Middle States Geographer, 1995, 28:53-58

IDENTIFYING CRIME CLUSTERS: THE SPATIAL PRINCIPLES

Sanjoy Chakravorty
Department of Geography and Urban Studies
Temple University
Philadelphia, PA 19122

ABSTRACT: *There is increasing attention being given to the spatial analysis of crime, particularly on the identification of clusters, or hot spots. I point out the weaknesses of current cluster identification methods based on quadrat analysis, especially the tendency to find clusters in random distributions. I also suggest that (1) tests for spatial autocorrelation be carried out, (2) using non-uniform polygon bases like census block groups, and (3) incorporating methods to identify local instabilities. These principles are utilized in the identification of hot spots of aggravated assaults in central Philadelphia.*

INTRODUCTION

The distribution of crime is known to have a spatial dimension; that is, a map of point locations of crime often reveals spatial patterns, or clusters. This phenomenon is partly explained by the fact that population is not homogeneously distributed over space (i.e., neither the density of population, or characteristics like age, income, etc.)--it is expected that crime, too, will not be homogeneously distributed. Other explanations, like proximity to bars or night clubs, may also have a spatial aspect. Despite this clear link, spatial analytical tools are not yet commonly used in crime analysis. On the other hand, such tools are being used more often now--analytical geography is being used to create spatial profiles of serial criminals, and increasingly digital mapping is becoming part of the crime analyst's everyday toolbox (Hirschfield, 1994; Maltz, et al., 1991). Lately we have also witnessed the emergence of spatial algorithms for the detection of crime clusters, or hot spots.

The logic and principles of hot spot identification appear to have a number of serious problems, so many that results based on such methods are questionable. I argue that the current generation of hot spot identification software is flawed because it ignores two important geographic principles: First, programs do not incorporate any methodology to differentiate between random distributions and clustered distributions. That is, the crime distributions analyzed are always assumed to be clustered--and, hence, whether or not clusters exist, some are identified. Second, the notion of environmental heterogeneity is not incorporated. Population space in general, and urban space in particular, is not homogeneous. Nevertheless, when identifying clusters, geographical space is explicitly assumed to be homogeneous.

In this paper I discuss some fundamentals of point pattern analysis as they apply to the identification of crime clusters. I begin with a discussion of the basic principles of hypothesis testing, extend this to incorporate the idea of spatial autocorrelation, and introduce a family of measurement and mapping techniques that incorporate spatial principles. Finally I illustrate the use of these techniques.

TESTING FOR THE PRESENCE OF HOT SPOTS

Analytical interest in point distributions, and techniques to identify patterns within them began about sixty years ago with the work of plant ecologists and botanists (see Boots and Getis, 1988 for a review).[1] Generally, the initial assumption is that of Complete Spatial Randomness (or CSR) under homogeneous planar Poisson point process conditions. Therefore, the null hypothesis is that there are no clusters in a given distribution. Two

specific assumptions are made: (1) each location in the study area has an equal chance of receiving a point (uniformity), and (2) the selection of location for a point does not influence the location of any other point (independence). These are clearly restrictive assumptions, and can be questioned when a given distribution is found to be clustered; that is, when the null hypothesis has to be rejected.

The usual method of testing the null hypothesis is by using quadrat analysis. In this method a grid is superimposed on the study area and the number of points falling in each cell of the grid are counted. By calculating the difference between the distribution of expected values (in a CSR distribution) and the values actually found (sometimes by using a sample of cells in the grid), it is possible to determine whether clustering exists. There is a serious problem with this methodology. The results, since they are obtained using values within cells irrespective of adjoining cell values, are immutable with respect to spatial distributions of the cells in the grid. White (1983) has termed this the "checkerboard problem" with reference to segregation research; no matter how the cells are arranged the same value is obtained (see Morril, 1991 and Wong, 1993 for approaches to solve this problem with reference to segregation).

The literature and methodologies associated with spatial autocorrelation deal with precisely this problem. Spatial autocorrelation refers to the tendency of events to cluster; or like values to be proximate to each other (being similarly influenced by similar processes). Positive spatial autocorrelation exists when like values are clustered, negative spatial autocorrelation exists when unlike values are clustered (see Odland, 1987). The most popular measures of spatial autocorrelation are Moran's I and Geary's c, where the former is far the more popular. I is derived from the following equation:

$$I = \frac{\frac{n}{2a}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}(x_i - x_a)(x_j - x_a)}{\sum_{i=1}^{n}(x_i - x_a)^2}$$

Where,
n = number of quadrats or cells
d_{ij} is a measure of contiguity between cells, equalling 1 when cells are contiguous (including optional diagonal links, or Queen's case), and 0 when they are not
x_i is the number of points in quadrat i
x_j is the number of points in quadrat j
x_a is the number of points per quadrat (or average)
a is the number of joins (i.e., when d_{ij} is 1)

The expected value of I or $E(I) = -(n-1)^{-1}$
The calculated value of I can be tested for significance using standardized z scores, where z is given by

$$z = \frac{I - E(I)}{\sqrt{var(I)}}$$

If the value derived for z is statistically significant, a determination can be made whether spatial autocorrelation exists in a given distribution. If positive spatial autocorrelation exists we can say that there are clusters or hot spots in that distribution.

LOCATING HOTSPOTS

The simplest method of revealing the clusters is to create a thematic map of the area with the quadrats shaded in proportion to the number of points per quadrat. Here we run into the second problem--the assumption of spatial homogeneity. This assumption is particularly troubling in crime analysis because there will be a tendency to find crime clusters where there are population clusters. That is, tests of clustering will reveal only the cliche--crime in central city.

The source of this problem is the use of an uniform grid size to test for spatial autocorrelation. The uniform cell size carries the implicit assumption that all areas covered by each of these cells is also uniform--a one square mile quadrat in suburban Main Line is as likely to receive a crime point as one square mile of North Philadelphia. It is possible to get around this problem by using the street network as the grid; i.e., instead of using an arbitrary superimposed grid, one could use a real

existing grid. This could work because the density of the street network tends to be proportional to the density of population. However, in very densely settled areas this could result in a very fine grid which may tend to underestimate the intensity of crime.

A simpler solution may be to use census geography as the base map or grid. Digital data are available from the Census Bureau at various levels of spatial disaggregation. An urban block typically has about 250 people, a block group has around 1000 people, and a census tract usually contains about 4000 people. Using data at the block level may be time consuming, and, again, may create too fine a grid. The block group geography, typically, should serve our purpose. A geometrically uniform lattice should, therefore, be replaced by a spatially representative map/polygon base.

A third problem is also directly related to the heterogeneity of urban space. In large areas such as complete cities or counties, there may be local hot spots which are weak when compared to global statistics, and may not be identified by comparing to an areawide mean (which is what Moran's I does). To use the Main Line example again, there may be block groups in that area which have high crime values relative to its neighbors, but not compared to central city crime. These are called areas of local instability, or local clusters. Recently a few measures that have the ability to identify such local clusters, that usually are devised for other purposes, have become available.

One of the more promising of these new measures is Anselin's (1995) Local Indicator of Spatial Association (or LISA). Anselin devised a family of LISA measures, to be used in association with Moran's I or Geary's c. The local Moran form of Lisa is given below:

$$LISA(M)_i = (x_i - x_a) \sum_{j=1}^{n} d_{ij} (x_i - x_a) (x_j - x_a)$$

As before, d_{ij} is a measure of contiguity, and equals 1 for parcels with a common boundary. The subscript i denotes that this calculation is carried out for every parcel in the area; the values derived may be directly mapped (or the z scores may be mapped) and clusters identified.

Another measure, also recently devised, has been used in the measurement of spatial income disparity (Chakravorty, 1996). This measure, called the Neighborhood Disparity (or ND) index, can be disaggregated to the individual parcel level, such that the value derived is an indicator of the extent to which a parcel's value is different from the average of all its neighbors. The ND for an individual parcel is given by

$$ND_i = \frac{\left| \frac{\Sigma x_j}{\Sigma j} - x_i \right|}{x_i}$$

Both LISA and ND are contiguity based measures (though second order neighbors may be incorporated). Getis and Ord (1992) suggested a distance based measure, G, which is given by

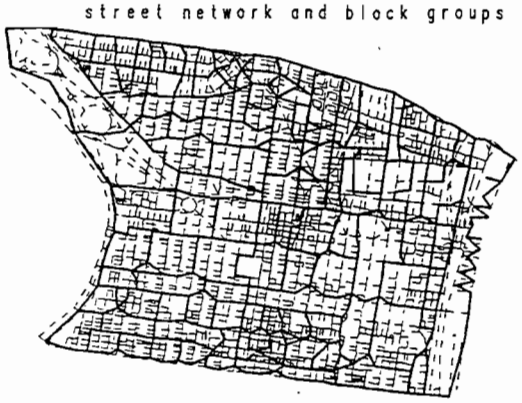
$$G_i(d) = \frac{\sum_{j=1}^{n} w_{ij}(d) x_j}{\sum_{j=1}^{n} x_j}$$

where $\{w_{ij}\}$ is a spatial weight matrix with value one for all links within distance d of a given i. The denominator is the sum of all x_j not including x_i.

A TEST: CRIME CLUSTERS IN CENTRAL PHILADELPHIA

The rationale and methodology outlined above, specifically the use of I and LISA statistics, was operationalized for a section of the city of Philadelphia, PA. I used 1990 crime data for the Sixth and Ninth Police Districts. The area covered is centered around downtown Philadelphia (including City Hall, Chinatown, Jewelers row, and the office complexes on Chestnut and Market Streets, and JFK Boulevard). To the north the area stretches to Poplar Street, and includes large sections of minority dominated residential districts. To the south the limit is South Street, the eastern end of which has been successfully gentrified. The total area covers around 3.95 square miles (Fig. 1).

Figure 1. Central Philadelphia: street network and block groups



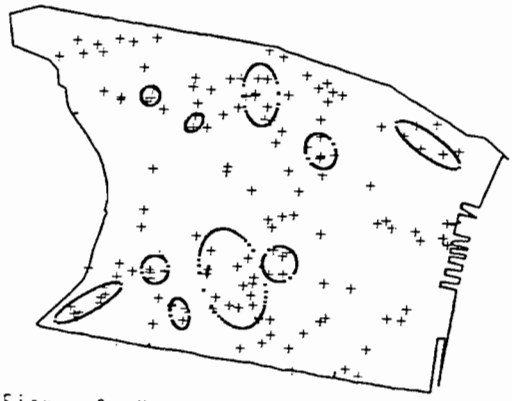
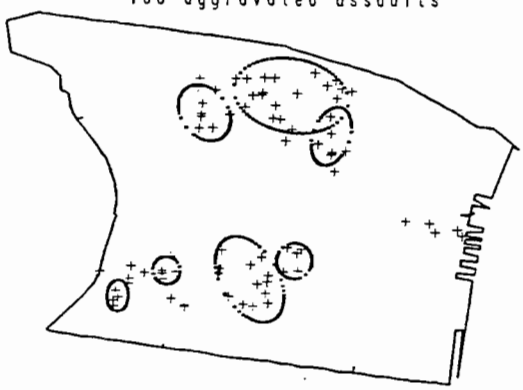


Figure 2. Hot Spots created by STACV4 238 aggravated assaults

Figure 3. Hot Spots created by STACV4, 100 aggravated assaults



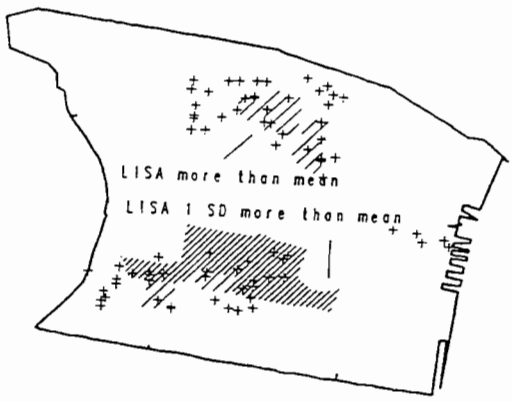


Figure 4. Hot Spot block groups identified with LISA

I ran a program called STAC to identify clusters (for details see Illinois Criminal Justice Information Authority, 1989). This program is now in widespread use in police departments, and its current (fourth) version is called STACV4. The program operates on the cartesian (state plane) coordinates of crime incidents. The software creates its own grid, based on search radius information input by the user, and encourages the use of its built in Nearest Neighbor Analysis subroutine to test for clusters. The program uses quadrat analysis to identify incident clusters, and then mathematically defines a standard deviational ellipse (a statistical method to locate the best fitting ellipse surrounding a group of coordinates) around its high value quadrats.

For this area only aggravated assaults were considered. There were 238 cases of aggravated assault in the area, as reported by the police. The spatial distribution of these incidents is shown in Figure 2 (each cross is one incident). I ran STACV4 using a search radius of 600 feet (which is about one and a half block length in this area). Twelve clusters were identified (as shown in Figure 2).

The identification of clusters by this method does not necessarily mean that the given distribution is clustered. To test for spatial autocorrelation I overlaid the study area with a grid representing census block groups.[2] Moran's I was calculated to be 0.009 for this distribution. The expected value of I or E(I) was -0.008; the z score was calculated to be 0.3233, implying that the calculated I was not significantly different from the value of I expected when the distribution was spatially random. In other words, the distribution of 238 aggravated assaults is random--there are no clusters present. It is important to note that though Moran's I indicates the absence of clusters, local clusters may still exist and ideally should be tested for.

Next, I randomly eliminated 138 incidents of aggravated assault from the dataset, so that visually there would appear to be some clustering in the distribution. First, I tested for the presence of spatial autocorrelation. The values derived were I = 0.1187, E(I) = -0.008, and z = 2.039. The calculated value of I was shown to be significantly different from the value of I expected under conditions of complete spatial randomness; in other words, the spatial distribution of these 100 incidents was clustered.

Next I ran STACV4 (with a search radius of 600 feet) on these 100 points, and ended up with eight ellipses (as shown in Figure 3). I also calculated LISA for each block group in the study region. High positive values of LISA should indicate areas of local instability, or identify parcels whose incident counts are considerably higher than their neighbors. In Figure 4 I have identified 12 such parcels (or block groups) out of 125 total parcels. Five of these parcels have LISA values higher than $LISA_{mean}$, seven of these parcels have LISA values one standard deviation or more higher than $LISA_{mean}$.

Not surprisingly, the hot spots identified by STACV4 and LISA are similar. STACV4 is over-inclusive in the upper half of the study area, while LISA appears to be over-inclusive in the lower half. There are some intuitive advantages of LISA. First, the areas identified are shaped in the form of aggregated city blocks, which is conceptually clearer than the ellipses generated by STACV4. Second, because the ellipses are mathematically generated rather than being the aggregation of high value grid cells, they tend to overlap (see both Figures 2 and 3), leading to difficulties in interpretation.

The most significant advantage of LISA (or ND or G, which have not been shown here) is not clear from the nature of the space used for the illustration. Typically, when large heterogenous areas are being studied, LISA (or ND or G) is more likely to identify small local clusters. The study area here is small, and relatively homogenous (at least in the sense that it is generally densely populated), and does not illustrate this advantage.

CONCLUDING COMMENTS

The use of point pattern analysis in criminal justice is likely to be a continuing and growing process. The methods currently in use, however, are fundamentally flawed. The need now is to integrate or customize some capabilities of Geographic Information Systems to handle and analyze crime data. This can be done easily as the geographical analysis machines now being designed

(Openshaw, et al., 1990) can be simply modified to incorporate crime data. A principal goal of spatial analysis should be to follow sound spatial principles. I have suggested here that for crime analysis these principles include: (1) tests for spatial autocorrelation to ensure that clusters actually exist in a given distribution; (2) the use of a parcel base that reflects that heterogeneity of population distribution (instead of assuming spatial homogeneity); and (3) the use of specialized measures like LISA and ND to help identify local clusters. In order to be considered geographically sound, any hot spot or cluster identification algorithm or software must incorporate these principles.

ENDNOTES

1. Upton and Fingleton (1985) who wrote an exhaustive survey of the field, found that, in their work, only one of the five most cited journals (Geographical Analysis, ranked third) was from geography; the other journals were Biometrics, Biometrika, Journal of the Royal Statistical Society (Series B), and Journal of Ecology. Many of these ideas are receiving much attention in the health literature (see Elliott et al., 1992).
2. As drawn in Figure 1, the block group in the upper left corner is incorrect. This block group represents a portion of Fairmount Park, and is actually about twice as large as shown here.

REFERENCES

Anselin, L. 1995. Local Indicators of Spatial Association--LISA. *Geographical Analysis* 27:93-115.

Boots, B. N., and Getis, A. 1988. *Point Pattern Analysis*. Newbury Park, Ca.: Sage Publications.

Chakravorty, S. 1996. A Measurement of Spatial Disparity: The Case of Income Inequality. *Urban Studies*. (forthcoming, Fall)

Elliott, P., Cuzick, J., English, D. and Stern, R. 1992. *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*. New York: Oxford University Press.

Getis, A. and Ord, J. K. 1992. The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis* 24:189-206.

Hirschfield, A. 1994. *Crime and the Spatial Concentration of Disadvantage in Northern Britain: An Analysis Using Geographical Information Systems*. The Urban Research and Policy Evaluation Regional Research Laboratory. University of Liverpool.

Illinois Criminal Justice Information Authority. 1989. *Spatial and Temporal Analysis of Crime: Users and Technical Manual*. Chicago: State of Illinois.

Morril, R. L. 1991. On the Measure of Segregation. *Geography Research Forum*. 11:25-36.

Maltz, M. D., Gordon, A. C., and Friedman, W. 1991. *Mapping Crime in its Community Setting: Event Geography Analysis*. New York: Springer-Verlag.

Odland, J. 1987. *Spatial Autocorrelation*. Newbury Park, Ca.: Sage Publications.

Openshaw, S., Cross. A., and Charlton, M. 1990. Building a Prototype Geographical Correlates Exploration Machine. *International Journal of Geographical Information Systems* 4:297-311.

Upton, G. and Fingleton, B. 1985. *Spatial Data Analysis by Example: Point Pattern and Quantitative Data*. Chichester: John Wiley and Sons.

White, M. 1983. The Measurement of Spatial Segregation. *American Journal of Sociology* 88:1008-19.

Wong, D. W. S. 1993. Spatial Indices of Segregation. *Urban Studies* 30:559-72.