Review of Cluster Analysis Software

by

Luc Anselin, Ph.D.

Anselin and Associates, LLC.
408 W. Nevada Street
Urbana, IL 61801
anselin@mac.com

September 27, 2004

The GIS Ad Hoc Committee determined that it would be useful to have a review of cluster analysis software programs for cancer registries. It enlisted Luc Anselin, PhD, of the University of Illinois at Urbana-Champaign to prepare this report, working in close consultation with the Committee. The Committee selected four software packages for spatial analysis of cancer clusters to be reviewed based on the following criteria: 1) they are free and/or open source, 2) they are up to date and under active development, 3) they come with a manual and documentation, 4) they are downloadable from the Internet, and 5) they work in a Microsoft Windows operating system. The report, "Review of Cluster Analysis Software," was accepted by the NAACCR Board of Directors on August 18, 2004. For cancer registries that are considering using software for cluster analysis, we believe that this report will be a valuable resource.

Executive Summary

This report reviews four free software packages (as of August 2003) that can be used in a spatial analysis of cancer clusters: *CrimeStat*, *GeoDa*, *SaTScan*, and packages developed in the open source R programming environment (specifically, the packages *spatial*, *splancs*, *spatstat* and *spdep*). The selection of these specific packages was carried out in consultation with and agreed upon by NAACCR and based on the following criteria: (i) they were free and/or open source; (ii) they are up to date and under active development; (iii) they come with a manual and documentation; (iv) they are downloadable from the internet; and (v) they work in a Microsoft Windows operating system.

To carry out an exploratory analysis of cancer clusters, an effective software environment should include (i) efficient data input; (ii) spatial data manipulation functions; (iii) descriptive spatial statistics; (iv) point pattern analysis; (v) spatial autocorrelation analysis; (vi) visualization of the results; and (vii) flexible program output. None of the four packages reviewed satisfies all these criteria, although they come quite close and to some extent they are complementary. It should be noted that there is no single commercial alternative that meets these criteria (commercial software was not included in the review).

Methods to detect cancer clusters should at least include Kulldorff's scan statistic and the Local Moran. Both of these allow for the assessment of significance as well as the identification of the location of clusters. The *SaTScan* software is the only one of the four reviewed that includes Kulldorf's scan statistic. However, this is also the only method implemented in *SaTScan*. Also, its data input and output capabilities are somewhat cumbersome. The other three software packages include the Local Moran. It is thus recommended that a basic software infrastructure for cancer cluster detection should consist of *SaTScan* and at least one of the other three packages.

The packages differ in terms of range of functionality, ease of use and statistical sophistication required from the user. At the introductory end of the spectrum are *CrimeStat* and *GeoDa*, which are relatively accessible, even to non-expert (non-statistician) users. *Crimestat* has more extensive point pattern analysis functionality, whereas *GeoDa* is geared towards interactive data exploration and spatial autocorrelation analysis. While there is some overlap between the two packages, they are largely

complementary. At the high end of the spectrum are the packages developed in the R software environment. They are the most demanding in terms of technical sophistication (both in terms of statistical background and with respect to programming skills), and only appropriate for power users. However, its extensibility and the existence of a large library of sophisticated statistical functions makes R a powerful platform to implement a strategy that goes beyond exploratory analysis. In terms of ease of use, *SaTScan* is somewhere in the middle, requiring a more than elementary understanding of the underlying scan statistic, but otherwise fairly straightforward to use.

Of the four packages, *GeoDa* is the only one that implements mapping and visualization and does not require external GIS software for this functionality. The other packages do not include explicit visualization tools, but export results in files for import into GIS or other mapping software. Without such GIS software, they are of limited usefulness in a *spatial analysis* strategy.

Postscript (3/2004)

All four packages are under active development at the time this review was written (August 2003). The R packages undergo continuous change and new ones appear all the time. In this respect, it may be worthwhile to note that in early 2004, the experimental R package DCluster was released, containing several cluster detection methods and scan statistics (http://sal.agecon.uiuc.edu/csiss/Rgeo/index.html). Also, SaTScan released Version 4.03 in February 2004, GeoDa released Version 0.95-i in February 2004, (which includes spatial regression), and CrimeStat is likely to release a version 3 by summer 2004. On the commercial side, ESRI's ArcGIS 9.0, to be released by mid 2004, will include a "geoprocessing toolbox" covering some methods of exploratory spatial data analysis and point pattern analysis.

Suggested citation: Anselin L. Review of Cluster Analysis Software. Springfield (IL): North American Association of Central Cancer Registries, September 2004, 21 pp.

Review of Cluster Analysis Software

Introduction

This report includes a review of selected software for the analysis of spatial clusters, with particular reference to the detection of cancer clusters. The software is assessed with respect to its capabilities to address three different questions:

- given a data set, where are the potential cancer clusters?
- given that there may be a cluster, what is its statistical significance?
- given a suspect location, is there a cluster found around it?

The software is evaluated with respect to a number of characteristics, such as the level of expertise required, the extent to which it depends on other software, and the statistical assumptions used.

There are several software packages available that carry out spatial cluster analysis in one form or another.[1] The current review is limited to four packages that have been developed specifically to implement methods for *spatial* data analysis: [2]

- *CrimeStat* 2.0 (5/02), developed by Ned Levine, with support form the National Institute of Justice,
  - available at http://www.icpsr.umich.edu/NACJD/crimestat.html;
- *GeoDa* 0.9.3.g (8/5/03), developed by Luc Anselin through the Center for Spatially Integrated Social Science at the University of Illinois, Urbana-Champaign, supported by the National Science Foundation,
  - available at http://sal.agecon.uiuc.edu/csiss/geoda.html;
- *SaTScan* 3.1.2 (7/16/03), developed by Martin Kulldorff, with support from the Alfred P. Sloan Foundation and the National Cancer Institute,
  - available at http://www.satscan.org;

---

[1] Links to the web pages of spatial data analysis software can be found at http://www.csiss.org/clearinghouse/.

[2] **Disclosure**: The author of this report (Luc Anselin) is the main developer of *GeoDa*, one of the packages discussed. He has also been a consultant to Ned Levine and Associates, with respect to the incorporation of LISA statistics in *CrimeStat*. In addition, he is an active participant in the *R-Geo* network to promote open source spatial statistical software in the R environment, and has contributed to the R *spdep* package (considered in the current review). He has no relationship with the *SaTScan* development group.

- and a collection of "packages" that can be used in the R open source environment for statistical computing (Version 1.7.1), specifically *spatial* (included in the VR package, Version 7.1-8 of 6/10/03), *spatstat* (Version 1.3-2 of 5/5/03), *spdep* (Version 0.1-13 of 8/6/03), and *splancs* (Version 2.01-9 of 7/29/02)
  - available at http://cran.r-project.org.

CrimeStat, *GeoDa* and *SaTScan* are freestanding, and can be installed directly after downloading (using the standard setup.exe type installer). R can be installed as a binary (using a standard installer), or compiled from the source code. The R "packages" with spatial data analysis functionality are not necessarily included in the base R installation and may need to be installed separately.[3] Of the four packages, R is cross-platform, and runs identically on Windows, Linux/Unix and MacOS platforms.

The selection of these specific software packages was carried out in consultation with and agreed upon by NAACCR and based on the following five criteria:

- the software had to be free and/or open source;
- the software had to be up to date and under active development;
- the software had to come with a manual and documentation;
- the software had to be downloadable from the internet;
- the software had to work in a Microsoft Windows operating system.

The four packages considered here meet these criteria. Other free software that addresses spatial cluster analysis includes the DOS-based Cluster 3.1, developed by the Agency for Toxic Substances and Disease Registry (http://www.atsdr.cdc.gov/HS/cluster.html). However, it no longer seems to be under development and was not available for downloading at the time of the writing of this report. Commercial statistical software to carry out spatial cluster analysis includes S+SpatialStats (http://www.insightful.com/) and ClusterSeer (http://www.terraseer.com), among others.

---

[3] On a MS Windows platform, this can be done by selecting an item on a menu, as well as through explicit R commands entered on the R command line. Alternatively, the R packages can be compiled from their source code. The binary installation is similar in order of "difficulty" (or, rather, ease) to the installation of the other software packages. However, installation from source can be challenging and requires a good understanding of the workings of C and Fortran compilers.

The remainder of the report starts with a brief and informal review of the methodological background, followed by an overview of software requirements for the implementation of cancer cluster analysis. Next is a brief overview of the main features for each of the packages. The report closes with a summary evaluation. A detailed, feature by feature comparison of the four software environments is given in an Appendix.

Brief Methodological Background

Although an extensive methodological review is beyond the scope of the current report, it is useful to outline some of the main issues in order to better put the different software packages into perspective. In the analysis of cancer clusters, a first distinction pertains to the type of *spatial data* that are analyzed. There is an important difference between a situation where data on individual events are available, such as the addresses of patients or deceased persons, represented as *points*, and the case where only aggregated data can be analyzed. The latter is the situation where individual data points have been summarized by administrative units, or *areal units*, and expressed either as a count of events, or as a rate or proportion,[4] in defined geographic areas such as census tracts, zip codes, municipalities, counties, etc.

While both situations ultimately pertain to event data (e.g., a treatment, a screening, a diagnosis, a death), the type of analysis and the interpretation of the results are slightly different. When the data are "points," interest focuses on the extent to which points are "closer" than they would be in a reference situation.[5] When the data are aggregated, a cluster typically refers to the situation where an areal unit is surrounded by other units that are more similar to it than would be the case randomly, which is termed *positive spatial autocorrelation*. More precisely, when an area with high risk is surrounded by other areas with high risk, more so than would be the case by chance (i.e., at a given significance level), this is referred to as a *cluster*. Strictly speaking, the "cluster" consists of the location and its neighbors, although the latter are not always

---

[4] The rate is an estimate of the underlying *risk*, obtained by dividing the number of events by the size of the population at risk, possibly adjusted for differences in age/sex distribution.

[5] Strictly speaking, one is equally interested in the situation where points are "further apart" than they would be in the reference distribution. However, this is less relevant in the analysis of cancer "clusters."

made explicit. A slightly different interpretation is to consider a "cluster" to be an areal unit (e.g., a county) where the estimated risk is much higher than in other locations, irrespective of the neighbors. Since the events happen within the same areal unit, they are "clustered," although the degree of intra-unit clustering is related to the spatial scale. Typically, this situation is not called a cluster, but rather referred to as an "outlier." A special case (and not necessarily the same location) is a *spatial outlier*, where an areal unit has a much higher or much lower risk than its "neighbors."

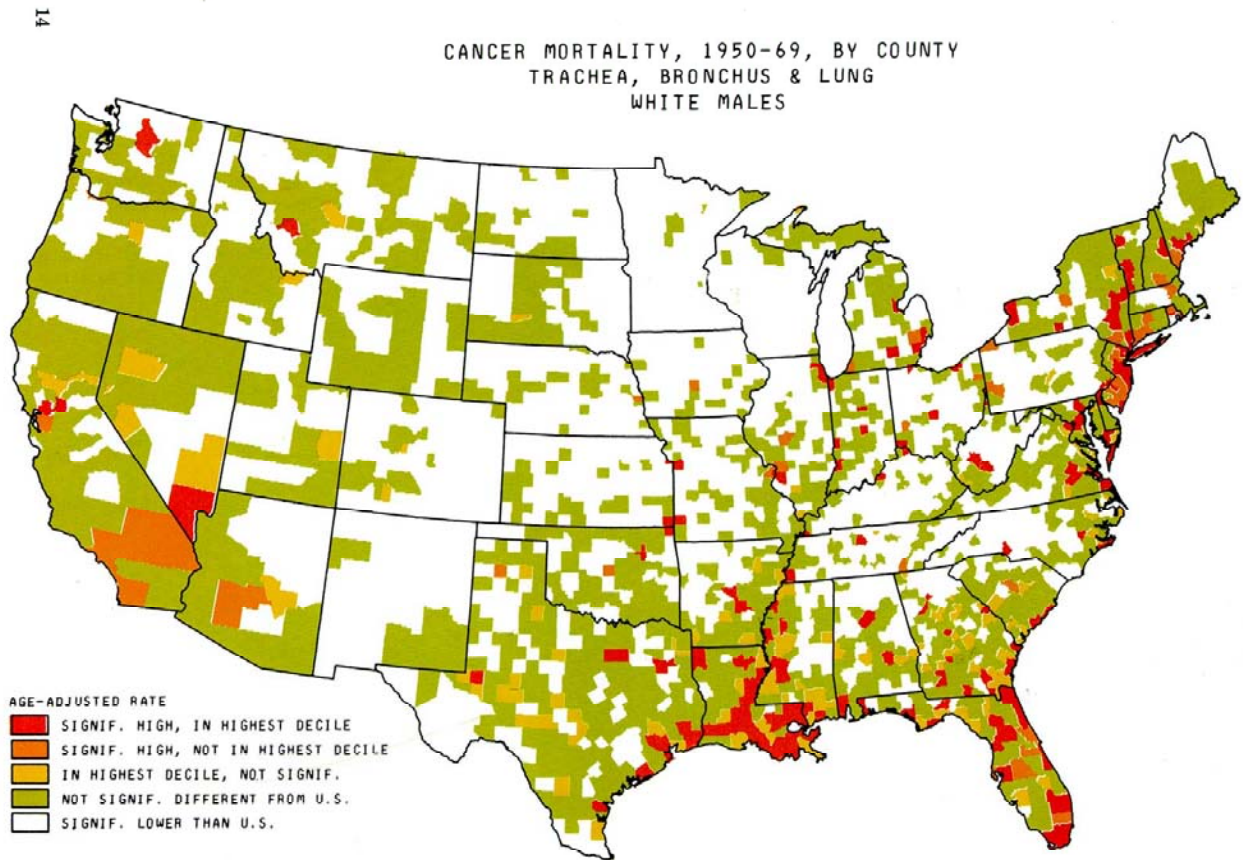**Figure 1. This county level map of lung cancer mortality in the U.S. shows a cluster in Louisiana, a spatial outlier in Montana, and high rates along coastal areas associated with asbestos exposure in shipyard workers during WWII.**

Source: Mason TJ, McKay FW, Hoover R, Blot WJ, Fraumeni, JF Jr. Atlas of Cancer Mortality for U.S. Counties: 1950-1969. DHEW Publ. No. (NIH) 75-780. Washington, DC, 1975.

The reference situation is *spatial randomness*, which means that location, as such, is irrelevant for the analysis, and points could be anywhere with equal probability, or values could occur at any location equally likely. This spatial randomness needs to be specified in a mathematical model, so that a decision can be made about the extent to which the observed values (pattern) are compatible with this null hypothesis. For points, the reference is traditionally a homogeneous Poisson process; for areas it is an uncorrelated normal distribution, or the assumption of equal probability.

In practice, even when there is no significant patterning, points may not *appear* equally likely in all locations, due to differences in population density, road networks, or natural features (lakes, rivers). Similarly, when the risk is estimated for areal units with a very different population at risk (e.g., both small counties and large counties), the precision of the estimate is affected, and "high" values may be spuriously labeled as extreme. Cluster tests should properly take this heterogeneity into account, either by comparing the observed pattern to a "control" group (the case-control approach) or by smoothing the observed rates to correct for the extra variability.

A second distinction is between tests for *clustering* and tests for *clusters*. The former are also referred to as *global* tests, whereas the latter are *local*. A test for *clustering* is designed to reject the null hypothesis of spatial randomness for the data set as a whole. In other words, the objective is to find evidence of significant patterning, but without designating particular locations as clustered. In contrast, a test for *clusters* is local, in the sense that they are designed to identify the locations of the clusters (or of the spatial outliers). Among the local tests, there is a further distinction when a test is *focused* with respect to a particular feature that may "cause" the patterning (e.g., clusters of cancer around a source of carcinogens).

Commonly used tests for clustering and clusters of point events tend to be either based on inter-point *distances* (or distances between the event locations and reference points), or on the number of events within a so-called *quadrat* (i.e., a geometric shape, such as a grid cell, a circle or ellipse). The distances, functions of the distances, or the density of points in a quadrat are compared to what they would be under a null hypothesis of spatial randomness. Significant clusters are indicated when the inter-point distances

tend to be shorter, or when there are more points in the quadrat than expected under randomness.

Cluster tests applied to areal units can be classified into two broad types. In one, the areal unit is simplified to a point (e.g., the center of gravity, or centroid of a county) and all events for the areal unit are associated with the point. This point and all its events can then be analyzed by means of a quadrat method (like a spatial scan statistic) as if it were a collection of individual points. Alternatively, clusters for data in areal units (whether represented as points or polygons) can be detected from spatial autocorrelation tests, which consist of a joint evaluation of locational similarity (neighbors) and attribute similarity (cross product correlation, squared difference). A crucial aspect of spatial autocorrelation statistics is the requirement to define the neighbors. This is typically based on geography, such as considering points to be neighbors when they are within a critical distance, or to take polygons as neighbors when they share a common boundary.

A core element of software implementations of cluster tests is therefore an effective use of data structures and algorithms to store and compute *inter-point distances* and *polygon contiguity* relationships. While all software packages considered here compute distances internally, they differ in the way they handle contiguity. A second core requirement for cluster analysis software is a method to compute or simulate the distribution of a test statistic under the null hypothesis of spatial randomness. The latter is typically implemented by means of a permutation or *randomization* procedure, a technique (also called Monte Carlo hypothesis testing) that is prominent in the nonparametric statistical literature and goes back more than forty years. Randomization procedures feature prominently in all four packages considered.
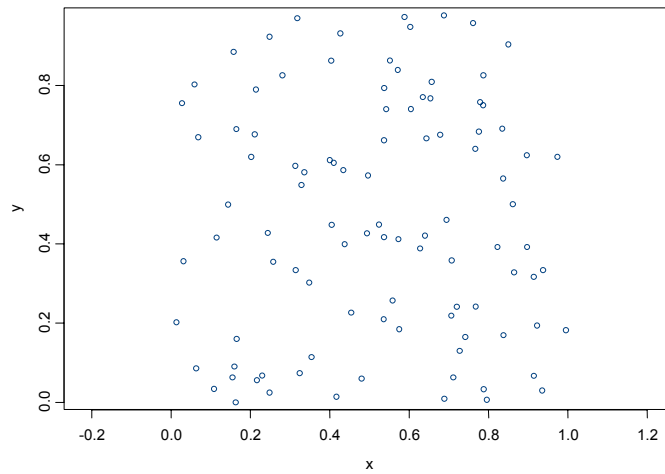
Figure 2. Random distribution, simulation generated in S+.
Source: L. Pickle, pers. comm.
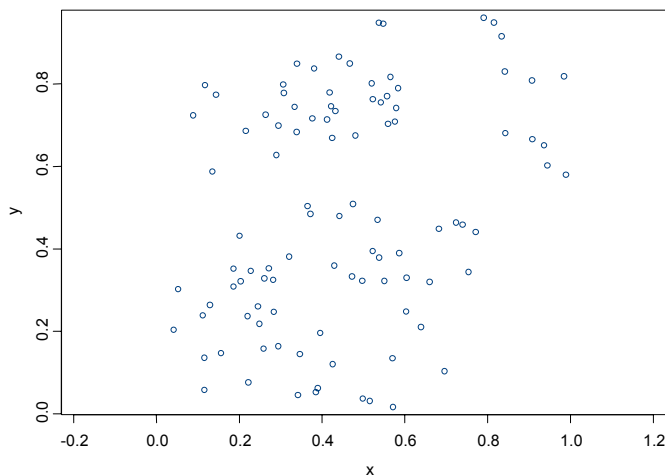
Figure 3. Spatially clustered pattern, Poisson, simulation generated in S+.
Source: L. Pickle, pers. comm.

Further References

Each of the three self-contained software packages considered (SaTScan, CrimeStat and

GeoDa) comes with extensive documentation in a manual, which also includes a

discussion of the underlying methodology. The documentation for the R packages is less

organized and must be found in several books (on R and/or S) as well as web sites.

Two general introductions to spatial data analysis are:

T. Bailey and A. Gatrell (1995). *Interactive Spatial Data Analysis* **(London: Longman)**

A.S. Fotheringham, C. Brunsdon and M. Charlton (2000), *Quantitative Geography, Perspectives on Spatial Data Analysis* **(London: Sage Publications).**

The formal background for the spatial scan statistic can be found in:

M. Kulldorff (1997). A spatial scan statistic, *Communications in Statistics, Theory and Methods* **26, 1481-1496.**

The local indicators for spatial association were introduced in:

L. Anselin (1995). Local indicators of spatial association – LISA, *Geographical Analysis* **27, 93-115.**

More specialized treatments, with specific emphasis on the analysis of disease clusters are contained in the texts by:

A. Lawson (2001). *Statistical Methods in Spatial Epidemiology* **(Chichester: John Wiley)**

L. Waller and C. Gotway (2004). *Applied Spatial Statistics for Public Health Data* **(New York: Wiley).**

In addition, there are several specialized edited volumes devoted to spatial data analysis for public health data, two such works are:

Disease Mapping and Risk Assessment for Public Health **(A. Lawson et al, editors, 1999, Chichester: John Wiley)**

Spatial Cluster Modeling **(A. Lawson and D. Denison, editors, 2002, Boca Raton: Chapman & Hall/CRC)**

Each of these sources has ample references to specific journal articles and books. Finally, several tutorials may be found at http://sal.agecon.uiuc.edu/stuff_main.php.

Cluster Software Requirements

Effective use of software for cancer cluster detection by analysts in cancer registries and other public health entities pre-supposes the existence of some basic GIS infrastructure, in which the data of interest are geo-coded as points on a map, or have been aggregated to

areal units. Cluster analysis software should have an efficient interface to the GIS, in the sense of providing straightforward means to extract the relevant data and to feed back results for map display (and/or addition to the GIS data base). Most commercial GIS (and several open source GIS as well) contains the functionality to convert data points from one projection to another and to "clean" polygons for overlaps and unwanted intersections. Statistical software, while sometimes containing this functionality, is typically less efficient at carrying out these operations.

Assuming that the geographic data have been previously cleaned in a GIS, a software tool for exploratory analysis of cancer clusters should then address the following essential requirements:

- *Effective data input*: x, y coordinates of locations (cases and controls), digital boundaries of polygons (areal units), events, population at risk, or, alternatively, pre-computed (age-standardized) rates as risk estimates.
- *Spatial information*: distance computations, spatial weights construction.
- *Descriptive statistics*: identification of "extreme" values, outlier detection, smoothing of rates to remove "spurious" outliers.
- *Point pattern analysis*: distance or quadrat (scan) based statistics and inference to identify locations that are more clustered than likely under the null hypothesis; distance or quadrat-based statistics to assess "elevated" incidence around a point of interest.
- *Spatial autocorrelation analysis*: measures of global and local spatial autocorrelation (LISA) and inference to identify areas with elevated risk, that are surrounded by similar areas, or to identify spatial outliers; correction of the spatial autocorrelation statistics for variance instability of rates (Empirical Bayes correction).
- *Visualization of the results*: maps and/or graphs indicating outliers and significant clusters.
- *Program output*: presentation of results in a form that can be integrated with other software, such as GIS and mapping packages and data base management systems.

None of the four packages reviewed here satisfies all these criteria, although they come quite close. This is considered in more detail in the following sections.

Before proceeding with the technical discussion, however, it is worthwhile to note that the basic functionality outlined above pertains to cluster *detection*, which is only the first step in an analysis. The broader issue of modeling the spatial variation in events or risk by relating them to explanatory variables (e.g., in spatial regression models, or hierarchical specifications) is beyond the scope of the current report. Suffice it to say, the degree of statistical and modeling know-how necessary to satisfactorily implement this second stage is significantly beyond what is required to carry out the exploratory analysis discussed here.

Summary Features of the Software Packages

CrimeStat

CrimeStat is a program originally developed for the analysis of crime incident locations. However, it is equally applicable to other point pattern data, such as patterns in the location of cancer incidence and mortality. *CrimeStat* deals with point data (no polygons) that can be read in from a wide range of formats. It is organized along four main sets of functions, each corresponding to a tab on the user interface: data setup, spatial description, spatial modeling, and options. The data input is by far the most flexible of the four packages considered and includes primary file (for cases or counts), secondary file (for controls or population) and reference grid file for scan and kernel density estimation, each available from a variety of input file formats.

The statistical functionality in *CrimeStat* is divided between "description" and "modeling." Of these, the description contains most of the functions relevant to cluster analysis, such as nearest neighbor statistics, Ripley's K, spatial autocorrelation indices and so-called "hot spot" analysis. The latter contains a number of clustering methods (i.e., partitioning the data), the Local Moran, as well as the STAC method of Block and Block (space time analysis of crime), which is similar in spirit to Kulldorff's scan statistic, although not based on a likelihood ratio criterion. The "Spatial Modeling" tab includes interpolation routines (such as kernel density estimation) and space-time analysis, with the Knox and Mantel tests. Both are essential tools in a space-time analysis of clustering of events. *CrimeStat* also contains a journey to crime function, a specialized technique used in crime analysis, and less applicable to cancer studies.

The analysis in *CrimeStat* proceeds in two stages. First, the analyst sets all the input information, options and type of analysis. In the second step, the actual computations are carried out and the results presented on the screen. Several options exist to export the results (such as ellipses indicating the locations of the clusters) to GIS and mapping packages. *CrimeStat* itself does not have a mapping or plotting capability, typically relying on the GIS for these functions.

The statistical models included in *CrimeStat* are standard for point pattern analysis. Some test statistics are based on analytical derivations for the inference, but most techniques use Monte Carlo randomization to assess significance. The software comes with an extensive manual (more than 450pp.) with several worked out case studies and examples. While some understanding of the cluster statistics is required, the manual and examples go a long way to explaining the necessary statistical background for the first time user. As mentioned above, *CrimeStat* does require a separate GIS software package if the results are to be visualized.

GeoDa

GeoDa is intended to be an introduction to spatial data analysis, aimed at users who desire some minimal mapping functionality, without having to use a full-fledged GIS. The emphasis is on visualization and interactive data analysis, using the technique of dynamically linked windows to implement linking and brushing of all statistical graphs and maps. *GeoDa* is primarily geared to the analysis of areal data, although points can be used to represent areas (such as areal centroids). Besides statistical graphs, maps and autocorrelation statistics, it includes several functions to manipulate spatial data, such as the computation of centroids, creation of Thiessen polygons and conversion of data from ASCII and dBase point input files to point shape files.

GeoDa is the only package of the four considered that has its own mapping functionality and does not require a GIS for visualization. It is built on ESRI's MapObjects library and uses shape files as the main input. These shape files can either be projected (in which case Euclidean distances are computed) or in decimal degree latitude and longitude (in which case the distance is great circle distance). Besides the standard choropleth maps, several outliers maps are included, such as a box map (a map equivalent

of a box plot), as well as cartograms and methods for smoothing rate maps. The maps can be saved as bitmap files for use in presentations.

The analytical focus of *GeoDa* is on Exploratory Spatial Data Analysis (ESDA), which includes a range of statistical graphs (histogram, box plot, scatter plot, and parallel coordinate plot) as well as the analysis of spatial correlation. The latter is implemented in the form of a Moran scatter plot (with randomization to assess significance) for the global Moran's I (including an Empirical Bayes correction for rate data), and as LISA significance and cluster maps for the Local Moran's I. The results for the Local Moran are also saved in a table, which can be exported for use in other software. The correlation statistics are extended to a bivariate case, which allows for space-time analysis. To carry out the spatial correlation analysis, extensive functionality is included in *GeoDa* to construct spatial weights, both based on contiguity between polygons as well as from the distance between points. These weights files are used by *GeoDa*, but can also be used by other software (e.g., the R *spdep* package contains functions to read the weights files created by *GeoDa*).

GeoDa is intended for a fairly general audience that need not be GIS experts, nor are the statistical prerequisites much beyond familiarity with descriptive statistics and basic statistical inference. It comes with an extensive user's guide and several sample data sets and tutorials.

R Packages

R is not a statistical software package in a narrow sense, but rather a high-powered programming environment for data manipulation, computation and graphics.[6] It is an extensible and customizable environment in which specific statistical procedures are implemented by writing "functions." This puts the burden of software development mostly on the user, and requires a relatively high degree of computational and statistical sophistication. A large user community has developed software applications in the form of "packages" that contain functions to carry out specific types of statistical analyses.[7]

---

[6] R is the open source counterpart to the commercial S and S-Plus software.
[7] These packages must be installed individually. The "base" R installation comes with a collection of installed packages, but the specialized functions for spatial data analysis are not necessarily included and must be installed explicitly.

The packages *spatial* (part of what is known as the VR package), *splancs*, and *spatstat* contain a range of methods for the analysis of point patterns (tests against clustering), while the package *spdep* is geared toward tests for spatial autocorrelation. They will be referred to as R-Geo, which is a recent initiative to consolidate and support spatial data analysis in R.

The R packages cannot be characterized as user-friendly, but they contain by far the widest range of the advanced methods of the four software packages considered here. Since the R environment is expandable, more specialized techniques can be incorporated as well, provided that the user (or the user community) does the programming. The data input and output facilities are rudimentary and in many situations some "tweaking" is required to get things to work. Statistical tests are carried out by entering an explicit function call (with the proper parameters and options) on a command line.[8] It is up to the user to make sure the correct options are selected. Results are not always presented in an obvious fashion, but are typically stored in an "object" which can be manipulated separately (e.g., to display a "summary" of the results). In practice, a statistical procedure requires several steps (entered in sequence on the command line), using a UNIX-like approach where complex tasks are split up in a series of smaller tasks. Each of these tasks typically corresponds to a function in R, and it is up to the user to construct the proper sequence.

Specific cluster tests included in the R-Geo packages consist of exploratory point pattern analysis techniques, such as the F, G, J and K functions. Spatial autocorrelation tests include Moran's I, Geary's c and the Local Moran. While R includes an extensive set of plotting functions, the mapping capabilities are (still) somewhat rudimentary.

The R environment is an effective tool for spatial data analysis in the hands of a power user, but it is not for the uninitiated. The packages come with documentation, which includes several sample data sets.

---

[8] R in a windows environment has a limited graphical user interface. Also, it is possible to develop a customized user interface, although this requires some fairly sophisticated programming.

SaTScan

SaTScan is a specialized package that implements Kulldorff's scan statistic for spatial, temporal and space-time cluster detection. This statistic is arguably the most widely used in the public health arena to detect disease clusters. The statistic is based on a combination of a quadrat-like counting of events in a circular area with a likelihood ratio test to identify the most likely cluster. Inference is based on Monte Carlo randomization. The test can be applied to individual event locations (cases and controls, using a Bernoulli statistical model), or to aggregate event counts associated with points (event count and population count, using a Poisson statistical model). While the test is not intended to be a focused cluster test, it can be made to serve that purpose by means of a special grid reference file (the grid points used to compute the counts in the scan circles).

The analysis proceeds in two steps. In the first, the user sets all the "parameters" for an analysis (these may be saved in a file for later use), such as the file names for the input data files, output files and various settings required for the statistical analysis. In the second step, the analysis is "executed" and the results appear on the screen (and are written to output files, if specified). GIS users may find it surprising that each aspect of the inputs requires a separate file (e.g., a separate file for cases, for controls, for geographic coordinates) and that the location coordinates of the events are stored separate from the attributes.[9] Also, the format of these files is quite rigid, especially in the ASCII form (there is also dBase format input through a conversion routine).

The results consist of a listing of the most likely cluster location, its likelihood and significance, as well as additional clusters. This information can be saved in both ASCII or dBase files for incorporation into a GIS, for example, as an "event theme" in ArcView.[10] There is no built-in visualization of either inputs or results.

SaTScan currently only implements the scan statistic and is rather limited in terms of its interface with GIS and other software. Users are assumed to be familiar with the statistical aspects of the scan statistic in order to set the analysis parameters appropriately and to interpret the nature of the clusters.

---

[9] The files include an ID variable, which is used to match the attributes to their location.
[10] Note, however, that the clusters are not necessarily significant, although they are included in the output file.

Summary Evaluation

Since the four software packages considered here are free, monetary constraints are not a criterion in their assessment. They differ primarily in terms of range of functionality, ease of use and statistical sophistication required from the user. No single package meets all the listed requirements, and to some extent, they are complementary. In order to implement an effective cancer cluster analysis capability in practice, it is therefore recommended to use a combination of several packages. It is also important to keep in mind that three of the four packages (GeoDa being the exception) require separate GIS or mapping software for visualization of the results.

Given the popularity of Kulldorff's scan statistic in applied cancer cluster analysis, it is suggested that the *SaTScan* package be part of any software collection, since the other packages (to date) do not contain this procedure. As mentioned before, the *SaTScan* software requires a solid understanding of the underlying test statistic, and its data input/output capabilities are somewhat cumbersome. At a minimum, it should be supplemented with functionality for the Local Moran statistic, which is included in all three other packages. In practice, it is also important to be able to address the intrinsic variance instability in rates and correct for this in autocorrelation statistics. This is implemented in *GeoDa* for both Global and Local Moran's I, and in the R package *spdep* for the Global Moran.

Both *CrimeStat* and *GeoDa* can be used by practitioners with a basic understanding of cluster statistics and spatial autocorrelation. *CrimeStat* has more extensive point pattern analysis functionality, whereas *GeoDa* is more geared towards interactive data exploration and visualization, and has a wider range of options for local autocorrelation statistics. While there is some overlap between the two packages, they are to a large extent complementary.

The R packages, while being the most demanding in terms of user sophistication (both in terms of statistical background as well as with respect to programming skills) are also the most advanced. If interest goes beyond an exploration and identification of clusters to the actual modeling and explanation of spatial patterns, the extensible environment provided by R is a powerful platform to implement such a strategy.

Postscript

Since all four packages are under active development, this review was dated the moment it was written (August 2003). For example, the R packages undergo continuous change and new ones appear all the time. In this respect, it may be worthwhile to note that in early 2004, the experimental R package DCluster was released, containing several cluster detection methods and scan statistics (http://sal.agecon.uiuc.edu/csiss/Rgeo/index.html). Also, SaTScan released Version 4.03 in February 2004, GeoDa released Version 0.95-i in February 2004, (which includes spatial regression), and CrimeStat is likely to release a version 3 by summer 2004. On the commercial side, ESRI's ArcGIS 9.0, to be released by mid 2004, will include a "geoprocessing toolbox" covering some methods of exploratory spatial data analysis and point pattern analysis.

Appendix

Table 1. Summary Overview of Cluster Software Functionality

	CrimeStat	GeoDa	R-Geo	SaTScan
Data Input Formats				
• ASCII (points)	x	x	x	x
• Dbf (points)	x	x		x
• Shape File (points)	x	x		
• Shape File (polygon)		x	$(x)^a$	
• MapInfo (points)	x			
• ODBC	x			
Parameter/Project File				x
Spatial Weights Construction				
• Contiguity		x	$(x)^b$	
• Distance Band		x	x	
• K Nearest Neighbors		x	x	
• Inverse Distance	x		x	
Visualization				
• Statistical Graphs		x	x	
• Point Plot		x	x	
• Kernel Density	x		x	
• Space-Time Kernel Density			x	
• Map		x	$(x)^c$	
Descriptive Statistics				
• Centrography	x			
• Outliers		x		
• Rate smoothing		x	x	
Point Clustering				
• Nearest neighbor statistics	x		x	
• Ripley's K	x		x	
Point Clusters				
• Scan statistic	$(x)^d$			x

---

[a] Shape file reading is not included in the four packages considered here, but can be carried out by means of the *shapefiles* package (Version 0.3, 7/23/03). The *maptools* package (Version 0.2-3 8/10/03) contains functions to convert shape file information into the format required by *spdep*.
[b] Contiguity weights can be created in *spdep* after polygon shape files are converted to the internal format used by *spdep*. This is done in the *maptools* package. The *spdep* package can also read contiguity weights created in *GeoDa*.
[c] Mapping in *spdep* is treated as a special case of plotting polygons.
[d] The scan statistic implemented in CrimeStat is not Kulldorff's statistic, but STAC (the space and temporal analysis of crime of Block and Block).

	CrimeStat	GeoDa	R-Geo	SaTScan
• Nearest Neighbor Clusters	x			
• Regionalization	$(x)^e$			
Areal Clustering				
• Global Moran	x	x	x	
• Global Moran + EB correction		x	x	
• Global Geary	x		x	
Areal Clusters				
• Scan Statistic	$(x)^d$			x
• Local Moran	x	x	x	
• Local Moran + EB correction		x		
• Getis-Ord Gi			x	
Space-Time Analysis				
• Space-Time Scan Statistic	$(x)^d$			x
• Space-Time Moran's I		x		
• Space-Time Local Moran		x		
• Knox Index	x			
• Mantel Index	x			
• Space-Time K Function			x	
Output				
• Screen	x	x	x	x
• ASCII Output File	x	x	x	x
• Dbf Output File	x	x		x
• Shape File Output File	x	x		
• MapInfo Output File	x			
• Atlas GIS Output File	x			
• Grid Files	x			

---

[e] Implemented as K-means clustering.

Table 2. Overview of Specific Cluster Statistics in Each Package

Cluster Statistics in CrimeStat

- Moran's I
- Geary's c
- Nearest neighbor index and K-th order nearest neighbor index
- Ripley's K
- Nearest neighbor hierarchical clustering
- STAC (spatial and temporal analysis of crime)
- K-means clustering
- Local Moran's I
- Knox Index (space-time)
- Mantel Index (space-time)

Cluster Statistics in GeoDa

- Outlier detection
- Global Moran's I
- Global Moran's I for rates with Empirical Bayes correction
- Local Moran's I
- Local Moran's I for rates with Empirical Bayes correction
- Bivariate (space-time) Moran's I
- Bivariate (space-time) Local Moran's I

Cluster Statistics in SaTScan

- Spatial scan statistic
    - Case-control (Bernoulli model)
    - Case-population (Poisson model)
- Space-time scan statistic
    - Same as above
- Temporal scan statistic
    - Same as above

Table 3. Cluster Statistics in R-Geo Packages

By package, with name of function listed

	spatial	splancs	spatstat	spdep
Nearest Neighbor F		Fhat	Fest	
Nearest Neighbor G		Ghat	Gest	
Multitype G			Gmulti	
Nearest Neighbor J			Jest	
Multitype J			Jmulti	
Ripley's K	Kfn	khat	Kest	
Inhomogeneous K			Kinhom	
Space-Time K		stkhat		
Multitype K			Kmulti	
Focused Test		tribble		
Moran				moran
Moran + EB				EBImoran
Geary				geary
Local Moran				localmoran
Getis-Ord Gi				localG