formulating point hypotheses, there is a heavy reliance on the prior identification of point locations assumed to be related or causally linked to the point pattern in some way which renders the process of hypothesis formulation biased against whatever point sources are used whilst ignoring all those which are not used.

In addition, there are several more general criticisms which can be applied. In particular, there are major problems in deciding what type of analysis is required. In the real world where the distribution of people, towns and points is not uniform but highly discontinuous, it hardly makes much sense to use statistics which try to summarize a complete map pattern for some arbitrary study region because of the high degree of abstraction from the geographical domain. Boundary effects, due to the inevitably subjective selection of a study region, may also influence the results. What value is there in averaging out all the information contained in a map pattern for a whole study region when it is the 'geography' of the map pattern within the area of interest and the locations of deviations from some overall measure of map pattern that are of most interest? So there needs to be some means of identifying both the existence and the location within the map where some general hypothesis or summary statistic breaks down rather than try to devise better descriptors of whole map patterns.

Finally, it is noted that the problem of *post hoc* hypothesis testing is particularly serious. Leamer (1978) discusses the difficulties caused when hypotheses are created or modified after examining the data on which they are to be tested. It seems that the spatial data should be viewed only as a means of generating, rather than testing, the very hypotheses they inspire. This is a major unresolved problem in spatial analysis where data and exploratory analysis usually precede more detailed subsequent studies, because it renders the results of inferential procedures for testing hypotheses invalid. What is needed is some means of extracting confessions from spatial data sets so as to be confident that the inferences are valid.

3. Building a Geographical Analysis Machine

Many of these problems result from selectivity in the hypotheses being tested and from fears of bias. Traditionally, only a small number of hypotheses could actually be formulated and tested, partly because of the nature of the scientific method widely employed as the basis for inference and partly because of the intellectual effort involved and the utter impracticality of listing all possible hypotheses before any data analysis. The increasing availability of databases in advance of any hypotheses suggests that the construction and modification of post-data models is emerging as an important problem as GIS start to provide vastly improved spatial data sets independent of the formulation of hypotheses and the specification of models. In short, a new era of *ad hoc* inference is dawning.

Openshaw (1987) suggests that the most general and least biased solution to these and other problems of creating knowledge by hypothetico-deductive means is simply to generate and test all possible geographical hypotheses relevant to a particular problem. Bias is excluded and prior knowledge rendered irrelevant because no selectivity is required. The universe of all possible hypotheses also contains the totality of all deductive knowledge about a given problem of spatial analysis, both known and unknown. It is now simply a matter of enumerating this universe and, armed with a procedure for assessing significance, tabulating the distribution of all meaningful or significant results that exist. This approach comes with guarantees that it is totally unbiased with respect to all knowledge both known and as yet undiscovered; it treats all locations equally; it is totally objective and capable of replication and it cannot fail to find any meaningful results if any exist because, within the design limits of the method, all relevant hypotheses have been examined.

This approach is termed a Geographical Analysis Machine (GAM). It was imagined initially that it might take 5–10 years to develop (Openshaw 1987). Six months later a prototype machine for point pattern analysis existed and was working (Openshaw *et al.* 1988). In principle, it is just another variant of an automated modelling system (Openshaw 1988) but this time applied to automating the process of hypothetico-deductive inference in the context of point data analysis. Another important difference is that there is no need for an elaborate search technique, provided the universe of all possible hypotheses is sufficiently small to be fully enumerated (as is usually the case).

The concept of a GAM is based on a very simple but computationally intensive type of statistical spatial analysis, yet it constitutes what is the most advanced hypothetico-deductively inferencing engine yet developed. There are four basic components to a GAM: (1) a spatial hypothesis generator, (2) a procedure for assessing significance, (3) a GIS to handle retrieval of spatial data and (4) a geographical display and map processing system.

3.1. *Generating a universe of all possible hypotheses*

The general hypothesis of interest here is whether there is an excess of observed points within x km of a specific location. A test statistic would be computed for this circular search area and some measure of significance obtained. The concept underlying the GAM is to take this generic hypothesis and generalize the locational aspects by examining circles of all sensible radii for all possible point locations in a given study region. In this way, the universe of spatial hypotheses of this general type can be defined and enumerated. This philosophy can be readily modified to handle different geometries of search area; for example, a wedge shape may be used to represent diffusion of atmospheric material. Here, attention is focused on a circular search region, which is the traditional geometry used in cancer studies. Squares could also be used but offer no significant computational benefits.

The following algorithm is used to generate a universe of all possible circle-based hypotheses.

Step 1

Define an initial two-dimensional grid lattice with the grid size (g). Define also a minimum circle radius, a maximum value and a radial size increment. Assume that a circle is to be located on each grid intersection, thereby completely covering the study region with a regular and even coverage of circles of some initial size. The lattice is sufficiently close-grained so that the circles overlap to a large degree. This requires that the spacing of the grid mesh (g) is some fraction of circle radius, namely, $g=z^*r$ where z is the circle overlap parameter and r the current circle radius.

Step 2

For each lattice point within the study region, retrieve the data needed to compute a spatial pattern measuring test statistic for circles of radius r. Store the results for all locations that pass a significance test set at a given threshold (see § 3.2).

Step 3

Increase the radius of the circle by a specified amount and change the grid mesh to reflect the increase in radius.

Step 4

Repeat steps *2* and *3* until all radii considered relevant have been examined.

In steps *1* and *3*, it is necessary to make the grid mesh a function of the radius of the circle so that the circles overlap. This is required so that the resulting sequence of circles provides a good discrete approximation to testing all possible point locations in two-dimensional space. The overlap between successive circles also serves another very important purpose. The overall effect in moving from one point on the grid lattice to another nearby is to perturb the circle centroids by an amount which is proportional to the radius of the circle. This is most useful because it provides a means of auto-sensitivity analysis which will identify the effects of changes in the data resulting from a slight but scaled shift in the location of circles and thus allows the GAM to take into account the effects of both possible data error and edge effects in the process of spatial data retrieval. This is necessary because few point data sets are both completely accurate and exist as zero-dimensional points in two-dimensional space. It is particularly relevant to data on the incidence of cancer which are typically locationally-referenced in the best practicable but still imprecise manner (for example, using 100 m point references for postcodes). This is a potential problem because uncertainties in the location of even a small number of cancers can produce spurious results.

Step 1 can also be modified to handle edge effects. Ideally, the study region needs to be surrounded either by an internal or by an external corridor equal in width to the radius of the largest circle. Circles located within the study region but overlapping the external boundary corridor will then be allocated the correct data values. This requires either that additional data are available for the corridor region outside the study region or that the analysis stops short of the boundary of the study region. The land–sea boundary offers no problem and can be ignored.

3.2. *A procedure for assessing significance*

For each of the circular search regions it is necessary to compute a test statistic and then decide whether it contains a statistically significant excess of observed points. The choice of test statistic is problematical. Previous epidemiological studies have used both incidence rates and Poisson probabilities (see, for example, Craft *et al.* 1985). Both statistics suffer from problems in assessing the populations at risk when the cancer data relate to a 10- or 20-year period and the census-based counts of the child population refer to one particular year. This problem can be solved only if better population data for small areas allow the adoption of a demographically-sound concept of population at risk (Rees and Wilson 1977). However, this is unlikely to be possible in the short term and the analysis has to proceed in the knowledge that some of the key data are almost certainly in error for perhaps the major part of whatever period is used. This is a problem that cannot be overcome at present although estimates of the possible magnitude of the effects can be obtained by sensitivity analysis, for instance, by exploring the effects of different denominator populations. The aim would be to identify the most robust results given the likely levels of uncertainty in estimates of the population at risk.

An additional difficulty is determining what might be a suitable threshold for measuring significance. One solution is to use Poisson probabilities although this assumes a particular null hypothesis, namely, that the point pattern is generated by a Poisson process. A more general approach is to use a Monte Carlo procedure for significance testing which shifts the decisions back in the direction of the specification of a particular null hypothesis regarding an underlying generating process. The preference is to use a Monte Carlo significance test based on a simple count of points within a circle. The number of observed points in a circle of any specific size based on a particular point location can be directly compared with the number that would be expected under a specific null hypothesis. This count statistic has the benefit of computational simplicity and it is easy to explain.

The standard Monte Carlo procedure for significance testing, used here to identify anomalously high counts of points, was developed by Hope (1968) and is widely used. Given a null hypothesis, it involves ranking the value of an observed test statistic, u_1, amongst a corresponding set of $n-1$ values generated by random sampling from the null distribution of u. When the test statistic is a real number, then the rank of the observed value u_1 amongst the complete set of values (u_i, $i=1,2,\ldots n$) determines an exact significance level for the test since, under the null hypothesis, each of the n possible rankings of u_1 is equally likely. When, as here, the distribution of u is discrete (i.e. integer counts) then tied ranks can occur. Following Besag and Diggle (1977), an estimate of the upper bound for the significance level can be obtained by choosing the most conservative ranking. Since it is not necessary to obtain a precise estimate of the null distribution function of the test statistic, the number of simulated samples (n) can be quite small, typically $n=100$. Indeed it is usually considered that a small value is sufficient (Besag and Diggle 1977), but whether this is so depends on the desired significance level. A value of $n=100$ offers a significance level only down to the 0·01 level. However, lower significance levels can always be obtained by increasing the value of n to 500 or more. This Monte Carlo procedure is most useful because it allows the use of any test statistic and any null hypothesis without being constrained by known distribution theory.

The choice of null hypothesis is very important. The purpose of the GAM is to identify locations where a particular hypothesis breaks down within the study region. The utility of the procedure depends on the extent to which it can handle a wide range of different null hypotheses. However, for purposes of exploratory study, significance in a spatial context has traditionally been based only on an assessment of the probability of obtaining a given level of pattern if the points were randomly distributed over those parts of the study region where data can exist. The suggestion is, therefore, that an initially useful general purpose null hypothesis for bench-mark purposes is to compare the observed pattern of points with what might be expected if they were generated by a Poisson process, and then to test for departures from this state. With this particular null hypothesis, all the Monte Carlo test of significance is doing is to provide an upper bound on the estimate of the Poisson probability of an observed number of points appearing in a particular circle. It would be far easier to calculate this statistic analytically rather than to estimate it indirectly by numerical methods; however, this would preclude the possibility of considering other types of null hypothesis. This limitation is best avoided as further research into the results obtained for different assumptions about types of process is likely to be important in the future.

A final issue is how to generate realistic random samples of point pattern data sets that reflect the chosen null hypothesis. It is not simply a matter of picking random sets of points anywhere within a study region, because the observed point patterns are selections from spatially discontinuous point populations at risk. It is necessary to take into account geographical variations in the distribution and density of whatever is considered to be the population at risk of becoming a member of the point data set under study. For the data on cancer in § 4, these random point distributions are produced by setting up a virtual vector of k elements, where k is the total number of children in the study region. For any fixed number of cancers (m) and under the null hypothesis used here, uniformly distributed random numbers between 1 and k can be used to select m children from this list and then assign them the locational reference of, in this case, the census enumeration district containing their houses. This procedure is repeated to provide n sets of m cancers generated under the null hypothesis of a Poisson generating process. It should be noted that the generation and storage of these simulated point data sets increase the size of the database by a factor of n.

The procedure for testing statistical significance can be summarized as follows. For a circle of specified radius drawn around a specific point, the number of observed points lying within it is obtained. This observed count is then compared with a reference set of 499 different and separately generated point data sets reflecting the null hypothesis. The rank of the observed test statistic gives some measure of the significance of the observed number of points. Each circle that is evaluated requires the retrieval of data from 500 different data sets. It is here that a good GIS is needed.

3.3. *Linking in a KDB tree GIS data structure*

Creating a conceptual GAM is straightforward. Implementing it on a computer is more difficult because of the amount of machine time that is required. Most of the computer time is spent in retrieving data for often millions of circular search regions from a large database containing both the observed data and perhaps several hundred sets of simulated data. For any reasonably large study region, this task of spatial data retrieval would be virtually impossible without a highly sophisticated geographical data structure that allows for very fast retrieval of spatial data.

There are several possibilities. The preferred choice will largely reflect the available hardware. The design objective used here was that the GAM should ultimately be capable of being run on a 32-bit microcomputer as well as on general purpose mainframes. It was therefore assumed that there were restrictions on the amount of memory that would be available to the GAM and that, as a consequence, the data would have to be held on disk. If the GAM was to be ported onto a supercomputer (see § 5.3) then a rather different approach could be adopted.

There are several hierarchical data structures which can be used for the efficient storage and retrieval of point data sets (Samet 1984), although not all are suitable for GAM applications. Probably the most efficient, and also the most neglected, is a multi-dimensional, multi-way search tree known as a KDB tree (Robinson 1981). This allows all the data (observed cancers and randomly generated cancers) to be stored on a single tree and retrieved simultaneously in a highly efficient manner. The KDB tree is attractive because it combines the search efficiency of the KD tree (Bentley 1975) with the I/O efficiency of the B tree (Bayer and McCreight 1972).

The GAM data are characterized by easting and northing coordinates, a simulation number and other basic information that is not currently used but which will be exploited in the future. The data are indexed by, at present, three keys rather than the more usual two. The KDB tree is used here because (i) it can deal with records with many keys and allows all the information of interest to be stored in a single integrated database of keys without affecting performance; (ii) it can perform efficient range searches based on those keys; (iii) it can cope efficiently with very large data sets which must reside on disk; (iv) the tree itself is optimally balanced in the sense that the length of path to any record is the same for all records and usually considerably shorter than in a KD tree and (v) prior knowledge of search requirements may be used to determine the 'shape' of a tree (i.e. depth and spread) when it is built and thus optimize retrieval times.

A KDB tree consists of a set of fixed-size pages, each with an identifier, and initial access to a tree is via its root page. Each page corresponds to a node of the tree and represents a (hyper-) rectangular region of the key space, with the root page representing the total domain of the key space. Pages corresponding to 'leaf' nodes of the tree are called point pages and contain index records, together with pointers to locations in the database. The remaining pages are called region pages. Each entry in a region page (1) defines a rectangular subregion of the region represented by that page and (2) has an associated pointer giving the identifier of the page which represents that subregion. The subregions within a region page are disjoint and their union is the region represented by that page.

Effectively then, a KDB tree produces a hierarchy of partitions of the domain of the key space. Each level in the tree completely covers the domain and the deeper the level, the finer the partition. The precise structure of a KDB tree is determined by the sequence of insertion of index records into the tree. Details of algorithms for insertion, splitting, deletion and range search are to be found in Robinson (1981) and a full description of the implementation of the KDB tree can be found in Wymer *et al.* (1988).

In the current application, there are three key variables, namely, the x and y coordinates of the point and a 'type' variable which distinguishes between observed cancers, child population and simulated cancers. The range search, to which the KDB tree is ideally suited, consistes of finding all index records whose key values lie within specified ranges, i.e., finding all points lying within a given hyper-rectangle. Thus to perform a search of a circular region, it is necessary first to perform a range search using the bounding square of the given circle and then to filter out those points found which do not lie within the required circle. Without use of the KDB tree or something equally efficient, the GAM would suffer from severe computational problems in terms of processing times and be restricted to small data sets. In principle, the KDB-tree-based GAM also opens up the prospect of dedicated micro-based systems some time in the future.

3.4. *A geographical display, evaluation and map processing system*

The GAM approach therefore, offers an examination of all geographical hypotheses of a particular type (in this case circular test statistics) with a built-in automated analyser of data sensitivity and a means of testing for statistical significance. The final step is to put the results back into an explicitly geographical context. This could hardly be simpler and is achieved merely by identifying all the locations with significant circles and presenting them as a map. The resulting map has the advantage of being easily communicated to third parties who do not possess detailed technical knowledge.

The distribution of significant circles provides a visual representation of all locations where the null hypothesis breaks down. This is the principal function offered by the GAM. Meaningful locations where such breakdowns occur are expected to be characterized by a large number of circles drawn around closely-spaced centroids covering a range of different radii. When the null hypothesis is spatial randomness, then a cluster can be visualized as a geographically localized and dense concentration of significant circles. By contrast, it is thought that any spurious circles that survive the significance test will occur in the form of a more scattered distribution of lower intensity.

The maps offer a means of identifying heavy concentrations of significant circles in localized areas which are unusual in their intensity, robust against uncertainty in the data and strong enough to stand out against a background of random noise. These qualities are desirable because the number of hypotheses being evaluated is sufficiently large for a possibly large number of significant circles to exist as type I errors. It is useful, therefore, to probe the patterns further by, for example, focusing only on certain sizes of radii and by raising the minimum number of points in those circles that are plotted. The changes in the resulting map patterns may provide indications of both the persistence and the strength of the resulting patterns of circles as well as some indication of possible scale effects and other clues about causal processes. This reliance on a system of eye-ball information may appear unscientific but the GAM is basically a descriptive technique designed for an exploratory purpose, that is, to identify areas of interest where further work will be necessary to either validate the findings or to test more specific hypotheses. The pattern of significant circles can be further processed to offer a means of validating the results.

One potential problem with the statistical analysis used in the GAM is that it involves testing multiple hypotheses based on overlapping circles with at least some of the data in common. This will influence the estimated significance levels. It is also very difficult to ascertain the total significance of the complete distribution of significant circles as shown on a map. For example, what is the probability of a given pattern of three or four dense concentrations of significant circles occuring by chance? The problem is that little is known as yet about what patterns may be discovered by the GAM in purely random data and this question also involves problems of map pattern detection. Other difficulties may be caused by the discreteness of the test statistic. The risks are thought to be small but further experimentation is needed before they can be quantified and the results of a GAM run considered as having been validated in a purely statistical sense.

A final aspect concerns various prospects of further processing to enhance the visual presentation. One useful technique would be to use a GIS to manipulate the pattern of significant circles, for example, to display only non-overlapping circles. Other forms of map presentation could also be investigated. Cluster analysis could be used to classify the two-dimensional patterns to further simplify the picture. Additionally, it is possible to ask the question 'Where are the most significant circles?' or 'How does this cluster of circles compare with that cluster of circles?'. Some of these questions can be answered by summarizing the map results for different sets of areal units which have connotations with place names. This might help to answer questions as to whether the patterns are influenced by rural–urban differences in the distribution of children. A major cluster in a rural area may survive over a wider range of circle radii than an equivalent cluster in an urban area where its effects may be quickly diluted by a greater density of population at risk. It would also be possible to alter the null hypothesis to take rual–urban differences into account. Various other solutions can be formulated to assess the relative strengths of the circles and thus obtain a quantified measure of local excess. One approach would be to compute a Poisson probability for the significant circles and display the top 10, 20, etc. per cent most significant ones on a map. Another would be to add a third dimension, in the form of a Poisson probability, to the distribution of circles and view the results as a surface or as some other type of three-dimensional display.

4. An application to cancer data for northern England

4.1. *Data*

The development of the GAM concept as a practical tool was largely a reaction to problems with more traditional statistical approaches to point pattern analysis when applied to data on cancer. With these methods, the basic questions as to whether there are any 'real' as distinct from 'illusory' clusters of cancer and their geographical locations still remain largely unanswered. It seems particularly appropriate therefore, to demonstrate the utility of a GAM on some data on cancer for northern England. The study region consists of the Newcastle and Manchester cancer registries and the observed data on cancer relates to the location of residence at the time of diagnosis for all 0–15 year olds in the period 1968–1985. The child population in this study region was 1 544 963 in 1981 and this is used at the level of the census enumeration district (see Rhind 1983) as the basis for generating the random data sets needed by the procedure for significance testing. Both data sets are regarded as two-dimensional point data, the cancers being given postcodes and then converted to 100 m grid references whilst the census enumeration districts are already coded by 100 m point references for their centroids.

The null hypothesis of interest here is that of spatial randomness. Two types of cancer are examined; acute lymphoblastic leukaemia, which is thought to cluster, and Wilms' tumour, which has not previously been reported as showing strong (if any) tendencies for clustering.

4.2. *Search parameters*

The GAM algorithm outlined in § 3 has a number of parameters that have to be set before it can be run. Optimal settings have yet to be determined and it is hoped to do so soon based on simulation experiments with a supercomputer version of GAM. Meanwhile, best estimates are made on the basis of experience so far. It is not thought that these settings are too critical but they might conceivably be data-dependent and further experimentation is needed to understand their effects on the GAM's power of detection.

The minimum and maximum radii of circles need to be specified and these could lie anywhere between 0·1 km (determined by the spatial resolution of the particular data used) and 75 km (determined by size of study region). In fact it seemed reasonable to run GAM using only radii between 1 and 20 km. The lower limit of 1 km and the 1 km increment in radial size used here are due to the large amounts of computer time that are required to run GAM should smaller values be used. It should also be pointed out

Table 1. Estimates of the number of circles to be evaluated for different search parameters.

Number of hypotheses	Radial size increment (km)	Circle overlap parameter
12271889	0·2	0·1
4502960	0·5	0·1
3271971	1·0	0·1
2476478	2·0	0·1
2830474	0·2	0·2
1263889	0·5	0·2
812993†	1·0	0·2
615995	2·0	0·2
1163490	0·2	0·3
527496	0·5	0·3
358497	1·0	0·3
272498	2·0	0·3
641994	0·2	0·4
306498	0·5	0·4
198999	1·0	0·4
152499	2·0	0·4

The total number of hypotheses given here is greater than the number that the GAM would actually evaluate since zero population and zero observed point circles need not be examined.

† Results for these values reported later.

that circles of large radii are of relatively little interest and that the increment in radial size is probably not too critical because a smaller value would duplicate the effect of circle overlap. This parameter was set to 0·2, again mainly to reduce computer time, although this figure is thought to be adequate for a prototype system. Another parameter is the minimum count of observed points in a circle; this is set to 1 but can be changed during the mapping of the results. Finally, it is necessary to set the number of random simulations to be used for the Monte Carlo significance test. Here, *n* is set to 500; a larger value was considered unneccessary.

All these decisions affect the number of hypotheses that are evaluated and the spatial resolution of the results. Table 1 indicates the number of hypotheses that would need to be evaluated for different settings for two of the key GAM search parameters; increment in size of radius and degree of overlap. Currently, it is not known whether the optimal settings for the search parameters will tend towards the smaller values in table 1, or whether the results will prove to be fairly insensitive across a broad range of values. It should also be noted that the total numbers of circles generated are about three times larger than shown in table 1, but that only those circles with centroids located within the study region are currently evaluated. Moreover, only those circles with non-zero counts of children and one or more observed cancers are subjected to the statistical significance test. At present, no use is made of boundary corridors (see § 3.1).

4.3. *Results*

Table 2 reports some statistics relating to the building of the KDB tree used by the GAM with the cancer data for northern England. These times could have been reduced by sorting the data prior to input. The other noteworthy aspect is the very shallow nature of the tree; this is one of the reasons why it is so efficient at spatial data retrieval.

Table 3 gives details of the numbers of hypotheses tested and the run times for the three data sets used here. The computer times are not excessive but this is largely a

Table 2. KDB tree build statistics.

Cancer	Total points	Page size	Total pages	Tree depth	I/O activity	Processing time (seconds)
Leukaemia	412925	400	1595	3	543000	835
Wilms' tumour	93853	87	1714	3	37000	87

Time on Amdahl 5860.

Table 3. GAM run statistics.

Cancer	Total cancers	Number of hypotheses	Processing time (seconds)
Leukaemia	853	812993	22758
Wilms' tumour	163	812993	5595

Time on Amdahl 5860.

reflection of the GAM search parameters that were used. One run with an increment of circle size of 0·2 km and a parameter of circle overlap of 0·1 required 26 hours of CPU time with the same leukaemia data.

One problem with the KDB tree version of GAM is that disk I/O activity rapidly increases once there is a large disparity between sizes of point page region and circle radius. Table 4 shows this effect for a typical GAM run. It begins as highly efficient and then deteriorates as the size of circle increases. Fortunately, there is a simple solution. It is possible to design an adaptive KDB tree which rebuilds itself with a size of point page region appropriate to the size of search areas being used once it predicts that the improvement in efficiency of future retrieval is greater than the cost of reorganization. A simple procedure for monitoring should be able to determine when to rebuild because the GAM's spatial retrieval activity is predictable and the decision rule is that of a classic discounted cost-benefit analysis.

Table 5 summarizes the number of circles that were found to be statistically significant departures from the expected Poisson pattern for radii in the range 1–20 km.

Table 4. Some aspects of KDB tree performance in a typical GAM run.

Circle search radius (km)	Number of search circles	Total number of disk I/Os	Number of disk I/Os per circle
1	465801	19968	0·0429
5	54099	53290	0·985
10	8571	258682	30·2
15	4705	349023	74·2
20	2485	466185	188·0
25	1097	281055	256·0

A disk I/O is a measure of system disk activity.

Table 5. Distribution of significant circles.

Circle radius (km)	Total circles generated	Acute lymphoblastic leukaemia		Wilms' tumour	
		$p=0{\cdot}01$	$p=0{\cdot}002$	$p=0{\cdot}01$	$p=0{\cdot}002$
1	510367	549	164	430	49
2	127572	338	142	298	24
3	56719	311	153	220	17
4	31910	302	139	221	16
5	20428	298	116	171	7
6	14195	273	81	110	2
7	10423	238	53	68	1
8	7983	202	45	36	
9	6304	165	33	29	
10	5112	142	30	20	
11	4207	120	32	12	
12	3557	97	32	10	
13	3028	92	30	10	
14	2602	90	27	12	
15	2269	88	27	9	
16	1993	94	26	11	
17	1766	94	25	7	
18	1580	83	26	4	
19	1408	81	29	5	
20	1280	74	31	2	

There are clear differences between the cluster-prone leukaemia and the other data in terms of the persistence of the distribution of significant circles. However, it should be noted that these results will vary slightly every time the GAM is run because of variability in sampling in the Monte Carlo significance test.

Figures 1 and 2 show plots of those circles significant at the $p=0{\cdot}002$ level for each of the three data sets. This lower-than-usual significance level is the smallest that can be achieved by 499 simulations (namely, $0{\cdot}002=1/500$). It is also thought useful in order to reduce the presence of type I errors in the maps and to focus attention on the most significant circles. The results are remarkable for their clarity. The maps for leukaemia show evidence of intense clustering in a small number of locations whilst, as expected, the Wilms' tumour map patterns are considerably weaker. There are only very slight indications of a few localized clusters. The qualitative differences between these maps are very large indeed, confirming the power of the GAM as a means of detecting departures from a Poisson distribution.

These maps are unique in that this is the first time that the complete set of (nearly) all possible geographical hypotheses of a particular type have been tested and the significant locations displayed. The pattern of significant circles of acute lymphoblastic leukaemia in figure 1 shows the well-known cluster at Seascale but also identifies one other cluster that appears even stronger (Gateshead) but which was not previously identified, probably because of boundary effects in the small-area studies previously used and because attention was focused mainly on the Sellafield area (see Craft *et al.* 1985, Craft and Openshaw 1987). The weaker, less dense clusters at Sedbergh, Whittingham, Bishop Auckland and Macclesfield are probably spurious while the speculatively-identified Springfield cluster is missing. It is thought that the weaker

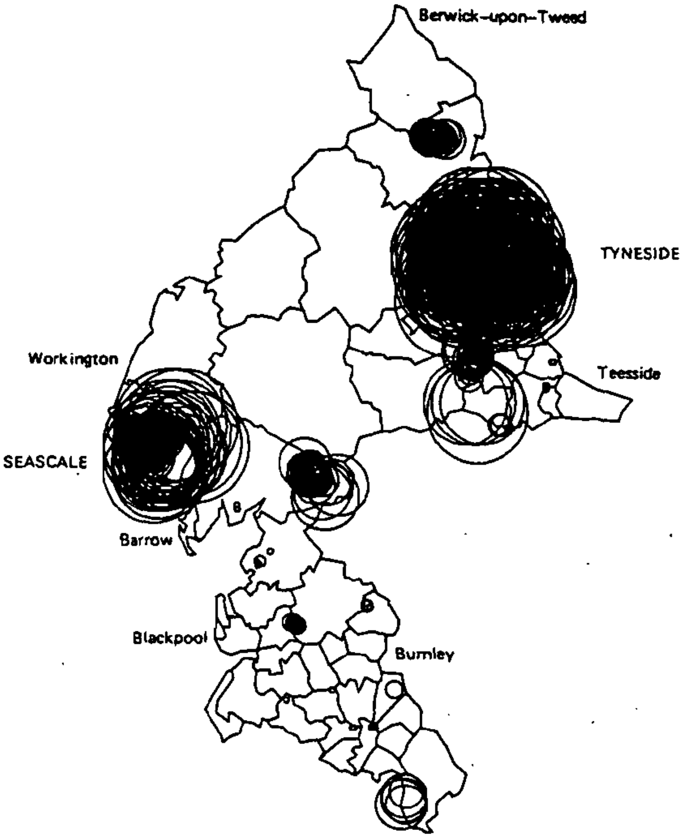


Figure 1. Significant circles at $p=0{\cdot}002$ for acute lymphoblastic leukaemia.

clusters that survive the significance testing may well prove to be sensitive to the different assumptions used to generate the null hypothesis data sets so that, without further research into GAM power levels, it might be unwise to put too much emphasis on these particular areas. However, what is far more remarkable here is the lack of significant circles in both Cleveland and Manchester. This would indicate that there might be some value in running the GAM in reverse so as to identify areas with the greatest significant deficiency of cancer.

The results are surprising, not because they confirm the existence of the cluster at Seascale, but because of the much larger cluster that seems to exist in Tyneside focused on Gateshead. There is no known local major source of low-level radiation in this area so that, for the first time, there would appear to be a possible link with some other form of environmental pollution. Indeed, it might even be that a common non-radiation link might be responsible for both Seascale and Gateshead. This is a matter for further research. However, the significance of finding a major new leukaemia cluster in an area where there are no known local discharges of radiation, at a time when seemingly low-level discharges of radiation from nuclear installations are being blamed for all leukaemia clusters, is considerable.

One of the advantages of a map-based display system is that various filters can be imposed on the data to reveal different aspects of the patterns. Table 6 shows the effect

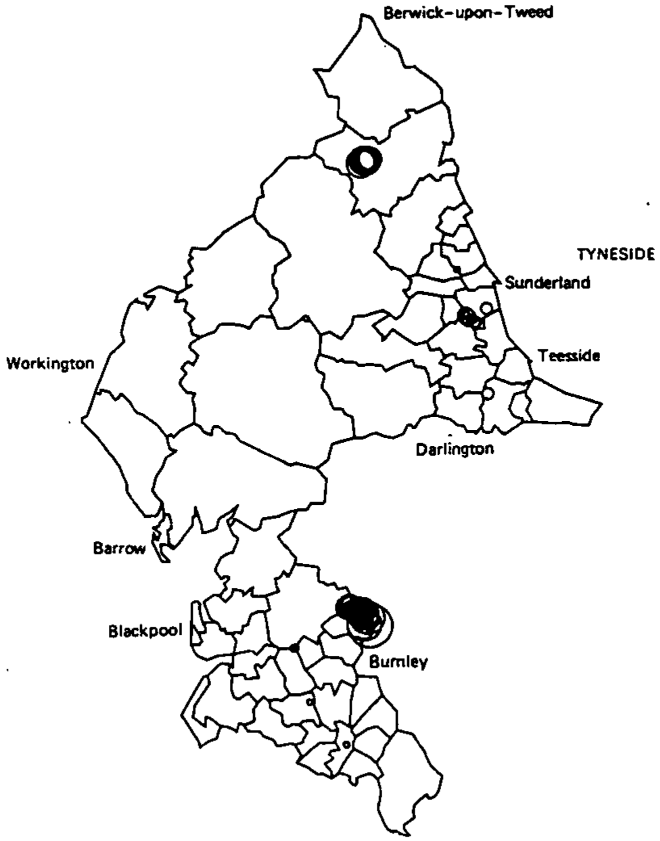
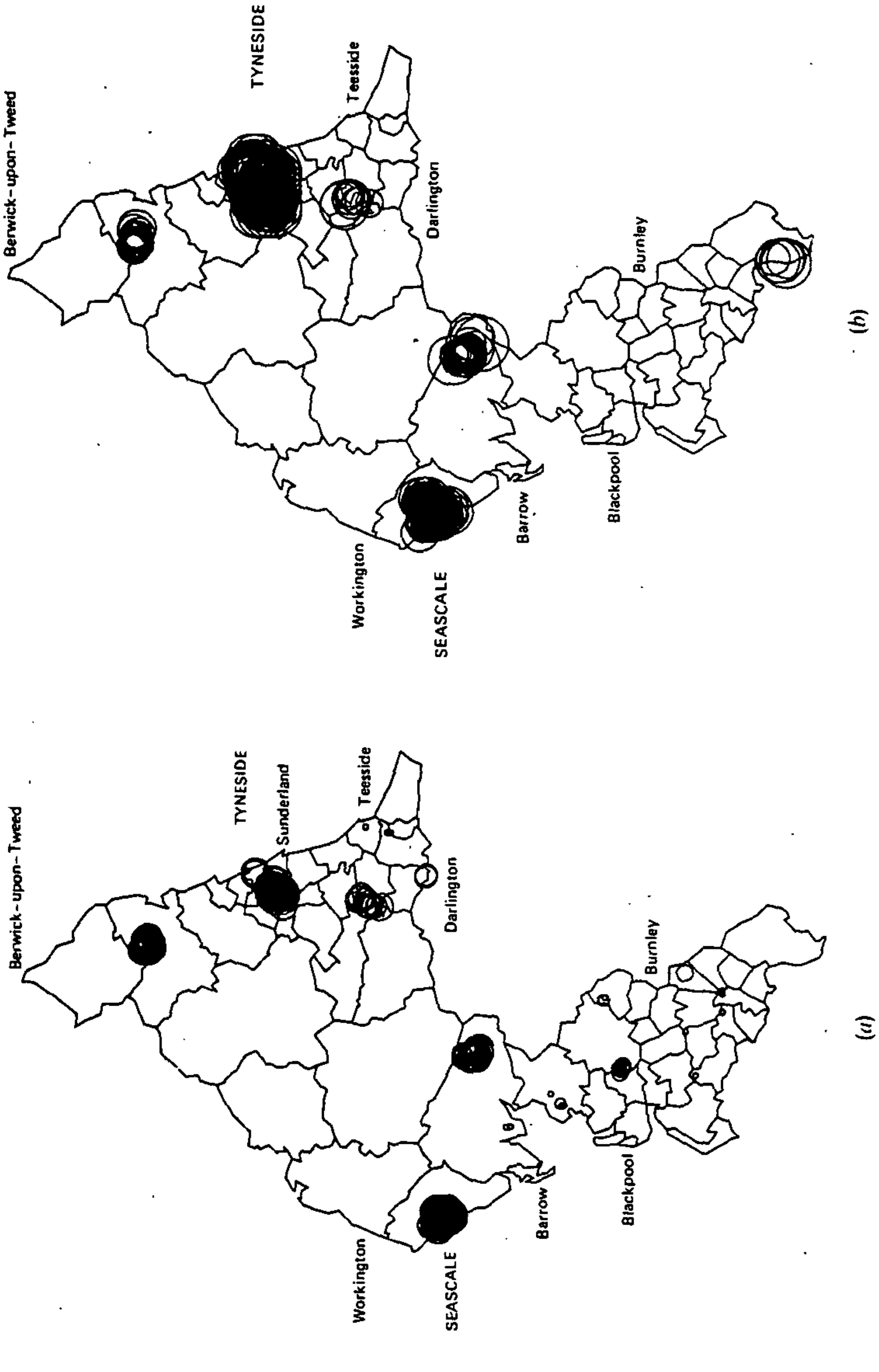


Figure 2. Significant circles at $p = 0{\cdot}002$ for Wilms' tumour.

of deleting all circles with an observed cancer count of less than two. This choice can be justified on the grounds that clusters of cancer need to have at least two observed cases. The counter argument is that any such constraints tend to discriminate against rural areas. The effects are small for the $p = 0{\cdot}002$ level of significance.

Figure 3 shows the effects on the patterns of leukaemia of disaggregation by size of radii; namely, (*a*) 1–5 km, (*b*) 6–10 km, (*c*) 11–15 km and (*d*) 16–20 km. The idea here is that a 'real' cluster will be sufficiently robust to persist across a range of different radii, whereas spurious ones that survive a significance test will be much less able to do so. This is certainly the case since the Wilms' tumour circles are not persistent across many changes of circle radii. The more localized nature of the Seascale cluster, compared with that on Tyneside may well reflect the different scale of possible causes and also differences in the distribution of children.

A final analysis involves trying to measure the relative strength of the various circles. In the light of the discussion in § 3.4, figures 4 (*a*), (*b*), (*c*) and (*d*) show the top 50, 100, 500 and 750 circles from figure 1 respectively when ranked by the Poisson probabilities of the observed number of cancers occuring purely by chance. These maps tend to emphasize the existence of a dense localized core at both Seascale and Gateshead and that parts of both areas have similar levels of elevated risk.


Berwick-upon-Tweed
TYNESIDE
Teesside
Darlington
Burnley
Blackpool
Barrow
Workington
SEASCALE
(b)
Berwick-upon-Tweed
TYNESIDE
Sunderland
Teesside
Darlington
Burnley
Blackpool
Barrow
Workington
SEASCALE
(a)

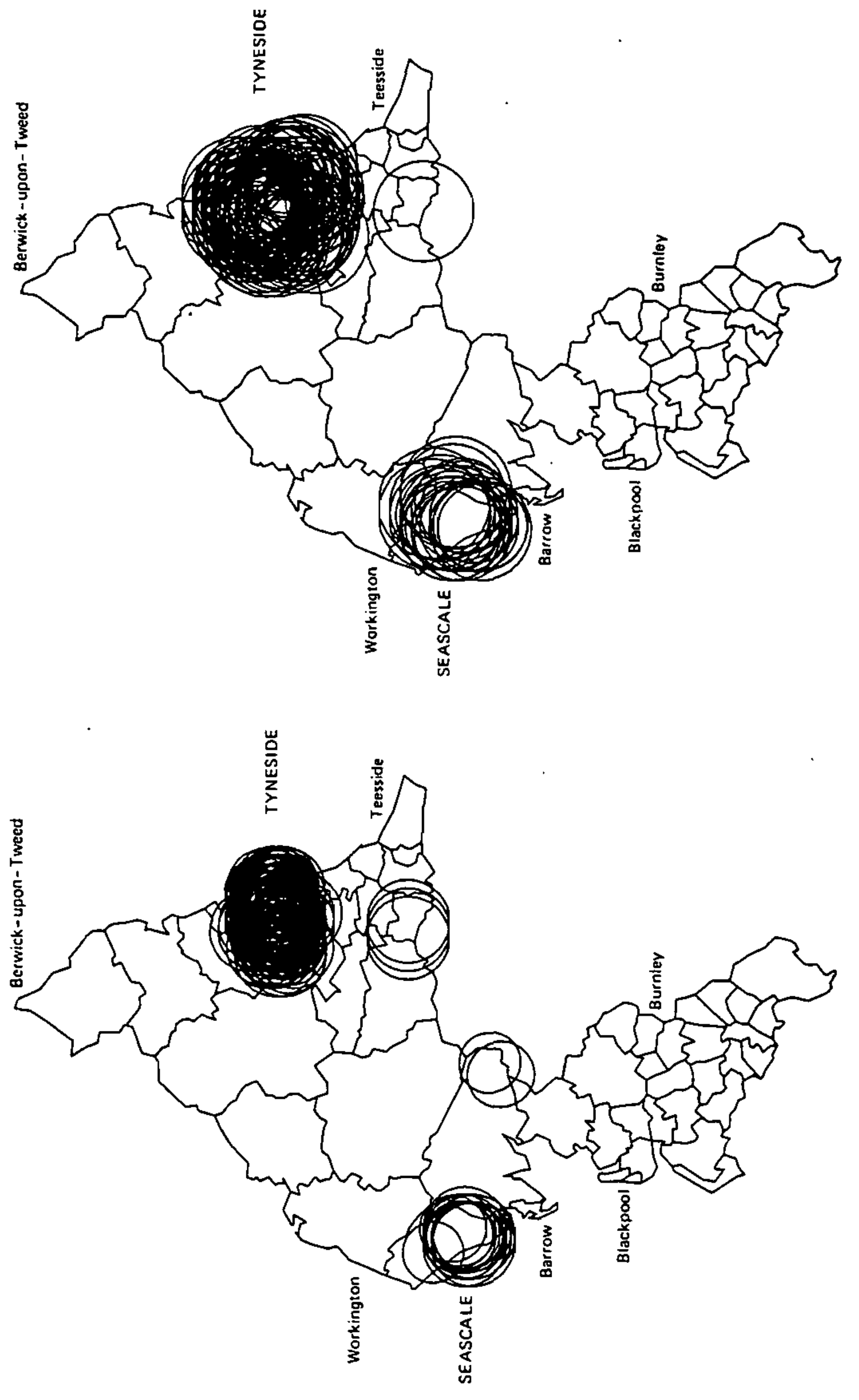
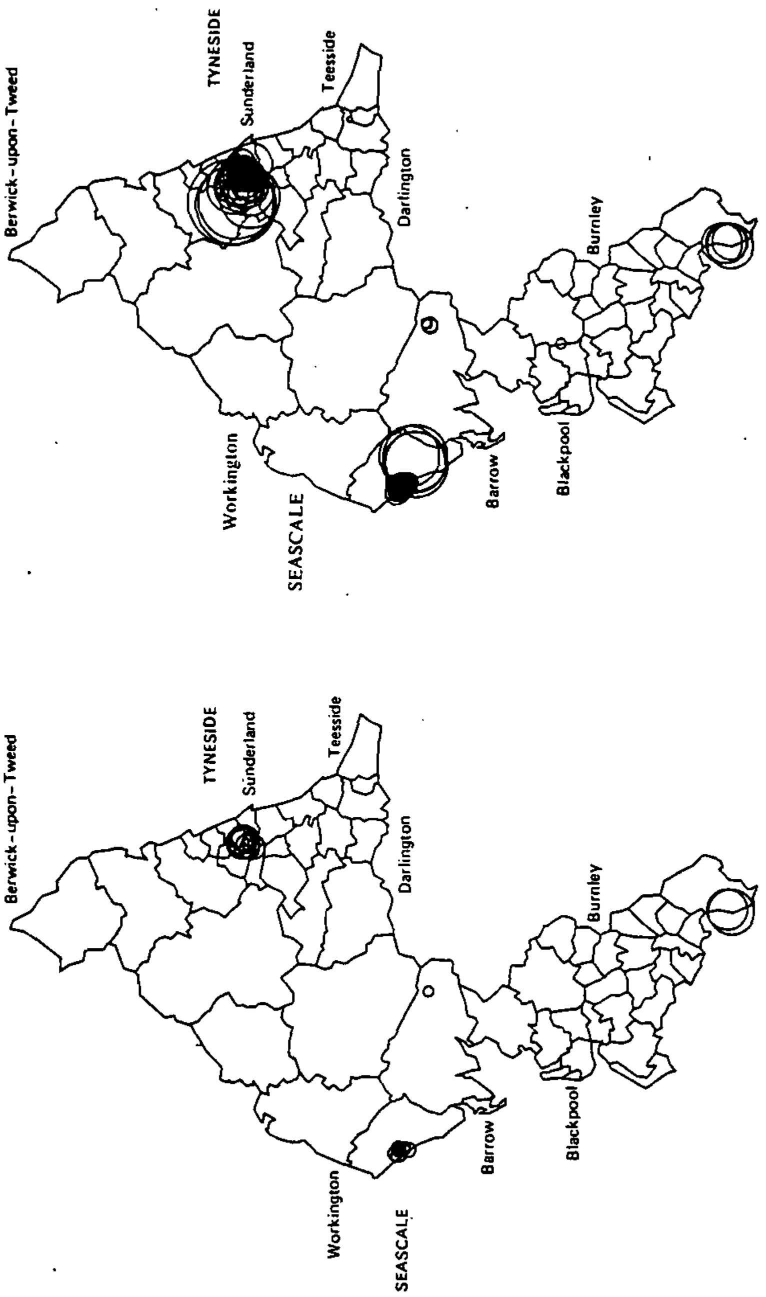


Figure 3. Leukaemia circles significant at $p=0{\cdot}002$ for four different sets of radii. (*a*) 1–5 km, (*b*) 5–10 km, (*c*) 11–15 km and (*d*) 16–20 km.


Berwick-upon-Tweed
TYNESIDE
Sunderland
Teesside
Darlington
Burnley
Workington
SEASCALE
Barrow
Blackpool
(b)
Berwick-upon-Tweed
TYNESIDE
Sunderland
Teesside
Darlington
Burnley
Workington
SEASCALE
Barrow
Blackpool
(a)

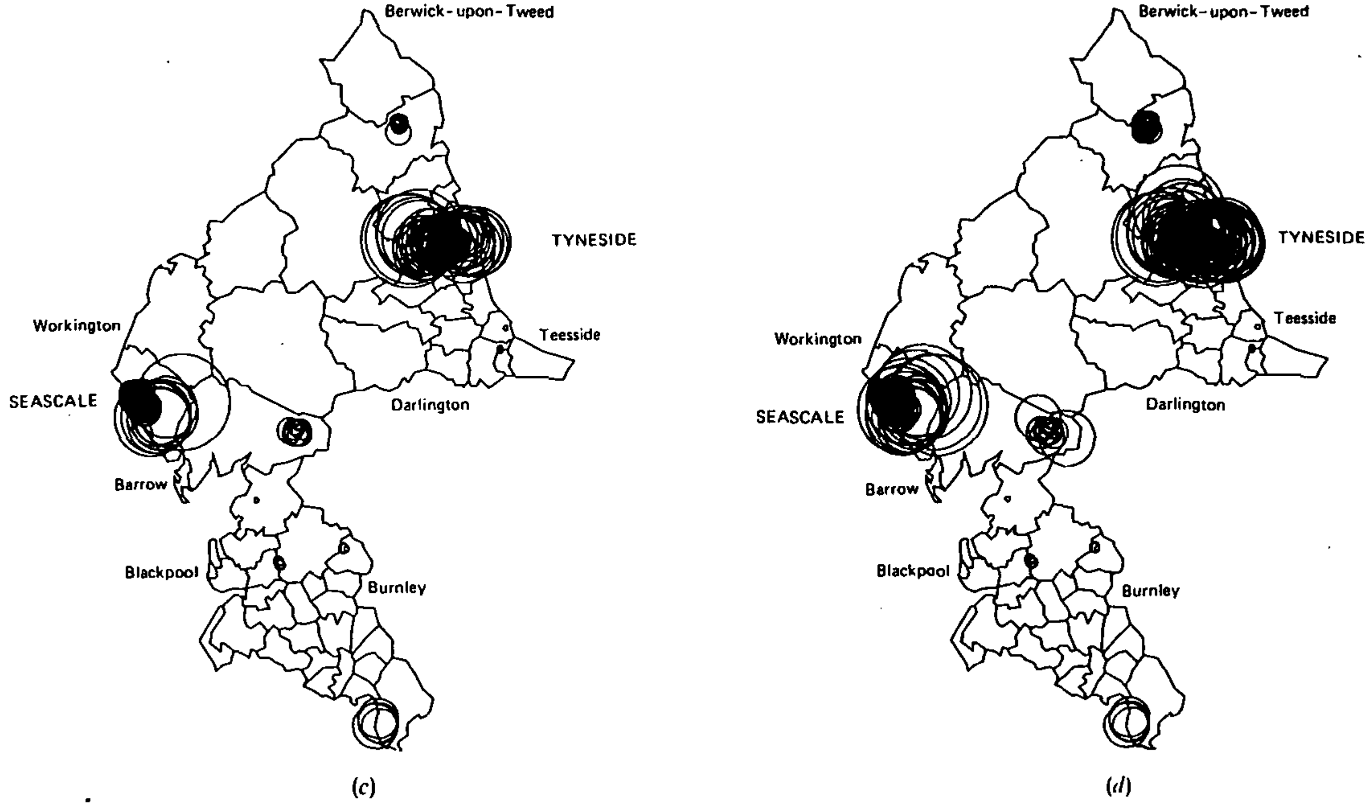


Figure 4. Location of top 50, 250, 500 and 750 circles showing greatest deviations from a Poisson distribution. (*a*) Top 50, (*b*) top 250, (*c*) top 500 and (*d*) top 750.

Table 6. Distribution of significant circles with two or more observed cancers.

Circle radius (km)	Total circles generated	Acute lymphoblastic leukaemia		Wilms' tumour	
		$p=0{\cdot}01$	$p=0{\cdot}002$	$p=0{\cdot}01$	$p=0{\cdot}002$
1	510367	452	153	238	47
2	127572	255	141	123	22
3	56719	281	153	62	17
4	31910	296	137	44	10
5	20428	294	116	25	4
6	14195	270	81	12	2
7	10423	238	53	10	1
8	7983	202	45	3	
9	6304	165	33	6	
10	5112	142	30	4	
11	4207	120	32	1	
12	3557	97	32	3	
13	3028	92	30	5	
14	2602	90	27	9	
15	2269	88	27	8	
16	1993	94	26	10	
17	1766	94	25	7	
18	1580	83	26	4	
19	1408	81	29	5	
20	1280	74	31	2	

5. Further developments of GAM technology

This case study has demonstrated only one aspect of GAM. It concentrated on a data set with no disaggregation and this example does not fully illustrate the versatility and power of this method of point pattern analysis. The GAM can be extended to handle most other aspects of a research design and to include a search for clustering over time. There are no obvious restrictions on the magnitude of the problems or data sets it can handle other than those enforced by computer hardware, disk space and the user's tolerance of long run times. Uniquely, GAM has properties that make it suitable for use on both a supercomputer for processing national data sets, where the explicit parallelism in the grid search and possibilities for fully-vectorized data structures can be exploited, and with continuously-running microcomputers as database scavengers, forever on the alert for the appearance of new disease clusters in local frequently updated databases.

5.1. *Handling permutations of data categorization*

When GAM is run on all the data without disaggregation, it is in effect concentrating on overall geographical effects. It is, of course, possible to run it on various disaggregations of the data; for example by age group and time period when the interest is.in retrospective explorations of historic data. A more disaggregated approach might also greatly reduce the effects of errors in estimates of the populations at risk used in the simulations and greatly improve both the sensitivity of GAM and its ability to suggest further testable hypotheses.

In the case study, the data refers to 0–15 year olds and to an 18-year period. Obviously, the least biased way of handling disaggregations would be to examine all possible permutations of age group and period. Unfortunately there are too many, although it would be possible to examine samples. However, if it is assumed that it is necessary only to split both the period and age variables into sub-groups of contiguous years, then there would be a more manageable number of permutations. In fact, there would be a maximum of $(m/2)^*(m+1)^*(t/2)^*(t+1)$ possible subsets of contiguous age groups and years, where m is the number of year age groups ($m = 15$) and t the number of periods ($t = 18$). If run separately, there would be a requirement for 20 520 GAM runs. This would not take 20 520 times as long to run because the data sets are smaller that those used here, but total computer times would be considerably greater than at present. Fortunately, there is a more effective solution. Nearly all the computer time is associated with data retrieval. There is no reason why all the subset hypotheses cannot be evaluated simultaneously, requiring only a single pass through the data. However, there would now be about 16 682 million hypotheses to evaluate for the standard search parameters used here. Clearly this would be an exercise for a supercomputer but it would seem to be within the ability of current technology to handle this scale of problem.

5.2. *Identifying space–time clustering*

The search for space–time clustering requires only a slightly different approach. There are two different definitions of a space–time cluster. One definition is to look for a significant excess of cancers or points in the same circle over two or more contiguous periods. This implies only a very limited degree of contagion, one that does not extend outside a fixed circular boundary, for example, connected with a point source. If the effect of spatial contagion due, for instance, to a virus is considered, then the radii of circles may well increase for each subsequent period and, of course, in keeping with the GAM philosophy, all permutations of radius would need to be examined. These search models could be programmed for simultaneous evaluation with or without the other forms of search.

5.3. *Running GAM on a supercomputer*

The GAM could also be run on a supercomputer, primarily as a means of reducing run times in order to analyse a given data set more quickly, to allow for disaggregated searches by age group and period, to reduce development times associated with the fine tuning and extensions of the technology, to investigate its power in handling synthetic situations and allow large parts or all of the United Kingdom to be processed in a single run. Initially, a key requirement is for further experimentation. One topic of particular relevance is a study of the effects of different sizes of Monte Carlo samples and other GAM search parameters in order to build up more experience about the GAM's powers of detection and discrimination quickly. Another area of investigation is the power of the test statistic being used. It is most important to know what level or strength of deviation from a Poisson (or some other distribution) can be reliably detected. Only if this is known can greater confidence be placed in the results of epidemiological research. One of the advantages of GAM is that it is possible to measure these detection levels by experimentation and this is an in-built feature of this approach. These experiments can most readily be performed on a supercomputer, leaving more leisurely production runs with optimized search parameter settings for mainframe and micro environments.

It is useful to note briefly those aspects of the GAM that make it intrinsically suitable for use on a supercomputer. In particular, the search process on a grid is suitable for multi-processor machines (such as the Cray XMP) and (even more so) array processors. On a Cray XMP the database would be held in memory rather than on disk and parts of the KDB tree retrieval process are thought to be vectorizable. Research is also under way to develop an alternative, intelligent, totally vectorizable GIS data structure; indeed, initial experiments with a nine-level nested two-dimensional structure are quite promising. The idea is that the data structure monitors itself and adaptively switches the retrieval strategy whenever the current method becomes sub-optimal.

5.4. *Running GAM on a microcomputer*

At the same time, the availability of fast, cheap, powerful 32-bit microcomputers with a hundred megabytes or so of hard-disk storage opens up the prospect of developing dedicated GAMs that concentrate on data analysis for specific areas, e.g. for regional health authorities in the United Kingdom. This is feasible with the KDB tree version of GAM because (1) the needs for memory and disk storage are not excessive; (2) the GIS data structure is portable; (3) the data structure can be tuned for optimal performance on any given hardware; (4) computer times are a function of size of data set and it is far quicker to handle data for a single health region that for larger areas and (5) the supercomputer-dependent disaggregated searches are far less relevant. GAM was also meant to be an automatic system. It was originally developed as a 'run-and-forget' system and its principal applied uses in the future are thought to be as database scavengers. For example, it might be programmed to explore a cancer database for a study area, taking into account all relevant aggregations or disaggregations of the data. It would run for 24 hours a day, 7 days a week. Once a month, its findings would be checked and, when it had finished, it would be loaded with more recent data and re-run. The run times are a function of size of study region, speed of microcomputer and the search parameters used. They could always be reduced by the simple expedient of sharing the total search over two or more machines or by setting the GAM search parameters to slightly cruder values.

This sort of 'database trawler', programmed to be on the lookout for interesting and significant results, offers many advantages. If there are, for example, localized and persistent environmental causes for certain diseases, then the patterns might be spotted quickly enough for them to be of some practical prescriptive use rather than, as is often the case at present, being of only retrospective and historical value.

5.5. *Identifying relationships and testing other hypotheses*

A further application is to use the GAM output as a basis for formulating and testing more specific and detailed hypotheses relating to possible explanatory causes. The main purpose of the GAM is to detect areas of deviation from a null hypothesis. It cannot generate and test new hypotheses simultaneously. However, it can indicate where to concentrate subsequent research effort and it can provide clues for the formulation of new, location-specific, hypotheses, for example, by using GIS to overlay the centroids of significant circles on top of other geographical information and to look for interesting and recurrent relationships with, for example, different types of land use. The results, of course, constitute only circumstantial evidence but they do offer a very focused geographical basis for further research. This ability to point others in the right direction is a useful exploratory function.

6. Conclusions

This paper has outlined the design of a prototype geographical analysis machine, designed specifically for point pattern analysis. The implications of a working GAM may be quite profound in that it is necessary to consider whether all previous analyses of point pattern data in epidemiology, geography and other disciplines should now be re-examined using a GAM, so as to be certain that the conclusions that were drawn were not in fact spurious or biased and that important patterns did not remain undiscovered.

The GAM offers a radically new approach to spatial analysis that combines geostatistical thinking with GIS and a computational philosophy. Its great strength lies in its sophistication of spatial analytical technology. It is seemingly able to detect clusters worthy of further investigations and it is unlikely that any major clusters will be missed. As such, it is the first application of a largely post-statistical technique in a geographical context. However, it is also emphasized that the GAM is offered as a Mark I system and, no doubt, other variants of the basic technology will be developed which are able to offer even more confident results. It would seem to suggest that the alternative to simple science is a far more complex, machine-based science with computational procedures reducing the importance of human imagination, statistical theory and knowledge as the basis for inference. The objective that the GAM addresses is how to improve dramatically the power and usefulness of spatial analytical technology in both exploratory and confirmatory modes of operation. It answers most of the problems raised in § 2 and offers some prospect of a radically new approach to spatial analysis which is very relevant to a GIS environment.

Acknowledgments

The authors wish to acknowledge the helpful comments made by R. Flowerdew, J. M. Birch, R. Wakeford, D. Wilkie, M. Croasdale and K. Binks. Terry Coppock and an anonymous referee also made a number of very useful suggestions. The data came from two cancer registries supported by the North of England Children's Cancer Research Fund and the Cancer Research Campaign. Finally, the research into building GAM was sponsored partly by the Department of Health and Social Security and partly by the Economic and Social Research Council's Northern Region Research Laboratory at Newcastle University.

References

BAYER, R., and MCCREIGHT, E., 1972, Organisation and maintenance of large ordered indexes. *Acta Informatica*, **1**, 173.

BENTLEY, J. L., 1975, Multidimensional binary search trees used for associative searching. *Communications of the Association for Computing Machinery*, **18**, 509.

BESAG, J., and DIGGLE, P. J., 1977, Simple Monte Carlo tests for spatial pattern. *Applied Statistics*, **26**, 327.

CRAFT, A. W., OPENSHAW, S., and BIRCH, J. M., 1985, Childhood cancer in the Northern Region, 1968–82: incidence in small geographical areas. *Journal of Epidemiology and Community Health*, **39**, 53.

CRAFT, A. W., and OPENSHAW, S., 1987, Children, radiation, cancer and the Sellafield nuclear reprocessing plant. In *Nuclear Power in Crisis*, edited by A. Blowers and D. Pepper (London: Croom Helm).

CROASDALE, M. R., and WHITE, A. A. L., 1987, A critical review of statistical evaluations of the clustering of rare diseases with particular reference to the frequency of cancers around nuclear sites in Great Britain. Presented at the British Nuclear Energy Society Conferences held in London in May 1987 (mimeo).

Department of the Environment, 1987, *Handling Geographic Information.* Report of the Committee of Enquiry chaired by Lord Chorley (London: Her Majesty's Stationery Office).

Diggle, P. J., 1983, *Statistical Analysis of Spatial Point Patterns* (New York: Academic Press).

Hope, A. C. A., 1968, A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society,* B, **30,** 582.

Leamer, E. E., 1978, *Specification Searchers:* ad hoc *Inference with Nonexperimental Data* (New York: Wiley).

Openshaw, S., 1986, Modelling relevancy. *Environment and Planning,* A, **18,** 143.

Openshaw, S., 1987, An automated geographical analysis system. *Environment and Planning* A, **19,** 431.

Openshaw, S., 1988, Building an automated modelling system to explore a universe of spatial interaction models. *Geographical Analysis* (in the press).

Openshaw, S., Charlton, M. E., and Craft, A. W., 1988, Searching for cancer clusters using a Geographical Analysis Machine. *Papers and Proceedings of the Regional Science Association* (in the press).

Openshaw, S., and Goddard, J. B., 1987, Some implications of the commodification of information and the emerging information economy for applied geographical analysis in the United Kingdom. *Environment and Planning* A, **19,** 1423.

Rees, P. H., and Wilson, A. G., 1977, *Spatial Population Analysis* (London: Arnold).

Rhind, D., 1983, *A Census Users Handbook* (London: Methuen).

Ripley, B., 1981, *Spatial Statistics* (Chichester, Sussex: Wiley).

Robinson, J. T., 1981, The KDB Tree: A search structure for large multidimensional indexes. Research report CMU-CS-81-106, Carnegie-Mellon University, Pittsburgh, U.S.A.

Samet, H., 1984, The quadtree and related hierarchical data structures. *Association for Computing Machinery Computing Surveys,* **16,** 187.

Upton, G., and Fingleton, B., 1985, *Spatial Data Analysis by Example* (London: Wiley).

Wymer, C., Charlton, M. E., and Openshaw, S., 1988, An implementation of the KDB tree and its application to the GAM. Northern Regional Research Laboratory research report, Centre for Urban and Regional Development Studies, University of Newcastle upon Tyne, England (in preparation).