Detecting Hot Spots Using Cluster Analysis and GIS

Tony H. Grubesic

Center for Urban and Regional Analysis and Department of Geography
The Ohio State University
1036 Derby Hall
154 North Oval Mall
Columbus, OH 43210

Email: grubesic.1@osu.edu

Alan T. Murray
Department of Geography
The Ohio State University
1036 Derby Hall
154 North Oval Mall
Columbus, OH 43210

Email: murray.308@osu.edu

Detecting Hot spots Using Cluster Analysis and GIS

Abstract

One of the more popular approaches for the detection of crime hot spots is cluster analysis. Implemented in a wide variety of software packages, including CrimeStat, SPSS, SAS, and SPLUS, cluster analysis can be an effective method for determining areas exhibiting elevated concentrations of crime. However, it remains a particularly challenging task to detect hot spots using clustering techniques because of the uncertainty associated with the appropriate number of clusters to generate as well as establishing the significance of individual clusters identified. This paper highlights the potential utility of several diagnostics for resolving such issues.

Introduction

Crime mapping and analysis have evolved significantly over the past 30 years. In the beginning, many agencies utilized city and precinct maps with colored pins to visualize individual crime events and crime plagued areas. Today, with the rapid advancement of technology, computer-based techniques for exploring, visualizing, and explaining the occurrences of criminal activity have been essential. One of the more influential tools facilitating exploration of the spatial distribution of crime has been GIS (Ratcliffe and McCullagh, 1999; Harries, 1999). As Murray et al. (2001) note, it is the ability to combine spatial information with other data that makes GIS so valuable. Furthermore, the sheer quantity of information available to most analysts necessitates an intelligent computational system, able to integrate a wide variety of data and facilitate the identification of patterns with minimal effort.

Fundamental to the explanation of criminal activities in a spatial context are certain environmental factors, such as the physical layout of an area, proximity to various services, and land use mixes - all of which are likely to influence criminal behavior (Greenburg and Rohe, 1984). Issues of access, exposure, opportunity, and the availability of targets are also important elements in helping explain crime from an environmental perspective (Cohen and Felson, 1979; Brantingham and Brantingham, 1981). Not surprisingly, research indicates that certain areas are more prone to higher concentrations of crime. Widely labeled as 'hot spots', such areas are often targets of increased manpower from law enforcement agencies in an effort to reduce crime. Where resources are concerned, the identification of hot spots is helpful because most police departments are understaffed. As such, the ability to prioritize intervention through a geographic lens is appealing (Levine, 1999a).

Operationally, the delineation of hot spot boundaries is somewhat arbitrary. As Levine (1999a) notes, crime density is measured over a continuous area. Therefore, the boundaries separating hot spots of crime from areas without enough activity to merit the label hot spot are perceptual constructs. Moreover, depending on the scale of geographic analysis, a hot spot can mean very different things (Harries, 1999).

Recent studies by the Crime Mapping Research Center at the National Institute of Justice categorize hot spot detection and analysis methods. These techniques have been classified as follows (Jefferis, 1999; Harries, 1999): *visual interpretation, choropleth mapping, grid cell analysis, spatial autocorrelation, and cluster analysis*. Further, twelve different variations on the five classes of hot spot identification techniques were systematically documented and evaluated, yielding several important results. Although there are a variety of methods for detecting hot spots in crime event data, no single approach was found to be superior to others.

What does become clear in previous work on hot spot detection is that combining cartographic visualization of crime events with statistical tools provides valuable insight for detecting areas of concern. Results of the CMRC (1998) study suggest that a good approach for detecting hot spots are tests of spatial autocorrelation. Implemented in a variety of packages, including *CrimeStat* 1.1, *SpaceStat*, and *Splus Spatial Statistics*, and

SAGE, both global and local tests of spatial autocorrelation assist in crime analysis. As demonstrated by Szakas (1998) the implementation of the Getis-Org statistic (G_i statistic) in *SpaceStat* provided very good measures of crime hot spots for Baltimore County. The utility of spatial autocorrelation and the G_i statistic for hot spot analysis is further supported in the work of Craglia et al. (2000).

Considering the success of statistically grounded tests for hot spot detection, such as the G_i statistic for spatial autocorrelation, it is unfortunate that other well-established statistical tests, such as cluster analysis, are generally viewed to be less useful (Chainey and Cameron, 2000). Gordon (1999) suggests that cluster analysis is one of the most useful methods for exploratory data analysis, especially in large multivariate data sets. If this is the case, why has it failed to help crime analysts in hot spot detection?

Statistical approaches for cluster analysis are widely available in a number of software packages, including *CrimeStat*, *SAS*, *SPSS*, *Systat*, and *SPlus*. However, the evaluation, documentation, and implementation of cluster analysis algorithms, particularly non-hierarchical versions commonly used in crime analysis such as *k*-means, are not clear nor do they give direction for useful application (Murray and Estivill-Castro, 1998; Murray and Grubesic, 2002). The gap between what has been developed and what is actually needed for hot spot detection exists for several reasons. First, crime hot spots are spatial phenomena. Therefore, in order to identify elevated concentrations of crime in a geographic area, tools that treat space appropriately are critical. Second, existing approaches for cluster analysis are not necessarily ideal when applied to spatially referenced data. This is best reflected by the relatively poor performance of the *k*-means algorithm, as implemented in leading statistical packages, for spatial data analysis (Murray and Grubesic, 2002). Given that non-hierarchical clustering approaches have proven fruitful in other research areas, it is premature to deem such techniques too complex or poorly performing for crime analysis as done by Chainey and Cameron (2000). The failure to date of non-hierarchical techniques is a product of how they are being used and supported.

The purpose of this paper is to explore two of the problematic aspects of cluster analysis for hot spot detection. First, we examine the difficulties in determining the appropriate number of clusters, *p*, to generate. Second, we highlight several statistical methods that have the potential for establishing the significance of clusters identified as hot spots.

The remainder of this paper is organized as follows. Section 2 outlines the differences between hierarchical and partitioning techniques for cluster analysis. Section 3 examines the problems associated with identifying the appropriate number of clusters. Included is a discussion of several approaches that have the potential to make the identification of the number of clusters more statistically grounded. Section 4 explores the issues of attaching significance to the clusters generated in hot spot detection. Section 5 contains a brief discussion and closing remarks.

2. Clustering Approaches

2.1 Hierarchical

Broadly defined, cluster analysis is a method of classification that places objects in groups based on the characteristics they possess. Bailey and Gatrell (1995) note that all clustering techniques begin in the same fashion. Namely, each method begins with the calculation of a (n x n) matrix, $\boldsymbol{D}$, of dissimilarities between every pair of observations. In most cases, a Euclidean metric is used as the measure of dissimilarity. Based on ($\boldsymbol{D}$), cluster analysis breaks observations into groups, linking the most similar observations together in clusters. For example, hierarchical clustering techniques begin with all observations in separate groups and proceed to join the most similar observations (or groups of observations) according to some pre-specified criteria ($\boldsymbol{D}$).[1] In hierarchical clustering, nearest neighbor distance is frequently used as the dissimilarity measure (Bailey and Gatrell, 1995). The nearest neighbor measure is a comparison of the distances between two points (or groups of points) with the average distance between all points. If the distance meets the *a priori* criterion (usually the calculated probabilities of a threshold distance between observations occurring by chance), observations are linked to form a new cluster. This process is repeated until all points have been assigned to a first-order cluster.[2] First-order clusters are then tested for second-order clustering in the same manner. Levine (1999a) notes that this process is repeated until all sub-clusters have converged into a single cluster or the threshold distance criterion fail.

Although hierarchical clustering allows analysts to examine the concentration of crime events in smaller geographical areas and the links between crime cluster hierarchies (e.g. first-order to second-order), the problems associated with hierarchical clustering techniques can outweigh the benefits. As Bailey and Gatrell (1995, 233) note, "although hierarchical clustering optimizes a criterion at each step, there is no guarantee that, if one ends up with p groups, this is the partition of the observations which would optimize this same criterion over all possible partitions of the observations into p groups." In other words, hierarchical clustering procedures frequently generate *local* rather than *global* optima. There are also problems associated with the threshold distances used in hierarchical clustering. Levine (1999a) suggests that crime distributions with many incidents (burglary) typically have lower threshold distances than distributions with fewer incidents (murder). As such, hierarchical clustering does not treat space appropriately, producing inconsistent "hot spots". Finally, the implementation of a minimum number of observations rule for clusters is arbitrary at best. How many points constitute a meaningful cluster? Ten? Twenty? This effectively eliminates any objectivity in an analysis as the definition of cluster size is likely to vary between users.

---

[1] Alternatively, one can begin with all observations in a single group and break them up into separate clusters.

[2] Clusters can also be subject to a criterion that defines the minimum number of points to be included in a group.

2.2 Partitioning

Partitioning or optimization clustering techniques *do* attempt to split observations into a pre-specified number of groups, *p*, where the specified criterion is optimized globally over all possible splits. Clearly, the disadvantage to this approach is the requirement that the number of groups must be specified *a priori*.

One of the most prominent statistical techniques for cluster analysis has been the *k*-means approach proposed in Fisher (1958). This technique is based upon multivariate analysis of variance in the evaluation of homogeneity among entities (Estivill-Castro and Murray, 2000a). Specifically, the scatter matrix of similarity between entities may be evaluated by its trace (Aldenderfer and Blashfield, 1984). Homogeneity is then measured for a grouping of entities using the sum of squares loss function (Rousseeuw and Leroy, 1987).

Other non-hierarchical clustering approaches have also been developed and utilized. Some alternatives are detailed in Kaufman and Rousseeuw (1990). In the context of spatial application, a review of approaches is given in Murray and Estivill-Castro (1998). More specific to the analysis of crime, a discussion may be found in Murray et al. (2001). If we are intent on identifying areas or entities that are strongly related in some predefined sense, then many non-hierarchical clustering techniques may potentially be useful. This is important and significant because a user could select alternative clustering approaches to conduct analysis. However, it is also possible that a user selects an *inappropriate* approach because they are not aware of associated biases and inaccuracies. That is, all clustering approaches are not equivalent. Recent geographical research has focused on appropriateness issues in the use and application of non-hierarchical clustering techniques (Murray 1999, 2000a; Murray and Grubesic, 2002). What has been found is that substantial variation exists in the structure and quality of identified clusters. For example, Murray and Grubesic (2002) found that the *k*-means solutions identified in several leading statistical packages such as SPSS, SAS, and SPlus deviate by as much as 30% from the best-known solutions (Figure 1).

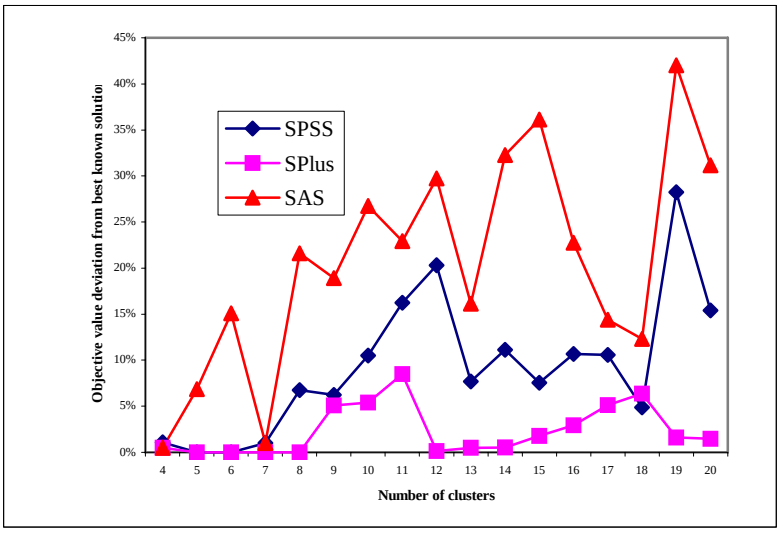

Figure 1: Performance of Commercial Statistical Packages (*k*-means)
Source: Murray and Grubesic (2002)

3. How Many Clusters?

The ability to identify the appropriate number of clusters for a given set of crime events is one of the most fundamental shortcomings of non-hierarchical techniques for hot spot detection. Levine (1999a) suggests that both the strength and weakness of the k-means procedure is the ability for the user to define the number of clusters to be generated for a given set of observations. Although most software packages allow the user to specify *k* groups, it is certainly *not* a strength in practice. While local knowledge and experience can play a role in hot spot analysis, user defined parameters such as *k* groups builds significant subjectivity into analysis. Furthermore, implicit to most discussions of the *k*-means approach in crime analysis is the notion that there are no established methods for determining the optimal number of clusters (Levine, 1999a,b). In fact, there are numerous methods outlined in the statistics literature detailing potential methods for detecting the appropriate number of clusters (Gordon 1996; Podani, 1996; Lozano et al., 1996; Milligan and Cooper, 1985). As an example, Milligan and Cooper (1985) assessed the ability of thirty different stopping rules to predict the correct number of clusters in randomly generated data sets. Although some of these rules performed poorly, others performed quite well. More importantly, Milligan and Cooper (1985) suggest that stopping methods developed for hierarchical cluster analysis are easily modified for optimization-based approaches.

One of the more effective procedures for determining the number of clusters in a data set is the *cubic clustering criterion* (CCC). CCC is the test statistic provided by the SAS package. Developed by Sarle (1983), inflection points in the CCC column of SAS output should be analyzed from *n* groups to 1 group. These inflection points are indicative of appropriate cluster groupings for the data. Moreover, there may be more than a single inflection point. Alternatively, graphical plots of CCC values may be utilized for analysis. Peaks greater than 2 or 3 on such plots suggest good clustering. Peaks between 0 and 2 suggest potential clusters, but must be interpreted cautiously (Sarle, 1983). The CCC values can also be used in conjunction with pseudo F (PSF) and t^2 statistics in SAS. Both measures provide additional information, with large PSF values suggesting a good stopping point. Inflections in the t^2 statistic also suggest possible cluster stops.

A second approach Milligan and Cooper (1985) determined to be effective was the Calinski and Harabasz (1974) index. The index is computed as follows:

$$[\text{trace } B/k - 1)]/[\text{trace } W/n - k)$$

Trace refers to the sum of squared Euclidean distances between entities and their centroids. Further, *n* and *k* are the total number of entities and number of clusters in the solution, respectively. The *B* and *W* terms are the between and pooled within clusters sum of squares and cross products matrices. Better performing than the CCC, the Calinski and Harabasz index performed consistently for varying cluster group simulations (Milligan and Cooper, 1985).

Regardless of the stopping technique selected for cluster detection, varying spatial scales of analysis must also be considered. For example, clusters that exist at the neighborhood level might be insignificant at the city level. Thus, Gordon (1998) indicates that stopping rules for cluster analysis have both a global and local component. Global rules are based on the complete data set, typically seeking an optimal index value that compares both within-group and between-group variation (Gordon, 1998). Local rules generally consider whether or not a single cluster should be subdivided into two sub clusters (Gordon, 1998). On the surface, global and local perspectives for cluster detection appear to resemble global/local aspects of spatial autocorrelation. Clearly, more research and experimentation is needed to determine potential links, if they exist.

4. Cluster Significance

An essential need in the use of non-hierarchical approaches for the detection of hot spots is to better understand what makes certain clusters more significant in the context of hot spot analysis. In other words, what makes a hot spot 'hot' and what makes other clusters 'cold'. A body of research does exist for approaching this issue, though its use in the context of hot spot detection is uncultivated. In assessing partitions, Gordon (1999) suggests one must address several issues. First, does a specified partition (*k*) provide compact and isolated clusters? This is closely linked to research in determining the appropriate number of clusters. The implementation of both global and local stopping rules, as outlined in the previous section will undoubtedly provide additional insight into this process. Second, is it possible to address and validate the internal structure of clusters? The work of Arnold (1979) and Milligan and Mahajan (1980) suggests that Monte Carlo tests of partition validity and significance are potentially useful.

In addition to statistical tests of cluster significance, geometric properties of cluster partitions may also prove useful. Perhaps the most basic indicator of cluster significance is the number of entities in a partition. At face value, groups with a relatively larger number of events are certainly indicative of more criminal activity. However, this does not include any measure of spatial dispersion relative to other clusters. Thus, the generation of a minimum-bounding polygon (convex hull) can provide additional insight into the spatial extent of identified clusters (Figure 2). Basic calculations of the area covered by the hull give some indication to cluster compactness. As illustrated in Figure 2, cluster partition 3 is associated with the most compact hull. It is also possible to make basic calculations that consider the number of entities in a partition as they relate to convex hull size. In this case, partition 3 (12 members) has the most compact member/area ratio (.3904). A final indicator of cluster compactness and potential significance is the average distance between cluster members and the cluster center. As displayed in Figure 2, cluster partition 3 (12 total entities) has the smallest average distance (278.87) between its members and its cluster center. This suggests a more compact cluster group.

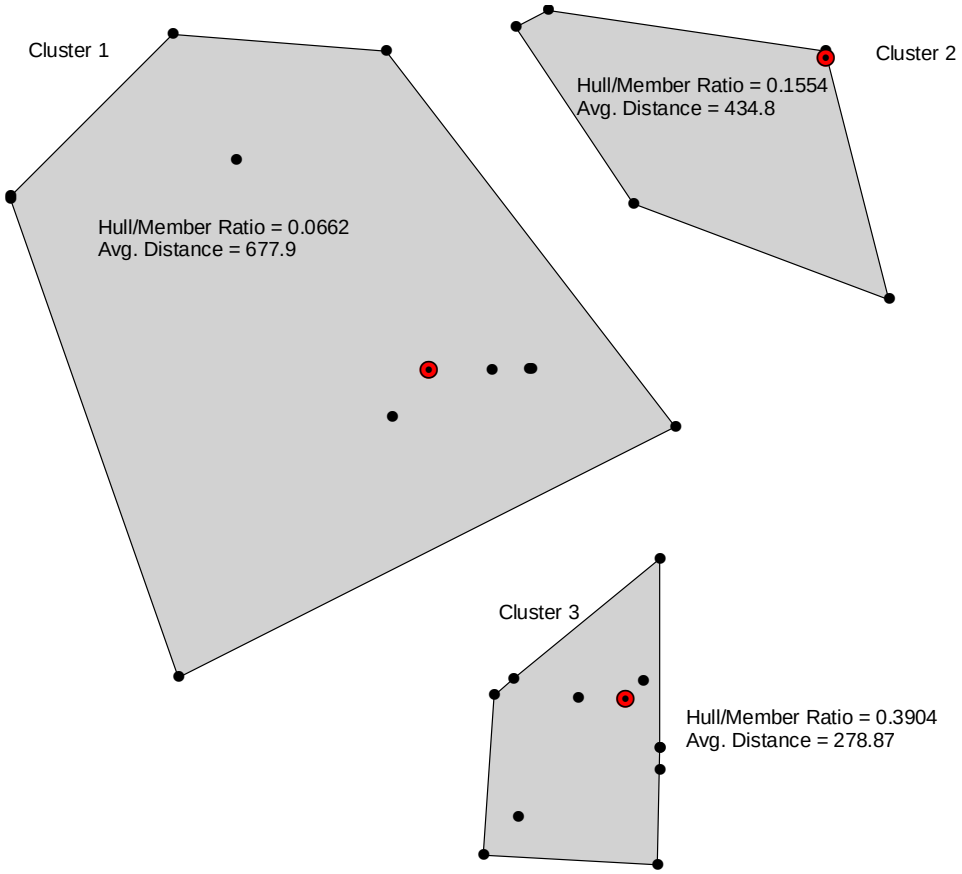

Figure 2: Geometric Properties for Determining Cluster Significance

It is important to note that all of the tests previously outlined must be implemented across a series of cluster partitions, 1, 2, …*p*, to better evaluate the spatial characteristics of the crime events or study area. More importantly, although these techniques are promising starting points for future research, additional work is needed to make more robust geometric and statistical measures for determining cluster significance.

5. Conclusion

This paper has outlined several problematic aspects of optimization based cluster analysis for crime hot spot detection. Rather than simply dismissing cluster analysis as being too complex for hot spot detection, additional research effort should be directed toward adapting existing statistical and geometric techniques make cluster detection more intuitive and useful for crime analysts.

References

Aldenderfer, M. and R. Blashfield. 1984. *Cluster Analysis*. Beverly Hills: Sage Publications.

Arnold, F. J. 1979. "A test for clusters." *Journal of Marketing Research*. 16: 545-551.

Bailey, TC., and AC Gatrell. 1995. *Interactive Spatial Data Analysis*. London: Longman Scientific and Technical.

Brantingham, P. and P. Brantingham. 1981. *Environmental Criminology*. Beverly Hills: Sage.

Calinski, RB., J. Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics, 3*. pp 1-37.

Chainey, S. and J. Cameron. 2000. "Understanding Hot Spots." Presentation prepared for 2000 CMRC Conference: *Wheredunit? Investigating the Role of Place in Crime and Criminality*. San Diego, CA.

CMRC [Crime Mapping Research Center]. 1998. Crime Mapping Research Center "Hot Spot" Project.
URL: http://www.ojp.usdoj.gov/cmrc/whatsnew/hotspot/toc.html

Cohen, L. and M. Felson. 1979. "Social Change and Crime Rate Trends: A Routine Activity Approach". *American Sociological Review*. 44: 588-608.

Craglia, M., R. Haining, and P. Wiles. 2000. "A comparative evaluation of approaches to urban crime pattern analysis." *Urban Studies*. 37(4): 711-729.

Fisher, W. 1958. "On grouping for maximum homogeneity." *Journal of the American Statistical Association*. 53: 789-798.

Gordon, A.D. 1998. How many clusters? An investigation of five procedures for detecting nested cluster structure. In, *Data Science, Classification, and Related Methods*, edited by C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. Bock, and Y. Baba. Tokyo: Springer-Verlag.

_____. 1999. *Classification*. New York: Chapman and Hall/CRC.

Greenburg, S. and W. Rohe. 1984. "Neighborhood Design and Crime." *Journal of the American Planning Association*. 50: 48-61.

Harries, K. 1999. *Mapping Crime: Principle and Practice*. Washington DC: National Institute of Justice (NCJ 178919).

Jefferis, E. 1998. "A Multi-Method Exploration of Crime Hot Spots." Paper presented at *1998 Academy of Criminal Justice Sciences (ACJS) Annual Conference*. URL: www.ojp.usdoj.gov/cmrc/whatsnew/hotspot/intro.pdf

Levine, N. 1999a. *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations*, version 1.1. Washington DC: Ned Levine & Associates / National Institute of Justice.

_____. 1999b. "Hot spot analysis using both the Systat k-means routine and a risk assessment." URL: http://www.ojp.usdoj.gov/cmrc/whatsnew/hotspot/kmeans.pdf

Lozano, J.A., P. Larranaga, and M. Grana. 1996. "Partitional cluster analysis with genetic algorithms: Searching for the number of clusters." In, *Data Science, Classification, and Related Methods*, edited by C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. Bock, and Y. Baba. Tokyo: Springer-Verlag.

Milligan, G.W. and M.C. Cooper. 1985. "An examination of procedures for determining the number of clusters in a data set." *Psychometrika*. 50(2): 159-179.

Milligan, G.W. and V. Mahajan. 1980. A note on procedures for testing the quality of a clustering of a set of objects. *Decision Sciences*. 11: 669-677.

Murray, A.T. 1999. "Spatial analysis using clustering methods: evaluating the use of central point and median approaches." *Journal of Geographical Systems*. 1: 367-383.

_____. 2000a. "Spatial characteristics and comparisons of interaction and median clustering models." *Geographical Analysis*. 32: 1-19.

Murray, A.T. and T.H. Grubesic. 2002. "Identifying Non-hierarchical Clusters." *International Journal of Industrial Engineering*. To appear.

Murray, A.T., I. McGuffog, J.S. Western, and P. Mullins. 2001. "Exploratory spatial data analysis techniques for examining urban crime." *British Journal of Criminology*. 41: 309-329.

Podani, J. 1996. "Explanatory variables in classifications and the detection of the optimum number of clusters." In, *Data Science, Classification, and Related Methods*, edited by C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. Bock, and Y. Baba. Tokyo: Springer-Verlag.

Ratcliffe, J.H. and M.J. McCullagh. 1999. "Hotbeds of crime and the search for spatial accuracy." *Journal of Geographical Systems*. 1: 385-398.

Rousseeuw, P. and A. Leroy. 1987. *Robust Regression and Outlier Detection.* New York: John Wiley.

Sarle, WS. 1983. "Cubic Clustering Criterion." SAS Technical Report A-108. Cary, NC: SAS Institute Inc.

Szakas, J. 1998. "A Multi-Method Exploration of Crime Hot Spots Software Evaluation: SpaceStat." URL: http://www.ojp.usdoj.gov/cmrc/whatsnew/hotspot/spacestat.pdf