Chapter 8:

Hot Spot Analysis of Points: II

Richard Block
Dept. of Sociology
Loyola University
Chicago, IL

Carolyn Rebecca Block
Illinois Criminal Justice
Information Authority
Chicago, IL

Ned Levine
Ned Levine & Associates
Houston, TX

Table of Contents

Table of Contents (continued)

Chapter 8:
Hot Spot Analysis of Points: II

This chapter continues the discussion of hot spots. Two additional routines are discussed: the STAC routine and the K-Means routine. Figure 8.1 displays the Hot Spot Analysis II page. The first of these routines, the Spatial and Temporal Analysis of Crime (STAC), was developed by the Illinois Criminal Justice Information Authority and integrated into *CrimeStat* in version 2. The second routine - K-Means, is a partitioning technique. We will start first with STAC.

Spatial and Temporal Analysis of Crime (*STAC*)

The amount of information available in an automated pin map can be enormous. When geographic information systems were first introduced into policing, there were few ways to summarize the huge reservoir of mapped information that was suddenly available. In 1989, police departments in Illinois asked the Illinois Criminal Justice Information Authority to develop a technique to identify Hot Spot Areas, the densest clusters of points on a map (Block, 1994; Block & Block, 1999; Block & Block, 1995). The result was STAC, the first crime hot spot program.[1] Through the years, bells and whistles have been added to STAC, but the algorithm has remained essentially the same. STAC is a quick, visual, easy-to-use program for identifying Hot Spot Areas.

The STAC Hot Spot Area routine in *CrimeStat* searches for and identifies the densest clusters of incidents based on the scatter of points on the map. The STAC Hot Spot Area routine creates areal units from point data and identifies the major concentrations of points for a given distribution. It then represents each dense area by either a standard deviational ellipse or a convex hull.

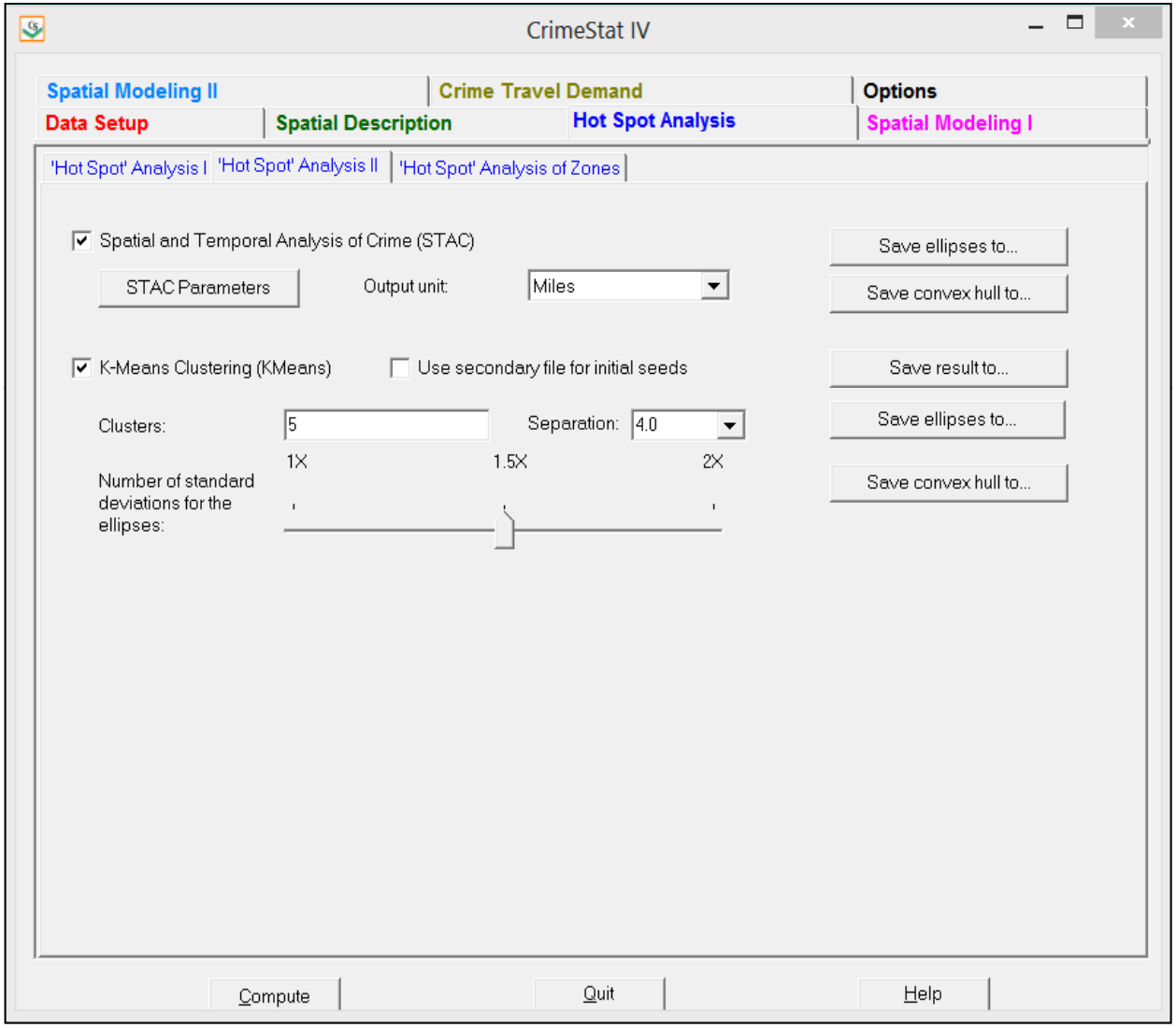
STAC is a scan-type clustering algorithm in which a circle is repeatedly laid over a grid and the number of events within the circle are counted (Openshaw, Charlton, Wymer and Craft, 1987; Openshaw, Craft, Charlton, and Birch, 1988; Turnbull, Iwano, Burnett, Howe, and Clark, 1990; Kulldorff, 1997). It, thus, shares with those other scan routines the property of multiple tests, but it differs in that the overlapping clusters are combined into larger cluster until there are no longer any overlapping circles. Thus, STAC clusters can be of differing sizes.

---

1 STAC is an abbreviation for Spatial and Temporal Analysis of Crime. The temporal section of the program was superceded by several other programs and was not updated for the millennium. Because many law enforcement users refer to STAC ellipses, we have retained that name.

Figure 8.1:

Hot Spot Analysis II Screen



The STAC Hot Spot Area routine in *CrimeStat* searches for and identifies the densest clusters of incidents based on the scatter of points on the map and then creates areal units from point data. It does this by identifying major concentrations of points for a given distribution and represents each dense area by either a standard deviational ellipse or a convex hull, or both (see Chapter 4). The boundaries of the ellipses or convex hulls can easily be displayed as mapped layers by standard GIS software.

STAC is not constrained by artificial or political boundaries, such as police beats or census tracts. This is important, because clusters of events and places (such as drug markets, gang territories, high violence taverns, or graffiti) do not necessarily stop at the border of a police beat.[2] Also, shading over an entire area may make it seem that the whole neighborhood is high-crime (or low-crime), even though the area may contain only one or two dense pockets of crime. Therefore, area-shaded maps could be misleading. In contrast, STAC Hot Spot Areas are based on the actual clusters of events or places on the map.

STAC is designed to help the crime analyst summarize a vast amount of geographic information so that practical policy-related issues can be addressed, such as resource allocation, crime analysis, beat definition, tactical and investigation decisions, or development of intervention strategies. An immediate concern of a law enforcement user of crime points on a GIS is the identification of areas that contain especially dense clusters of events. These pockets of crime demand police attention and can indicate different things for various crimes. For instance, a grouping of Criminal Damage to Property offenses could indicate gang activity. If motor vehicle thefts consistently cluster in one section of town, it could point to the need to change patrol patterns and procedures.

To take an example, Figure 8.2 shows the location of the seven densest Hot Spot Areas of street robbery in 1999 in Chicago. Four of the seven span the boundaries of police districts and two cover only a small part of a larger district. In a shaded area map, these dense clusters of robbery might be not easily identifiable. An area that is really dense might appear to be low-crime because it is divided by an arbitrary boundary. Using a shaded areal map aggregating the data within each district would give a general idea of the distribution of crime over the entire map, but it would not tell exactly where the clusters of crime are located.

For example, Figure 8.3 zooms in on Hot Spot Area 4 (the northernmost Hot Spot Area in Figure 8.2). Hot Spot Area 4 covers parts of two districts (shown by a pink boundary line in Figure 8.2). There are also four beats (shown by blue boundary lines). The shaded map indicates

2 However, there may be inadequate or, even, a lack of data on the other side of a border so that a hot spot is not fully defined.

Figure 8.2: STAC Hot Spots for 1999 Street Robberies

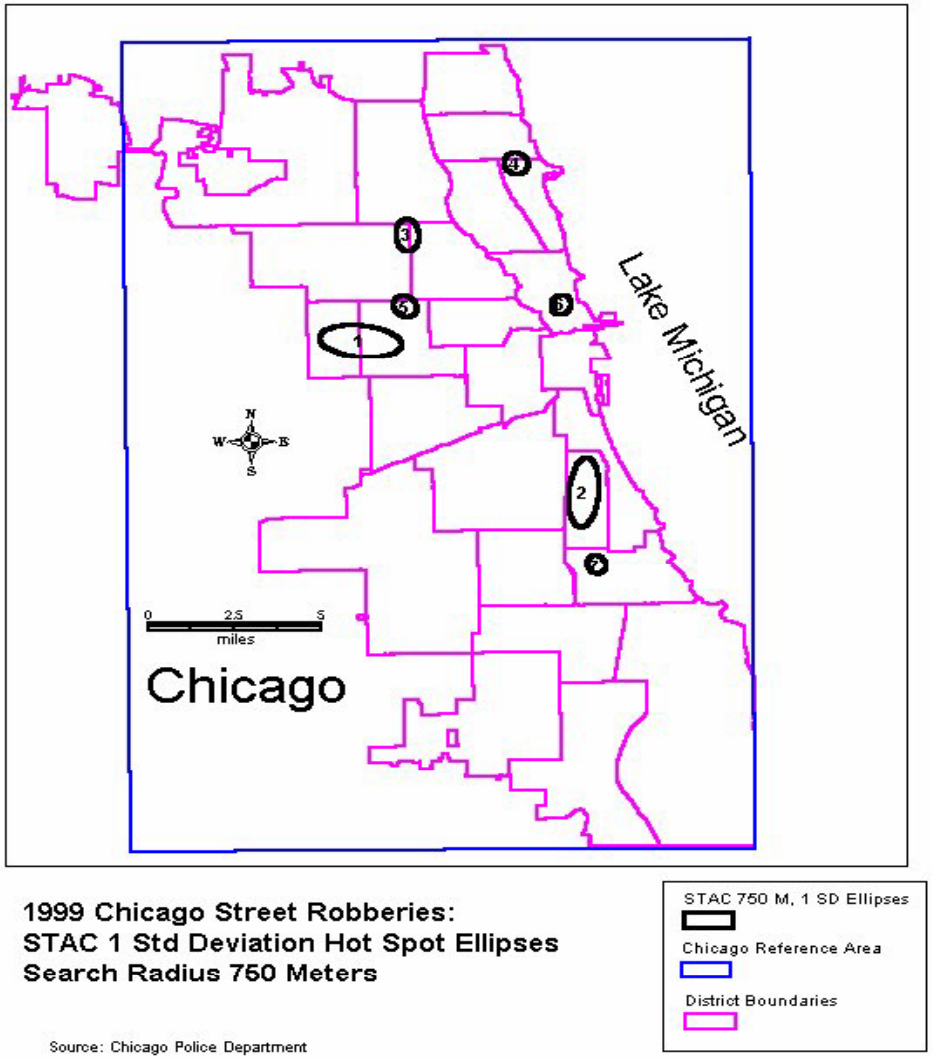
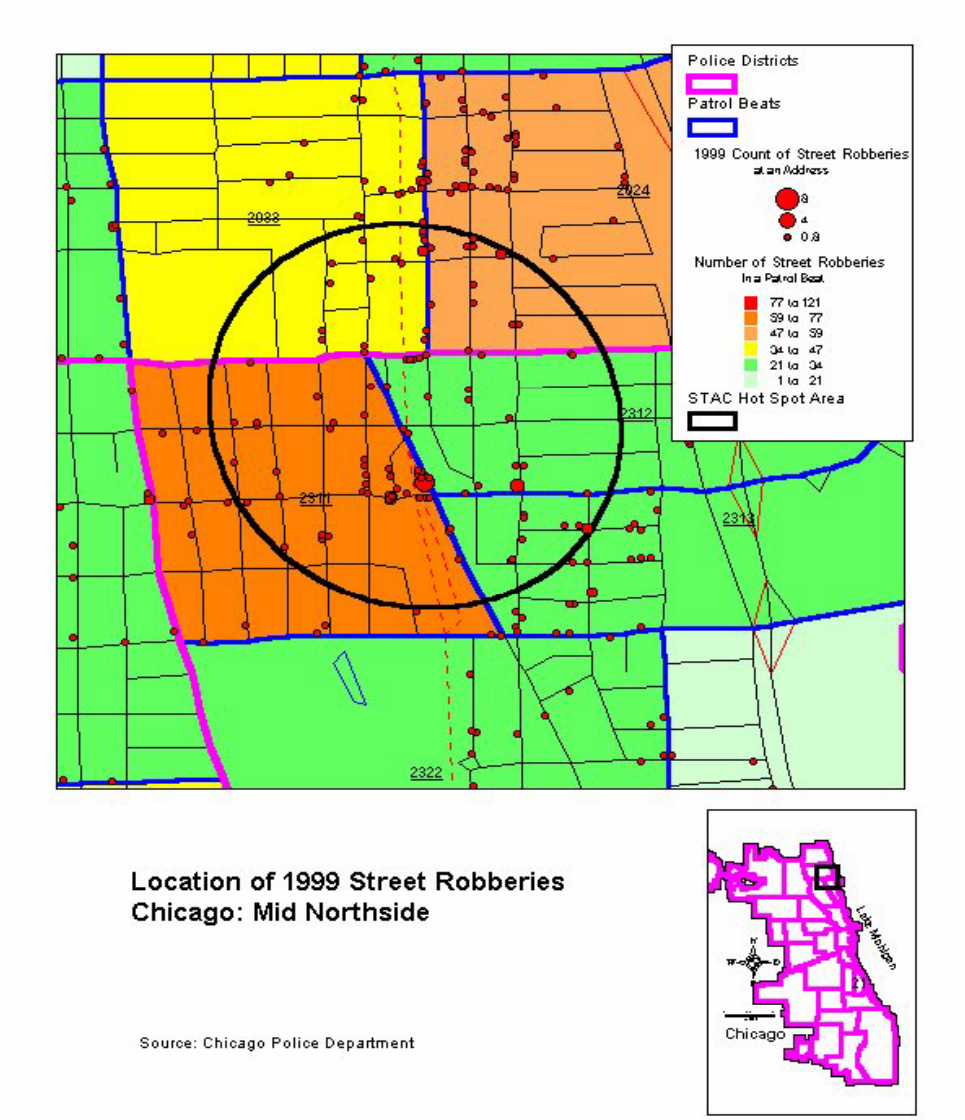


Figure 8.3: STAC 1999 Street Robbery Hot Spot Area 4



many incidents in beat 2311, but few in beats 2312, and 2313.[3] The incident distribution indicates that, while few incidents occurred overall in 2312 and 2313, most of the incidents that did occur were near to beat 2311. Incidents in beat 2311 mainly occurred on its eastern boundary. Portions of the beat were relatively free from street robbery. The Hot Spot Area identifies this clustering that spans beats and districts. Hot Spot Areas that overlap beat and district boundaries might suggest that patrol officers in these neighboring areas should coordinate their efforts in combating crime.

How STAC Identifies Hot Spot Areas

The following procedures identify hot spots in STAC. The program implements a search algorithm, looking for Hot Spot Areas:

1. STAC lays out a 20 x 20 grid structure (triangular or rectangular, defined by the user) on the plane defined by the area boundary (defined by the user on the Measurement Parameters page).

2. At each intersection of grid lines, there is a node. STAC places a circle on every node of the grid with a radius equal to 1.414 (the square root of 2) times the specified search radius. Thus, the circles overlap.

3. STAC counts the number of points falling within each circle, and ranks the circles in descending order. Multiple events can be counted at the same location.

4. STAC records all circles with at least two data points along with the number of points within each circle up to a maximum of 25 circles,. The X and Y coordinates of any node with at least two incidents within the search radius are recorded along with the number of data points found for each node.

5. These circles are then ranked in descending order according to the number of points and the top 25 search areas are selected.

6. If a point belongs to two different circles, the points within the circles are combined. This process is repeated until there are no overlapping circles. This routine avoids the problem of data points belonging to more than one cluster, and the additional problem of different cluster arrangements being possible with the same points. The result is called Hot Clusters.

3 The first two digits of a beat number designate the District.

Using the data points in each Hot Cluster, the routine calculates the standard deviational ellipse or convex hull (see Chapter 4). These are called *Hot Spot Areas*. Because the standard deviational ellipse is a statistical summary of the Hot Cluster points, it may not contain every Hot Cluster point. It also may contain points that are not in the Hot Cluster. On the other hand, the convex hulls will create a polygon around all points in the cluster.

The user can specify different search radii and re-run the routine. Given the same area boundary, different search radii will often produce different numbers of Hot Clusters. A search radius that is either too large or too small may fail to produce any. Experience and experimentation are needed to determine the most useful search radii.

Steps in Using *STAC*

STAC is available on the Hot Spot Analysis II tab under Spatial Description (see Figure 8.1). A brief summary of the steps is as follows:

1. *STAC* requires a primary file and a reference file (see Chapter 3). Optionally, STAC will use coverage area (on the Measurement Parameters tab) for simulation runs. Note: while *STAC* runs quite quickly, it runs more quickly with a Euclidean coordinate system such as UTM or State Plane.

2. Define the reference file (see Chapter 3). While *CrimeStat* does not include a data base manager or query system, a user can carry out analysis of different areas of a jurisdiction by using the boundaries of several reference areas. For example, define all of Chicago as a reference area and define each of the twenty-five police districts as additional reference areas. Hot Spot Areas can be identified for the city as a whole and for each district. In other words, the same incident file may be used for analysis of different map areas by using multiple reference files.
3. Define the search radius. Generally, a two-stage analysis is best. Start with a larger search radius and then analyze Hot Spot Areas with a smaller search radius. A search radius of more than one mile may not yield useful results in an area the size of Chicago (230 square miles).

4. Set the output units to miles or kilometers.

5. Specify the file output name for the ellipses or convex hulls.

6. Click on the *STAC* parameters button.

The object of *STAC* is to identify hot spots and display them with ellipses or convex hulls. Its key function is visual. Save the ellipses or hulls in the form most appropriate for the

system (e.g., *ArcGIS, MapInfo*). Because the ellipses or convex hulls are generated as polygons, they can be used for selections, queries, or thematic maps in a GIS. In addition to the ellipses and convex hulls, a table is output with all the information on density and location for each ellipse. It can be saved to a 'dbf' file, which can then be read by any spreadsheet program. The ellipses and convex hulls are numbered in the same order as the printed output.

STAC Parameters

The two most important parameters for running STAC are the boundary of the study area (reference area) and the search radius. A detailed discussion of the parameters follows. Figure 8.4 shows the *STAC* parameters screen.

Search radius

The search radius is the key setting in *STAC*. In general, the larger the search radius, the more incidents that will be included in each Hot Cluster and the larger the ellipse that will be displayed. Smaller search radii generally result in more ellipses of a smaller size.

A good strategy is to initially use a larger radius and then re-analyze areas that are 'hot' with a smaller radius. In Chicago, we have found that a 0.5 mile radius is appropriate for the city as a whole and a 0.25 mile search radius for one of the 25 districts. It will be necessary to experiment to determine an appropriate search radius.

Units

Specify the units for the search radius. The default is miles and the default search radius is 0.5 miles.

Minimum points per cluster

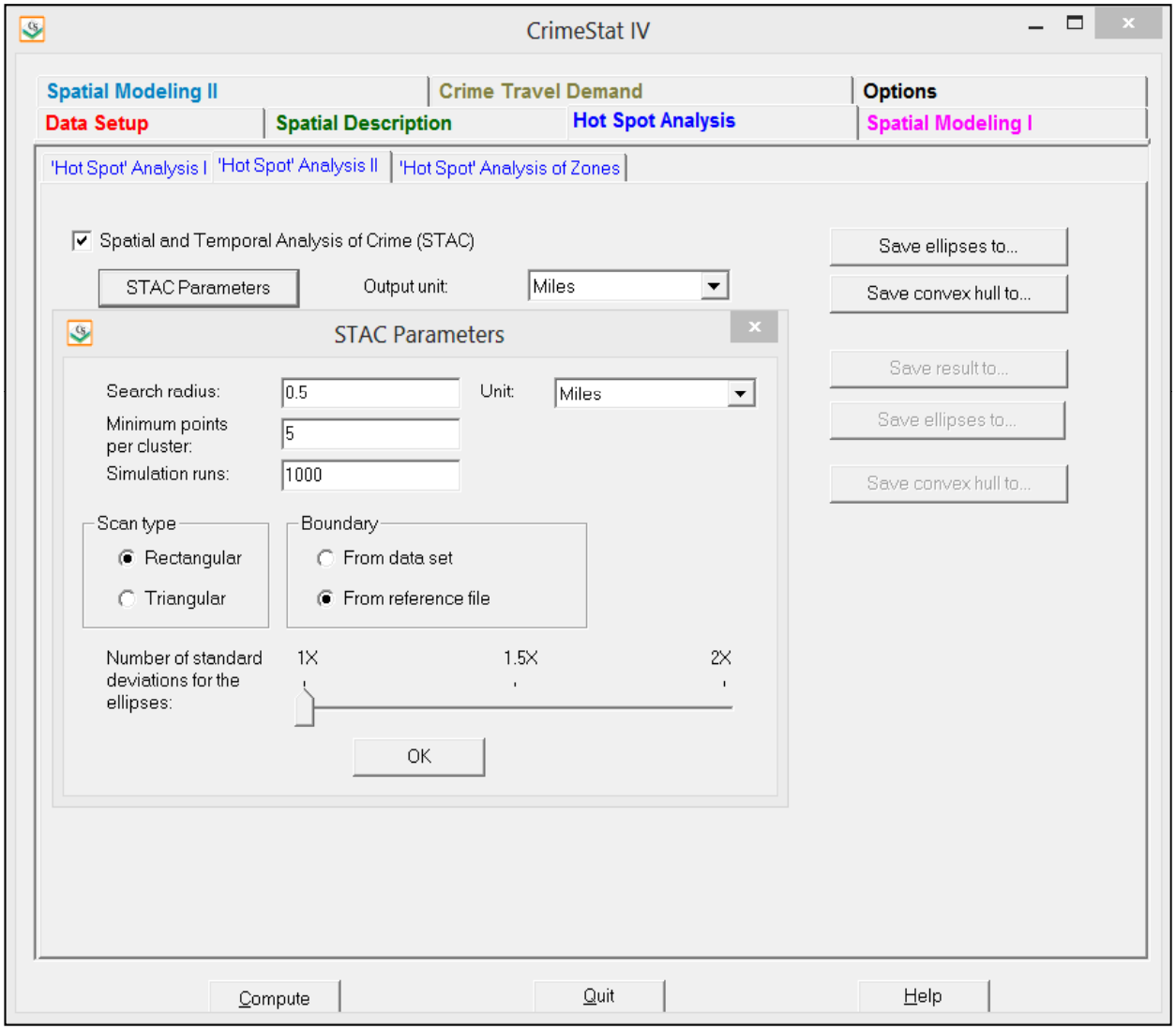
Specify the minimum number of points to be included in a Hot Cluster. The limit for the minimum points in a Hot Cluster is two. The usual choice is to use a minimum of 10.

Boundary

Select the reference file to be used for the analysis. The user can choose the boundary from the data set (i.e., the minimum and maximum X/Y values) or from the reference boundary.

Figure 8.4:

STAC Parameters Setup



In our opinion, the choice of the reference boundary is best. If the data set is used to define the reference boundary, the rectangle defined by the minimum and maximum X and Y coordinates will be used.

Scan type

Select the scan type for the grid. Choose Rectangular if the analysis area has a mostly grided street pattern. Chose Triangular if the analysis area generally has an irregular street pattern.

Graphical output files

Select whether the graphical output will be displayed as standard deviational ellipse or as convex hulls, or both (see Chapter 4). For ellipses, select the number of standard deviations for the ellipses. One (1X), 1.5X, and 2X standard deviational ellipses can be selected.

One standard deviational ellipse should be sufficient for most analysis. While 1X standard deviational ellipses rarely overlap, 1.5X and 2X standard deviational ellipses often do. A larger ellipse will include more of the Hot Cluster points; a small ellipse will produce a more focused Hot Cluster identification. The user will have to work out a balance between defining a cluster precisely compared to making it so large as to be unclear.

Simulation runs

Specify whether any simulation runs are to be made. To test the significance of *STAC* clusters, it is necessary to run a Monte Carlo simulation (Dwass, 1957; Barnard, 1963). *CrimeStat* includes a Monte Carlo simulation routine that produces approximate confidence intervals (called *credible intervals*) for the particular STAC model that has been run. Essentially, the Monte Carlo simulation assigns N cases randomly to a rectangle with the same area as specified on the Measurement Parameters tab and evaluates the number of clusters according to the defined parameters (i.e., search radius). The simulation routine repeats the random clustering *K* times, where *K* is defined by the user (e.g., 100, 1,000, 10,000).

By running the simulation many times, the user can assess credible intervals for the particular number of clusters and density of clusters. The default is zero simulation runs.. If a simulation run is selected, the user should identify the area of the study region on the Measurement Parameters tab. It is better to use the jurisdictional area rather than the reference area if the jurisdiction is irregularly shaped. For those jurisdictions, using the area defined by the

reference file coordinates (minimum X/Y and maximum X/Y) may result in identifying areas as hot spots that are not.

To compare the STAC output with the Monte Carlo simulation, there are two criteria that can be used – the number of clusters and cluster density (incidents per unit area). However, these tend to have contrary trends which depend on the search circle. Since STAC works by, first, counting incidents that fall within a search circle and, second, by aggregating overlapping search circles, a larger search circle will tend to show fewer, but higher density, clusters than would be expected on the basis of chance. The difference between the density of incidents in *STAC* ellipses in a spatially random data set and the *STAC* ellipses in the actual data set is a test of the strength of the clustering detected by *STAC*. Alternatively, a smaller search circle will tend to identify more clusters than would be expected on the basis of chance. In general, for citywide planning purposes, use a larger search circle (e.g., 0.5 miles) while for neighborhood planning purposes, use a smaller search circle (e.g., 0.1 miles or 0.25 miles).

Output

Ellipses or convex hulls

The ellipses are output with a prefix of 'St' before the output file name while the convex hulls are output with a prefix of 'Cst'before the output file name. *ArcGIS* 'shp' files can be opened as themes and can also be added as a *MapInfo* layer using the Universal Translator Tool. *MapInfo* Mif/Mid files must be imported using the command 'Table Import'. Both *MapInfo* and *ArcGIS* files are polygons and can be used for queries and thematic mapping. Google Earth 'kml' file can be displayed in that program.

Printed Output

Table 8.1 shows the printed output. Be sure to record the file name and the reference file (if any that is used). The output includes:

1. The first section of the output documents parameter settings and file size. Sample size indicates the number of points in the file specified in the setup.

2. Measurement Type indicates the type of distance measurement, direct or Indirect (Manhattan).

3. Scan Type indicates a rectangular or triangular grid specified in the setup.

Table 8.1:
Printed Output for STAC
1999 Street Robberies on Chicago's Northeast Side

```
Spatial and Temporal Analysis of Crime:
---------------------------------------

    Sample size ...........:  1181
    Measurement type ......:  Direct
    Scan type....... ......:  Rectangular
    Input units .... ......:  Degrees
    Output units ... ......:  Miles, Squared Miles, Points per Squared Miles
    Standard Deviations ...:  1
    Search radius..........:  804.672000
    Boundary...............:  -76.83302,39.23274 to -76.38390,39.59103
    Points inside boundary.:  1179
    Simulation runs .......:  1000
```

Cluster	Mean X	Mean Y	Rotation	X-Axis	Y-Axis	Area	Points	Ellipse Density
1	-76.44915	39.31484	89.41867	1.04768	0.25053	0.82460	106	128.546688
2	-76.73681	39.28658	69.91502	0.22142	0.88202	0.61354	63	102.682109
3	-76.57098	39.38499	37.10812	0.34793	0.82213	0.89863	61	67.880882
4	-76.77129	39.35987	11.26360	0.94336	0.26216	0.77695	61	78.511958
5	-76.51830	39.26019	8.37773	0.43717	0.25497	0.35017	43	22.796997
6	-76.60231	39.40086	14.84392	0.17969	0.29466	0.16634	36	16.423811
7	-76.73087	39.34246	41.07812	0.31007	0.25885	0.25215	35	38.806566
8	-76.75451	39.31110	74.78196	0.19154	0.31572	0.18998	24	26.326405

Distribution of the number of clusters found in simulation (percentile):

Percentile	Clusters	Area	Points	Density
min	12	0.01113	5	4.673554
0.5	13	0.02389	5	4.924993
1.0	13	0.03587	5	4.977644
2.5	14	0.05081	5	5.236646
5.0	14	0.06177	5	5.505124
95.0	19	1.24974	14	82.281060
97.5	19	1.39923	16	101.053102
99.0	20	1.58861	17	140.078387
99.5	20	1.67065	19	209.279368
max	20	2.08665	23	449.401912

4. Input Unit indicates the units of the coordinates specified in the setup, degrees (if latitude/longitude) or meters or feet (if projected).

5. Output Units indicate the unit of density and length specified in the setup for the output and ellipses. Output Units are generally, miles or kilometers.

6. Search Radius is the units specified in the setup. In Figure 8.2 above, this is meters.

7. Boundary identifies the coordinates of the lower left and upper right corner of the study area.

8. Points inside the boundary count the number of points within the reference file. This may be fewer than the number of points in the total file when a smaller area is being used for analysis (see Table 8.1).

9. Simulation Runs indicate the number of runs, if any specified in the setup.

10. Finally, STAC printed output provides summary statistics for each Hot Spot Area:

    A. Cluster identification number for each ellipse. This corresponds to their order in a table view in *ArcGIS* or the browser in *MapInfo*.

    B. Mean X and Mean Y - Coordinates of the mean center of the ellipse.

    C. Rotation- the degrees the ellipse is rotated (0 is horizontal; 90 is vertical).

    D. X-axis and Y-axis - the length (in the selected output units) of the x and y axis. In the example, the length of the x axis of ellipse 1 is 1.04768 miles.

    E. Area - the area of the ellipse in square units. Ellipses are ordered according to their size. In the example, Ellipse 1 is 0.8246 square miles.

    F. Points - the number of points in the Hot Cluster. In the example, there are 61 points in cluster 3.

    G. Cluster Density - the number of points per square unit. The largest cluster is not necessarily the densest. In this output, cluster eight is the smallest, but its density is higher than two other clusters.

H. The distribution of the simulations (if specified).

Note that the number of actual clusters in the example (8) is smaller than the number that would be expected if the data were randomly distributed at the 95 percentile (19). The reason for this is that STAC aggregates smaller clusters that are close to each other, that is where their search circles overlap. Hence, with a large search circle, as in this output – 0.5 miles, will generally lead to fewer clusters than a Monte Carlo simulation. On the other hand, the cluster density indicates that two of the clusters (1 and 2) have a higher density than the 95 percentile density. These clusters are most likely real clusters, rather than random collections, and should be the focus of further analysis.

The best way to print or save *CrimeStat* printed output is to place the cursor inside the output window and *Select all,* then copy and paste the selection into a word processing document in landscape mode. Make sure to adequately annotate the file, especially the type of incidents, the reference boundary, and the name of the output file. This can be very important for future reference.

Example: A STAC Analysis of 1999 Chicago Street Robberies

STAC Hot Spot Areas were calculated for all street (or sidewalk or alley) robberies occurring in Chicago in 1999 (n=13,009).[4] There were 13,007 within the search boundary. The search radius was set for 750 meters (approximately 0.5 mile), and the ellipses were set to one standard deviation. Ten was the minimum number of incidents per cluster.

In Figure 8.2 (shown earlier), STAC detected seven ellipses. The areas of the seven ellipses ranged from 5 square kilometers to 0.7 square kilometers, and the number of incidents in an ellipse ranged from 760 to 153. The smallest ellipse (number 7 in Figure 8.2) was the densest, 222 robberies per square kilometer. Of the 13,007 street robberies, 2,375 were in a cluster. Therefore, 18 percent of all of Chicago's street robberies in 1999 occurred in 6% of its 233 square mile area.

To map the results, the ellipse boundaries were imported into *MapInfo* as a mif/mid file and overlaid on a map of police districts. The large blue rectangle in Figure 8.2 designates the search boundary (reference file). O'Hare Airport was excluded because exact geo-coding is not possible for the few street robberies that occurred there. At a city-wide scale, the map is interesting, but is mainly useful for confirming what is already known. Ellipse 1, on the west

4 The Chicago Police Department made available the incidents in this analysis to Richard Block for the evaluation of the Chicago Alternative Police Strategy (CAPS).

side, has had a high level of violence for many years. Ellipses 2 and 6 are centered on areas where high rise public housing projects are gradually being abandoned. Overall, these ellipses are not very useful for tactical purposes. However, they point out that four Hot Spot Areas cross District boundaries, and that the large number of street robberies in these areas might be lost in separate district reports.

A Neighborhood STAC Analysis

The presence of Ellipse 4 (the northernmost ellipse in Figure 8.2) might be unexpected to many Chicagoans. The mid-Northside, near the Lake Michigan, is generally considered to be a relatively affluent and safe neighborhood. However, the neighborhood around Ellipse 4 has had a high level of crime for many years. It was an entertainment center in the Roaring Twenties, and several institutions of that era remain. Today it is an area with multiple, often conflicting, uses. A more detailed analysis of the neighborhood with the help of STAC may point to specific areas that need increased patrol or prevention activities.

The second step of STAC analysis was to define a focused search boundary area around Ellipse 4. This was done easily by creating a new map layer in MapInfo and drawing a rectangle around the desired study area. Clicking on the study area gave the required *CrimeStat* reference boundary maximum and minimum coordinates. Using this more focused boundary, STAC was run a second time with a 200 meter search radius and the same file of 13,009 cases. The search boundary (reference file) now contained 442 incidents. STAC detected three ellipses that contained 231 incidents. The STAC ellipses were then imported into *MapInfo* and mapped (Figure 8.5).

As the area covered by a map grows smaller, detailed information about crime patterns and the community can be added. In this map, the STAC ellipses were overlaid on the locations of incidents (sized according to the number occurring at each location) and streets.[5] Much of the area is relatively crime-free. The most frequent locations for street robbery do not coincide with main streets. Street robbery incidents tend to cluster near rapid transit stations and the blocks immediately surrounding them. For example, Argyle Street, between Broadway and Sheridan, is the site of 'New China Town'. It is an area with a number of street robberies and is a destination area for 'Northsiders' who want an inexpensive Chinese or Vietnamese meal.

---

5 In general a designated main surface street occurs every mile on Chicago↘s grid, and there are eight blocks to the mile. In this map, Lawrence and Ashland are main Grid streets. In this area, there are also several diagonal main streets that either follow the lake shore or old Indian trails.

There is a particularly risky area in the neighborhood of Broadway and Wilson adjacent to Truman Community College. In a previous analysis of the Bronx, Fordham University was shown to be a similar attractor for robbery incidents. Colleges supply good targets for street robbery. Also, authority for security is split between the college and the city police. The area around Broadway and Wilson has been risky for many years. Ninety years ago, it was the northern terminus of rapid transit, and the site of several very inexpensive hotels, two of which still existed. Today the area has several pawn shops and currency exchanges. There is an ATM located in the EL station. In 1999, the area looked dangerous and dirty. Finally, the area has many blind corners and alleys that could serve as sites for robbery; this is unusual for Chicago.

The census block that includes the northwest corner of Broadway and Wilson ranked fifth among Chicago's 21,000 census blocks in number of street robberies in 1999.[6]

Changes need to be made to reduce the risk of street robbery in this area. Mapping identifies a problem with street robberies, but to investigate possible changes it is necessary to go beyond mapping. Aside from changes in patrol practices, what physical changes might aid in crime reduction? The campus has very little parking. The administration assumes that students take public transportation, but many do not. A secure parking garage that could serve both the elevated station and the school could be constructed (vacant land is available). In addition, increased police patrol in the area between the school and the el station could be implemented.

Advantages of STAC

STAC has a number of advantages as a clustering algorithm:

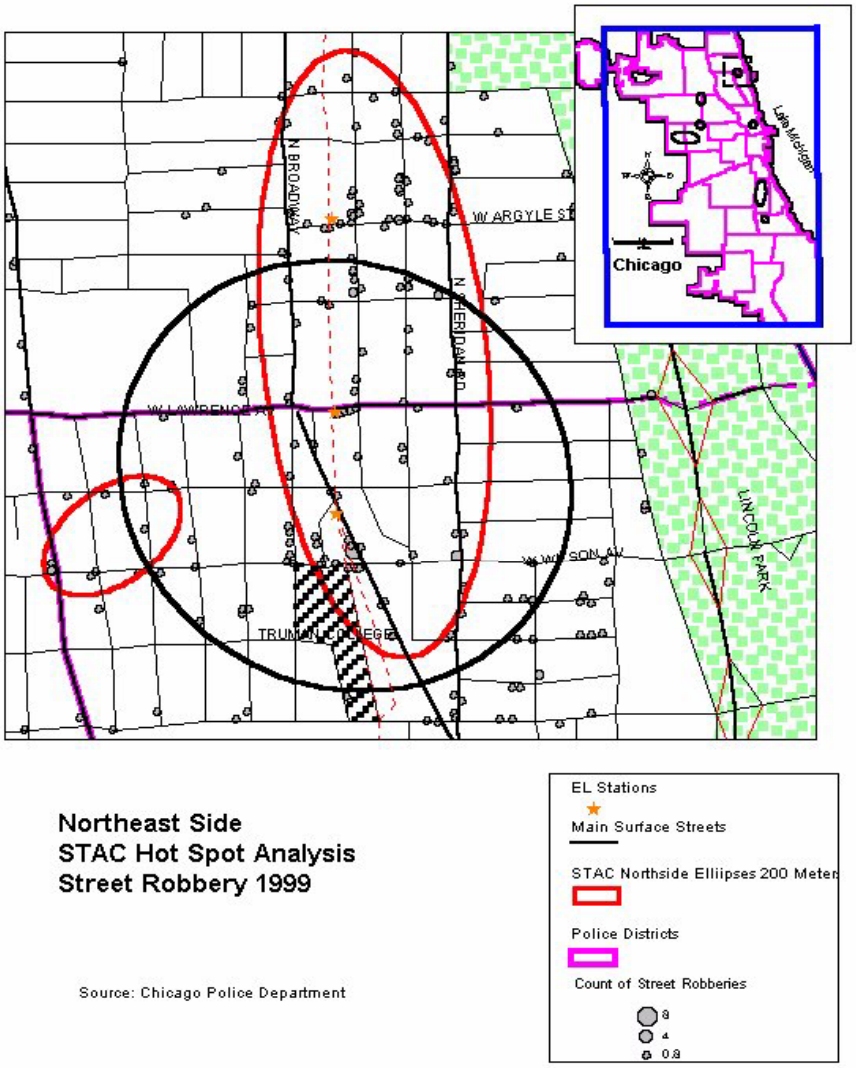
1. The routine can analyze a very large number of cases quickly. It is very fast using a Euclidean projection such as UTM or State Plane, but not quite as fast using spherical coordinates (latitude/longitude).

2. The user can control the approximate size of the ellipses through the search radius, the minimum number of points per hot spot, and the study area. These features allow for a brod search for Hot Spot Areas over an entire city and a second search concentrating on a smaller area and more focused Hot Spot Areas for local tactical use.

---

6 This example was originally conducted with CrimeStat II. In subsequent years, many of these suggestions were implemented and the area is no longer a hot spot.

Figure 8.5: STAC Hot Spots for Northeast Side Street Robberies



3. STAC and Nnh hierarchical clustering (discussed in Chapter 7) are complimentary. The Nnh first derives small ellipses and then aggregates to larger ones. The recommended STAC procedure is to first derive large area ellipses and then break these down into smaller areas for tactical analysis. There should be some convergence between the two approaches.

4. The visual display of STAC ellipses or convex hulls is quite intuitive, especially for area-wide interventions (e.g., patrol beats)

5. Hot spots need not be limited to a single kind of crime, place or even. For example, ellipses of drug crime can be overlaid on those for burglary. Some causal factors are also analyzable with STAC ellipses. For example, ellipses of street robbery can be compared to those for liquor licenses.

6. STAC is a density search clustering method that adapts itself to the size of the clusters. Essentially, it looks for areas of common, high density.

7. Unlike the Nnh routine, which has a constant threshold (search radius), STAC can create clusters of unequal size because overlapping clusters are combined until there is no overlap.

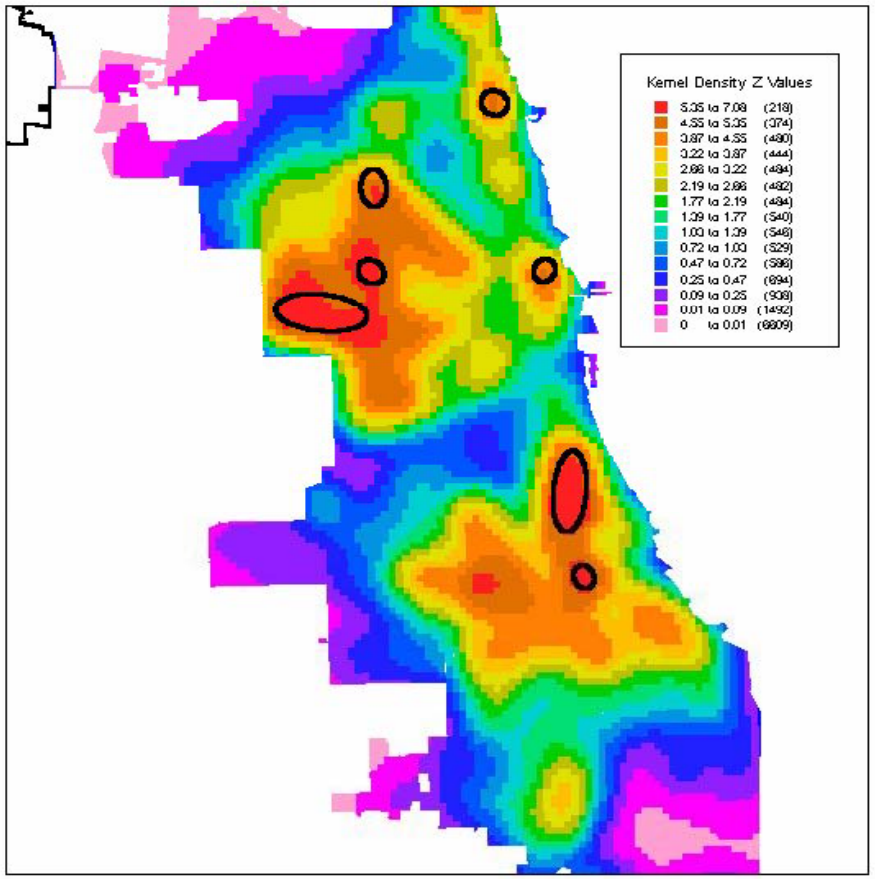
Limitations of STAC

There are also some limitations to using STAC:

1. The distribution of incidents within clusters is not necessarily uniform. The user should be careful not to assume that it is. A mapped theme of the Mode routine (see Chapter 7) according to number of incidents or the single kernel density interpolation (see Chapter 10) overlaid with STAC ellipses are good ways to overcome this problem (see Figure 8.5 above and Figure 8.6 below).

2. STAC tends to create larger clusters than the Nnh. The reason is that it combines points from overlapping search circles. It is unable to identify smaller clusters that are part of a larger grouping (a hierarchy) and, instead, tends to choose the larger grouping. The result is the density of events in STAC clusters are not as intense as in Nnh first-order clusters, but are more similar to Nnh second-order clusters (Chainey, Thompson & Uhlig, 2008; Levine, 2008).

Figure 8.6: STAC Robbery Hot Spots and Kernel Density Estimation



Chicago Street Robbery 1999 :
Comparison of STAC and
Single Kernel Density Estimation

For example, with a 1996 Baltimore County burglary file of 6,051 incidents, the default settings for STAC (0.5 mile search radius and a minimum of 5 points per cluster) produced 8 clusters compared to 158 for the Nnh with its default settings (random nearest neighbor distance and a minimum of 10 points per cluster). Depending on the purpose of the clustering, this can be an advantage or a disadvantage. STAC clusters can identify areas for patrol beats but are less able to identify very small areas where there is an intense concentration of events and which require geographically-specific interventions (e.g., improving street lighting, setting up block-wide security strategies).

3. Small changes in the STAC study area boundary can result in quite different depictions of the ellipses, even with the same study area measurement. Retaining the same reference file for repeated analyses alleviates this problem. The analysis should also be documented for the analysis parameters.

4. Because STAC aggregates overlapping search circles, it tends to miss identifying smaller clusters that are close to each other. This is particularly true when a larger search circle is used. Thus, the method tends to increase Type II statistical errors (failing to reject a false null hypothesis). The use of smaller search circles can minimize this problem. While there are definite uses in a larger search circle, for example in identifying patrol areas or multi-neighborhood crime hot spots, the user needs to be aware of how the search circle can affect the number of clusters identified and the potential for missing clusters that are actually separate yet close to each other.

5. STAC is based on the distribution of events. Neither land use nor risk factors is accounted for. It is up to the analyst to identify the characteristics that make a Hot Spot 'hot'.

Nevertheless, if used carefully, STAC is a useful tool for detecting clusters and can allow an analyst to experiment with varying search radii and reference boundaries.

K-Means Partitioning Clustering

The *K-Means* clustering routine (Kmeans) is a partitioning procedure where the data are grouped into *K* groups defined by the user. A specified number of seed locations, *K*, are defined by the user (Fisher, 1958; MacQueen, 1967; Aldenderfer and Blashfield, 1984; Systat, 2008). The routine tries to find the best positioning of the *K* centers and then assigns each point to the center that is nearest. Like the nearest neighbor hierarchical (Nnh) routine, the Kmeans assigns

points to one, and only one, cluster. However, unlike the Nnh procedure, all points are assigned to clusters. Thus, there is no hierarchy to the assignment; that is there are no second- or higher-order clusters. It is part of a family of cluster methods called *supervised clustering* (Finley & Joachins, 2005; Eick, Zeidat & Zhao, 2004).

The technique is useful when a user want to control the grouping. For example, if there are 10 police precincts in a jurisdiction, an analyst might want to identify the 10 most compact clusters, one for each precinct. Alternatively, if a previous analysis has shown there were 24 clusters, then an analyst could check whether the clusters have shifted over time by also asking for 24 clusters. By definition, the technique is somewhat arbitrary since the user defines how many clusters are to be expected. Whether a cluster should be considered a hot spot or not should depend on the extent to which a user wants to replicate hot spots.

The theory of the K-Means procedure is relatively straightforward. The implementation is more complicated. K-Means represents an attempt to define an optimal number of *K* locations where the sum of the distance from every point to each of the *K* centers is minimized. It is a variation of the old location theory paradigm of how to locate *K* facilities (e.g., police stations, hospitals, shopping centers) given the distribution of population (Haggett, Cliff, and Frey, 1977). That is, how does one identify *supply* facilities in relation to the location of *demand*? In theory, solving this question is an empirical solution, what is frequently called *global optimization*. One tries every combination of *K* objects where *K* is a subset of the total population of incidents (or people), *N*, and measures the distance from every incident point to every one of the *K* locations. The particular combination which gives the minimal sum of all distances (or all squared distances) is considered the best solution. In practice, however, solving this is computationally almost impossible, particularly if *N* is large. For example, with 6000 incidents grouped into 20 partitions (clusters), one cannot solve this with any normal computer since there are:

$$\frac{6000!}{20!5980!} = 1.456x10^{57} \tag{8.1}$$

combinations. No computer can solve that number and few spreadsheets can calculate the factorial of *N* greater than about 127.[7] In other words, it is almost impossible to solve computationally.

Practically, therefore, the different implementations of the K-Means routine make initial guesses about the *K* locations and then optimize the seating of this location in relation to the

7 The total number of ways for selecting K distinct combinations of *N* incidents, irrespective of order, is $\frac{N!}{K!(N-K)!}$ (Burt and Barber, 1996, 155).

nearby points. This is called *local optimization*. Unfortunately, each K-Means routine has a different way to define the initial locations so that two K-Means procedures will usually not produce the same results even if *K* is identical (Everitt, 2011; Systat, 2008; Everitt, Landau & Leese, 2001).

CrimeStat K-Means Routine

The K-Means routine in *CrimeStat* also makes an initial guess about the *K* locations and then optimizes the distribution locally. The procedure that is adopted makes initial estimates about location of the K clusters (seeds), assigns all points to its nearest seed location, re-calculates a center for each cluster which becomes a new seed, and then repeats the procedure all over again. The procedure stops when there are very few changes to the cluster composition (see endnote *i*).

The default K-Means clustering routine follows an algorithm for grouping all point locations into one, and only one, of these *K* groups. There are two general steps: 1) the identification of an initial guess (seed) for the location of the *K* clusters, and 2) local optimization which assigns each point to the nearest of the *K* clusters. First, a grid is overlaid on the data set and the number of points falling within each grid cell is counted. The grid cell with the most points is the initial first cluster and the centroid of the cell becomes the initial seed location.

The second initial cluster is the grid cell with the next most points that are separated by at least:

$$Separation = t * 0.5\sqrt{\frac{A}{N}} \tag{8.2}$$

where t is the Student's t-value for the .01 significance level (2.358), A is the area of the region, and N is the sample size. Again, the centroid of the grid cell becomes the initial second seed location. A third initial cluster is then selected which is the grid cell with the third most points and which is separated from the first two grid cells by at least the separation factor defined above. This process is repeated until *K* initial seed locations are chosen.

The algorithm then conducts *local optimization*. It assigns each point to the nearest of the initial K seed locations to form an initial cluster. For each of the initial clusters, the routine then calculates the center of minimum distance and re-assigns all points to the center of minimum distance to which it is closest. This becomes the second iteration of clusters with the center of minimum distance being the second seed location.

The routine repeats this process (assigning each point to the nearest seed location, re-calculating the center of minimum distance for each cluster to form a new seed location, and then re-assigning all points to the nearest new seed location) until no points change clusters. Finally, for each cluster, the routine outputs to the screen the statistics for a 1X standard deviational ellipse and can also output the results graphically as either standard deviational ellipses (1X, 1.5X, or 2X) or as convex hulls.

Control over Initial Selection of Clusters

Changing the separation between clusters

One problem with this approach is that in highly concentrated distributions, such as with most crime incidents in a metropolitan area, the separation between clusters may not be sufficiently large to detect clusters farther away from the concentration; the algorithm will tend to sub-divide concentrated groupings of incidents into multiple clusters rather than seek clusters that are less concentrated and, usually, farther away. To increase the flexibility of the routine, *CrimeStat* allows the user to modify the initial selection of clusters since this has a large effect on the final grouping (Everitt, 2011).

There are two ways the initial selection of cluster centers can be modified. First, the user can increase or decrease the ***separation factor***. Formula 8.2 is still used to separate each of the initial clusters, but the user can either select a t-value from 1 to 10 from the drop down menu or write in any number for the separation, including fractions, to increase or decrease the separation between the initial clusters. The default separation is set at 4. The effect of this is to modify the grid cell sizes for the initial cluster so as to force larger or small distances between the clusters.

Figure 8.7 shows a simulation of eight clusters in Baltimore County, four of which have higher concentrations than the other two. Figure 8.8 shows the results of running the K-Means clustering routine twice, both of which requested K=8 groupings. However, in one of the partitions there was a separation of 4 (the default separation shown as dashed green ellipses) while the other partition had a separation of 18 (solid blue ellipses). As seen, the partition with the larger separation captures the eight clusters better. With the smaller separation (4), the routine subdivided the dense cluster in the west into three separate clusters while combining one of these with the grouping of points directly to the north. Similarly, it combined two groupings in the northern part of the study area into a single cluster. The effect of increasing the separation was to produce a better visual fit with the groupings of the points.

Figure 8.7:

Separated Data and K-Means Clustering

Data Grouped into Eight Clusters

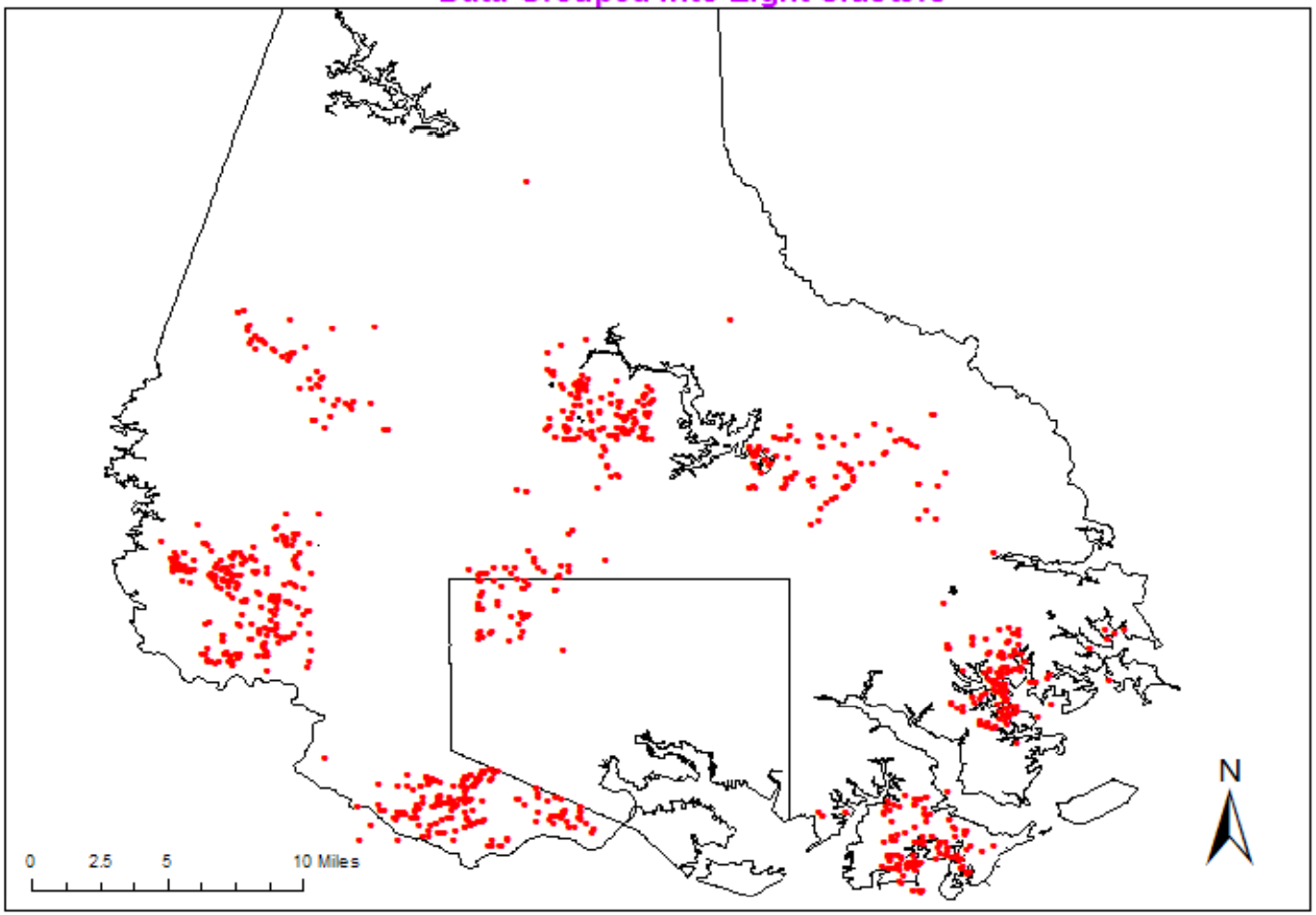
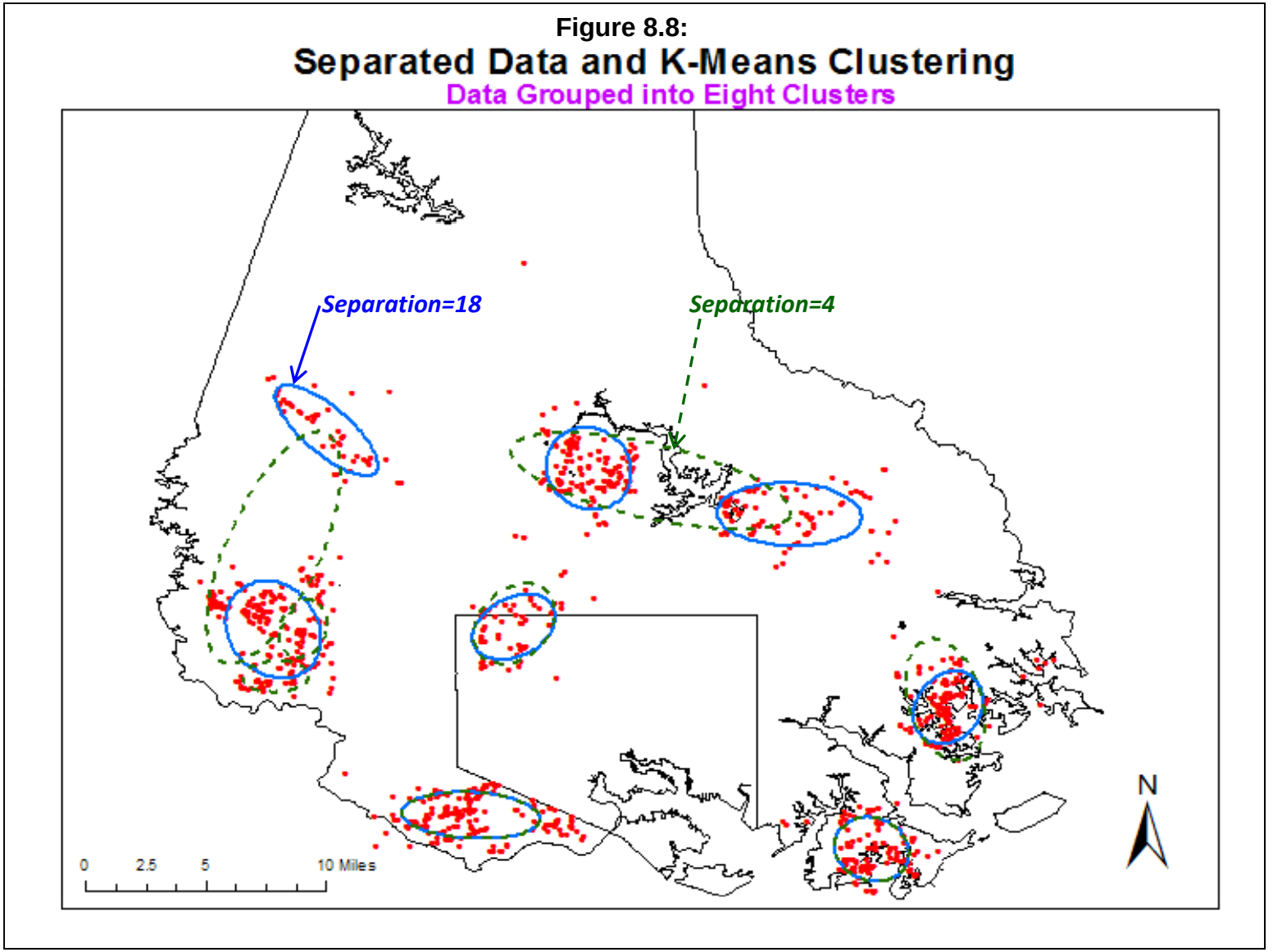


Figure 8.8:


Separated Data and K-Means Clustering
Data Grouped into Eight Clusters
Separation=18
Separation=4
0
2.5
5
10 Miles
N

One has to be careful tweaking the cluster structure, however. For example, as we increased the separation beyond 18, the number of clusters actually decreased. A separation of 20 produced only 7 clusters while a separation of 30 produced only 3. The algorithm could not solve for 8 clusters with such a large separation between them being required.

Selecting the initial seed locations

A second way to control the initial selection of clusters is that the user can define the actual locations for the initial cluster centers. This approach was used by Friedman and Rubin (1967) and Ball and Hall (1970). In *CrimeStat*, the user-defined locations are entered with the secondary file which lists the location of the initial clusters. The routine uses the number of points in the secondary file as *K* and the X/Y coordinates of each point as the initial seed locations. It then proceeds in the same way with local optimization.

When eight points that were approximately in the middle of the eight clusters in Figure 8.7 were input as the secondary file, the K-Means routine immediately identified the eight clusters (results not shown). Again, depending on the purpose the user can test a particular clustering by requiring the routine to consider that model, at least for the initial seed location. The routine will conduct local optimization for the rest of the clustering, as in the above method.

K-Means Screen Output

The K-Means output has both screen and graphical output. The screen output includes the parameters for the 1X standard deviational ellipse of each cluster in the table. In addition, the routine can output graphically the clusters as standard deviational ellipses (1X, 1.5X, or 2X) or convex hulls. The convex hull draws a polygon around all the points in a cluster (see Chapter 4). Hence it is a literal description of the extent of the cluster. The ellipse, on the other hand, is an abstraction for a cluster and may be arranged in an irregular manner. For a small area, a 1X standard deviational ellipse or a convex hull would be a good way to display the ellipses but may not be very visible with a regional view. The user has to balance the need to accurately display the cluster compared to making it easier for a viewer to understand its location.

Mean squared error

In addition, the output for each cluster lists two additional statistics:

$$Sum\ of\ squares\ of\ cluster\ C = SSE_C = \sum_{i=1}^{N_C}[(X_{iC} - \bar{X}_C)^2 + (Y_{iC} - \bar{Y}_C)^2] \qquad (8.3)$$

$$Mean\ squared\ error\ of\ Cluster\ C = MSE_C = \frac{SSE_C}{(N_C - 1)} \quad (8.4)$$

where X_{iC} is the X value of a point that belongs to cluster C, Y_{iC} is the Y value of a point that belongs to cluster C, $MeanX_C$ is the mean X value of cluster C (i.e., of only those points belonging to C), $MeanY_C$ is the mean Y value of cluster C, and N_C is the number of points in cluster C.

There is also a total sum of squares and a total mean square error which is summed over all clusters:

$$Total\ sum\ of\ squares = SS = \sum_{C=1}^{K} SSE_C \quad (8.5)$$

$$Total\ mean\ squared\ error = MSE = \frac{SS}{(N - K - 1)} \quad (8.6)$$

where SSE_C is the sum of squares for cluster C, N is the total sample size, and K is the number of clusters. The sum of squares is the squared deviations of each cluster point from the center of minimum distance while the mean squared error is the average of the squared deviations for each cluster corrected for degrees of freedom.

The sum of squares (or sum of squared errors) is frequently used as a criterion for identifying 'goodness of fit' (Everitt, 2011; Everitt, Landau & Leese, 2001; Aldenderfer & Blashfield, 1984; Gersho & Gray, 1992). In general, for a given number of clusters, *K*, partitions with a smaller sum of squares and, correspondingly, a smaller mean square error are better defined than partitions with a larger sum of squares and larger mean squared error. Similarly, a K-Means solution that produces a smaller overall sum of squares is a tighter grouping than a grouping that produces a larger overall sum of squares.

But, there can be exceptions. If there are points which are outliers, that is which do not obviously fall into one cluster or another, re-assigning them to one or another cluster can distort the sum of squares statistics. Also, in highly concentrated distributions, such as with crime incidents, a smaller sum of squares criteria can be obtained by splitting the concentrations rather than clustering less central and less dense groups of incidents (such as in Figure 8.7). The result, while minimizing the sum of squared errors from the cluster centers, will be less desirable because the peripheral clusters are ignored. Thus, these statistics are presented for the user's information only. In assigning points to clusters, *CrimeStat* still uses the distance to the nearest seed location, rather than a solution that minimizes the sum of squared distances.

K-Means Graphical Output

Finally, the K-Means clustering routine (Kmeans) can output clusters graphically as either ellipses or convex hulls, similar to the other clustering routines. For the ellipses, the user can choose between 1X, 1.5X, and 2X standard deviations. The ellipses are output with the prefix 'KM' before the file name. It should be noted, however, that the ellipses are an abstraction of the cluster. The clusters are *not* necessarily arranged in ellipses. They are for visualization purposes only. For the convex hull, the routine draws a polygon around the points in each cluster. The graphical convex hulls are output with the prefix 'CKM' before the file name.

Naming convention for K-Means clusters

The naming convention for the K-Means outputs is:

Km<*username*>	[for the ellipse]
Ckm<*username*>	[for the convex hull]

where *username* is the name of the file provided by the user. Within the file, each cluster is named

KmEll<N><*username*>	[for the ellipse]
CkmHull<N><*username*>	[for the convex hull]

where *N* is the cluster number and *username* is the name of the file provided by the user. For example,

KmEll3robbery

is the third ellipse for the file called 'robbery' and

CkmHull12burglary

is the 12th convex hull for the file called 'burglary'.

For the ellipses, a slide-bar allows ellipses to be defined for 1X, 1.5X, and 2X standard deviations and can be output in *ArcGIS* '.shp', *MapInfo* '.mif' or various Ascii formats. The convex hulls, on the other hand, draw a polygon around the clustered points.

Example: K-Means Clustering of Baltimore County Street Robberies

In *CrimeStat*, the user specifies the number of groups to sub-divide the data. Using the 1996 robbery incidents for Baltimore County, the data were partitioned into 10 groups with the K-Means routine (Figure 8.9). As can be seen, the clusters tend to fall along the border with Baltimore City. But there are three more dispersed clusters, one concentrated in the central eastern part of the county and two north of the border with the City. Because these clusters are very large, a finer mesh clustering was conducting by partitioning the data into 34 clusters (Figure 8.10). Thirty-five clusters were requested but the routine only found 34 seed location. Consequently, it outputted 34 clusters, which are displayed as ellipses. Though the ellipses are still larger than those produced by the nearest neighbor hierarchical procedure (see figure 7.7 in Chapter 7), there is some congruency; clusters identified by the nearest neighbor procedure have corresponding ellipses using the K-Means procedure.

Figure 8.11 shows a section of southwest Baltimore County with four full clusters and three partial clusters visible. They are displayed as convex hulls. Looking at the distribution, several clusters make intuitive sense while a couple of others do not. For example, the two clusters at the top of the map highlight a concentration along a major arterial (U.S. Highway 40). Similarly, the cluster in the middle right appears to capture incidents along two arterial roads. However, the other three full clusters do not appear to capture meaningful patterns and appear somewhat arbitrary.

Other uses of the K-Means algorithm are possible. For example, one problem that affects most police departments is the need to allocate personnel throughout a city in a balanced and fair way. Too often, some police precincts or districts are overburdened with Calls for Service whereas others have more moderate demand. The issue of re-drawing or re-assigning police boundaries in order to re-establish balance is a continual one for police departments. The K-Means algorithm can help in defining this balance, though there are many other factors that will affect particular boundaries. The number of groupings, K, can be chosen based on the number of police districts that exist or that are desired. The locations of division or precinct stations can be entered in a secondary file in order to define the initial 'seed' locations. The K-Means routine space. Once an agreed upon solution is found, it is easy to then re-assign police beats to fit the new arrangement.

Advantages and Disadvantages of the K-Means Procedure

In short, the K-Means procedure will divide (partition) the data into the number of groups specified by the user, *K*. Whether these groups make any sense or not will depend on how carefully the user has selected clusters. Choosing too many will lead to defining patterns that do

Figure 8.9:

Baltimore County Robbery Hot Spots: 1996

K-Means Clustering with K=10

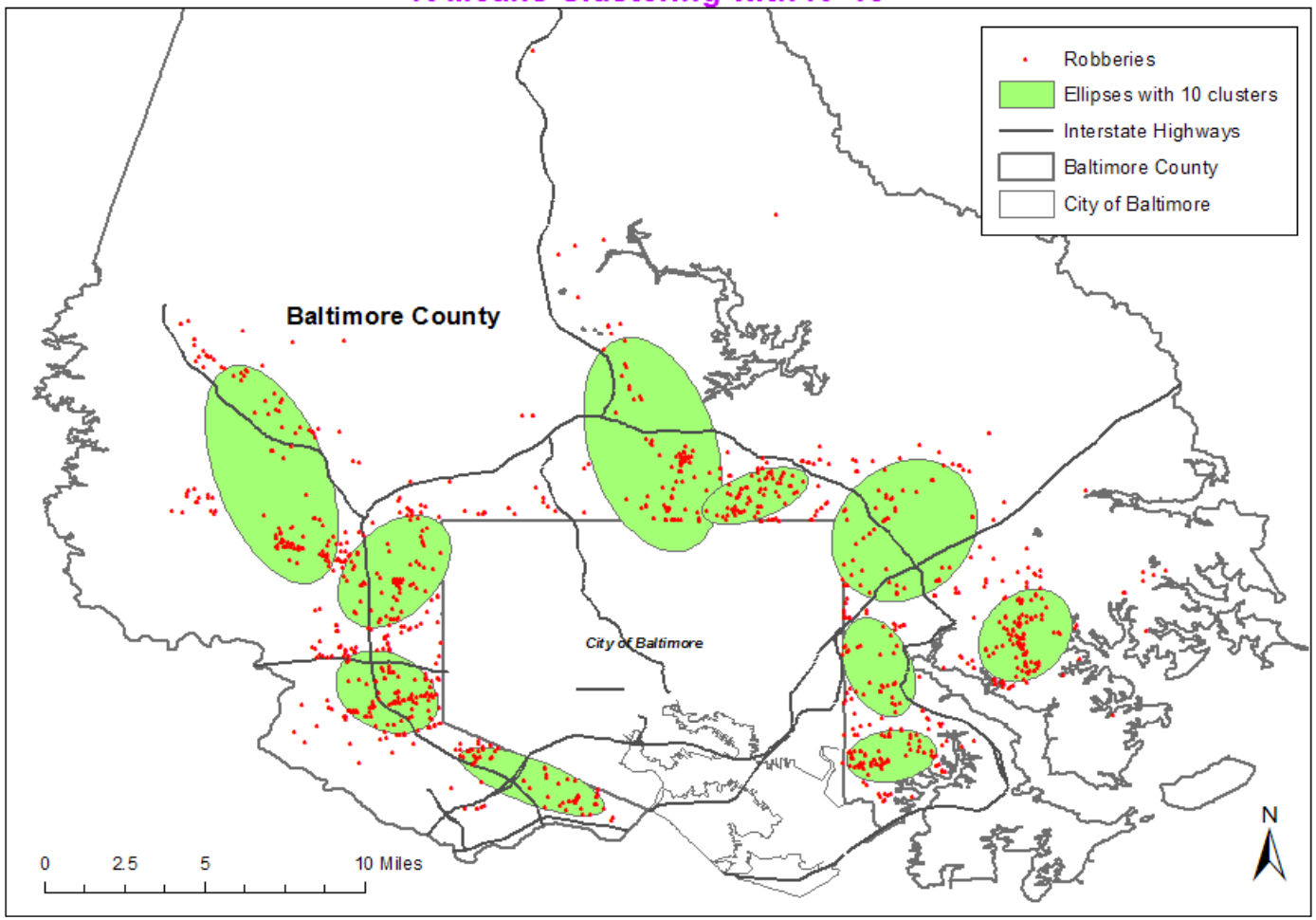


Figure 8.10:

Baltimore County Robbery Hot Spots: 1996

K-Means Clustering with K=34

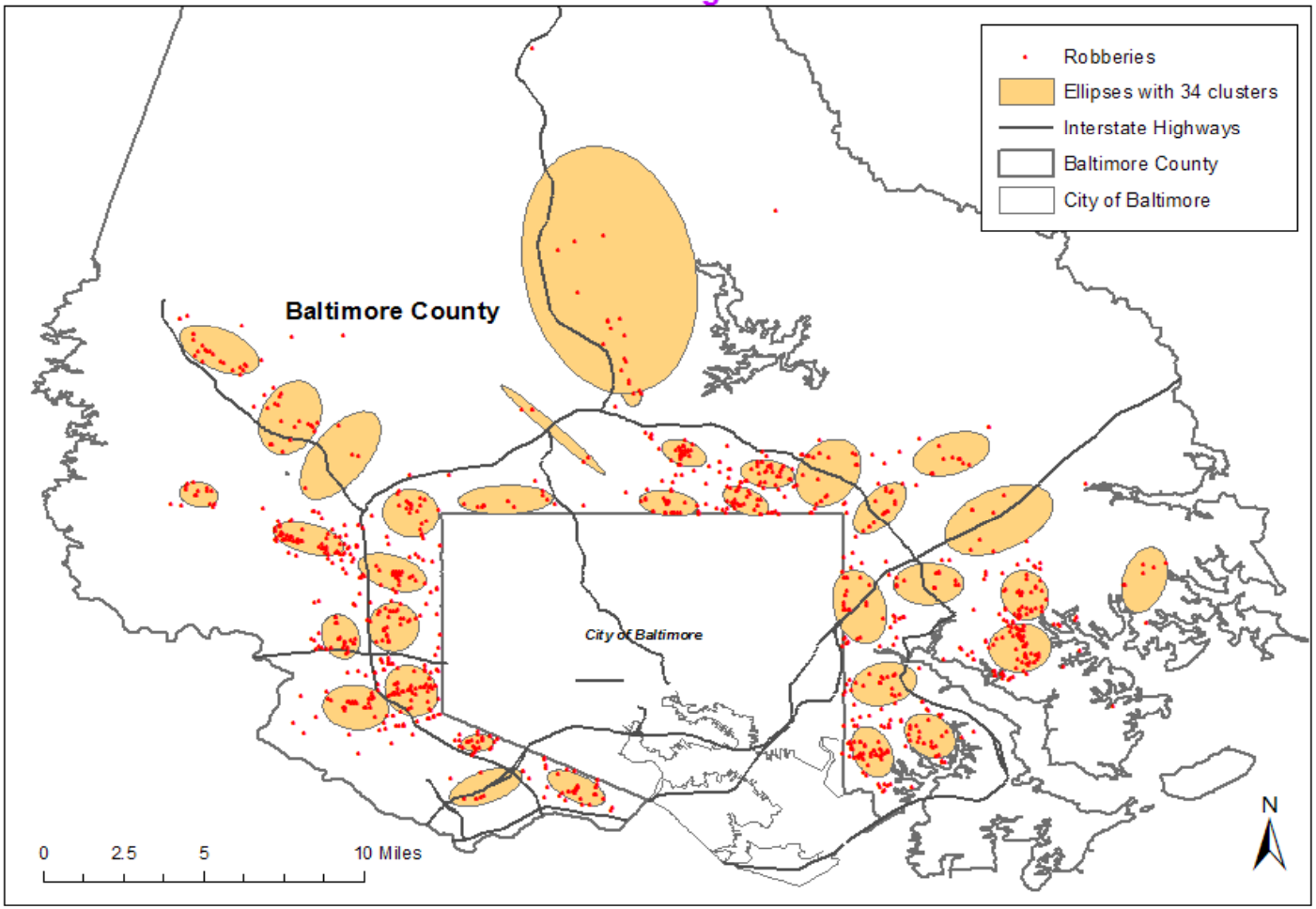
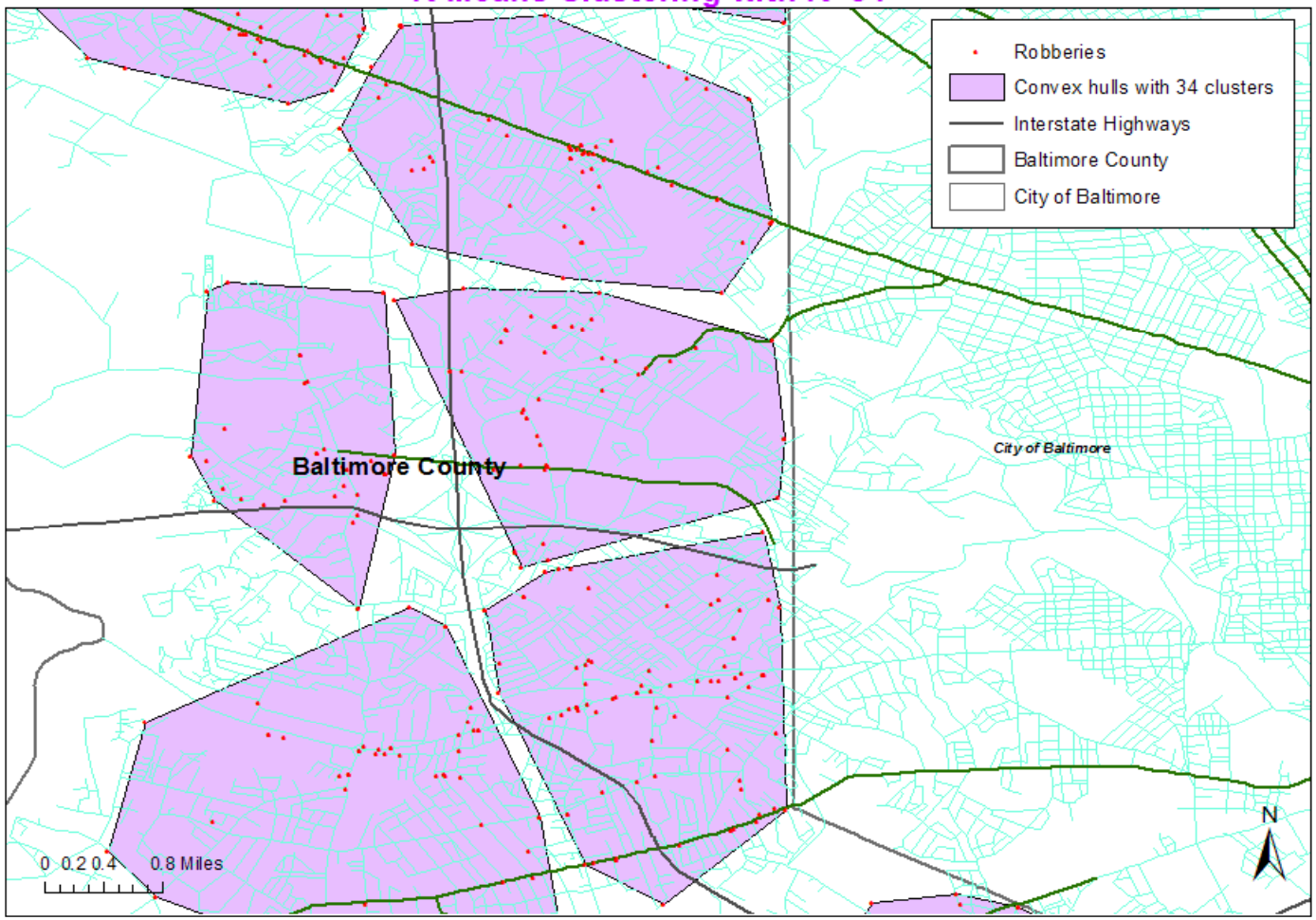


Figure 8.11:

Southwest Baltimore County Robbery Hot Spots: 1996

K-Means Clustering with K=34



not really exist whereas choosing too few will lead to poor differentiation among neighborhoods that are distinctly different.

This choice is both a strength and weakness of the technique. The K-Means procedure provides a great deal of control for the user and can be used as an exploratory tool to identify possible hot spots. Whereas the nearest neighbor hierarchical method produces a solution based on geographical proximity with most clusters being very small and STAC identifies autonomous areas of high density, the K-Means can allow the user to control the size of the clusters. In terms of policing, the K-Means is better suited for defining larger geographical areas than the nearest neighbor method, perhaps more appropriate for a patrol area than for a particular hot spot. Again, if carefully used, the K-Means gives the user the ability to fine tune a particular model of hot spots, adjusting the size of the clusters (vis-a-via the number of clusters selected) as well as their separation in space in order to fit a particular pattern which is known.

Yet it is this same flexible characteristic that makes the technique potentially difficult to use and prone to misuse. Since the technique will divide the data set into *K* groups, there is no assumption that these *K* groups represent real hot spots or not. A user cannot just arbitrarily put in a number and expect it to produce meaningful results. A more extensive discussion of this issue can be found in Murray and Grubesic (2002). Grubesic and Murray (2001) present some newer approaches in the K-Means methodology.

The technique is, therefore, better seen as both an exploratory tool as well as a tool for refining a hot spot search. If the user has a good idea of where there should be hot spots, based on community experience and the reports of beat officers, then the technique can be used to see if the incidents actually correspond to the perception. It also can help identify hot spots which have not been perceived or identified by officers. Alternatively, it can identify hot spots that do not really exist and which are merely by-products of the statistical procedure. Experience and sensitivity are needed to know whether an identified hot spot is real or not.

Some Thoughts on the Concept of Hot Spots

Advantages of the Concept

The six techniques discussed in this and the last chapter have both advantages and disadvantages. Among the advantages are that they attempt to isolate areas of high concentration of incidents and can, therefore, help law enforcement agencies focus their resources on these areas. One of the powerful uses of a hot spot concept is that it is focused. It can provide new information about locations that police officers or community workers may not recognize

(Rengert, 1995). Given that most police departments are understaffed, a strategy that prioritizes intervention is very appealing. The hot spot concept is imminently practical.

Another advantage to the identification of hot spots is that the techniques systematically implement an algorithm. In this sense, they minimize bias on the part of officers and analysts since the technique operates somewhat independently of preconceptions. As has been mentioned, however, these techniques are not totally without human judgment since the user must make decisions on the number of hot spots and the size of the search radius, choices that can allow different users to come to different conclusions. There is probably no way to get around subjectivity since law enforcement personnel may not use a result unless it partly confirms what they already know. But, by implementing an algorithm, it forces users to at least go through the steps systematically.

A third advantage is that these techniques are visual, particularly when used with a GIS. The mode and fuzzy mode routines output the results as a dbf file, which can be displayed in a GIS as a proportional circle. The Nnh, Rnnh, Stac, and Kmeans routines can output the results directly as graphical objects, either as standard deviational ellipses or as convex hull, which can be displayed directly in a GIS. Visual information can help crime analysts and officers to understand the distribution of crime in an areas, a necessary step in planning a successful intervention. We should never underestimate the importance of visualization in any analysis.

Limitations of the Concept

However, there are also some distinct limitations to the concept of a hot spot, some technical and some theoretical. The choice involved in a user making a decision on how strict or how loose to create clusters allows the potential for subjectivity, as has been mentioned. In this sense, isolating clusters (or hot spots) can be as much an art as it is a science. There are limits to this, however. As the sample size goes up, there is less difference in the result that can be produced by adjusting the parameters. For example, with 6,000 or more cases, there is very little difference between using the 0.1 significance level in the nearest neighbor clustering routine and the 0.001 significance level.[8] Thus, the subjectivity of the user is more important for smaller samples than larger ones.

8 On one test of 6,051 burglaries with a minimum cluster size requirement of 10 incidents, for example, we obtained 100 first-order clusters, 9 second-order clusters, and no third-order clusters by using a 0.1 significance level for the nearest neighbor hierarchical clustering routine. When the significance level was reduced to 0.001, the number of clusters extracted was 97 first-order clusters, 8 second-order clusters, and no third-order clusters.

A second problem with the hot spot concept is that it is usually applied to the volume of incidents and not to the underlying risk. Clusters (or hot spots) are defined by a high concentration of incidents within a small geographical area, that is, on the volume of incidents within an area. This is an implicit *density* measure - the number of incidents per unit of area (e.g., incidents per square mile). But higher density can also be a function of a higher population at risk.

For some policing policies, this is fine. For example, beat officers will necessarily concentrate on high incident density neighborhoods because so much of their activity revolves around those neighborhoods. From a viewpoint of providing concentrated policing, the density or volume of incidents is a good index for assigning police officers (Sherman and Weisburd, 1995). From the viewpoint of ancillary security services, such as access to emergency medical services, neighborhood watch organizations, or residential burglar alarm retail outlets, areas with higher concentrations of incidents may be a good focal point for organizing these services.

But for other law enforcement policies, a density index is not a good one. From the viewpoint of crime prevention, for example, high incident volume areas are not necessarily unsafe and that effective preventive intervention will not necessarily lead to reduction in crime. It may be far more effective to target high risk areas rather than high volume areas. In high risk areas, there are special circumstances which expose the population to higher-than-expected levels of crime, perhaps particular concentrations of activities (e.g., drug trading) or particular land uses that encourage crime (e.g., skid row areas) or particular concentrations of criminal activities (e.g., gangs). A prevention strategy will want to focus on those special factors and try to reduce them.

Risk, which is defined as the number of incidents relative to the number of potential victims/targets, is only loosely correlated with the volume of incidents. Yet, hot spots are usually defined by volume, rather than risk. The risk-adjusted hierarchical nearest neighbor clustering routine, discussed in Chapter 7, is the only tool among these that identifies risk, rather than volume. It is clear that more tools will be needed to examine hot spot locations that are more at risk.

The final problem with the hot spot concept is more theoretical. Namely, given a concentration of incidents, how do we explain it? To identify a concentration is one thing. To know how to intervene is another. It is imperative that the analyst discover some of the underlying causes that link the events together in a systematic way. Otherwise, all that is left is an empirical description without any concept of the underlying causes. For one thing, the concentration could be random or haphazard; it could have happened one time, but never again. For another, it could be due to the concentration of the population *at risk*, as discussed above.

But, it could also be due to the concentration of activities that attract offenders along with victims. In Chapter 14 and, again, in Chapter 28, we examine locations where offenders are attracted. Many of these are shopping malls, which is where a lot of crime occurs. Thus, the hot spot could be a destination as much an origin variable. Finally, the concentration could be circumstantial and not be related to anything inherent about the location.

The point here is that an empirical description of a location where crime incidents are concentrated is only a first step in defining a real 'hot spot'. It is an *apparent* 'hot spot'. Unless the underlying vector (cause) is discovered, it will be difficult to provide adequate intervention. The causes could be environmental (e.g., concentrations of land uses that attract attackers and victims) or behavioral (e.g., concentrations of gangs). The most one can do is try to increase the concentration of police officers. This is expensive, of course, and can only be done for limited periods. Eventually, if the underlying vector is not dealt with, incidents will continue and will overwhelm the additional police enforcement. In other words, ultimately, reducing crime around a 'hot spot' will need to involve many other policies than simply police enforcement, such as community involvement, gang intervention, land use modification, job creation, the expansion of services, and other community-based interventions. In this sense, the identification of an empirical 'hot spot' is frequently only a window into a much deeper problem that will involve more than targeted enforcement.

References

Aldenderfer, M. & Blashfield, R. (1984). *Cluster Analysis*. Sage: Beverly Hills, CA.

Ball, G. H. & Hall, D. J. (1970). A clustering technique for summarizing multivariate data. *Behavioral Science*, 12, 153-155.

Barnard, G. A. (1963). Comment on 'The Spectral Analysis of Point Processes' by M. S. Bartlett, *Journal of the Royal Statistical Society*, Series B, 25, 294.

Block, C. R. (1994). STAC hot spot areas: a statistical tool for law enforcement decisions. In *Proceedings of the Workshop on Crime Analysis Through Computer Mapping*. Criminal Justice Information Authority: Chicago, IL.

Block, R. & Block, C. R. (1999) Risky places: a comparison of the environs of rapid transit stations in Chicago and the Bronx in John Mollenkopf (ed), *Analyzing Crime Patterns: Frontiers of Practice*, Sage Publishing: Beverly Hills, CA.

Block, R. & Block, C. R. (1995). Space, place and crime: hot spot areas and hot places of liquor-related Crime,. In Eck, J. E. & Weisburd, D. (eds.), *Crime and Place*. Crime Prevention Studies, Volume 4. Criminal Justice Press: Monsey, NY, 147-185.

Chainey, S., Thompson, L. & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, **21**, 4-28.

Dwass, M (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28, 181-187.

Eick, C. F., Zeidat, N. & Zhao, Z. (2004). Supervised clustering: Algorithms and applications. proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI04) , Boca Raton, FL>

Everitt, B. S. (2011). *Cluster Analysis* (5th edition). J. Wiley: London.

Everitt, B. S., Landau, S. & Leese, M. (2001). *Cluster Analysis*. 4th Edition. Oxford University Press: New York.

References (continued)

Finley, T. & Joachims, T. (2005). Supervised clustering with support vector machines. *Proceedings of the 22nd International Conference on Machine Learning*. Bonn, Germany.

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

Friedman, H. P. & Rubin, J. (1967). On some invariant criteria for grouping data, *Journal of the American Statistical Association*, **62**, 1159-1178.

Gersho, A. & Gray, R. (1992). *Vector Quantization and Signal Compression*. Kluwer Academic Publishers: Dordrecht, Netherlands.

Grubesic, T. H. & Murray, A. T. (2001). Detecting hot spots using cluster analysis and GIS. Paper presented at Annual Conference of the Crime Mapping Research Center, Dallas, TX. http://www.ojp.usdoj.gov/cmrc.

Haggett, P., Cliff, A. D. & Frey, Allan (1977). *Locational Analysis in Human Geography* (2nd edition). Edward Arnold: London.

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

Levine, N. (2008). "The 'hottest' part of a crime hotspot: Comments on "The utility of hotspot mapping for predicting spatial patterns of crime" by Spencer Chainey, Lisa Tompson, and Sebastian Uhlig". *Security Journal*, 21, 295-302.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *5th Berkeley Symposium on Mathematics, Statistics and Probability*. Vol 1, 281-298.

Murray, A.T. & Grubesic, T. H. 2002. Identifying Non-hierarchical Clusters. *International Journal of Industrial Engineering*, 9, 86-95.

Openshaw, S. A., Craft, A. W., Charlton, M., & Birch, J. M. (1988). Invetigation of leukemia clusters by use of a geographical analysis machine, *Lancet*, 1, 272-273.

References (continued)

Openshaw, S. A., Charlton, M., Wymer, C. & Craft, A. W. (1987). A mark 1 analysis machine for the automated analysis of point data sets, *International Journal of Geographical Information Systems*, 1, 335-358.

Rengert, G. F. (1995). Comparing cognitive hot spots to crime hot spots. In Carolyn Rebecca Block, C. R., Dabdoub M. & Fregly, S. *Crime Analysis Through Computer Mapping*. Police Executive Research Forum: Washington, DC, 33-47.

Sherman, L. W. & Weisburd, D. (1995). General deterrent effects of police patrol in crime hot spots: a randomized controlled trial. *Justice Quarterly*. 12, 625-648.

Systat, Inc. (2008). *Systat 13: Statistics I*. SPSS, Inc.: Chicago.

Turnbull, B. W., Iwano, E.J., Burnett, W. S., Howe, H. L. & Clark, L. C. (1990). Monitoring for clusters of disease: application to leukemia incidence in upstate New York, *American Journal of Epidemiology*, 132, S136-S143.

Endnotes

i. The steps are as follows:

Global Selection of Initial Seed Locations

A. A 100 x 100 grid is overlaid on the point distribution; the dimensions of the grid are defined by the minimum and maximum X and Y coordinates.

B. A separation distance is defined, which is:

$$Separation = t * 0.5\sqrt{\frac{A}{N}}$$

where *t* is the Student's t-value for the .01 significance level (2.358), *A* is the area of the region, and *N* is the sample size. The separation distance was calculated to prevent adjacent cells from being selected as seeds.

C. For each grid cell, the number of incidents found are counted and then sorted in descending order.

D. The cell with the highest number of incidents found is the initial seed for cluster 1.

E. The cell with the next highest number of incidents is temporarily selected. If the distance between that cell and the seed 1 location is *equal to or greater than* the separation distance, this cell becomes initial seed 2.

F. If the distance is less than the separation distance, the cell is dropped and the routine proceeds to the cell with the next highest number of incidents.

G. This procedure is repeated until *K initial seeds* have been located thereby selecting the remaining cell with the highest number of incidents and calculating its distance to all prior seeds. If the distance is equal to or greater than the separation distance, then the cell is selected as a seed. If the distance is less than the separation distance, then the cell is dropped as a seed candidate. Thus, it is possible that *K* initial seeds cannot be identified because of the inability to locate *K* locations greater than the threshold distance. In this case, *CrimeStat* keeps the number it has located and prints out a message to this effect.

Local Optimization of Seed Locations

H. After the *K* initial seeds have been selected, all points are assigned to the nearest initial seed location. These are the initial cluster groupings.

I. For each initial cluster grouping in turn, the center of minimum distance is calculated. These are the second seed locations.

J. All points are assigned to the nearest second seed location.

K. For each new cluster grouping in turn, the center of minimum distance is calculated. These are third seed locations.

L. Steps J and K are repeated until no more points change cluster groupings. These are the final seed locations and cluster groupings.

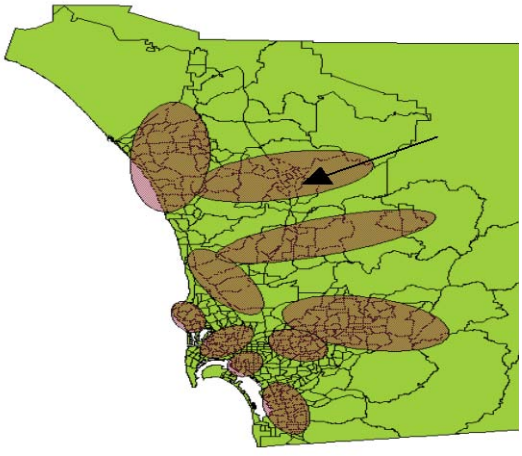
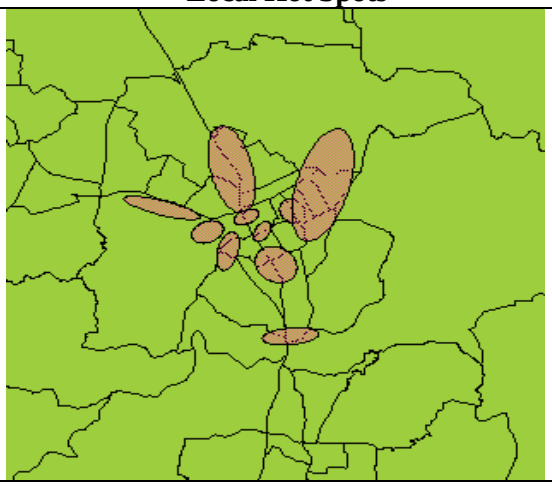
Attachments

K-Means Clustering as an Alternative Measure of Urban Accessibility

Richard J. Crepeau
Department of Geography and Planning
Appalachian State University
Boone, NC

The relationship between land use and the transportation system is an important issue. Many planners recognize that transportation policies, practices and outcomes affect changes in land use, and vice versa, but there is disagreement as to how best to describe this phenomenon. Traditional methods include measures of accessibility via a matrix of zones (tracts, traffic analysis zones, etc.). However, there are limits to the way interaction and accessibility is described with such discrete units.

Through the use of K-Means clustering, an alternate measure of accessibility can be calculated. Rather than relying on census geography, the left map shows ten retail clusters in San Diego County (1995) as calculated by *CrimeStat*'s K-Means clustering technique (using 1x standard deviational ellipse). The retail hot spots were calculated using a geocoded point file of retail establishments in the county. These clusters are not bound by census geography and allow a more realistic appraisal about the attractiveness of specific regions within the county. An analyst can then determine if residential location within a hot spot has an effect on travel patterns, or if there is a relationship between proximity to a hot spot and travel behavior. While this example illustrates a measure of regional retail attractiveness, the flexibility of *CrimeStat* allows an analyst to evaluate these relationships on a local level, thus allowing a scope of inquiry from regional to local accessibility (as shown in right map, which uses the same parameters as the left figure, but limiting its sample to retail in a sub-region of San Diego County noted by the arrow).

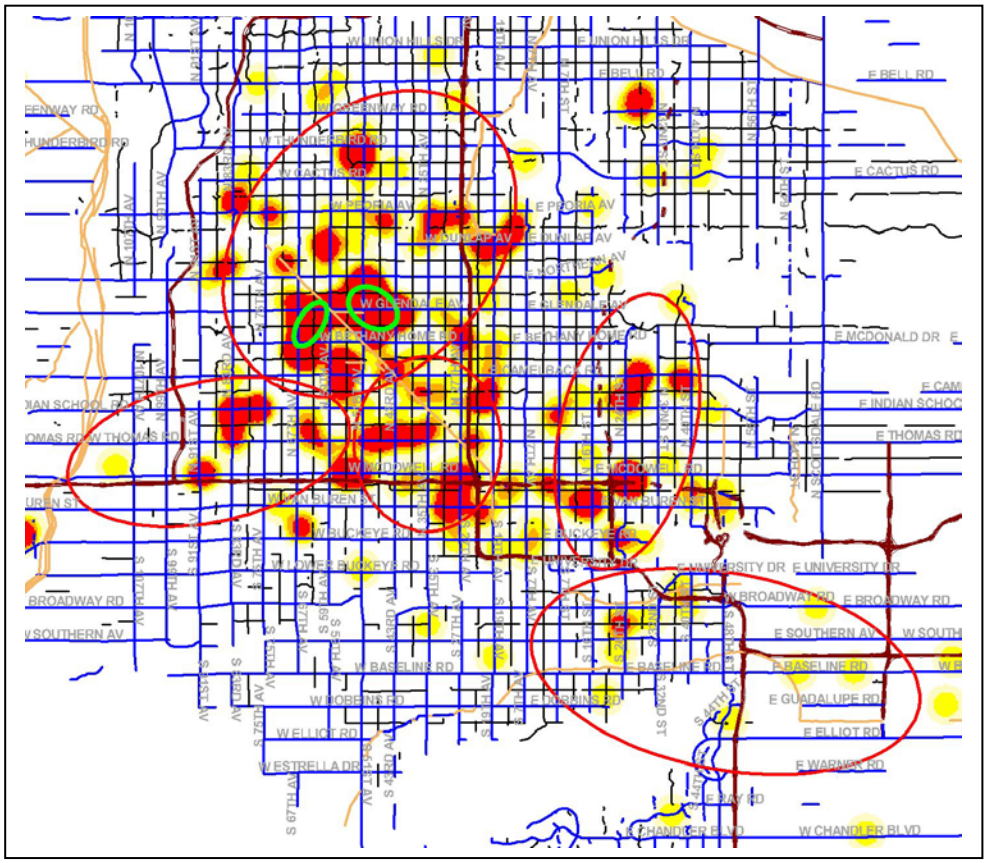
Regional Hot Spots	Local Hot Spots
	

Hot Spot Verification in Auto Theft Recoveries

Bryan Hill
Glendale Police Department
Glendale, AZ

We use *CrimeStat* as a verification tool to help isolate clusters of activity when one application or method does not appear to completely identify a problem. The following example utilizes several *CrimeStat* statistical functions to verify a recovery pattern for auto thefts in the City of Glendale (AZ). The recovery data included recovery locations for the past 6 months in the City of Glendale which were geocoded with a county-wide street centerline file using *ArcView*.

First, a spatial density "grid" was created using *Spatial Analyst* with a grid cell size of 300 feet and a search radius of 0.75 miles for the 307 recovery locations. We then created a graduated color legend, using standard deviation as the classification type and the value for the legend being the *CrimeStat* "Z" field that is calculated.



In the map, the K-means (red ellipses), Nnh (green ellipses) and *Spatial Analyst* grid (red-yellow grid cells) all showed that the area was a high density or clustering of stolen vehicle recoveries. Although this information was not new, it did help verify our conclusion and aided in organizing a response.