CrimeStat IV

Part III: Hot Spot Analysis

Chapter 7:

Hot Spot Analysis of Points: I

Ned Levine
Ned Levine & Associates
Houston, TX

Table of Contents

Table of Contents (continued)

Table of Contents (continued)

Chapter 7:
Hot Spot Analysis of Points: I

In this and the next two chapters, we describe ten tools for identifying clusters of crime incidents. Six of the tools apply to points while four apply to zones. The discussion has been divided into three chapters primarily because of the length of the discussion. This chapter discusses the concept of a *hot spot* and four hot spot techniques: the mode, fuzzy mode, nearest neighbor hierarchical clustering, and risk-adjusted nearest neighbor hierarchical clustering. Chapter 8 discusses STAC and the K-means algorithm. Chapter 9 discusses Anselin's Local Moran, the Getis-Ord Local "G", the zonal nearest neighbor hierarchical clustering algorithm, and the risk-adjusted zonal nearest neighbor hierarchical methods. However, the ten techniques should be seen as a continuum of approaches towards identifying hot spots.

Hot Spots

Typically called *hot spots* or *hot spot areas*, these are concentrations of incidents within a limited geographical area that appear over time (Braga & Weisburd, 2010). Police have learned from experience that there are particular environments that attract crimes in larger-than-expected concentrations, so-called *crime generators*. Sometimes these hot spot areas are defined by particular activities (e.g., drug trading; Weisburd & Green, 1995; Weisburd, Maher, & Sherman, 1992; Sherman, Gartin & Buerger, 1989; Maltz, Gordon, & Friedman, 1989), other times by specific concentrations of land uses (e.g., skid row areas, bars, adult bookshops, itinerant hotels), and sometimes by interactions between activities and land uses, such as thefts at transit stations or bus stops (Block & Block, 1995; Levine, Wachs & Shirazi, 1986). Whatever the reasons for the concentration, they are real and are known by most police departments.

While there are some theoretical concerns about what links disparate crime incidents together into a cluster, nonetheless, the concept is very useful (Chainey, Thompson, & Uhlig, 2008; Levine, 2008). Police officers patrolling a precinct can focus their attention on particular environments because they know that crime incidents will continually reappear in these places. Crime prevention units can target their efforts knowing that they will achieve a positive effect in reducing crime with limited resources (Sherman & Weisburd, 1995). In short, the concept is very useful.

Nevertheless, the concept is a perceptual construct. Hot spots do not exist in reality, but are areas where there is sufficient clustering of certain activities (in this case, crime) such that they get labeled such. There is not a border around these incidents, but a gradient where people draw an imaginary line to indicate the location at which the hot spot *starts*. In reality, any variable that is measured, such as the density of crime incidents, will be continuous over an area, being higher in some parts and lower in others. Where a line is drawn in order to define a hot spot is somewhat arbitrary.

Statistical Approaches to the Measurement of Hot Spots

Unfortunately, measuring a hot spot is also a complicated problem. There are literally dozens of different statistical techniques designed to identify hot spots (Everitt, Landau and Leese, 2001). Many, but not all, of the techniques are typically known under the general statistical label of *cluster analysis*. These are statistical techniques aimed at grouping cases together into relatively coherent clusters. All of the techniques depend on optimizing various statistical criteria, but the techniques differ among themselves in their methodology as well as in the criteria used for identification. Because hot spots are perceptual constructs, any technique that is used must approximate how someone would perceive an area. The techniques do this through various mathematical criteria.

Types of Cluster Analysis (Hot Spot) Methods

Several typologies of cluster analysis have been developed as cluster routines typically fall into several general categories (Everitt, 2011; Can and Megbolugbe, 1996):

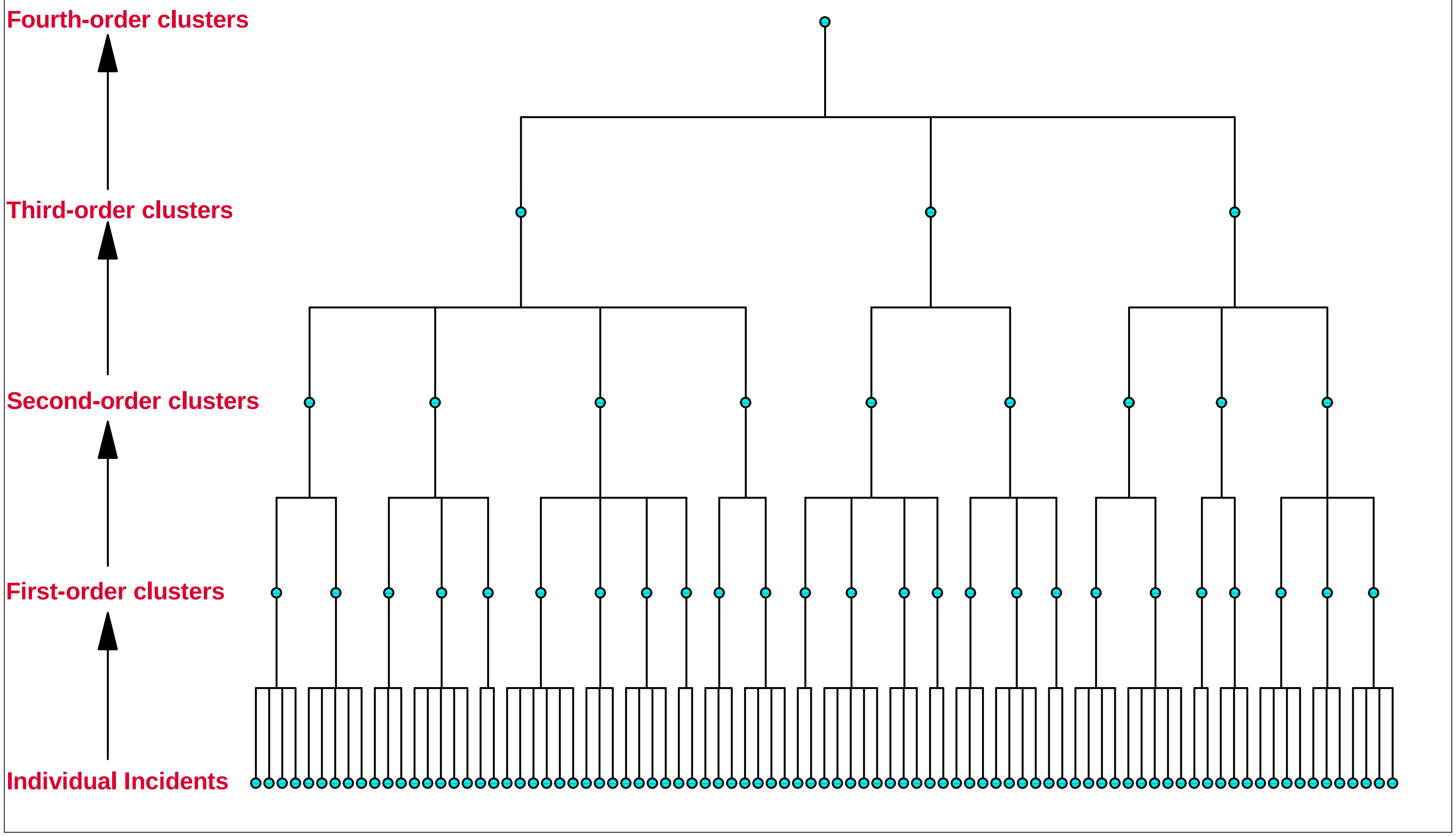
1. *Point locations*. This is the most intuitive type of cluster involving the number of incidents occurring at different locations. Locations with the most number of incidents are defined as hot spots. *CrimeStat* includes two point location techniques: the Mode and Fuzzy Mode;

2. *Hierarchica*l techniques (Sneath, 1957; McQuitty, 1960; Sokal & Sneath, 1963; King, 1967; Sokal & Michener, 1958; Ward, 1963; Hartigan, 1975) are like an inverted tree diagram in which two or more incidents are first grouped on the basis of some criteria (e.g., nearest neighbor). Then, the pairs are grouped into second-order clusters. The second-order clusters are then grouped into third-order clusters, and this process is repeated until either all incidents fall into a single cluster or else the grouping criteria fail. Thus, there is a hierarchy of clusters that can be displayed with a dendogram (an inverted tree diagram).

   Figure 7.1 shows an example of a hierarchical clustering where there are four orders (levels) of clustering; the visualization is non-spatial in order to show the linkages. In this example, all individual incidents are grouped into first-order

Figure 7.1:

Hierarchical Clustering Technique



clusters that, in turn, are grouped into second-order clusters that, in turn, are grouped into third-order clusters and which all converge into a single fourth-order cluster. Many hierarchical techniques, however, do not group all incidents or all clusters into the next highest level. *CrimeStat* includes four hierarchical techniques: Nearest Neighbor Hierarchical Clustering (Nnh) routine and Risk-adjusted Nearest Neighbor Hierarchical Clustering (Rnnh) routines in this chapter and Zonal Nearest Neighbor Hiearchical Clustering (Znnh) and Risk-adjusted Nearest Neighbor Zonal Hierarchical Clustering (RZnnh) routines in Chapter 9.

3. *Partitioning* techniques, frequently called the K-means technique, partition the incidents into a specified number of groupings, usually defined by the user (Thorndike, 1953; MacQueen, 1967; Ball and Hall, 1970; Beale, 1969). Thus, all points are assigned to one, and only one, group. Figure 7.2 shows a partitioning technique where all points are assigned to clusters and are displayed as ellipses. *CrimeStat* includes one partitioning technique, a K-means partitioning technique;

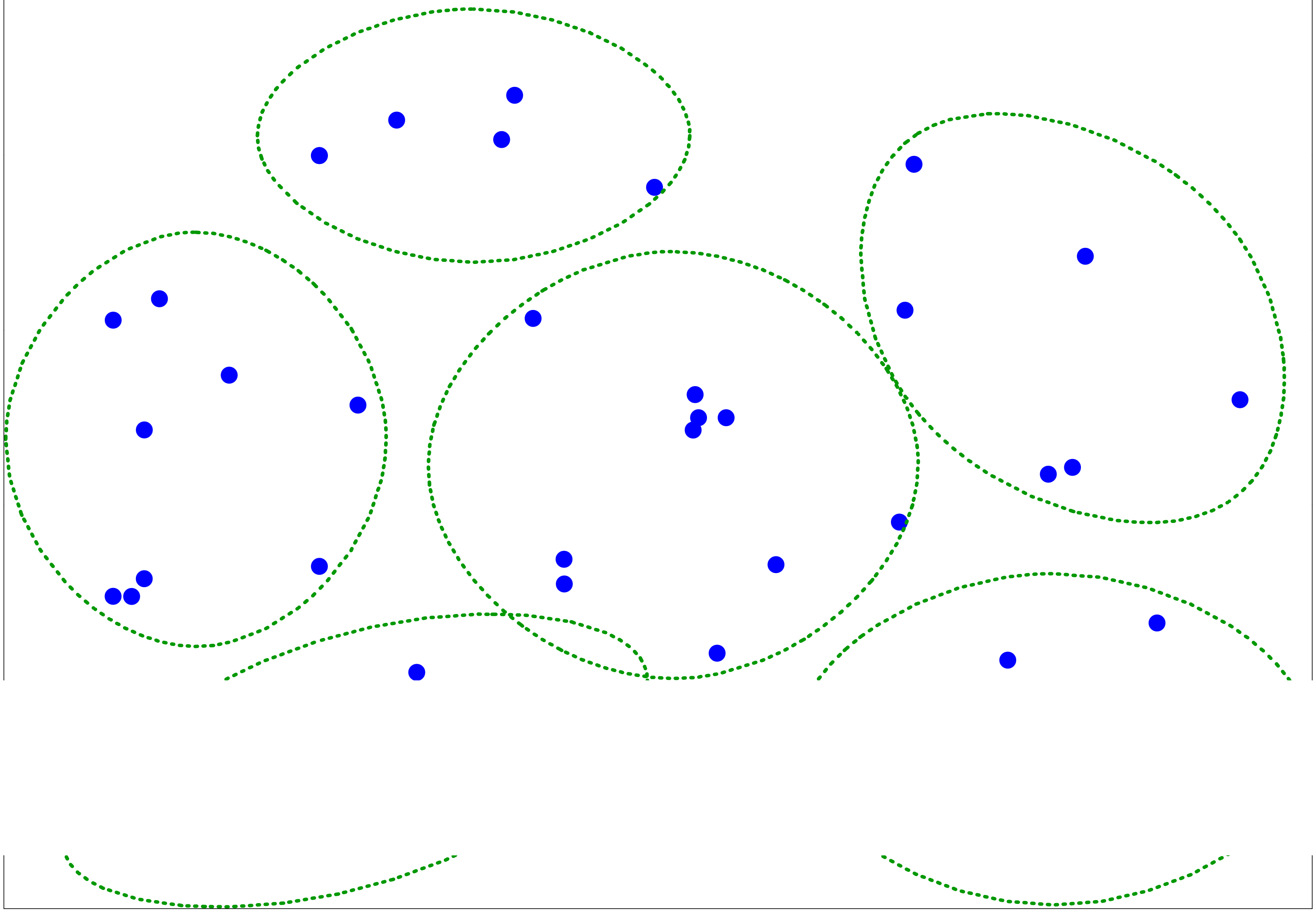
4. *Scan statistics* that apply a search circle uniformly throughout the study area, either to each point or to each node of a reference grid (Block & Block, 1999; Kulldorff, 1997; Block & Block, 1995; Block, 1994; Openshaw, Craft, Charlton, & Birch, 1988; Openshaw, Charlton, Wymer, & Craft, 1987.

5. *Density* techniques identify clusters by searching for dense concentrations of incidents (Bailey & Gattrell, 1995; Silverman, 1986; Gitman & Levine, 1970; Weishart, 1969; Carmichael, George, & Julius, 1968; Cattell & Coulter, 1966). CrimeStat has two density search routines: a Single-kernel Density (K) method and a Dual-kernel Density Interpolation (Dk); this is discussed in chapter 10;

6. *Clumping* techniques involve the partitioning of incidents into groups or clusters, but allow overlapping membership (Jones & Jackson, 1967; Needham, 1967; Jardine & Sibson, 1968; Cole & Wishart, 1970);

7. *Risk-based* techniques identify clusters in relation to an underlying base ‘at risk’ variable, such as population, employment, or active targets (Jefferis, 1998; Kulldorff and Nagarwalla, 1995). *CrimeStat* includes three risk-based techniques - a Risk-adjusted Nearest Neighbor Hierarchical Clustering routine, discussed in this chapter; a Zonal Risk-adjusted Nearest Neighbor Hierarchical Clustering routines discussed in Chapter 9; and a Dual Kernel Density method, discussed in Chapter 10).

Figure 7.2: PartitioningCl usteringT echnique



8. *Zonal* clustering techniques identify contiguous zones that have either high levels or similar levels of an attribute variable or (Getis & Ord, 1996; Anselin, 1995). *CrimeStat* includes four zonal clustering methods: Anselin's Local Moran; the Getis-Ord Local "G"; Zonal Nearest Neighbor Hierarchical Clustering; and Zonal Risk-adjusted Nearest Neighbor Hierarchical Clustering.

9. *Miscellaneous* techniques are other methods that are less commonly used (Everitt, 2011).

Many of these methods are hybrids of these classes. For example, the *Risk-adjusted Nearest Neighbor Hierarchical Clustering* routine is primarily a risk-based technique but involves elements of clumping while *STAC* is primarily a partitioning method but with elements of hierarchical grouping.

Optimization Criteria

In addition to the different types of cluster analysis, there are different criteria that distinguish techniques applied to space (Everitt, 2011). Among these are:

1. The *definition* of a cluster - whether it is a discrete grouping or a continuous variable; whether points must belong to a cluster or whether they can be isolated; whether points can belong to multiple clusters.

2. The *choice of variables* in addition to the X and Y coordinates - whether weighting or intensity values are used to define similarities.

3. The measurement of *similarity and distance* - the type of geometry being used; whether clusters are defined by closeness or not; the types of similarity measures used.

4. The *number* of clusters - whether there are a fixed or variable number of clusters; whether users can define the number or not.

5. The geographical *scale* of the clusters - whether clusters are defined by small or larger areas; for hierarchical techniques, what level of abstraction is considered optimal.

6. The *initial selection* of cluster locations ('seeds') - whether they are mathematically or user defined; the specific rules used to define the initial seeds.

7. The *optimization routines* used to adjust the initial seeds into final locations - whether distance is being minimized or maximized; the specific algorithms used to readjust seed locations.

8. The *visual display* of the clusters, once extracted - whether drawn by hand or by a geometrical object (e.g., an ellipse, a convex hull); the proportion of cases represented in the visualization.

This is not the place to provide a comprehensive review of cluster techniques (see Everitt, 2011 for such a review). Nevertheless, it should be clear that with the several types of cluster analysis and with the many criteria that can be used for any particular technique provides a large number of different techniques that could be applied to an incident data base. It should be realized that there is not a single solution to the identification of hot spots, but that different techniques will reveal different groupings and patterns among the groups. A user must be aware of this variability and must choose techniques that can complement other types of analysis. It would be very naive to expect that a single technique can reveal the existence of hot spots in a jurisdiction that are unequivocally clear. In most cases, analysts are not sure why there are hot spots in the first place. Until that is solved, it would be unreasonable to expect a mathematical or statistical routine to solve that problem.

Cluster Routines in *CrimeStat*

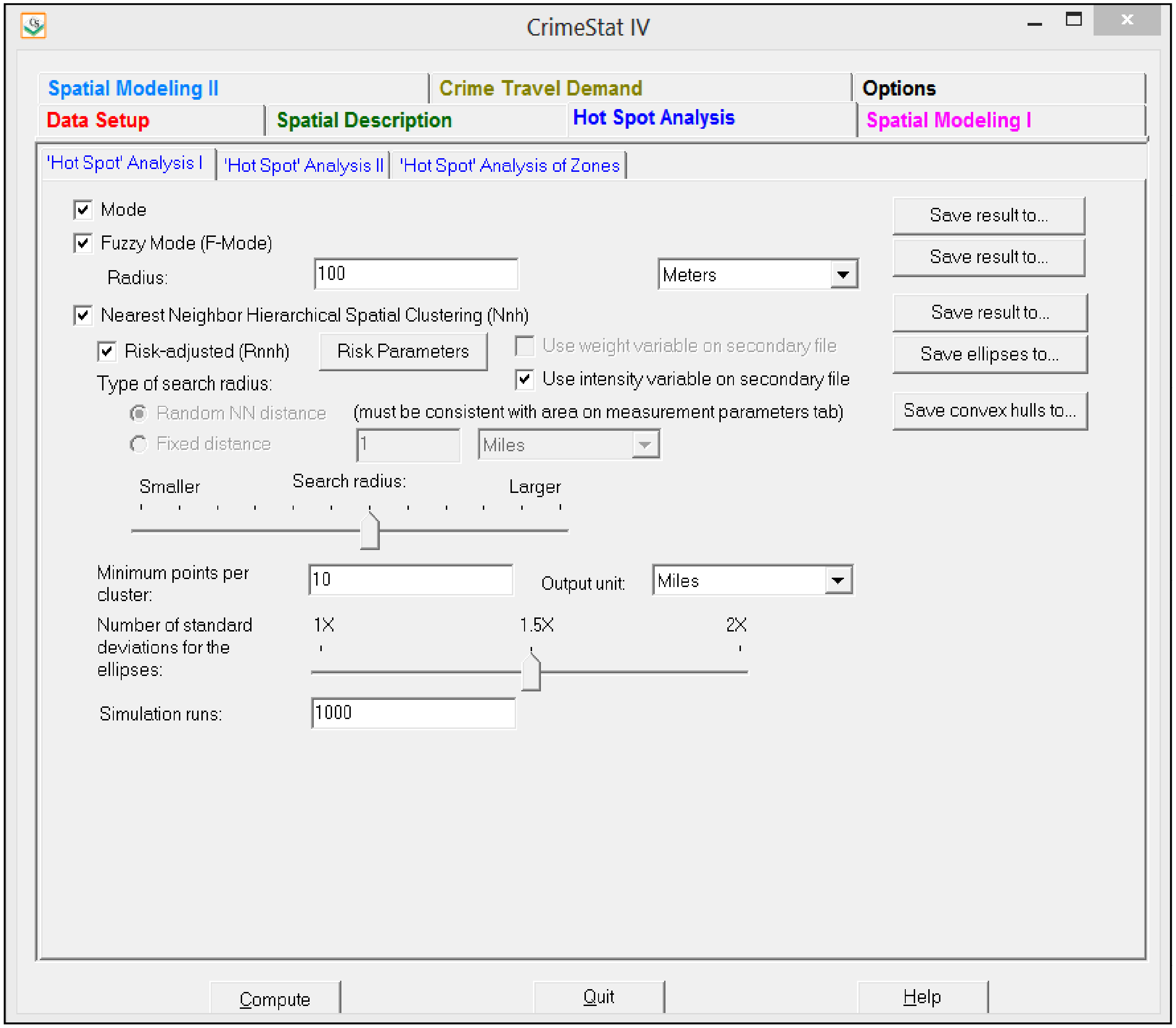
Figure 7.3 shows the Hot Spot Analysis I page. Because of the variety of cluster techniques, *CrimeStat* includes ten techniques that cover the range of techniques that have been used:

1. The Mode
2. The Fuzzy Mode
3. Nearest neighbor hierarchical clustering
4. Risk-adjusted nearest neighbor hierarchical clustering
5. The Spatial and Temporal Analysis of Crime (*STAC*) module
6. K-means clustering
7. Anselin's Local Moran
8. Getis-Ord Local "G"
9. Zonal nearest neighbor hierarchical clustering
10. Zonal risk-adjusted nearest neighbor hierarchical clustering

These are not the only techniques, of course, and analysts should use them as complements to other types of analysis. Because of the number of routines, these routines have

Figure 7.3:

Hot Spot Analysis I Screen



☑ Mode

☑ Fuzzy Mode (F-Mode)

Radius: 100 Meters

☑ Nearest Neighbor Hierarchical Spatial Clustering (Nnh)

☑ Risk-adjusted (Rnnh)

□ Use weight variable on secondary file

☑ Use intensity variable on secondary file

Type of search radius:

● Random NN distance (must be consistent with area on measurement parameters tab)

○ Fixed distance 1 Miles

Search radius: Smaller Larger

Minimum points per cluster: 10

Output unit: Miles

Number of standard deviations for the ellipses: 1X 1.5X 2X

Simulation runs: 1000

been allocated to two different setup tabs in *CrimeStat* called Hot Spot' Analysis I and Hot Spot Analysis II. However, they should be seen as one collection of similar techniques. This chapter will discuss the first four of these and the next two chapters the remaining ones.

Mode

The *mode* is the most intuitive type of hot spot. It is the location with the largest number of incidents. The *CrimeStat* Mode routine calculates the frequency of incidents occurring at each unique location (a point with a unique X and Y coordinate), sorts the list, and outputs the results in rank order from the most frequent to the least frequent.

Only locations that are represented in the primary file are identified. The routine outputs a 'dbf' file that includes four variables:

1. The rank order of the location with 1 being the location with the most incidents, 2 being the location with the next most incidents, 3 being the location with the third most incidents, and so forth until those locations that have only one incident each;

2. The frequency of incidents at the location. This is the number of incidents occurring at that location;

3. The X coordinate of the location; and

4. The Y coordinate of the location.

To illustrate, Table 7.1 presents the formatted output for the ten most frequent locations for 14,853 motor vehicle thefts that occurred within the City of Baltimore or Baltimore County in 1996.[1] Figure 7.4 maps the ten locations with the most vehicle thefts (two were tied for rank three and two were tied for rank nine). The map displays the locations with a round symbol, the size of which is proportional the number of incidents. Also, the number of incidents at the location is displayed. These vary from a high of 43 vehicle thefts at location number 1 to a low of 15 vehicle thefts at location numbers 9 and 10. In order to know what these locations represent, the user will have to overlay other GIS layers over the points. In the example, of the ten locations, eight are at shopping centers, one is the parking lot of a train station, and one is the parking lot of a large organization.

[1] The output in Table 7.1 has been formatted. *CrimeStat* only outputs an Ascii file.

Figure 7.4:

Metropolitan Baltimore Vehicle Thefts: 1996

10 Most Frequent Vehicle Theft Locations

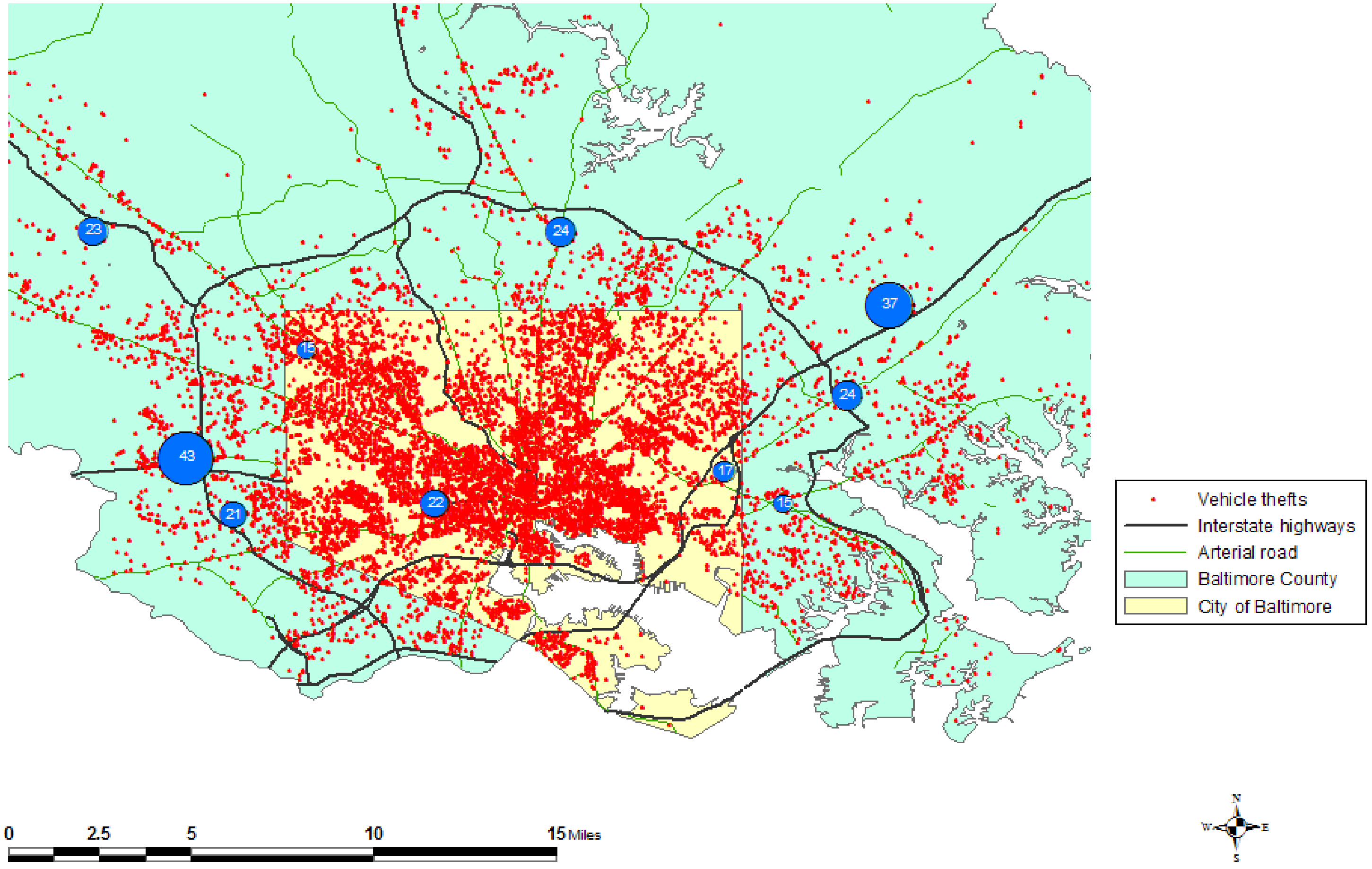


Table 7.1:
Mode Output for
Most Frequent Locations for Motor Vehicle Thefts
City of Baltimore and Baltimore County: 1990
(ONLY 10 SHOWN)

```
Mode:
-------
          N = 14,853

          Rank                Freq       X             Y
          -------             ------  -------------  ------------
          1                   43      -76.7507       39.3115
          2                   37      -76.4710       39.3741
          3                   24      -76.4880       39.3372
          4                   24      -76.6015       39.4042
          5                   23      -76.7877       39.4046
          6                   22      -76.6517       39.2927
          7                   21      -76.7319       39.2880
          8                   17      -76.5363       39.3060
          9                   15      -76.7026       39.3560
          10                  15      -76.5128       39.2927
          Etc.
```

The mode is a very simple measure, but one that can be very useful. In the example, it is clear that most vehicle thefts occur at institutional settings, where there are a collection of parked vehicles. In the case of the shopping centers, the Baltimore County Police Department are aware of the number of vehicles stolen at these locations and work with the shopping center management offices to try to reduce the thefts. It also turns out that shopping centers are the most frequent locations for stolen vehicle retrievals, so it works both ways.

Fuzzy Mode

The usefulness of the mode, however, is dependent on the degree of resolution for the geo-referencing of incidents. In the case of the Baltimore vehicle thefts, thefts locations were assigned a single point at the address. Thus, all thefts occurring at any one shopping center are assigned the same X and Y coordinates. However, there are situations when the assignment of a coordinate will not be a good indicator of the hot spot location. For example, assigning the vehicle theft location to a particular stall in a parking lot will lead to few, if any, locations coming up more than once. In this case, the mode would not be a useful statistic at all. Another example is assigning the vehicle theft location for the parking lot of a multi-building apartment complex to the address of the owner. In this case, what is a highly concentrated set of vehicle thefts become dispersed because the owners live at different addresses within the complex.

Consequently, *CrimeStat* includes a second point location hot spot routine called the *Fuzzy Mode*. This allows the user to define a small search radius around each location to include events that occur *around* or near that location. For example, a user can put a 50 yard or 100 meter search radius and the routine will calculate the number of incidents that occur at each location *and* within a 50 yard or 100 meter radius.

The aim of the statistic is to allow the identification of locations where a number of incidents may occur, but where there may not be precision in measurement.[2] For example, if several apartment complexes share a parking lot, any vehicle theft in the lot may be assigned to the address of the owner, rather than to the parking lot. In this case, the measurement is imprecise. Plotting the location of the vehicle thefts will make it appear that there are multiple locations, when, in fact, there is only approximately one.

Another example would be the measurement of motor vehicle crashes that all occur at a single intersection. If the measurement of the location is very precise, the crashes could be assigned to slightly different locations when, in fact, they occurred at more or less the same location. In other words, the fuzzy mode allows a flexible classification of a location where the analyst can vary slightly the area around a location.

The fuzzy mode output file is also a 'dbf' file and, like the mode, also includes four output variables:

1. The rank order of the location with 1 being the location with the most incidents, 2 being the location with the next most incidents, 3 being the location with the third most incidents, and so forth until only those locations which have only one incident each;

[2] In the statistical literature, this type of statistic is known as a spatial scan with a fixed circular window (Kulldorff, 1997; Kulldorff and Nagarwalla, 1995). However, our emphasis here is on defining approximate point locations where there is either measurement error or very small locational differences. In this sense, the term 'fuzzy' is more similar to the classification literature where imprecise boundaries exist and an incident can belong to two or more groups (Bezdek, 1981; McBratney and deGruijter, 1992; Xie and Beni, 1991).

2. The frequency of incidents at the location. This is the number of incidents occurring at that location;

3. The X coordinate of the location; and

4. The Y coordinate of the location.

Note that allowing a search radius around a location means that incidents are counted multiple times, one for each radius they fall within. If used carefully, the fuzzy mode can allow the identification of high incident locations more precisely than the mode routine. But, because of the multiple counting of incidents that occurs, the frequency of incidents at locations will change, compared to the mode, as well as possibly the hierarchy.

To illustrate this, Figure 7.5 maps the top 13 locations for vehicle thefts identified by the fuzzy mode routine using a search radius of 300 feet (four were tied for number 2 and eight were tied for number 5). The 13 locations are displayed by a single magenta triangle and are compared to the 10 locations identified by the mode (blue circle). Notice that two of the 13 locations are clustered at the same places as those identified by the mode, but the other two triangles are different locations. Two of these locations have multiple fuzzy modes. The most southeastern triangle in Baltimore County actually includes three fuzzy modes while the one triangle within the City of Baltimore actually includes eight fuzzy modes.

Figure 7.6 zooms in to display the eight clustered locations within the City of Baltimore, each of which has a fuzzy mode count of 29 vehicle thefts. The eight fuzzy mode locations are actually eight parking lots within the Mondawin Shopping Mall. Since the parking lots are within 300 feet of each other, each has a cumulative count of 29 vehicle thefts. In other words, the fuzzy mode has identified a general location where there are multiple sub-locations in which vehicle thefts occur.

Uses of the Fuzzy Mode

The fuzzy mode routine can be useful because it allows the identification of small hot spot areas, rather than exact locations. Any one location may not have a sufficient number of incidents that occur at that location, but because it is close to other locations that have incidents occurring, the cumulative count may actually be quite high. Additional examples when it might be useful are in identifying multiple parking lots in parks or in identifying common parking areas for multi-unit buildings (e.g., large apartment complexes).

Figure 7.5:

Metropolitan Baltimore Vehicle Thefts: 1996

13 Most Frequent Vehicle Theft Locations within 300 Feet Search Radius

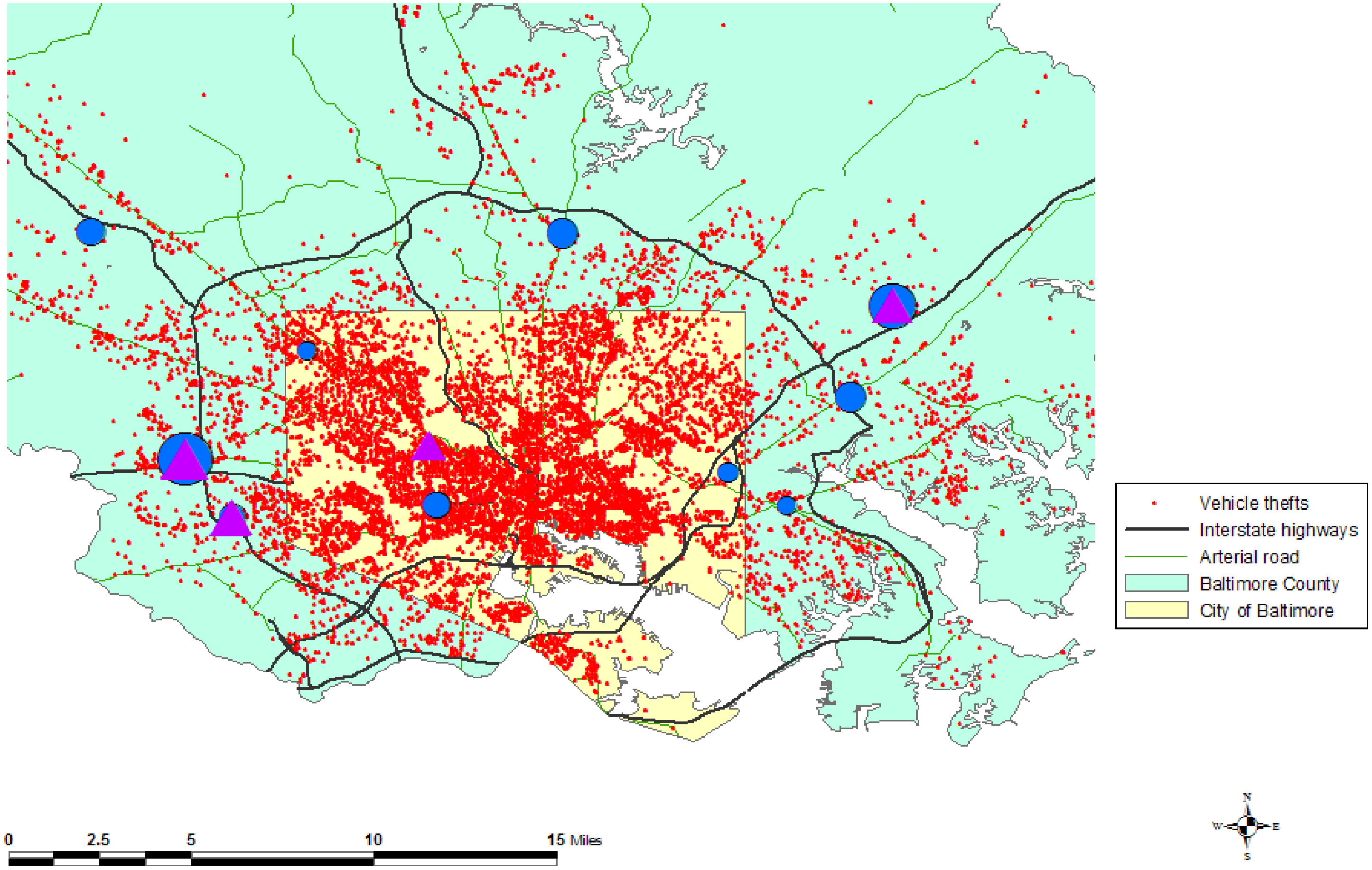
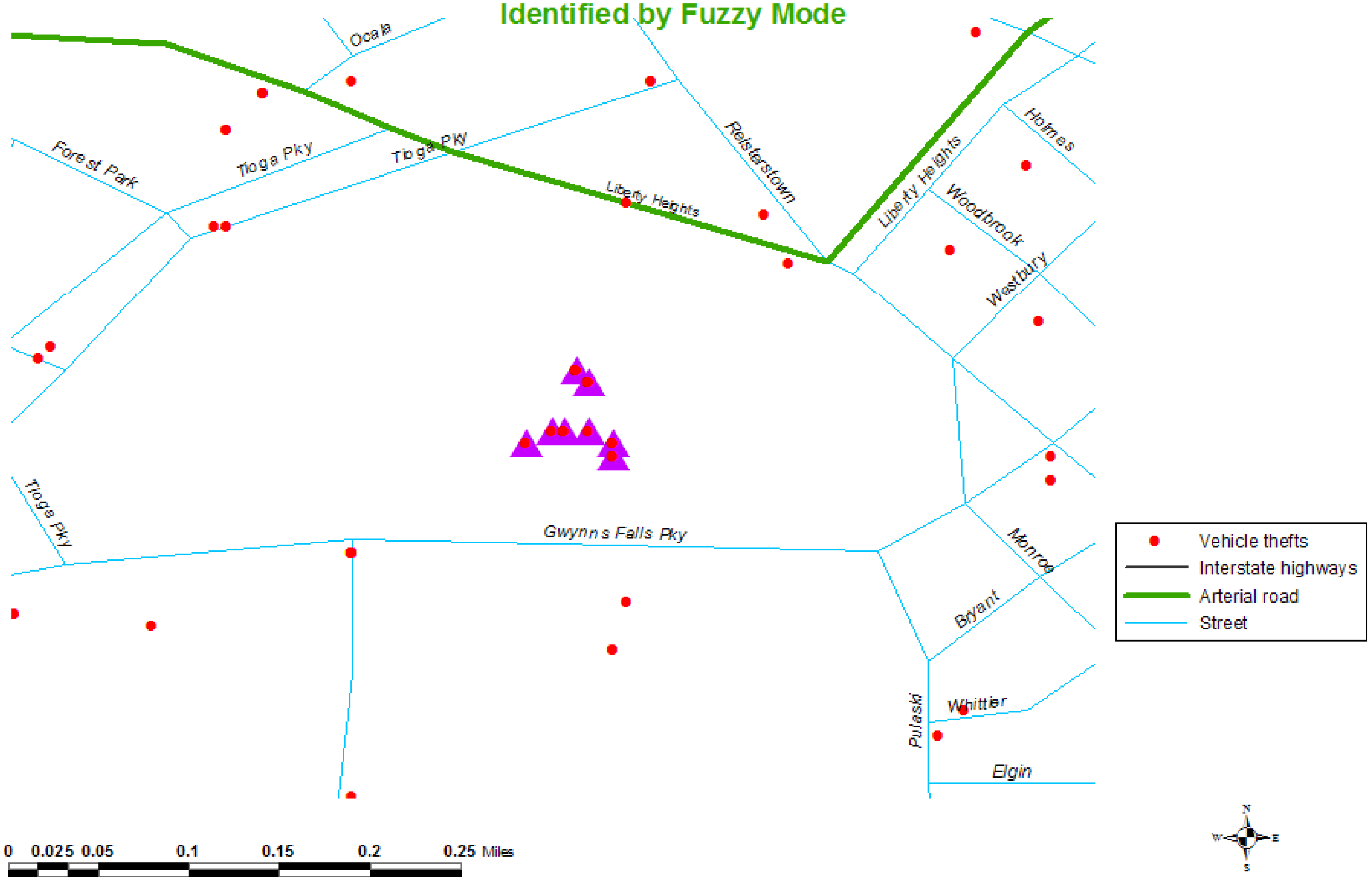


Figure 7.6:

Metropolitan Baltimore Vehicle Thefts: 1996

8 Concentrated Vehicle Theft Clusters within 300 Feet Search Radius

Identified by Fuzzy Mode



The method would also be useful for identifying hot spots when exact coordinates are specified for each incident. For example, in the parking lot example above, if each vehicle theft were identified by a stall number, as opposed to a single coordinate for the entire parking lot, few vehicle thefts would occur in exactly the same location. Allowing a search radius around the coordinates (the fuzzy part of the frequency count) allow a number of events to be grouped together whereas exact locations might not identify that grouping.

Limitations of the Fuzzy Mode

On the other hand, the fuzzy mode does involve duplicate counts points that are close to each other will be counted multiple times. This can allow distortion. By changing the search radius, the number of incidents counted for any one location changes as well as it's order in the hierarchy. For example, when a quarter mile search radius was used, all top locations occurred within a short distance of each other (not shown). In short, the user must be careful in using the fuzzy mode for analysis.

Nearest Neighbor Hierarchical Clustering

We now turn to methods that identify hot spot areas, as opposed to individual points that are clustered or are the center of a cluster. The *nearest neighbor hierarchical clustering* (or Nnh for short) routine in *CrimeStat* identifies groups of incidents that are spatially close. It is a hierarchical clustering routine that clusters points together on the basis of several criteria. The clustering is repeated until either all points are grouped into a single cluster or else the clustering criteria fail. Hierarchical clustering methods are among the oldest cluster routines (Everitt, Landau and Leese, 2001; King, 1967; Systat, 2008). Among the clustering criteria that have been used are the nearest neighbor method (Johnson, 1967; D'andrade. 1978), farthest neighbor, the centroid method (King, 1967), median clusters (Gowers, 1967), group averages (Sokal and Michener, 1958), and minimum error (Ward, 1967).

The *CrimeStat* Nnh routine is a variation on this approach but has its own unique algorithm. It uses a method that defines a *threshold distance* and compares the threshold to the distances for all pairs of points. Only points that are closer to one or more other points than the threshold distance are selected for clustering. In addition, the user can specify a minimum number of points to be included in a cluster. Only points that fit both criteria - closer than the threshold and belonging to a group having the minimum number of points, are clustered at the first level (first-order clusters).

The routine then conducts subsequent clustering to produce a hierarchy of clusters. The first-order clusters are themselves clustered into second-order clusters. Again, only clusters that are spatially closer than a threshold distance (calculated anew for the second level) are included.

The second-order clusters, in turn, are clustered into third-order clusters, and this re-clustering process is continued until either all clusters converge into a single cluster or, more likely, the clustering criteria fails.

Criterion 1: Threshold Distance

The first criterion in identifying clusters is whether points are closer than a specified threshold distance. There are two alternatives in selecting the threshold distance: 1) a random nearest neighbor distance (the default); or 2) a fixed distance.

Random nearest neighbor distance

The default alternative is to use the expected random nearest neighbor distance for first-order nearest neighbors. The user specifies a *one-tailed* confidence interval around the random expected nearest neighbor distance. The t-value corresponding to this probability level, t, is selected from the Student's t-distribution under the assumption that the degrees of freedom are at least 120.[3] This selection is controlled by a slide bar under the routine (see Figure 7.3). From chapter 6, the mean random distance was defined as:

$$d_{NN(ran)} = 0.5\sqrt{\frac{A}{N}} \qquad \text{repeat 6.2}$$

where A is the area of the region and N is the number of incidents and the standard error of the mean random distance is:

$$SE_{d(ran)} \cong \sqrt{\frac{(4-\pi)A}{4\pi N^2}} = \frac{0.26136}{\sqrt{\frac{N^2}{A}}} \qquad \text{repeat 6.5}$$

where A is the area of the region and N is the sample size (number of incidents). The confidence interval around that distance is defined as:

$$\text{..........} \qquad = \; d_{NN(ran)} \pm t * SE_{d(ran)} \qquad (7.1)$$

where t is the t-value associated with a probability level in the Student's t-distribution.

The approximate lower limit of this confidence interval is:

[3] This is the next highest degree of freedom in the Student's t-table below infinity.

$$\text{.....} \quad limit\ of \qquad\qquad interval = \ d_{NN(ran)} - t * SE_{d(ran)}$$

$$\cong 0.5\sqrt{\frac{A}{N}} - t\sqrt{\frac{(4-\pi)A}{4\pi N^2}} = \frac{0.26136}{\sqrt{\frac{N^2}{A}}} \tag{7.2}$$

and the upper limit of this confidence interval is:

$$Upper\ limit\ of \qquad\qquad ce\ interval = \ d_{NN(ran)} + t * SE_{d(ran)}$$

$$\cong 0.5\sqrt{\frac{A}{N}} + t\sqrt{\frac{(4-\pi)A}{4\pi N^2}} = \frac{0.26136}{\sqrt{\frac{N^2}{A}}} \tag{7.3}$$

The confidence interval defines a probability for the distance between any *pair* of points. For example, for a specific *one-tailed* probability, *p*, fewer than *p*% of the incidents would have nearest neighbor distances smaller than this selected limit *if* the distribution was spatially random. *If* the data were spatially random and if the mean random distance is selected as the threshold criteria (the default position on the slide bar), approximately 50% of the pairs will be closer than this distance. For randomly distributed data, if a $p \leq .05$ level is taken for t (two steps to the left of the default or the fifth in from the left), then only about 5% of the pairs would be closer than the threshold distance. Similarly, if a $p \leq .75$ level is taken for t (one step to the right of the default or the fifth in from the right), then about 75% of the pairs would be closer than the threshold distance.

In other words, the threshold distance is a probability level for selecting any *two* points (a pair) on the basis of a chance distribution. The slide bar has 12 levels and is associated with a probability level for a t-distribution from a sample of 120 or larger. From the left, the p-values are approximately (Table 7.2):

Taking a broader conception of this, if there is a spatially random distribution, then for all distances between unique pairs of points, of which there are

$$Combinations = \frac{N(N-1)}{2} \tag{7.4}$$

fewer than *p*% will be shorter than this threshold distance.

Table 7.2:
Approximate Probability Values Associated with Threshold Scale Bar

Position	Scale Bar Probability	Description
1	0.00001	Far left point of slide bar
2	0.0001	Second from left
3	0.001	Third from left
4	0.01	Fourth from left
5	0.05	Fifth from left
6	0.1	Sixth from left
7	0.5	Sixth from right (default value)
8	0.75	Fifth from right
9	0.9	Fourth from righ
10	0.95	Third from righ
11	0.99	Second from righ
12	0.999	Far right point of slide bar

This does not mean, however, that the probability of finding a cluster is equal to this probability. It only indicates the probability of selecting two points (a pair) on the basis of a chance distribution. If additional points are to be included in the cluster, then the probability of obtaining the cluster will be less. Thus, the probability of selecting three points or four points or more points on the basis of chance will be much smaller.

Area must be defined correctly

Note that it is *very* important that area be defined correctly for this routine to work. If the user defines the area on the measurement parameters page (see chapter 3), the Nnh routine uses that value to calculate the threshold distance. If the user does not define the area on the measurement parameters page, the routine calculates the area from the minimum and maximum X/Y values (the bounding rectangle), which will usually be a larger area. In either case, the routine will be able to calculate a threshold distance and run the routine.

However, if the area units are defined incorrectly on the measurement parameters page, then the routine will certainly calculate the threshold distance wrongly. For example, if data are in feet but the area on the measurement parameters page are defined in square miles, most likely the routine will not find any points that are farther apart the threshold distance since that distance is defined in miles. In other words, it is essential that the area units be consistent with the data for the routine to properly work.

Fixed distance

The second alternative for selecting a threshold distance is to choose a fixed distance (in miles, nautical miles, feet, kilometers, or meters). The user checks the "Fixed distance" box and selects a threshold distance. The main advantage in this approach is that the search radius can be specified exactly. This is useful for comparing the number of clusters for different distributions (e.g., the number of robbery hot spots compared to burglary hot spots using a search radius of 0.5 miles). The main disadvantage of this method is that the choice of a threshold is subjective. The larger the distance that is selected, the greater the likelihood that clusters will be found by chance. Of course, this can be tested using a Monte Carlo simulation (see below).

Criterion 2: Minimum Number of Points

Whichever method is used for selecting a threshold distance, a second clustering criterion is the minimum number of points that are required for each cluster. This criterion is used to reduce the number of very small clusters. With large data sets, hundreds, if not thousands, of clusters can be found if only pairs of points are selected as being closer than a threshold distance. To minimize numerous very small clusters as well as reduce the likelihood that clusters could be found by chance, the user can set a minimum number restriction. The default is 10. This decision does not affect the selection of the clusters, only the number that are output. By decreasing this number, more clusters are output; conversely, by increasing this number, fewer clusters are output. The routine will only include points in the final clustering that are part of clusters in which the minimum number is found.

First-order Clustering

Using these criteria, *CrimeStat* constructs a first-order clustering of the points (see endnote .). For each first-order cluster, the center of minimum distance is output as the cluster center, which can be saved as a '.dbf' file.

Second and Higher-order Clusters

The first-order clusters are then tested for second-order clustering. The procedure is similar to first-order clustering except that the cluster centers (the center of minimum distance for each) are now treated as 'points' which themselves are clustered (see endnote *ii*). The process is repeated until no further clustering can be conducted. Either all sub-clusters converge into a single cluster, the threshold distance criterion fails, or there are fewer than four seeds in the higher-order cluster.

Visualizing the Cluster Output

To identify the approximate cluster location, *CrimeStat* allows the cluster to be output as either as an ellipse, a convex hull, or both.

Ellipse output

A standard deviational ellipse is calculated for each cluster (see chapter 4 for the definition). The user can choose between 1 standard deviation (the default), 1.5 standard deviations, or 2 standard deviations (indicated on the interface by 1X, 1.5X, and 2X). Typically, one standard deviation will cover more than 50% of the cases, one and a half standard deviations will cover more than 90% of the cases, and two standard deviations will cover more than 99% of the cases, although the exact percentage will depend on the distribution. The user specifies the number of standard deviations to save as ellipses in *ArcGIS* '.shp', *MapInfo* '.mif', *Google Earth* 'kml' (if the data are in spherical coordinates), or various Ascii formats.

Be careful as standard deviations can create an exaggerated view of the underlying cluster. The ellipse, after all, is an abstraction from the points in the cluster that may be arranged in an irregular manner. For example, for a regional view, a 1 standard deviational ellipse may not be very visible while for a small area, a 2 standard deviational ellipse may be too big. The user has to balance the need to accurately display the cluster compared to making it easier for a viewer to understand its location.

Convex hull output

A convex hull is calculated for each cluster (see chapter 4 for definition). The convex hull draws a polygon around the points in the cluster. It is a literal definition of the cluster, as opposed to the ellipse which is an abstraction. The convex hull can be saved in *ArcGIS* '.shp', *MapInfo* '.mif', *Google Earth* 'kml', or various Ascii formats.

Ellipse or convex hulls?

With the choice of an ellipse or a convex hull, the user can visualize clusters in two different ways. There are advantages and disadvantages of each approach. The convex hull has the advantage of being a polygon that corresponds exactly to the cluster. For neighborhood level analysis, it is probably preferable to the ellipse, which is an abstraction. On the other hand, any convex hull is based on a sample (e.g., this year's robberies compared to last year's robberies) and like any sample will vary from one instance to another. It may not capture all the space associated with the hot spot. The shape of a convex hull is often un-intuitive, following the outline of the incidents. An ellipse, on the other hand, is more general and will usually be more stable from year to year. It usually looks better on a map or at least users seem to understand it better; it is a more familiar graphical object than an irregular polygon. The biggest disadvantage to an ellipse is that it forces a certain shape on the data, whether there are incidents in every part of it or not. So, in extreme cases, one finds ellipses that go outside of study area boundaries or extend into reservoirs or lakes or other features that are logically impossible to have incidents. At the same time, the ellipses may not include locations that are actually part of the hot spot..

In short, the user needs to balance the generality and visual familiarity of an ellipse with the limits of the actual hot spot. Probably for a small scale, regional perspective, the ellipses are preferable since a viewer can quickly see where the hot spots are located. For detailed neighborhood-level work, however, the convex hull is probably better since it shows where the incidents actually occurred.

Abstraction of incidents with second- and higher-order clusters

One thing to note is that second- and higher-order clusters can be visually misleading. The second-order clusters may visually encompass points that were not clustered in the first-order but they only are calculated using the centroids of the first-order clusters. Thus, in a GIS, one could select all incidents that fall within the boundaries of the second-order cluster (whether defined by an ellipse or a convex hull) and the number will generally be more than the points that were accumulated from the first-order clusters. A user needs to be aware of this as second- and higher-order clusters are abstractions from first- and earlier-order clusters.

Guidelines for Selecting Parameters

In the Nnh routine, the user has to define three parameters - the threshold distance, the minimum number of points, and the visual output of the hot spots. For a fixed threshold distance, the user has to choose a distance that is meaningful. For crime incidents, probably the threshold distance should not be more than 0.5 miles and, preferably, smaller.

If the random nearest neighbor distance is used as a threshold, the p-value is selected with a likelihood slider bar (see Figure 7.3). This bar indicates a range of p-values from 0.00001 (i.e., the likelihood of obtaining a pair by chance is 0.001%) to 0.999 (i.e., the likelihood of obtaining a pair by chance is 99.9%). The slider bar actually controls the value of *t* in equation 7.3, which varies from -3.719 to +3.090. The smaller the t-value, the smaller the threshold distance. With smaller threshold distances, fewer clusters are extracted and are typically smaller (although not always). Thus, they are more likely to be *not* due to chance.

If only pairs of points were being grouped, then the threshold distance would be critical. For example, with the default p≤.5 value, then about half the pairs would be selected by chance if the data were truly random. However, since there are a minimum number of points that are specified, the likelihood of finding a cluster with the minimum number of points is much smaller. The larger the minimum number selected, the smaller the likelihood of obtaining a cluster by chance.

Therefore, one can think of the slide bar as a filter for grouping points. One can make the filter smaller (moving the slide bar to the left) or larger (moving the slide bar to the right). There will be some effect on the final number of clusters, but the likelihood of obtaining a cluster by chance will be generally low. Statistically, there is more certainty with small threshold distances than with larger ones using this technique. Thus, a user must trade off the number of clusters and the size of an area that defines a cluster with the likelihood that the result could be due to chance.

This choice will depend on the needs of the user. For interventions around particular locations, the use of a small threshold distance may actually be appropriate; some of the ellipses seen in Figure 7.7 below cover only a couple of street segments. These define micro-neighborhoods. On the other hand, for a patrol route, for example, a cluster the size of several neighborhoods might be more appropriate. A patrol car would need to cover a sizeable area and having a larger area to target might be more appropriate than a 'micro' environment. However, there will be less precision with a larger cluster size in this type of area.

A second criterion is the minimum number of points that are required to define a cluster. If a cluster does not have this minimum number, *CrimeStat* will ignore the seed location. Without this criterion, the Nnh routine could identify clusters of two or three incidents each. A hot spot of this size is usually not very useful. Consequently, the user should increase the number to ensure that the identified cluster represents a meaningful number of cases. The default value is 10, but the user can type in any other value.

The user may have to experiment with several runs to get a solution that appears right. As a rule of thumb, start with the default settings. If there appears to be too many clusters, tighten up the criteria by selecting a lower probability for grouping a pair by chance (i.e., shifting the threshold distance to the left) or by increasing the minimum number of points required to be defined as a cluster (e.g., from 10 to 20). On the other hand, if there appears to be too few clusters, loosen the criteria by selecting a higher probability for grouping pairs by chance (i.e., shifting the threshold distance to the right) or decreasing the minimum number of points in a cluster (e.g., from 10 to 5). Then, once an appropriate solution has been found, the user can fine tune the results by slight changes.

In general, the minimum number of points criterion is more critical for the number of clusters than the threshold distance, though the latter can also influence the results. For example, with the 1996 Baltimore County robbery data set (N=1181 incidents), a minimum of 26 and a maximum of 28 clusters were found by changing the threshold distance from the minimum p-value ($p \leq 0.00001$) to the maximum p-value ($p \leq 0.999$). On the other hand, changing the minimum number of points per clusters from 10 to 20 reduced the number of clusters found (with the default threshold distance) from 26 to 11.

The third criterion is the visual display of the clusters. The convex hull is literal; it will draw a polygon around the points in the cluster. The ellipse, on the other hand, requires a decision by the user on the number of standard deviations to be displayed. The choices are one (the default), one and a half, and two standard deviations. Typically, one standard deviation will cover more than 50% of the cases; one and a half standard deviations will cover more than 90% of the cases, while two standard deviations will cover more than 99% of the cases although the exact percentage will depend on the distribution.

In general, I recommend using a 1.5 as the default as 1 standard deviation will be often be too small while 2 standard deviations can create an exaggerated view of the underlying cluster. The user has to balance the need to accurately display the cluster compared to making it easier for a viewer to understand its location.

Nnh Output Files

The Nnh routine has six outputs: First, for each cluster that is identified, the hierarchical order and the cluster number; Second, for each cluster that is calculated, the mean center of the cluster. Only 45 of the seed locations are displayed on the screen. The user can scroll down or across by adjusting the horizontal and vertical slider bars and clicking on the *Go* button. This can be saved as a '.dbf' file; Third, the standard deviational ellipses of the clusters is shown, whether the graphical output is an ellipse or a convex hull. The size of the ellipses is determined by the number of standard deviations to be calculated (see above); Fourth, the number of points in the cluster; Fifth, the area of the ellipse; and, Sixth, the density of the cluster (number of points divided by area).

The ellipses and convex hulls can be saved in *ArcGIS* '.shp', *MapInfo* '.mif', *Google Earth* 'kml', or various Ascii formats. Because there are also orders of clusters (i.e., first-order, second-order, etc.), there is a naming convention that distinguishes the order.

Naming conventions for ellipses

For the ellipses, the convention is

Nnh<O><*username*>

where *O* is the order number and *username* is a name provide by the user. Thus,

Nnh1robbery

are the first-order clusters for a file called 'robbery' and

Nnh2NightBurglaries

are the second-order clusters for a file called 'NightBurglaries'. Within files, clusters are named

Nnh<O>Ell<N><*username*>

where*O* is the order number, *N* is the ellipse number and *username* is the user-defined name of the file. Thus,

Nnh1Ell10robbery

is the tenth ellipse within the first-order clusters for the file 'robbery' while
Nnh2Ell1NightBurglaries

is the first ellipse within the second-order clusters for the file 'NightBurglaries'.

For the convex hulls, the name will be output with a 'CNNH1' prefix for the first-order clusters, a 'CNNH2' prefix for the second-order clusters, and a 'CNNH3' prefix for the third-order clusters. Higher-order clusters will index only the number.

In other words, names of files and features can get complicated. The easiest way to understand this, therefore, is to import the file into one of the GIS packages and display it.

Example 1: Nearest Neighbor Hierarchical Clustering of San Antonio Robberies

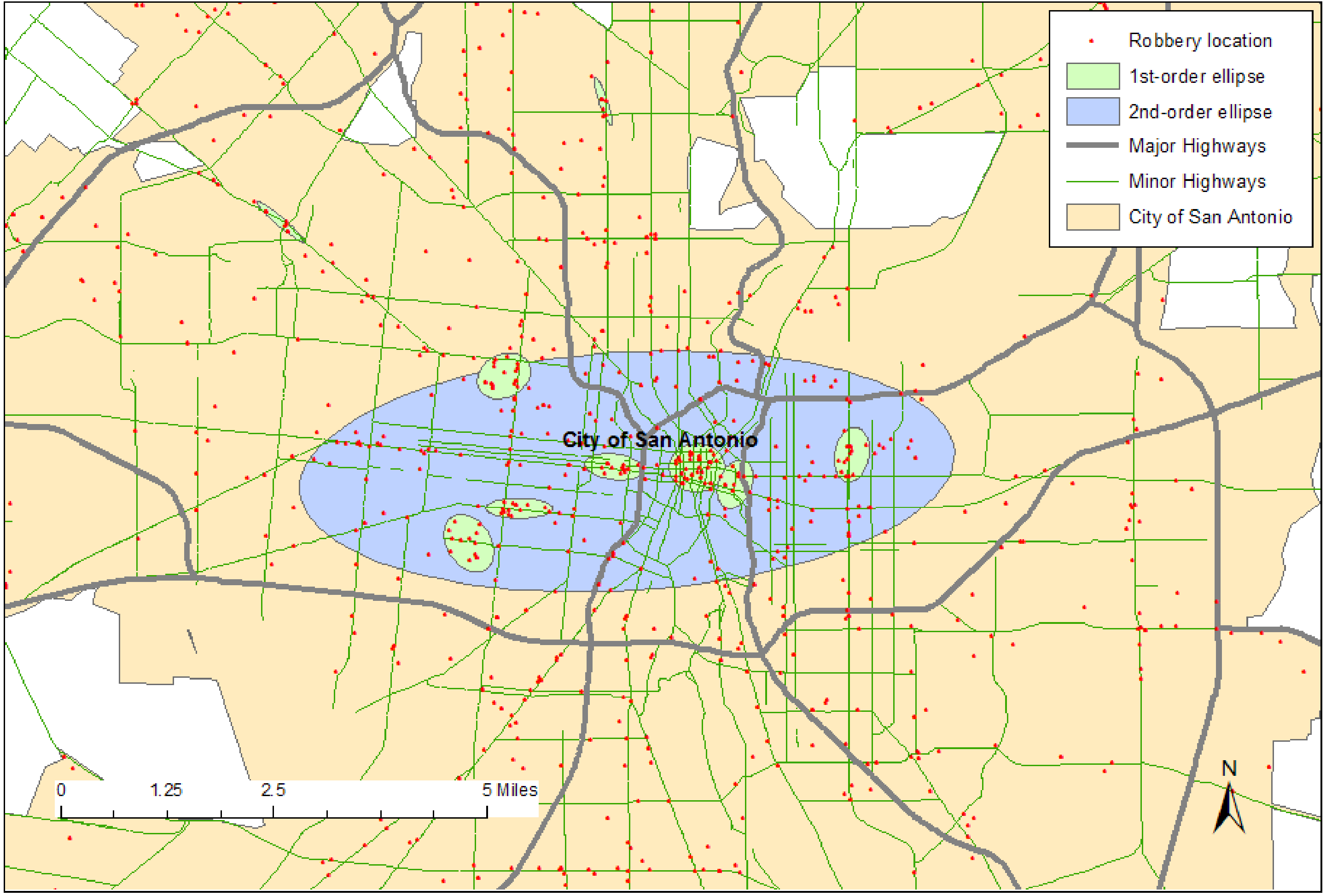
The Nnh routine was applied to 1,116 robberies that occurred in 2003 in San Antonio, TX. A default one-tailed probability level of .05 (or 5%) was selected for the threshold distance and each cluster was required to contain a minimum of 10 points (the default). Using these criteria, *CrimeStat* returned 9 first-order clusters and one second-order cluster. The 9 first-order clusters varied from 37 incidents for one cluster to 7 incidents for two clusters. Figure 7.7 shows the first-order clusters and the second-order cluster displayed as 1.5 standard deviational ellipses.

Since the criteria for clustering is the lower limit of the mean random distance, the distances involved are very small, as can be seen. Note, the standard deviational ellipse is

Figure 7.7:

San Antonio Robberies: 2003

Ellipses of 1st-order and 2nd-order Hot Spots



defined by the points in the cluster but is an abstraction, rather a literal definition. Thus, there is not a one-to-one match between the ellipse boundaries and the points included. For example, the top cluster had 37 points yet the 1.5 standard deviational ellipse included only 36 of those points.

Figure 7.8 shows the shows the same clusters as in Figure 7.7 but the clusters are displayed as convex hulls rather than ellipses. As seen, the convex hulls are irregular in shape and more limited in geographical spread; they show only the incidents that are clusters. Notice how the one second-order cluster defined by the convex hull is much more constrained than the ellipse definition of it.

Note also that the second-order cluster includes incidents that were not clustered in the first-order clusters. Thus, the area included in the second-order cluster is much greater than the sum of the first-order clusters from which it was derived. This may lead to a wider definition of a larger hot spot which may be real or not. One has to keep in mind that the second- and higher-order clusters are abstractions of the first-order clusters, and are not clusters by themselves.

Figure 7.9 zooms and compares the seven central clusters in terms of the ellipses and the corresponding convex hulls. Notice how the convex hulls are much more compact. Also, how the convex hulls 'stick out' beyond the ellipses for four of the clusters. Again, this is because the ellipse is a mathematical abstraction whose central axes are defined by the points, whereas the convex hull is defined by a polygon that defines an outer boundary.

From a policing viewpoint, a convex hull is probably more useful in that it shows where the hot spot incidents are actually located. As mentioned above, the polygons created by the convex hulls are irregular and are, therefore, less familiar to most people. Consequently, for presentations of crime patterns at a regional level or even neighborhood-level for non-specialists, the ellipses may convey better where the hot spots are located.

Simulating Statistical Significance

Testing the significance of clusters from the Nnh routine is complex. Conceptually, using the random nearest neighbor distance for the threshold distance defines the probability that two points could be grouped together on the basis of chance. The test is for the confidence interval around the first-order nearest neighbor distance for a random distribution. If the probability level is p%, then approximately p% of all pairs of points would be found under a random distribution. Under this situation, we would know whether the number of clusters (pairs) that were found were significantly greater than would be expected on the basis of chance.

The problem is, however, that the routine is not just clustering pairs of points, but clustering as many points as possible that fall within the threshold distance since there is an

Figure 7.8:

San Antonio Robberies: 2003

Convex Hulls of 1st-order and 2nd-order Hot Spots

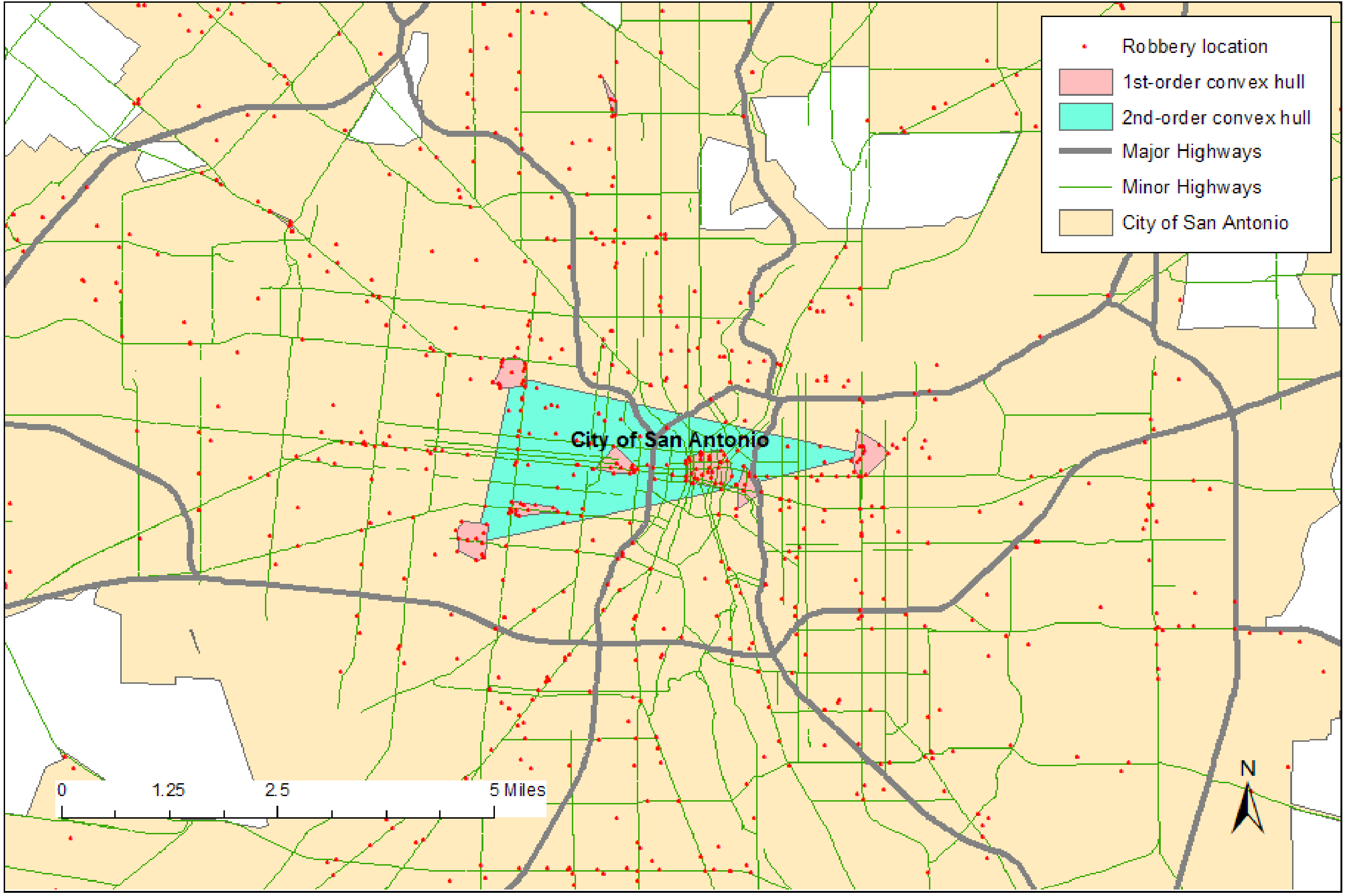
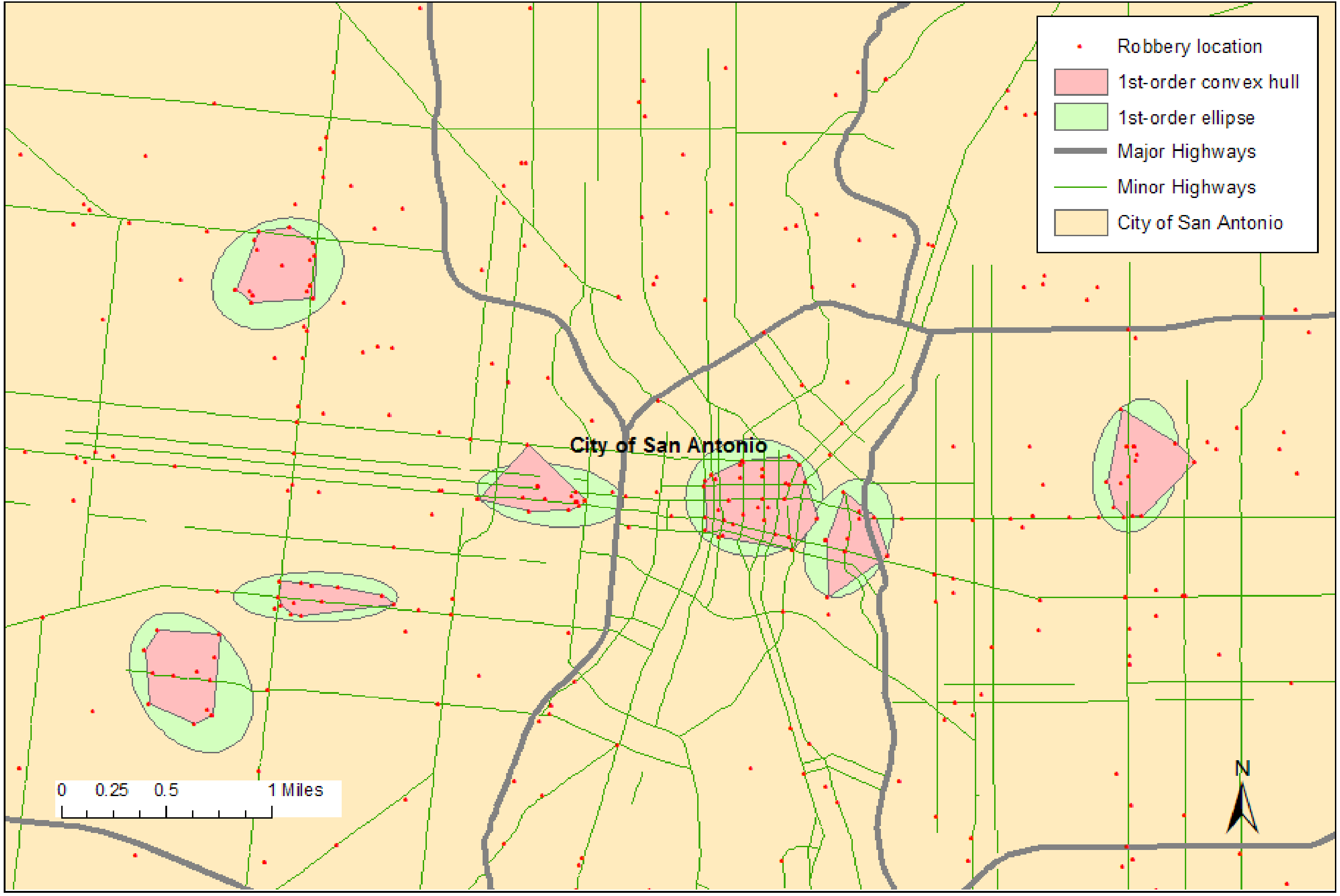


Figure 7.9:

San Antonio Robberies: 2003

Comparing Ellipses and Convex Hulls of 1st-order Hot Spots



additional requirement that there be a specified minimum number of points, with the minimum defined by the user. The probability distribution for this situation is not known. Consequently, there is a necessity to resort to a Monte Carlo simulation of randomness under the conditions of the Nnh test (Dwass, 1957; Barnard, 1963).

CrimeStat includes a Monte Carlo simulation routine that produces approximate confidence intervals (called *credible intervals*) for the first-order Nnh clusters that have been identified. Second- and higher-order clusters are not simulated since their structure depends on the first-order clusters. Essentially, the routine assigns *N* cases randomly to a rectangle with the same area as the defined study area, *A*, and evaluates the number of clusters according to the defined parameters (i.e., threshold distance and minimum number of points). It repeats this simulation *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 approximate credible intervals for the particular first-order Nnh.

The output includes five columns and twelve rows:

Columns:

1. The percentile,
2. The number of first-order clusters found for that percentile,
3. The area of the cluster for that percentile,
4. The number of points in the cluster for that percentile, and
5. The density of points (per unit area) for that percentile.

Rows:

1. The minimum (smallest) value obtained,
2. 0.5th percentile,
3. 1st percentile,
4. 2.5th percentile,
5. 5th percentile,
6. 10th percentile,
7. 90th percentile,
8. 95th percentile,
9. 97.5th percentile,
10. 99th percentile,
11. 99.5th percentile, and
12. The maximum (largest) value obtained.

The percentiles are calculated as follows. First, over all simulation runs (e.g., 1000), the routine calculates the number of first-order clusters obtained for each run, sorts them in ascending order, and defines the percentiles for the list. Thus, the minimum is the fewest number of clusters obtained over all runs, the 0.5 percentile is the lowest half of a percent for the number of clusters obtained over all runs, and so forth until the maximum number of clusters obtained over all runs. The routine does *not* calculate second- or higher-order clusters since those are dependent on the first order clustering. Second, within each run, the routine calculates the number of points per cluster, the area of each ellipse, and the density of each ellipse. Then, it groups all clusters together, over all runs, and sorts them into a list. The percentiles for individual clusters are then calculated. Note that the points refer to the cluster whereas the area and density refer to the ellipses, which is a geometrical abstraction from the cluster.

When a Monte Carlo simulation of 1000 iterations was run on the San Antonio robbery data, no clusters were found. That is, given the criteria that were used for clustering (the default random nearest neighbor distance and a minimum of 10 incidents per cluster), it would be very unlikely to find any clusters on the basis of chance!

To illustrate how a simulation which found random clusters looks, Table 7.3 presents an Nnh run that was conducted on a Baltimore County robbery data base (N=1181 incidents) using the default threshold distance ($p \leq .5$ for grouping a pair by chance) and a minimum number of points of at least five for each cluster. Then, 1000 Monte Carlo runs were conducted with simulated data. With the actual data, the Nnh routine identified 69 first-order clusters and 7 second-order clusters. Table 7.3 presents the parameters for the first ten first-order clusters.

In examining a simulation, one has to select percentiles as choice points. In this example, we use the 95th percentile. That is, we are willing to accept a one-tailed Type I error of only 5% since we are only interested in finding a greater number of clusters than by chance. For the simulation, look at each column of the simulation results in turn. Column 2 presents the number of clusters found in each simulation. Over the 1000 runs, there was a minimum of one cluster found (for at least one simulation) and a maximum of 7 clusters found (for at least one simulation). That is, running 1000 simulations of randomly assigned data only yielded between 1 and 7 clusters using the parameters defined in the particular Nnh run. The 95th percentile was 3. It is highly unlikely that the 69 first-order clusters that were identified would have been due to chance. That is, we would have expected at most three of them to have been due to chance. It appears that the robbery data is significantly clustered, though we have only tested significance through a random simulation.

Of course, the routine is not going to identify which three clusters could have been selected on the basis of chance. However, realistically the three clusters would be those with the lowest density, number of points per unit area (e.g., points per square mile; points per square

Table 7.3:
Simulated Confidence Intervals for Nnh Routine
Baltimore County Robberies: N=1181

```
Nearest Neighbor Hierarchical Clustering:
----------------------------------------
          Sample size..........................:              1181
          Likelihood of grouping  pair of points by chance....:   0.50000 (50.000%)
          Z-value for confidence  interval.............................:   0.000
          Measurement type..............:                    Direct
          Output units......................:                Miles, Squared Miles, Points per Squared Miles
          Clusters found....................:                76
          Simulation runs.................:                  1000

Displaying ellipses starting from 1 (ONLY 10 SHOWN)

Order    Cluster  Mean X       Mean Y       Rotation     X-Axis     Y-Axis     Area        Points   Density
-------  ---------  ---------   ----------    -----------  ---------   --------     -------     ----     -------------
1          1      -76.44927   39.31455    77.09164    0.28303    0.09636   0.08568       40    66.828013
1          2      -76.60219   39.40050    11.98132    0.11540    0.27452   0.09952       33     331.580616
1          3      -76.44601   39.30490    16.66988    0.21907    0.16239   0.11176       25    23.684859
1          4      -76.78123   39.36088    25.36983    0.27643    0.14530   0.12618       29    229.826284
1          5      -76.73103   39.34319    67.71617    0.19445    0.16058   0.09810       29    295.628310
1          6      -76.72945   39.28910    79.88383    0.16428    0.25957   0.13396       29    216.476166
1          7      -76.51486   39.25986    87.32563    0.19148    0.29428   0.17703       27    152.520725
1          8      -76.45374   39.32106    54.57635    0.15150    0.18261   0.08692        7      80.538112
1          9      -76.75368   39.31132    89.56994    0.19748    0.22914   0.14216       22    154.753006
1         10      -76.71641   39.29139    10.43857    0.15048    0.16879   0.07980       14    175.444372
Etc.

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

          Percentile        Clusters          Area              Points            Density
          --------------    -----------       -------           --------          ---------------
          min                  1              0.03845            5                  15.615111
          0.5                  1              0.04922            6                  16.608967
          1.0                  1              0.05603            6                  17.162252
          2.5                  1              0.06901            6                  18.570113
          5.0                  1              0.08243            6                  19.468353
          10.0                 1              0.10045            6                  21.256559
          90.0                 2              0.28706            7                  61.173748
          95.0                 3              0.31074            7                  73.463654
          97.5                 3              0.32442            7                  87.550868
          99.0                 4              0.35279            8                 115.460337
          99.5                 5              0.36489            8                 122.625375
           max                 7              0.38424            9                 156.056837
```

kilometer). Thus, the user could assume that the three clusters with the lowest density are less certain to be real than due to chance.

Column 3 shows the areas of clusters that were found over the 1000 runs using the ellipse as a definition for the clusters. For the individual clusters, the simulation showed a range from about 0.04 to 0.38. The 95th percentile was 0.31. In the actual Nnh, the area of clusters varied between 0.05 and 0.27, indicating that *all* first-order clusters were smaller than the smallest value found in the simulation. In other words, the real clusters are more compact than random clusters even though the random clusters were subject to the same threshold distance as the real data. This is not always true, but, in this case, it is.

Column 4 presents the number of points found per cluster in the simulation; these varied between 5 and 9 points per cluster. The 95th percentile was 7. With the actual data, the number of points varied between 5 and 40. Thus, some of the clusters could have been due to chance, at least in terms of the number of points per cluster. Analyzing the distribution (not shown), 27 of the 69 clusters had 7 or fewer points. In other words, about 39% had only as many points as might be expected on the basis of a chance distribution. Putting it another way, about 40% of the clusters had more points than would be expected on the basis of chance 95% of the time.

Finally, column 5 presents the density of points found per cluster. Since the output unit is squared miles, density is the number of points per square mile. The simulation presents a range from 15.6 points per square mile to 156.1 points per square mile. The 95th percentile was 73.4 points per square mile. The actual Nnh, on the other hand, finds a range of densities from 27.1 points per square mile to a very high number (11071821 points per square mile). Again, there is overlap between the actual clusters and what might be expected on the basis of chance; 26 out of 69 clusters have densities that are lower than the 95th percentile found in the simulation. Again, about 38% have densities are not different than would be expected on the basis of chance.

It should be clear that testing the significance of a cluster analysis is complex. In the example, some of the criteria chosen were definitely different than a chance distribution while other criteria were not very different. However, which of these criteria should be used to evaluate the actual distribution? We argue that it should be the number of incidents/points identified in the cluster, rather than the area or density by themselves since the area has to be defined by a polygon (ellipse or convex hull). The number of points is the relevant criterion since it is one of the criteria used for the clustering in the Nnh algorithm (the other being points that are closer than the threshold distance.

Uses of Hierarchical Clustering

There are four uses for the nearest neighbor hierarchical clustering technique. First, it can identify small geographical environments where there are concentrated incidents. This can be useful for specific targeting, either by police deployment or community intervention. There are clearly micro-environments that generate crime incidents and the Nhh technique tends to identify these small environments because the lower limit of the mean random distance is used to group the clusters. The user can, of course, control the size of the grouping area by loosening or tightening either the threshold distance or the minimum number of required points. Thus, the sizes of the clusters can be adjusted to fit particular groupings of points.

Second, the technique can be applied to any entire data set, such as for Baltimore County and Baltimore City, and need not only be applied to smaller geographical areas, such as precincts. This increases the ease of use for analysts and can facilitate comparisons between different areas without having to limit arbitrarily the data set.

Third, the linkages between several small clusters can be seen through the second- and higher-order clusters. Frequently, hot spots are located near other hot spots which, in turn, are located still near other hot spots.

In other words, the clustering of incidents, such as robberies, is hierarchical. With the San Antonio robbery data, we found two levels of grouping (first-order and second-order). With larger datasets, however, frequently third-order or, even, fourth-order hot spots can be found. Within these large areas, there are smaller hot spots and within some of those hot spots, there are even smaller ones. In other words, there are different scales to the clustering of points - different geographical levels, if you will, and the hierarchical clustering technique can identify these levels.

Typically, in cities as well as in small towns, there is a greater concentration towards the center of the settlement or city than at the periphery. This concentration necessarily means there will be more incidents (of any sort) towards the center than toward the periphery. The Nnh routine captures this logic very nicely because it seeks clusters systematically from the incident level upwards. More first-order clusters are going to be found in the center than in the periphery and this is also going to be true for second- and higher-order clusters since they build systematically on the first-order clusters. One can think of the first-order clusters as ‘building blocks’ for spatial autocorrelation. Thus, theoretically, hierarchical clusters capture the organization of a human settlement, particularly a city, in a way that no other clustering technique does.

Fourth, each of the levels implies different policing strategies. For the smallest level, officers can intervene effectively in small neighborhoods, as discussed above. Second-order clusters, on the other hand, are more appropriate as patrol areas; these areas are larger than first-order clusters, but include several first-order clusters within them. If third- or higher-order clusters are identified, these are generally areas with very high concentrations of crime incidents over a fairly large section of the jurisdiction. The areas start to approximate precinct sizes and need to be thought of in terms of an integrated management strategy - police deployment, crime prevention, community involvement, and long-range planning. Thus, the hierarchical technique allows different security strategies to be adopted and provides a coherent way of approaching these communities and gives flexibility to the analyst in order to choose an appropriate level of analysis. This depends, of course, on the need. For patrol cars covering an area, such as is common in the United States, larger hot spot areas are more appropriate. Police cars will drive around the area and will cover blocks and neighborhoods that don't necessarily have high crime in order to demonstrate their presence as well as make their behavior less predictable. For this use, second- or higher-order hot spots would be appropriate. Also, some of the techniques discussed in Chapters 8 and 10 are also appropriate for larger area analysis.

However, if the policing strategy involves working with businesses or even residents to develop, for example, a business- or neighborhood watch program, then the boundaries of the hot spot need to be defined fairly specifically, perhaps a block or two. Choosing a larger area may diffuse efforts and reduce the effectiveness of the intervention. Even more precise boundary definition are needed for public infrastructure improvements, such as improved lighting or closed circuit television systems (CCTV). The public works departments that install these improvements need to know exactly where to put the lights or CCTV cameras.

In other words, the analytical need is going to depend on the particular type of intervention or program that will be introduced and the hierarchical clusters provide a range of scales from which an appropriate one could be chosen.

Limitation to Hierarchical Clustering

There are also limitations to the technique, some technical and others theoretical. First, the method only clusters incidents (points); a weighting or intensity variable will have no effect. In Chapter 9, we introduce a variant of the Nnh that allows weighting incidents and can be applied to zonal data. The results are reasonable approximations to clusters of zones, but they lack the specificity of the incident data.

Second, the size of the grouping area is dependent on the sample size when the confidence interval around the mean random distance is used as the threshold distance criteria (see equation. 6.2). For crime distributions that have many incidents (e.g., burglary), the threshold distance will be a lot smaller than distributions that have fewer incidents (e.g., robbery). In theory, a hot spot is dependent on an environment, not the number of incidents. Thus, that approach does not produce a consistent definition of a hot spot area. Using a fixed distance for the threshold distance can partly overcome this. However, the fixed distance needs to be tested for randomness using the Monte Carlo simulation.

Third, there is some arbitrariness in the technique due to the minimum points rule. This implicitly requires the user to define a meaningful cluster size, whether the number of minimum points required is 5, 10, 15 or whatever. To some extent, this is how patterns are defined by human beings; with one or two incidents in a small area, people do not perceive any pattern. As soon as the number of incidents increases, say to 10 or more, people perceive the pattern. This is not a statistical way for defining regularity, but it is a human way. However, it can lead to arbitrariness since two different users may interpret the size of a hot spot differently. Similarly, the selectivity of the p-value, vis-a-via the Student's t-distribution, can allow variability between users.

In short, the technique produces a consistent result, but one subject to manipulation by users. Hierarchical techniques are, of course, not the only clustering procedures to allow users to adjust the parameters; in fact, almost all the cluster techniques have this property. But it is a statistical weakness in that it involves subjectivity and is not necessarily consistently applied across users.

Finally, there is no substantive theory or rationale behind the clusters. They are empirical derivatives of a procedure. Again, many clustering techniques are empirical groupings and also do not have any explanatory theory.[4] If one is looking for a substantive hot spot defined by a unique constellation of land uses, activities, and targets, the technique does not provide any insight into why the clusters are occurring or why they could be related. I will return to this point at the end of the next chapter, but it should be remembered that these are empirical groupings, not necessarily substantive ones.

Risk-Adjusted Nearest Neighbor Hierarchical Clustering

CrimeStat also includes a risk-adjusted nearest neighbor hierarchical clustering routine (Rnnh), which is a variation on the Nnh routine discussed above. It combines the hierarchical clustering capabilities of the Nnh routine with kernel density interpolation technique that is discussed in Chapter 10.

[4] A number of clustering techniques have a statistical theory behind them (e.g., Kulldorff, 1997), but not a substantive theory. While one can define consistent statistical criteria for identifying hot spots, this does not constitute an explanation for why the hot spots occurred. For this, other information is necessary.

The Nnh routine identifies clusters of points that are close together. That is, it will identify groups of points that are closer together than a threshold distance and in which the minimum number of points is greater than a user-defined value. Many of these clusters, however, are due to a high concentration of persons in the vicinity. That is, because the population is not arranged randomly over a plane, but is, instead, highly concentrated in population centers, there is a higher likelihood of incidents happening (whatever they are) simply due to the higher population concentration. In the above examples, many of the clusters for Baltimore burglaries or vehicle thefts were due primarily to a high concentration of households and vehicles in the center of the metropolitan area. In fact, one would normally expect a higher concentration of incidents in the center since there are more persons residing in the center and, certainly, more persons being concentrated there during the daytime through employment, shopping, cultural attendance, and other urban activities.

For many police purposes, the concentration of incidents is of sufficient interest in itself. Police have to intervene at high incidence locations irrespective of whether there is also a larger population at those locations. The demands for policing and responding to community emergency needs is population sensitive since there are more demands where there are more persons. From a service viewpoint, the concentration of incidents is what is important.

But for other purposes, the concentration of incidents relative to the baseline population is of interest. Crime prevention activities, for example, are aimed at reducing the number of crimes that occur for every area in which they are applied. For these purposes, the *rate* of decrease in the number of crimes is the prime focus. Similarly, after-school programs are aimed at neighborhoods where there is a high risk of crime, whether or not there is also a large population. In other words, for many purposes, the *risk* of crime or other types of incidents is of paramount importance, rather than the *volume* (i.e., absolute amount) of crime by itself. If the aim is to assess where there are high risk clusters, then the Nnh routine is not appropriate.

CrimeStat includes a Risk-adjusted Nearest Neighbor Hierarchical Clustering routine (or Rnnh) that defines clusters of points that are closer than what would be expected on the basis of a baseline population. It does this by dynamically adjusting the threshold distance in the Nnh routine according to the distribution of a second, baseline variable. Unlike the Nnh routine where the threshold distance is constant throughout the study area (i.e., it is used to pair points irrespective of where they are within the area), the Rnnh routine adjusts the threshold distance according to what would be expected on the basis of the baseline variable. It is a *risk* measure, rather than a volume measure.

Dynamic Adjustment of the Threshold Distance

To understand how this works, think of a simple example. In a typical metropolitan area, there are more people living towards the center than in the periphery. There are topographical and social factors that might modify this (e.g., an ocean, a mountain range, a lake), but in general population densities are much higher in the center than in the suburbs. If a different baseline variable were selected than population, for example, employment, one would generally find even higher concentrations since central city employment tends to be very high relative to suburban employment. Thus, if population or employment (or another variable that is correlated with population density) is taken as the baseline, then one would expect more people and, hence, more incidents occurring in the center rather than the periphery. In other words, all other things being equal, there should be more robberies, more burglaries, more homicides, more vehicle thefts, and more of any other type of event in the center than in the periphery of an urban area. This is just a by-product of urban societies.

Using this idea to cluster incidents together, then, intuitively, the threshold distance must be adjusted for the varying population densities. In the center, the threshold must be short since one would expect there to be more persons. Conversely, in the periphery - the far suburbs, the threshold distance must be a lot longer since there are far fewer persons per unit of area. In other words, *dynamic adjustment* of the threshold grouping distance means changing the distance inversely proportional to the population density of the location; in the center, a high density means a short threshold distance and in the periphery, a low density means a larger threshold distance.

Kernel Adjustment of the Threshold Distance

To implement this logic, *CrimeStat* overlays a standard grid and uses an interpolation algorithm, based on the kernel density method, to estimate the expected number of incidents per grid cell *if* the actual incident file was distributed according to the baseline variable. Chapter 10 discusses in detail the kernel density method and the reader should be familiar with the method before attempting to use the Rnnh routine. If not, the author highly recommends that Chapter 10 be read before reading the rest of this section.

Steps in the Rnnh Routine

The Rnnh routine works as follows:

1. Both primary and secondary files are required. The primary file is the basic file of incidents (e.g., robberies) while the secondary file is the baseline variable (e.g., population of zones; all crimes as a baseline; or another baseline variable). If the

baseline variable is identified by zones, the user must define both the X and Y coordinates as well as the variable assigned to the zone (e.g., population); the latter will typically be an intensity or weight variable (see Chapter 3).

2. A grid is defined in the reference file tab of the data setup section (see Chapter 3). The Rnnh routine takes the lower-left and upper-right limits of the grid, but uses a standard number of columns (50).

3. The area of the study is defined in the measurement parameters tab of the data setup section (see Chapter 3). If no area is defined, the routine uses the area of the entire grid.

4. The user checks the Risk-adjusted box under the Nnh routine. The risk variable is estimated with the parameters defined in the Risk Parameters box. These are the kernel parameters. Without going into detail, the user must define:

   A. The method of interpolation, which is the type of kernel used: normal, uniform, quartic, triangular, or negative exponential. The normal distribution is the default.

   B. The choice of bandwidth, whether a fixed or adaptive (variable) bandwidth is used. For a fixed bandwidth, the user must define the size of the interval (e.g., 0.5, miles; 2 kilometers). For an adaptive bandwidth, the user must define the minimum sample size to be included in the circle that defines the bandwidth. The default is an adaptive bandwidth with a minimum sample size of 100 incidents.

   C. The output units, which are points per unit of area: squared miles, squared nautical miles, squared feet, squared kilometers, or squared meters. The default is squared miles.

   D. Also, if an intensity or weight variable is used (e.g., the centroids of zones with population being an intensity variable), the intensity or weight box should be checked (be careful about checking both if there are both an intensity and a weight variable).

   Consult Chapter 10 for more detail about these parameters.

5. Once the baseline variable (the secondary file) is interpolated to the grid using the above parameters, it is converted into absolute densities (points per grid cell) and

re-scaled to the same sample size as the primary incident file. This has the effect of making the interpolation of the baseline variable the same sample size as the incident variable. For example, if there are 1000 incidents in the primary file, the interpolation of the secondary file will be re-scaled so that all grid cells add to 1000 points, irrespective of how many units the secondary variable actually represented. This creates a distribution for the primary file (the incidents) that is proportional to the secondary file (the baseline variable) if the primary file had the same distribution as the secondary file. It is then possible to compare the actual distribution of the incident variable with the expected distribution *if* it was similar to the baseline variable.

6. Once the risk parameters have been defined, the selection of parameters is similar to the Nnh routine with one exception.

   A. The threshold probabilities are selected with the scale bar. The probabilities are identical to those in Table 7.2.

   B. However, for each grid cell, a *unique threshold distance* is defined using formulas similar to equations 7.1 and 7.2. The difference is, however, that the formulas are applied to each grid cell with a unique distance for each grid cell (formulas 7.5-7.8):

$$Mean\ random\ distance \qquad grid\ cell\ i = d(ran)_i = 0.5\sqrt{\frac{A_i}{N_i}} \tag{7.5}$$

where A_i is the area of the grid cell and N_i is the *estimated number of points* from the kernel density interpolation. Thus, each grid cell has its own unique expected number of points, N_i, its own unique area, A_i (though, in general, all grid cells will have approximately equal areas), and, consequently, its own unique threshold distance.

$$.........\qquad for\ mean \qquad distance\ of\ grid\ cell\ i\ =$$

$$d(ran)_i \pm SE_{d(ran)_i} = 0.5\sqrt{\frac{A_i}{N_i}} \pm t\frac{0.26136}{\sqrt{\frac{N_i^2}{A_i}}} \tag{7.6}$$

where the Mean Random Distance of Grid Cell i, A_i and Ni are as defined above, t is the t-value associated with a probability level in the Student's t-distribution (defined by the scale bar).

The lower limit of this confidence interval is:

$$\text{.....} \qquad limit\ of \qquad\qquad interval\ for \qquad random\ distance \qquad grid\ cell\ i =$$

$$0.5\sqrt{\frac{A_i}{N_i}} - t\frac{0.26136}{\sqrt{\frac{N_i^2}{A_i}}} \tag{7.7}$$

and the upper limit of this confidence interval is

$$Upper\ limit\ of \qquad\qquad interval\ for \qquad random\ distance \qquad grid\ cell\ i =$$

$$0.5\sqrt{\frac{A_i}{N_i}} + t\frac{0.26136}{\sqrt{\frac{N_i^2}{A_i}}} \tag{7.8}$$

C. In addition, the user defines a minimum sample size for each cluster, as with the Nnh routine.

6. The actual incident points are then identified by the grid cell that they fall within and the unique threshold distance (and confidence interval) for that grid cell. For each pair of points that are compared for distance, there is, however, asymmetry since the threshold distance for each point may be different if they are in different grid cells. That is, the unique threshold distance for point A will not necessarily be the same as that for point B. The Rnnh routine, therefore, requires the distance between each pair of points to be the *shorter* of the two distances between the points.

7. Once pairs of points are selected, the Rnnh routine proceeds in the same way as the Nnh routine.

In other words, points are clustered together according to two criteria. First, they must be closer than a threshold distance. However, the threshold distance varies over the study area and is inversely proportional to the baseline variable. Only points that are closer together than would be expected on the basis of the baseline variable are selected for grouping. Second, clusters are required to have a minimum number of points with the minimum being defined by the user. The result is clusters that are more concentrated than would be expected, not just from chance but, from the distribution of the baseline variable. These are *high risk* clusters.

Guidelines for Selecting Parameters

The guidelines for selecting parameters in the Rnnh routine are similar to the Nnh except the user must also model the baseline variable using a kernel density interpolation. There are several guidelines that should be followed in developing the model.

Area must be defined correctly

First, it is essential that area be defined correctly for this routine to work. If the user defines the area on the measurement parameters page (see chapter 3), the Rnnh routine uses that value to calculate the area of each grid cell and, in turn, the grid-specific threshold distance. If the user does not define the area on the measurement parameters page, the routine calculates the total area from the minimum and maximum X/Y values (the bounding rectangle) and uses that value to calculate the area of each grid cell and, in turn, the grid-specific threshold distance. In either case, the routine will be able to calculate a threshold distance for each grid cell and run the routine.

However, if the area units are defined incorrectly on the measurement parameters page, then the routine will certainly calculate the grid cell-specific threshold distances wrongly. For example, if data are in feet but the area on the measurement parameters page are defined in square miles, most likely the routine will not find any points that are farther apart than any of the grid cell threshold distances since each distance will be defined in miles. In other words, it is essential that the area units be consistent with the data for the routine to properly work.

Use kernel bandwidths that produce stable estimates

Second, the bandwidth for the baseline variable must be defined in such a way as to produce a stable density estimate of the variable. Be careful about choosing a very small bandwidth. This could have the effect of creating clusters at the edges of the study area or very large clusters in low population density areas. For example, in low population density areas, there will probably be fewer persons or events than in more built-up areas. This will have the effect on the Rnnh calculation of producing a very large matching distance. Points that are quite far apart could be artificially grouped together, producing a very large cluster. Using a larger bandwidth will usually produce a more stable average.

The process is a little like tuning a shortwave radio, adjusting the dial until the signal is detected. We suggest that the user first develop a good density model for the baseline variable (see Chapter 10). The user has to develop a trade-off between identify areas of high and low population concentration to produce an estimate that is statistical reliable (stable).

One can think of two types of ‘fine tuning’ that must occur. One is the ‘background’ variation that has to be tuned (the baseline ‘at risk’ variable). This is done through the kernel density interpolation. If too narrow a bandwidth is selected, the density surface will have numerous undulations with small ‘peaks’ and ‘valleys’; this could produce unreal and unstable risk estimates. A grid cell with a very small density value could produce an extremely large threshold distance whereas a grid cell with a very low density could produce an extremely small threshold distance. Conversely, if too large a bandwidth is selected, the density surface will not differentiate very well and each grid cell will have, more or less, the same threshold distance. In this case, the Rnnh routine would yield a result not very different from the Nnh routine.

Another is the tuning of the clusters themselves through the threshold adjustment and minimum size criteria. If a large threshold probability is selected, too many incidents may be grouped; conversely, if a small threshold probability is selected, the result may be too restrictive. Similarly, if a small minimum sample size for clusters is used, there could be too many clusters whereas the opposite will happen if a large minimum sample size is chosen (i.e., zero clusters). The user must experiment with both these types of adjustment to produce a sensible cluster solution that captures the areas of high risk, but no more.

Example 2: Simulated Rnnh Clustering

To illustrate the logic of the Rnnh routine, a simulated example is presented. Two hundred points (incidents) were assigned to eight groups in the Baltimore metropolitan region (Figure 7.10). The figure shows the points in relation to year 2000 population density. Each group contained 25 individual points that were grouped exactly the same. However, three of the groups were placed in more dense areas of the region - one in central Baltimore, one in Towson to the north, and one is Reisterstown to the north east. The other five groups were placed in less populated areas. The Nnh and Rnnh routines were compared with these data. One would expect the Nnh routine to cluster the 200 points into eight groups whereas the Rnnh routine should identify only five groups in the low density areas. The reason for three of the groups not being clustered by the Rnnh is due to their higher population densities; all other things being equal, there should be more incidents in higher density areas than in lower density areas. Figures 7.11 and 7.12 show exactly this solution.

In other words, the Nnh routine clustered the points together irrespective of the distribution of the baseline population whereas the Rnnh routine clustered the points together relative to the baseline population (in this case, population). The specific parameter used were the default threshold distance (random nearest neighbor distance), a minimum of 15 points per cluster, and, for the Rnnh parameters, a normal kernel with a fixed interval of 0.5 miles.

Figure 7.10:

Incidents in Relation to Population Density

Baltimore: 1990

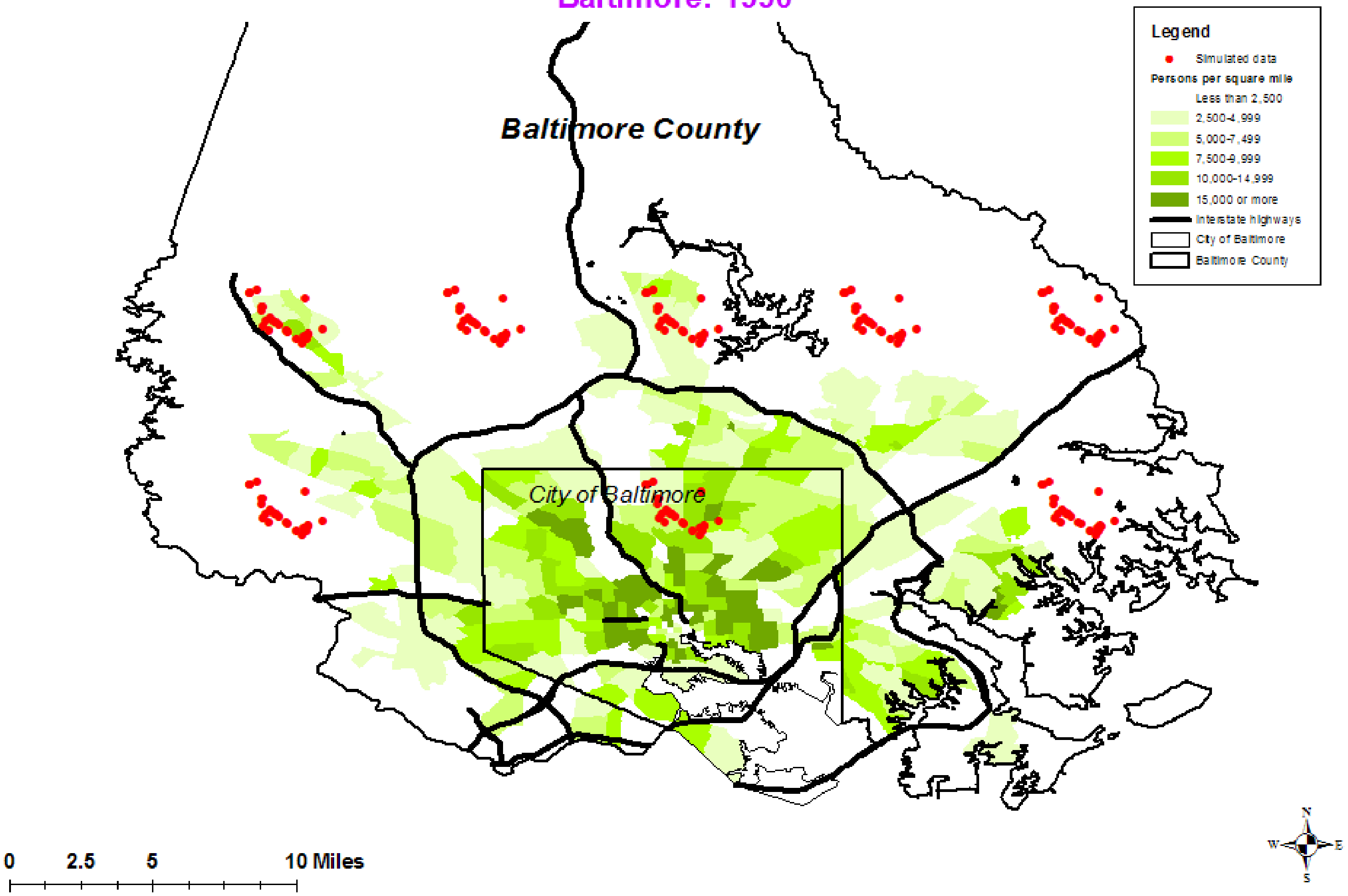


Figure 7.11:

Incidents in Relation to Population Density

Baltimore: 1990

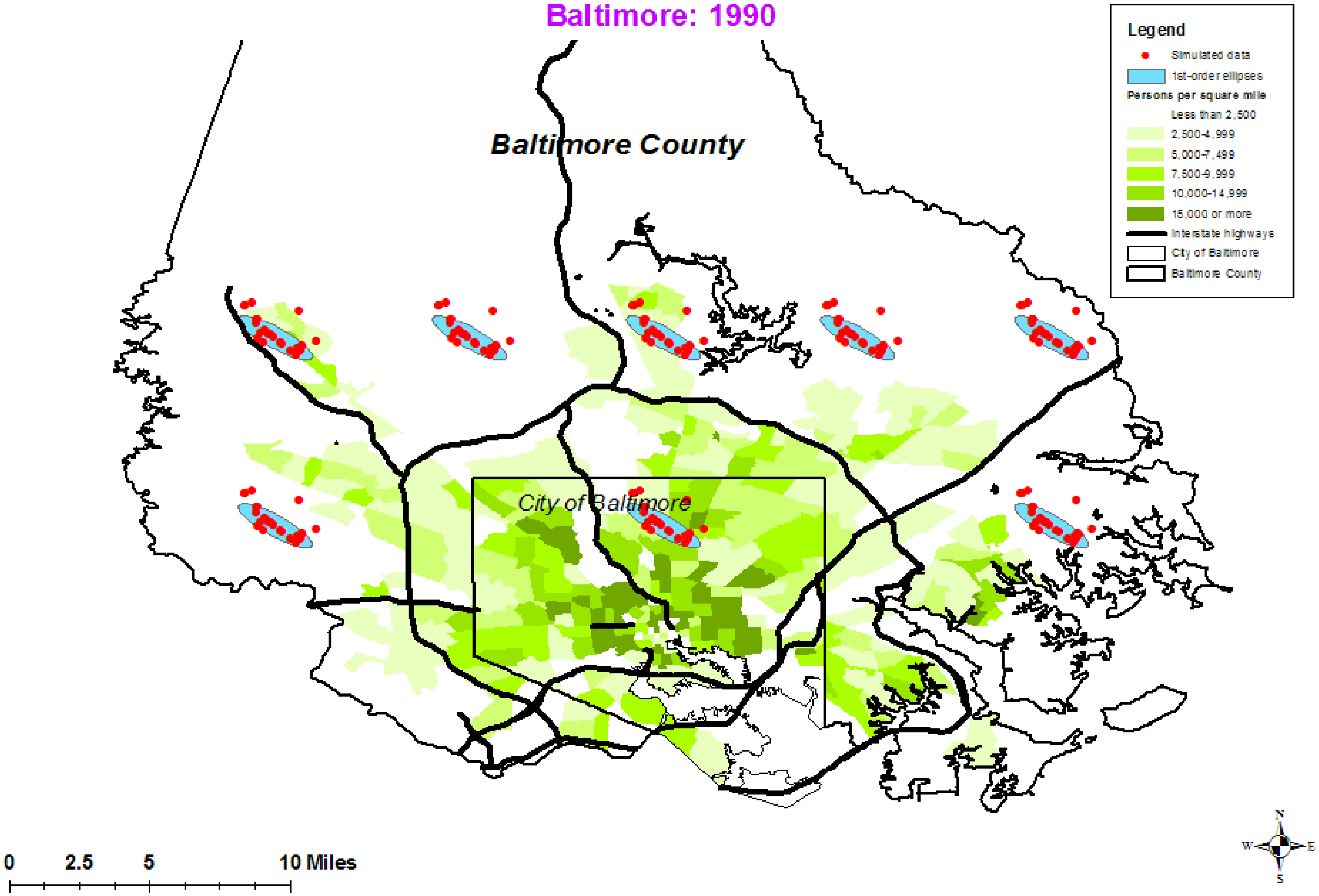
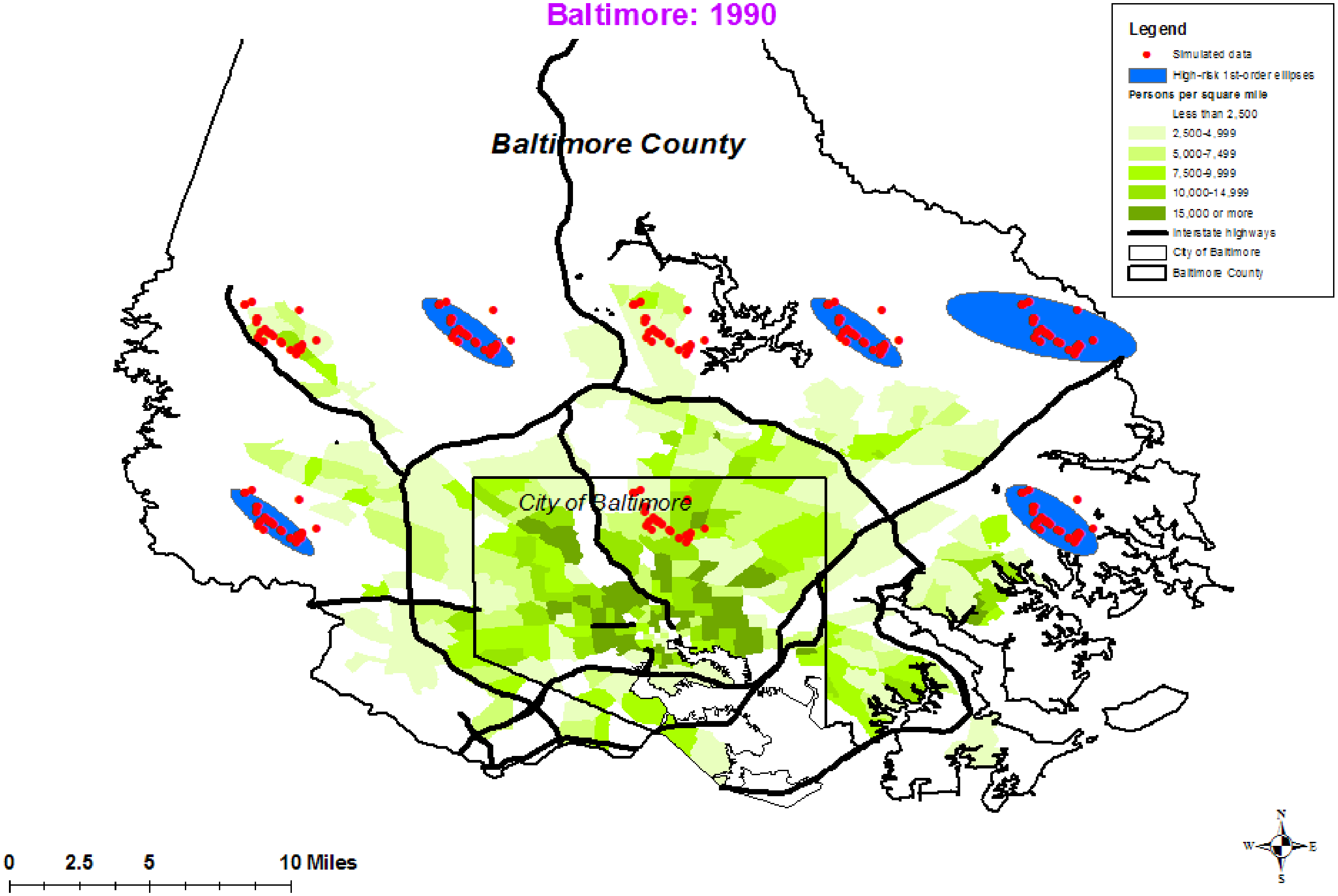


Figure 7.12:

Incidents in Relation to Population Density

Baltimore: 1990



Rnnh Output Files

The output files are similar to the Nnh routine. The Rnnh routine has three outputs. First, final seed locations of each cluster and the parameters of the selected standard deviational ellipse are calculated for each cluster. These can be output to a '.dbf' file or saved as a text ('.txt') file. Only 45 of the seed locations are displayed on the screen. The user can scroll down or across by adjusting the horizontal and vertical slider bars and clicking on the *Go* button.

Second, for each order that is calculated, *CrimeStat* calculates the mean center of the cluster. This can be saved as a '.dbf' file. Third, either standard deviational ellipses or convex hulls of the clusters can be saved in in *ArcGIS* '.shp', *MapInfo* '.mif', *Google Earth* 'kml' (if the coordinates are spherical), or various Ascii formats. Again, the convex hulls display polygons around the incidents whereas the ellipses are determined by the number of standard deviations to be calculated (see above). For small geographical area a 1X standard deviational ellipse may be appropriate since a 1.5X or 2X standard deviational ellipse can create an exaggerated view of the underlying cluster. On the other hand, for a regional view, a 1X standard deviational ellipse may not be very visible. The user has to balance the need to accurately display the cluster compared to making it easier for a viewer to understand its location.

As with the Nnh second- and higher-order clusters, these may cover incidents that were not clustered in the first-order. Thus, one has to be careful in interpreting second- and higher-order clusters. Essentially, these are abstractions made up of first-order clusters. In the routine, the first-order clusters are the primary clusters while the higher-order ones are ways to group the first-order clusters.

Naming conventions for ellipses

Because there are also orders of clusters (i.e., first-order, second-order, etc.), there is a naming convention that distinguishes the order.

For the ellipses, the convention is

Rnnh<O><*username*>

where *O* is the order number and *username* is a name provide by the user. Thus,

Rnnh1robbery

are the first-order clusters for a file called 'robbery' and

Rnnh2burglary

are the second-order clusters for a file called 'burglary'. Within files, clusters are named

Rnnh<O>Ell<N><*username*>

where *O* is the order number, *N* is the cluster number and *username* is the user-defined name of the file. Thus,

Rnnh1Ell10robbery

is the tenth cluster within the first-order clusters for the file 'robbery' while

Rnnh2Ell1burglary

is the first cluster within the second-order clusters for the file 'burglary'.

For the convex hulls, the cluster numbers are the same as the ellipses but the prefix name is output with a 'CRNNH1' prefix for the first-order clusters, a 'CRNNH2' prefix for the second-order clusters, and a 'CRNNH3' prefix for the third-order clusters. Higher-order clusters will index only the number.

Example 3: Rnnh Clustering of Vehicle Thefts

A second example is the clustering of 2003 San Antonio robberies relative to the 2000 population of census block groups. The test is for clusters of robberies that are more concentrated than would be expected on the basis of the population distribution.[5] Using the default threshold probabilities, a minimum sample size per cluster of 10, but a normal kernel function with a 0.5 mile fixed bandwidth, the Rnnh routine identified five first-order and one second-order cluster (Figure 7.13); the incidents are not shown.

Compare this distribution with the results of the Nnh on the same data, using the same parameters (Figure 7.14). The Nnh found 9 first-order clusters and one second-order cluster. To illustrate the differences in the baseline population, the ellipses of both the regular (Nnh) and risk-adjusted (Rnnh) clusters are overlaid on top of 2000 population density of census block groups. The cluster locations where there are both high volume (Nnh) and high-risk (Rnnh) involve two areas of low population density (just north of downtown) and one area of high

[5] It is not an exact risk test since we are comparing 2003 robberies with 2000 population. It is an approximate risk test.

Figure 7.13:

San Antonio Robbery Risk: 2003

Ellipses of 1st-order and 2nd-order Risk-adjusted Hot Spots

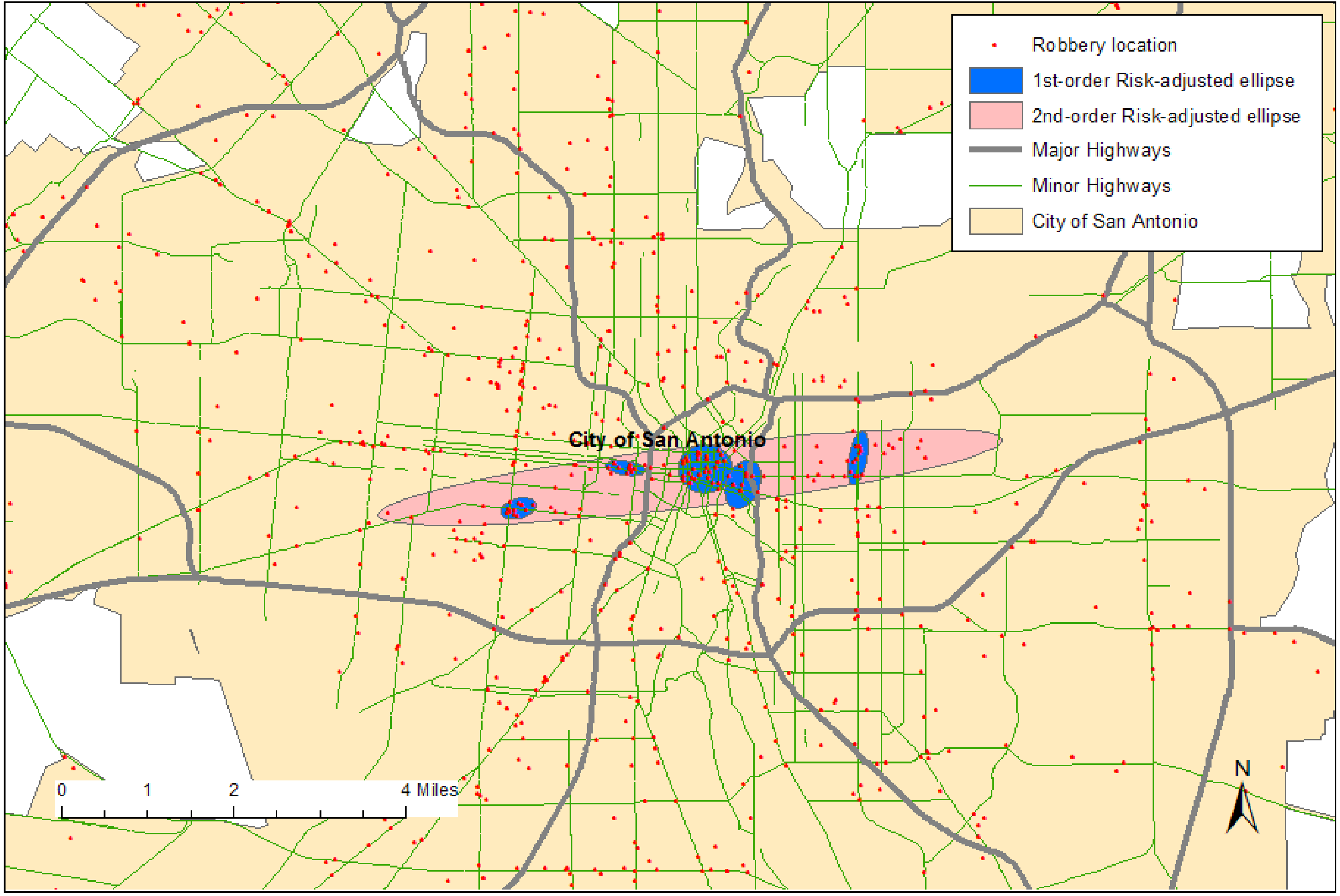
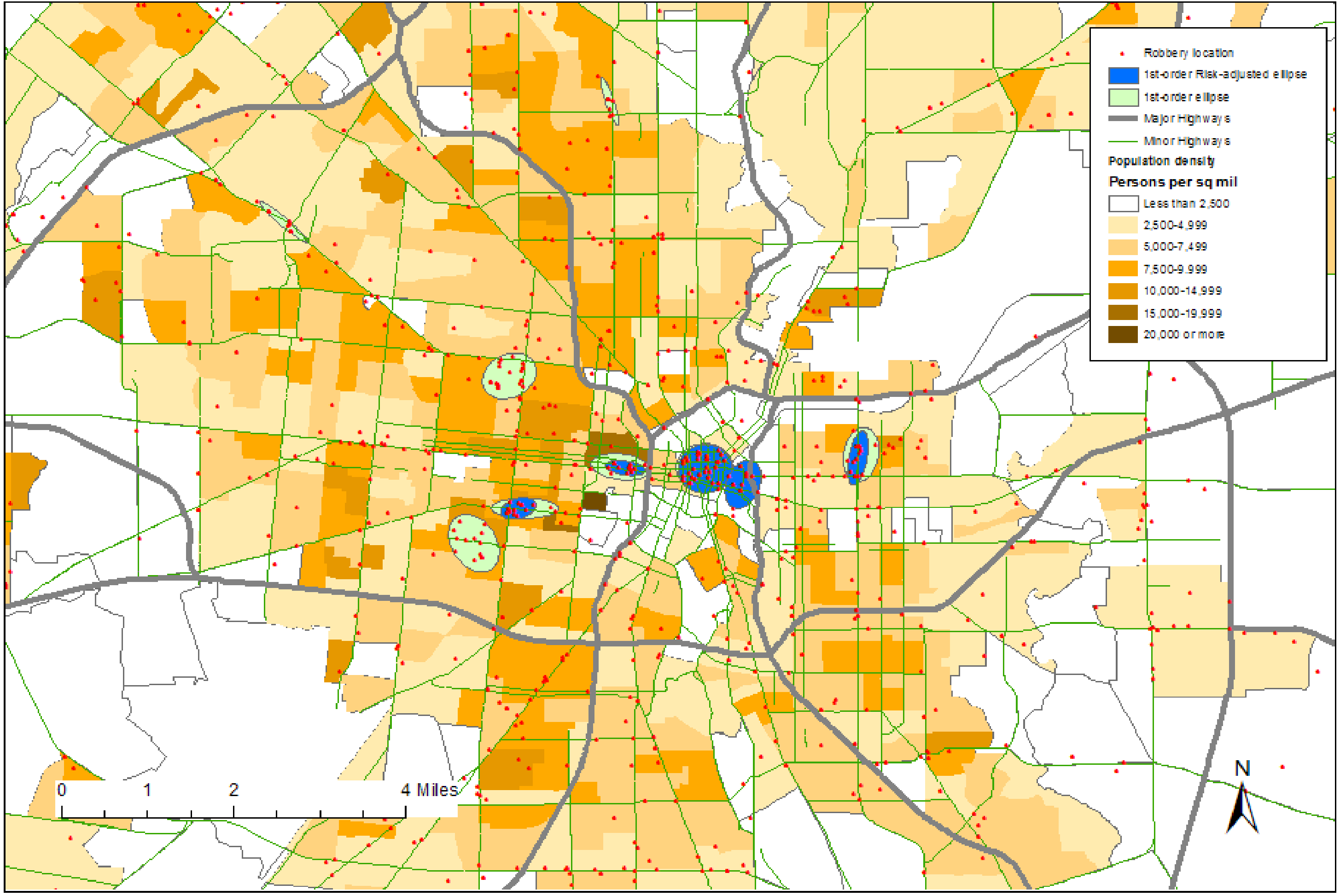


Figure 7.14:

San Antonio Robbery Risk: 2003

Comparison of Ellipses of Regular and Risk-adjusted Hot Spots



population density (just outside the downtown area); in this latter case, the number of robberies is so high that the area is both high volume and high risk. The fifth overlapping cluster is to the west of downtown and is an area of moderate population density. On the other hand, the four regular cluster locations that are only high volume (Nnh only) are in areas of low to moderate population density. In other words, the Rnnh routine identified areas of high *risk* for robberies whereas the Nnh routine identified areas of high *volume*.

Simulating Statistical Significance

Because the sampling distribution of the clustering method is not known, the Rnnh routine allows Monte Carlo simulations to approximate confidence intervals, similar to the Nnh routine (Dwass, 1957; Barnard, 1963). The output is identical to the Nnh routine. Essentially, it produces credible intervals for the number of first-order clusters, the area of clusters, the number of points in each cluster, and the density of each cluster. Second- and higher-order clusters are not simulated since their structure depends on the first-order clusters. The user can see whether the first-order cluster structure is different than that which is produced by a random distribution. See the notes above under Nnh for more details.

Uses of the Technique

The risk-adjusted nearest neighbor hierarchical clustering routine has several uses. First, like the high volume nearest neighbor hierarchical clustering (Nnh) routine, it allows a hierarchy of clusters to be identified, from first-order to second- or higher-order. As we see repeatedly with population dynamics, spatial clusters are frequently clustered together. One can think of them as small zones of concentrated events that are, in turn, close to other zones of concentrated events.

Second, unlike the Nnh, the Rnnh routine allows these clusters to be defined in terms of risk. Thus, it controls for the predominance of the *population at risk*. This is particularly important in epidemiological studies where the number of disease incidents is always related to the population at risk. The risk indicates a location where there are factors that are causing the disease to erupt. But, in crime analysis, too, analyzing incidents in relation to the number of potential victims can indicate problem neighborhoods where additional factors are triggering the outbreak (e.g., particular land uses that encourage disorder such as bars or pawn shops; poor social cohestion). Crime prevention efforts, in particular, often target neighborhoods of high risk and not just high volume of incidents. The Rnnh can be a valuable tool in the identification of such neighborhoods.

Third, the Rnnh routine goes beyond simply clustering events on the basis of proximity and frequency and applies a single variable that can account for the distribution. In other words, the baseline variable is the first step in developing a model for explaining the distribution of the incidents, in this case the baseline variable itself. In addition to focusing policing efforts on high volume or high risk neighborhoods, there needs to be an effort to build a statistical model of the phenomenon itself, both for prediction as well as for theory development.

Limitations of the Technique

However, as with all methods, there are some limitations of the technique that are partly shared with the Nnh routine. First, the method only clusters incidents (points); a weighting or intensity variable will have no effect. In Chapter 9, we will introduce a zonal variant of the Rnnh that allows a risk measure to be applied to zonal data. But, the Rnnh by itself is only applicable to individual point locations.

Second, the size of the grouping area is dependent on the sample size if the confidence interval around the mean random distance is used as the threshold distance criteria. However, since the threshold distance is adjusted dynamically, this has less effect than in the Nnh since it is now a relative comparison rather than an absolute distance.

Third, there is arbitrariness in the technique due to the minimum points rule. Different users could define the minimum differently, which could lead to different conclusions about the location of high risk clusters. Finally, unique to the Rnnh, the method requires both an incident file (the primary file) and a baseline file (the secondary file.

Nevertheless, the Rnnh routine is a useful technique for identifying clusters that are more concentrated than would be expected on the basis of the population distribution.

References

Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical Analysis*. 27, No. 2 (April), 93-115.

Bailey, T. C. & Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. Longman Scientific & Technical: Burnt Mill, Essex, England.

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.

Beale, E. M. L. (1969). *Cluster Analysis.* Scientific Control Systems: London.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press: New York.

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 John E. Eck & David Weisburd (eds.), *Crime and Place*. Crime Prevention Studies, Volume 4. Criminal Justice Press: Monsey, NY. 147-185.

Braga, A. & Weisburd, D. (2010). Policing Problem Places: Crime Hot Spots and Effective Prevention. Oxford: Oxford University Press.

Can, A. & Megbolugbe, I. (1996). The geography of underserved mortgage markets. Paper presented at the American Real Estate and Urban Economics Association meeting. May.

Carmichael, J. W., George, L.A. & Julius, R.S. (1968). Finding natural clusters. *Systematic Zoology*, 17, 144-150.

References (continued)

Cattell, R. B. & Coulter, M.A. (1966). Principles of behavioural taxonomy and the mathematical basis of the taxonome computer program. *British Journal of Mathematical and Statistical Psychology*, 19, 237-269.

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

Cole, A. J. & Wishart, D. (1970). An improved algorithm for the Jardine-Sibson method of generating overlapping clusters. *Comparative Journal*, 13, 156-163.

D'andrade, R. (1978). U-Statistic Hierarchical Clustering *Psychometrika*, 4,58-67.

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

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.

Getis, A. & Ord, J. K. (1996). Local spatial statistics: an overview. In Longley, P. & Batty, M. (eds), *Spatial Analysis: Modelling in a GIS Environment*. GeoInformation International: Cambridge, England, 261-277.

Gitman, I. & Levine, M. D. (1970). An algorithm for detecting uniomodal fuzzy sets and its application as a clustering technique. *IEE Transactions on Computers*, 19, 583-593.

Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley & Sons, Inc.: New York.

Jardine, N. & Sibson, R. (1968). The construction of hierarchic and non-hierarchic classifications. *Comparative Journal*, 11, 117-184.

Jefferis, E. (1998). A multi-method exploration of crime hot spots. Crime Mapping Research Center, National Institute of Justice: Washington, DC.

Johnson, S. C. (1967), Hierarchical Clustering Schemes *Psychometrika*, 2,241-254.

References (continued)

Jones, K. S. & Jackson, D. M. (1967). Current approaches to classification and clump finding at the Cambridge Language Research Unit. *Comparative Journal*, 10, 29-37.

King, B. F. (1967). Step wise clustering procedures. *Journal of the American Statistical Association*. 62, 86-101.

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

Kulldorff, M. & Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference, *Statistics in Medicine*, 14, 799-810.

Levine, N. (2008). “The ‘hottest’ part of a crime hotspot: Comments on “The utility of hotspot mapping for predicting spatial patterns of crime” by Chainey, S. Thompson, L. & Uhlig, S.”. *Security Journal*, 21, 295-302.

Levine, N., Wachs, M. & Shirazi, E. (1986). "Crime at Bus Stops: A Study of Environmental Factors". *Journal of Architectural and Planning Research*. 3 (4), 339-361.

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

McBratney, A. B. & deBruijter, J. J. (1992). A continuum approach to soil classification by modified fuzzy k-means with extragrades, *Journal of Soil Science*, 43, 159-175.

McQuitty, L. L. (1960). Hierarchical syndrome analysis. *Educational and Psychological Measurement*, 20, 293-304.

Maltz, M. D., Gordon, A. C., & Friedman, W. (1990). *Mapping Crime in Its Community Setting: Event Geography Analysis*. Springer-Verlag: New York.

Needham, R. M. (1967). Automatic classification in linguistics. *The Statistician*, 17, 45-54.

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. (1987). A Mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, 1, 335-358.

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.

Sherman, L. W.., Gartin, P. R. & Buerger, M. E. (1989). Hot spots of predatory crime: routine activities and the criminology of place. *Criminology*, 27(1), 27-56.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall: London.

Sneath, P. H. A. (1957). The application of computers to taxonomy. *Journal of General Microbiology*, 17, 201-226.

Sokal, R. R. & Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*. W. H. Freeman & Co.: San Francisco.

Sokal, R. R. & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, 1409-1438.

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

Thorndike, R. L. (1953). Who belongs in a family?. *Psychometrika*, 18, 267-276.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*. 58, 236-244.

Weisburd, D. & Green, L. (1995). Policing drug hot spots: the Jersey City drug market analysis experiment. *Justice Quarterly*. 12 (4), 711-735.

Weisburd, D., Maher, L.& Sherman, L. (1992). Contrasting crime general and crime specific theory: the case of hot-spots of crime. *Advances in Criminological Theory*, 4, 45-70.

Weishart, D. (1969). Mode analysis. In Cole, A. J. (ed), *Numerical Taxonomy*, Academic Press: New York.

References (continued)

Xie, X. L. & Beni, G. (1991). A validity measure for fuzzy clustering. IEEE Trans. Pattern Analysis Machine Intell., 13, 841-847.

Endnotes

i. The particular steps are as follows:

1. All distances between pairs of points are calculated, using either direct or indirect distance as defined on the measurements parameters page. The matrix is assumed to be symmetrical, that is the distance between A and B is assumed to be identical to the distance between B and A.

2. The mean expected random distance is calculated using formula 6.2 and the threshold distance (the confidence interval for the corresponding t) is calculated using formulas 7.2 and 7.3 depending on whether it is a lower or upper confidence interval. The particular interval is selected by the user on the slide bar.

3. All pairs that are separated by a distance smaller than the threshold distance are selected for clustering and placed in a *reduced matrix*. Any incident point that does not have another point within the threshold distance is not clustered.

4. In the reduced matrix, for each point the number of other points that are within the threshold distance are counted and are sorted in descending order.

5. The incident point with the largest number of below threshold distances is selected for the initial seed of the first cluster.

6. All other points that are within the threshold distance of the initial seed point are selected for the initial cluster 1 and temporarily removed from the reduced matrix.

7. The process is repeated for the remaining points in the reduced matrix (i.e., an initial seed is selected, all points within the threshold distance of that seed are clustered, and all the points are temporarily removed).

8. For each of the initial clusters that were identified, the center of minimum distance (CMD) is calculated to identify the cluster center.

9. The clustering process is repeated but using the CMD for each cluster to define each cluster. This process continues until no points change their cluster membership.

Endnotes (continued)

10. Once all the points in the reduced matrix have been initially clustered, the total number of points within each initial cluster is counted. If the number is equal to or greater than the minimum specified, then the cluster is kept. If the number is less than the minimum specified, then the cluster is dropped.

11. The final clusters are sorted in descending order of the number of points and the mean center of each is calculated to identify the cluster center.

12. The second- and higher-order clusters use the CMD of the first-order clusters as 'points' and follow the same algorithm.

ii. The particular steps are as follows:

1. Using the same p-values selected in the first-order, the mean random expected distance is calculated. However, the sample size is the number of first-order clusters identified, not the original number of points. Thus, the threshold distance is calculated by

$$Second - order\ threshold\ distance = \ d_{NN2(ran)} + t * SE_{d1(ran)} \qquad (7.8)$$

where $d_{NN2(ran)}$ is random nearest neighbor distance among the first-order clusters (i.e., with *M* first-order clusters rather than *N* points) and $SE_{d1(ran)}$ is the standard error of the random nearest neighbor distance among the first-order clusters. Thus, there is a different threshold distance for the second-order clustering. The t-value specified in the first-order clustering is maintained for second- and higher-order clustering.

2. All distances between first-order cluster centers are calculated and only those that are smaller than the second-order threshold distance are selected for second-order clustering.

3. If there are no distances between first-order cluster centers that are smaller than the second-order threshold distance, then the clustering process ends.

Endnotes (continued)

4. If there are distances between first-order cluster centers that are smaller than the second-order threshold distance, then the steps specified in endnote i above are repeated to produce second-order clusters. A minimum of four first-order clusters is required to allow a second-order cluster and four previous-order clusters to allow a higher-order cluster.

5. If there are second-order clusters, then this process is repeated to either extract third-order clusters or to end the clustering process if no distances between second-order cluster centers are smaller than the (new) third-order threshold distance or if there are fewer than four new seeds in the cluster.

6. The process is repeated until no further clustering can be conducted, either all sub-clusters converge into a single cluster or the threshold distance criteria fails or there are fewer than four seeds in the higher-order cluster.

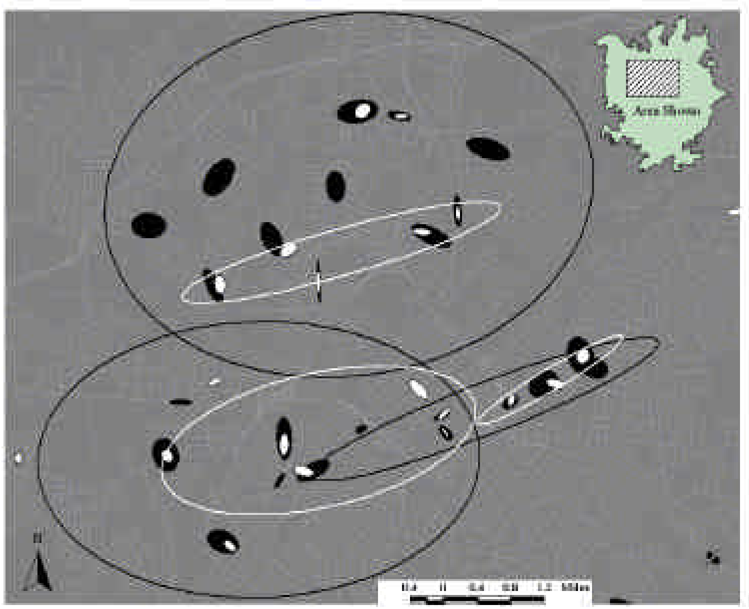
Attachments

Visualizing Change in Drug Arrest Hot Spots Using Nearest Neighbor Hierarchical Clustering: Charlotte, N.C. 1997 – 98

James L. LeBeau
Administration of Justice
Southern Illinois University at Carbondale

Stephen Schnebly
Criminology & Criminal Justice
University of Missouri – St Louis

The *CrimeStat* Nearest Neighbor Hierarchical clustering routine and GIS were used for defining, comparing, analyzing, and visualizing changes in drug arrest clusters between 1997 and 1998. Using a minimum cluster size of 25 arrests some of the emerging patterns or relationships include: 1) the overlapping of secondary clusters, but those emerging during 1998 were much larger, especially in the north because of new primary clusters; 2) many primary clusters during 1997 remaining static or increasing in area during 1998; and 3) the disappearing of some 1997 primary clusters during 1998, with new clusters emerging close by implying displacement.



	Clusters Primary	Secondary	*Total Arrests*
1997	○ N = 30	○ N = 4	4766
1998	● N = 29	○ N = 3	4802

Minimum Cluster Size 25

Source:CMPD

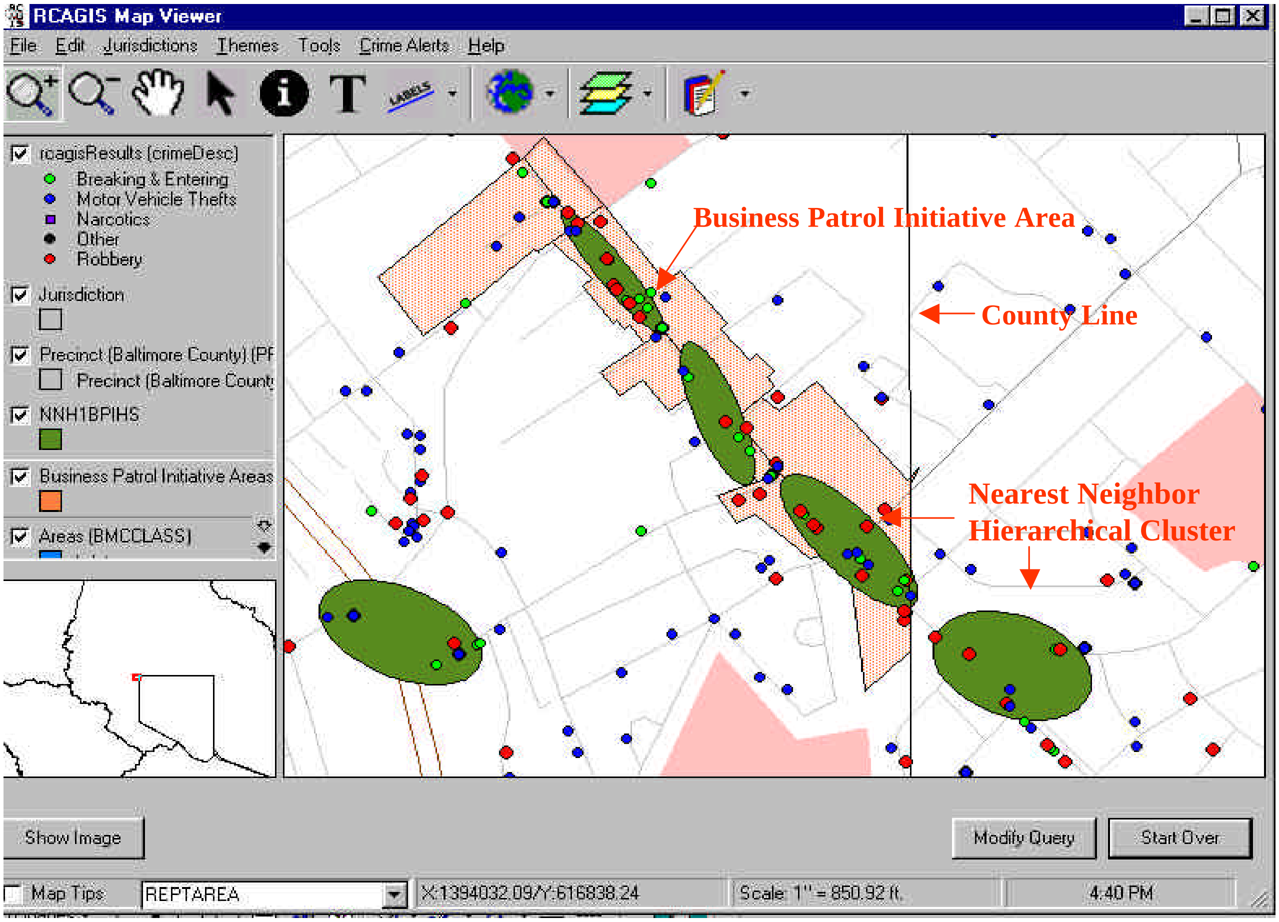
j.l.l.01

Using Nearest Neighbor Hierarchical Clustering to Identify High Crime Areas Along Commercial Corridors

Philip R. Canter
Baltimore County Police Department
Towson, Maryland

Robberies in Baltimore County had increased by 45% between 1990 and 199, and by 1997, were the highest on record. In 1997, 73% of all reported robberies in Baltimore County were occurring in commercial areas. The department wanted to target commercial districts with intensive patrol and outreach programs. These high crime commercial districts were identified as Business Patrol Initiative (BPI) areas. A total of 40 police officers working two 8-hour shifts were assigned to BPI areas. Robberies in the BPI areas declined by 26.7% during the first year of the program and another 13.8% one year following the BPI program.

Police analysts used *CrimeStat*'s Nearest Neighbor Hierarchical clustering (Nnh) method to identify high crime areas along commercial corridors. The Nnh routine was very effective in identifying commercial areas having the highest concentration of crime. The clustering also demonstrated that commercial crime was not restricted to county borders; rather, crime crossed municipal boundaries into neighboring jurisdictions. A neighboring jurisdiction was shown the crime cluster map, leading to their decision to implement a similar BPI program.

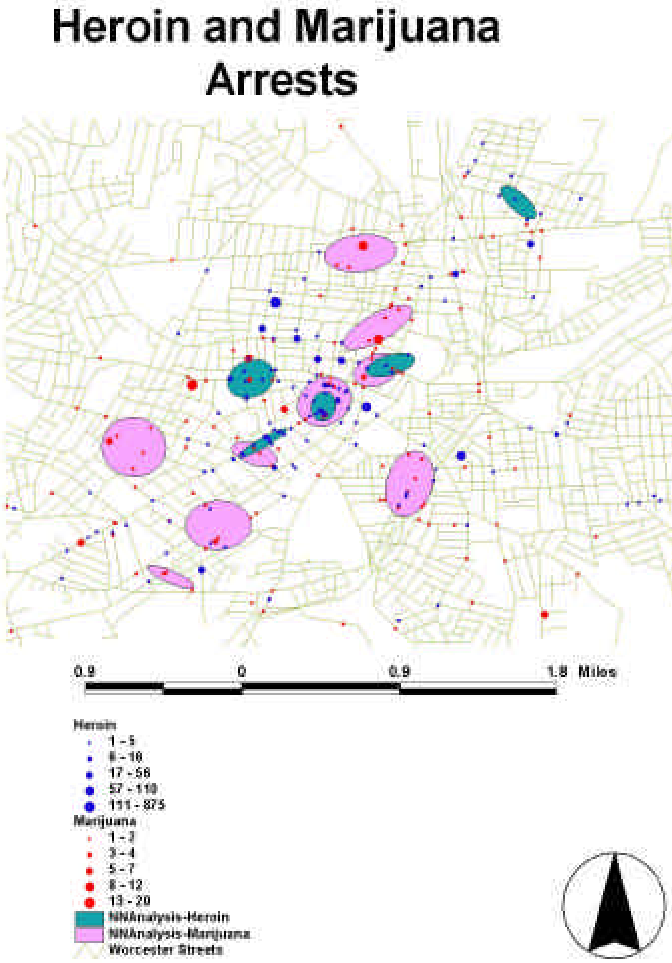


Arrest Locations as a Means for Directing Resources

Daniel Bibel
Massachusetts State Police
Crime Reporting Unit
Framingham, Massachusetts

The Massachusetts State Police is collecting incident addresses as part of its state-level implementation of the FBI's National Incident Based Reporting System (NIBRS). They intend to develop a regional and statewide crime mapping and analysis program. As an example of the type of analysis that can be done with the enhanced NIBRS database, the State Police's Crime Reporting Unit analyzed year 2000 drug arrests for one city in the Commonwealth, focusing on arrests for possession of heroin and marijuana. The arrest locations were plotted, with the size of points proportionate to the amount of drugs seized. A nearest neighbor clustering analysis was done of the data. It indicates that, while there is some small amount of overlap, the arrest locations for the two drug types are generally different.

This type of analysis can be very useful for smaller police agencies that do not have the resources to conduct their own analysis of crime data. It may also prove useful for crime problems with cross-jurisdictional boundaries.

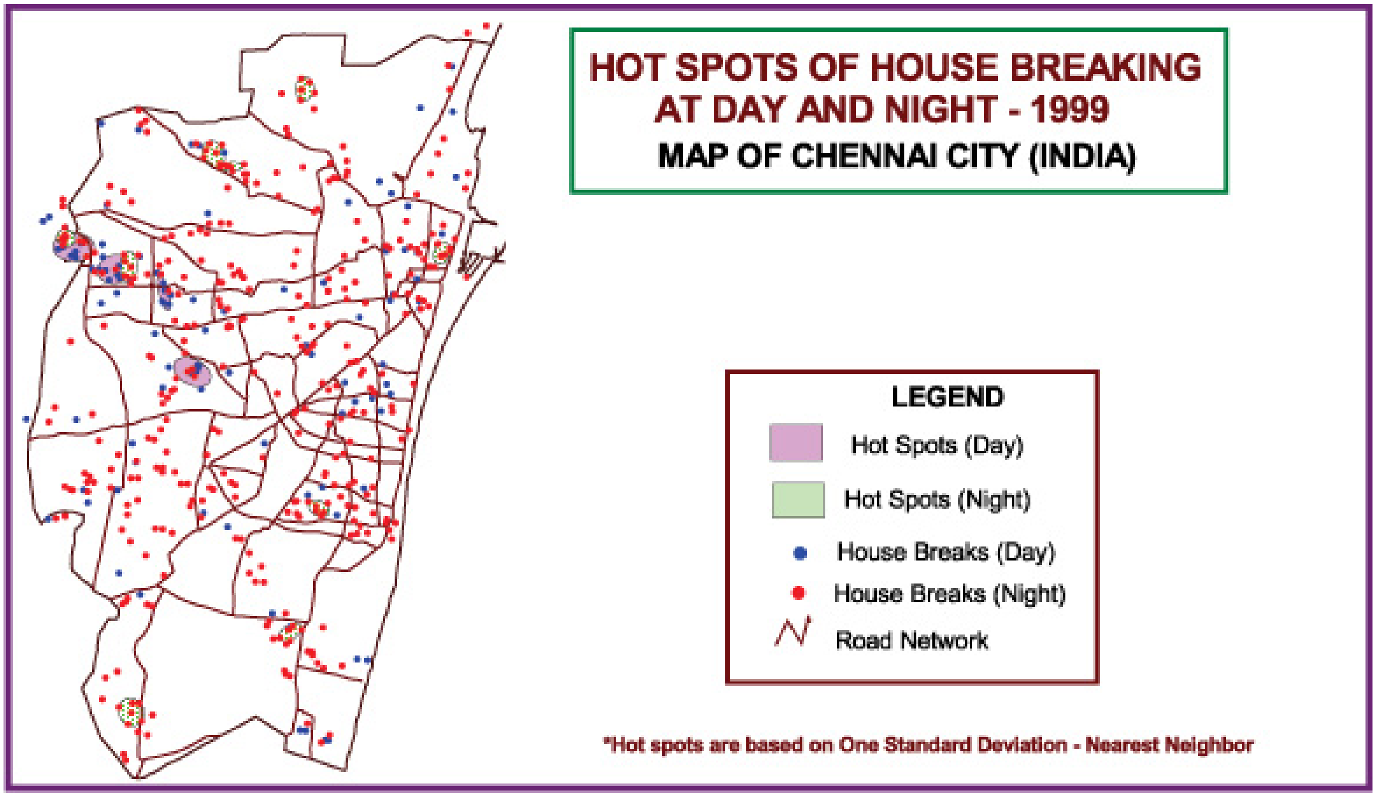


Use of *CrimeStat* in Crime Mapping in India: An Application for Chennai City Policing

Jaishankar Karuppannan
Department of Criminology & Criminal Justice
Manonmaniam Sundaranar University
Tamil Nadu, India

The present study was done as an implementation of GIS technology in Chennai (Madras), India. In the present study hotspot analysis was done with the help of *CrimeStat*. We converted the output to *Arcview* shape files.

When hotspot analysis examined changes over a period of time, the change seemed to be significant. There exists not only a change in the location of the hotspots, but also in their areal extent. The numbers of hotspots also differ over time. The map shows hotspots for residential burglary for both day and night. The hot spots for daytime house break-ins are confined to a smaller area in the west of the city, whereas the hot spots for nighttime residential break-ins are seen in all parts of the city. In particular, the Posh area of Anna Nagar is more prone to daytime burglaries. In this area, a higher proportion of couples work, which appears to make the homes in this neighborhood more open for burglaries.


HOT SPOTS OF HOUSE BREAKING
AT DAY AND NIGHT - 1999
MAP OF CHENNAI CITY (INDIA)
LEGEND
Hot Spots (Day)
Hot Spots (Night)
House Breaks (Day)
House Breaks (Night)
Road Network
*Hot spots are based on One Standard Deviation - Nearest Neighbor

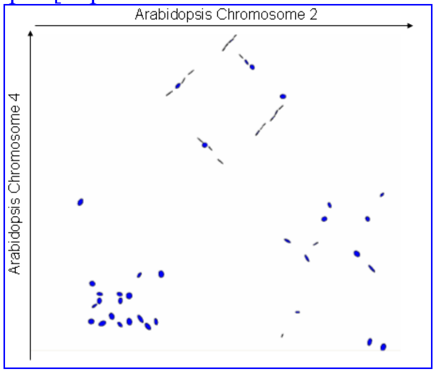
Identifying Duplications in Genomic Data
Using the *CrimeStat* Nearest Neighbor Hierarchical Spatial Clustering Routine

Nathalie Pavy and Jean Bousquet
Université Laval,G1K 7P4 Québec, QC, Canada, nathaliepavy@yahoo.fr

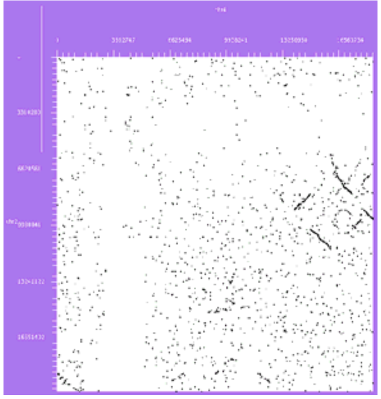
Sequencing projects provide the foundation for studying the organization of whole genomes. Comparisons of genomic sequences from related species provide a new insight into genome evolution for instance by showing locally conserved chromosomic segments. Detecting such conservation is far from trivial. Indeed, chromosome rearrangements, duplications and gene losses may hide traces of ancestry. The Nearest Neighbor Hierarchical Clustering routine (NNH) was applied to analyze regions duplicated between *Arabidopsis* chromosomes 2 and 4. These are well known for sharing similar series of genes derived from segmental duplication. Based on sequence similarities, each gene located on chromosome 2 was associated to one or several similar genes located on chromosome 4. Coordinates used as input for the NNH routine were the gene ranks along the chromosomes. A total of 53 clusters made of at least 6 similar genes were recovered. The significance of this finding was assessed with 1000 Monte Carlo simulations; only three clusters would be expected by chance alone (*P*>0.01). The gene clusters identified with the NNH approach were consistent with known duplicated chromosomic regions. The clusters found by using the NNH approach were vizualised with the GIS software CartoMap^{TM}. This graphical representation highlights in a visually comprehensive way the patterns of duplicated regions. The shape of the clusters and the relative positions of these reflect various evolutionary events that led to the structure of the present genome, as shown below (top-left): linear patterns indicate large segmental duplications with conserved gene order with or without inversion, and large dots indicate more condensed gene clusters.

Clusters of at least six genes found on *Arabidopsis* chromosome 2 and duplicated on chromosome 4.

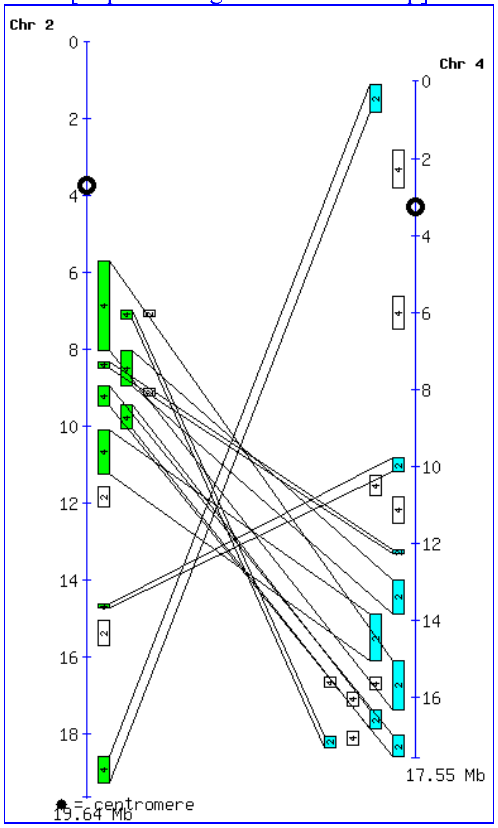
Clusters found with the NNH routine and visualized with CartoMap^{TM} [http://www.cartoworld.com/cartomap.htm]



Dot-Plot obtained by using DAGchainer [http://dagchainer.sourceforge.net/] [Haas et al., 2004, Bioinformatics 20:3643-3646]



Clusters extracted from the Paralogon database [http://wolfe.gen.tcd.ie/athal/dup]

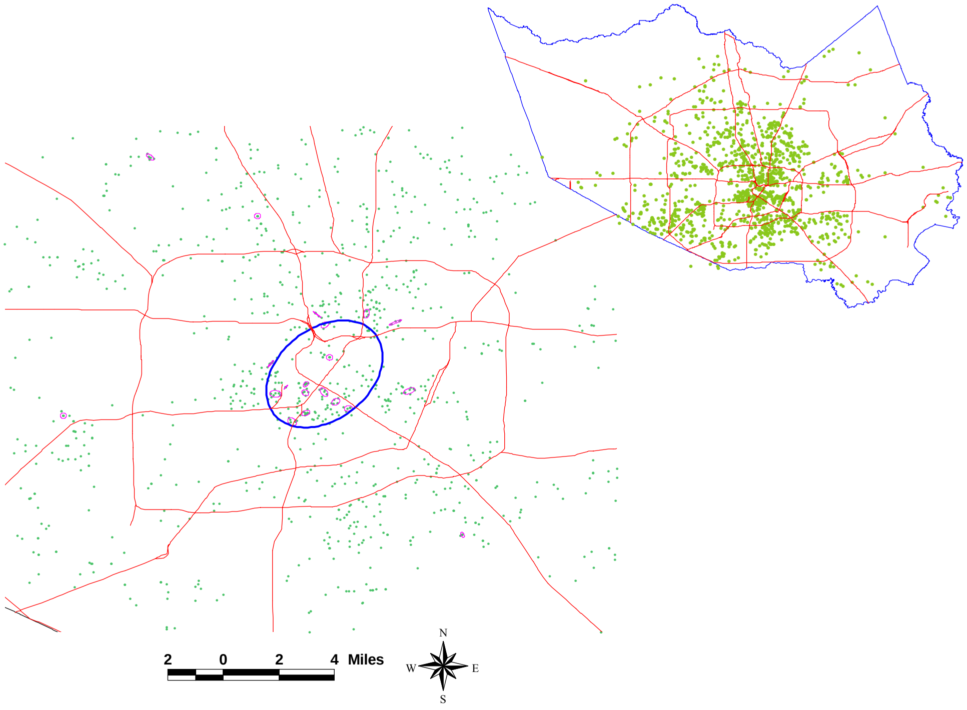


Risk Adjusted Nearest Neighbor Hierarchical Clustering of Tuberculosis Cases in Harris County, Texas: 1995 to 1998

Matthew L. Stone, MPH
Epidemiology and Program Evaluation Unit
University of California at San Francisco/California Department of Health Services
Sacramento, CA

Data was collected from an ongoing, population-based, active surveillance and molecular epidemiology study of tuberculosis cases reported to the City of Houston Tuberculosis Control Office from October 1995 to September 1998. During this time, 1774 cases of tuberculosis were reported and 1480 of those who participated in this study were successfully geocoded.

CrimeStat was used to make an initial survey of potential hot spot areas of tuberculosis cases where more focused TB control efforts could be implemented. Given a .05 level of significance for grouping a pair of points by chance and a minimum of five cases per cluster, 24 first-order clusters and one second-order cluster were detected after adjusting for the underlying population. Most first-order clusters were detected in the center of Harris County, including the metropolitan downtown area. By adjusting for the underlying population, the clusters identify areas with higher than average TB incidence. Some of these clusters are homeless shelters as many homeless persons are particularly prone to TB.



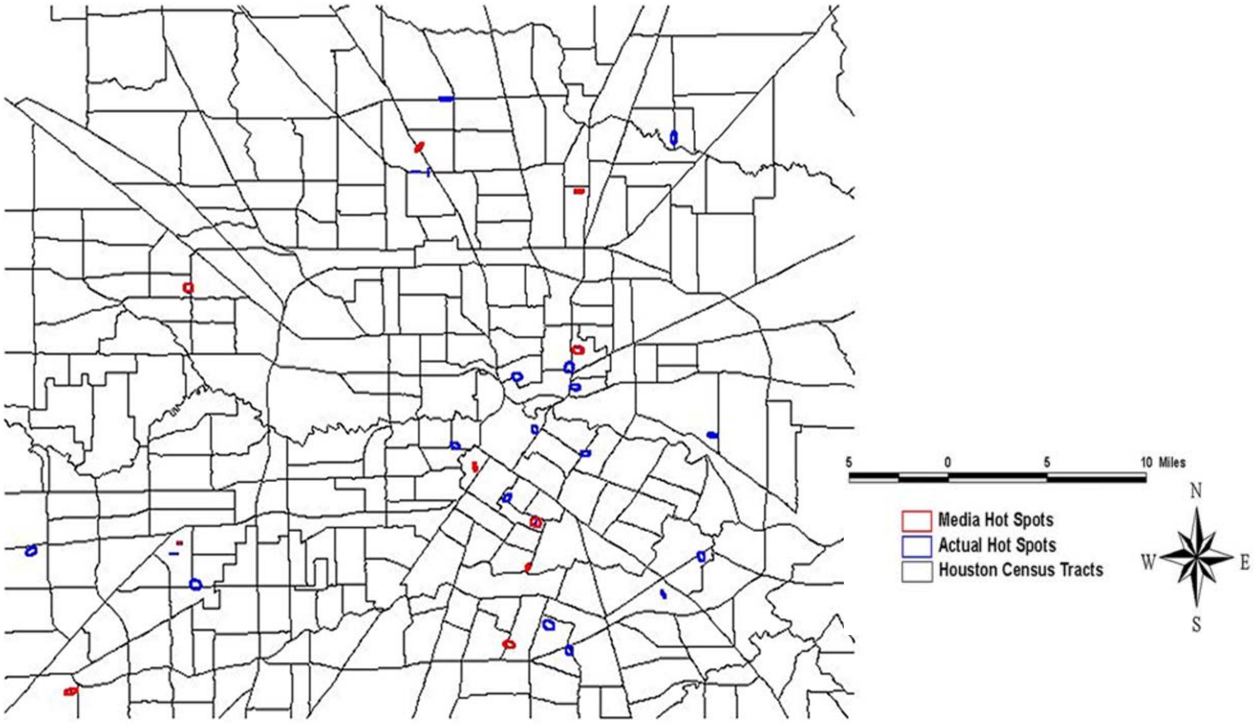
Using Risk Adjusted Nearest Neighbor Hierarchical Clustering to Compare Actual and Media Hotspots of Homicide

Derek J. Paulsen
Department of Criminal Justice and Police Studies
Eastern Kentucky University

Crimestat offers an excellent method for determining risk adjusted hot spots of crime incidents within a jurisdiction. Risk-adjusted nearest neighbor hierarchical spatial clustering (Rnnh) is a spatial clustering routine that groups points together based on both proximity to other points and the distribution of a baseline variable. In this example two different Rnnh analyses were conducted and compared for homicides in Houston, Texas. The first involves homicide incident locations adjusted for the population of each census tract, while the second involves incidents that were covered in the newspaper adjusted for the homicide rate of each census tract. The purpose of this analysis is to determine if there are differences in the spatial clustering of actual homicide incidents and those that are covered in the newspaper.

The preferences for the analysis were the same for both Rnnh analyses. For the primary file (homicide incidents & incidents covered in the newspaper) the pair probability search radius was set at .01, with a minimum of 10 points per cluster. For the secondary file (population & homicide rate), a quartic kernel density interpolation was used with an adaptive bandwidth and a minimum sample size of 100. Importantly, the analysis showed that media hot spots and actual hot spots do not coincide. Media coverage showed homicides to be concentrated in different areas than they are actually concentrated.

Actual Homicide Hot Spots vs. Media Coverage Hot Spots in Houston Texas



Seizures of Tiger Parts and Derivatives in India during 2000 – 2012

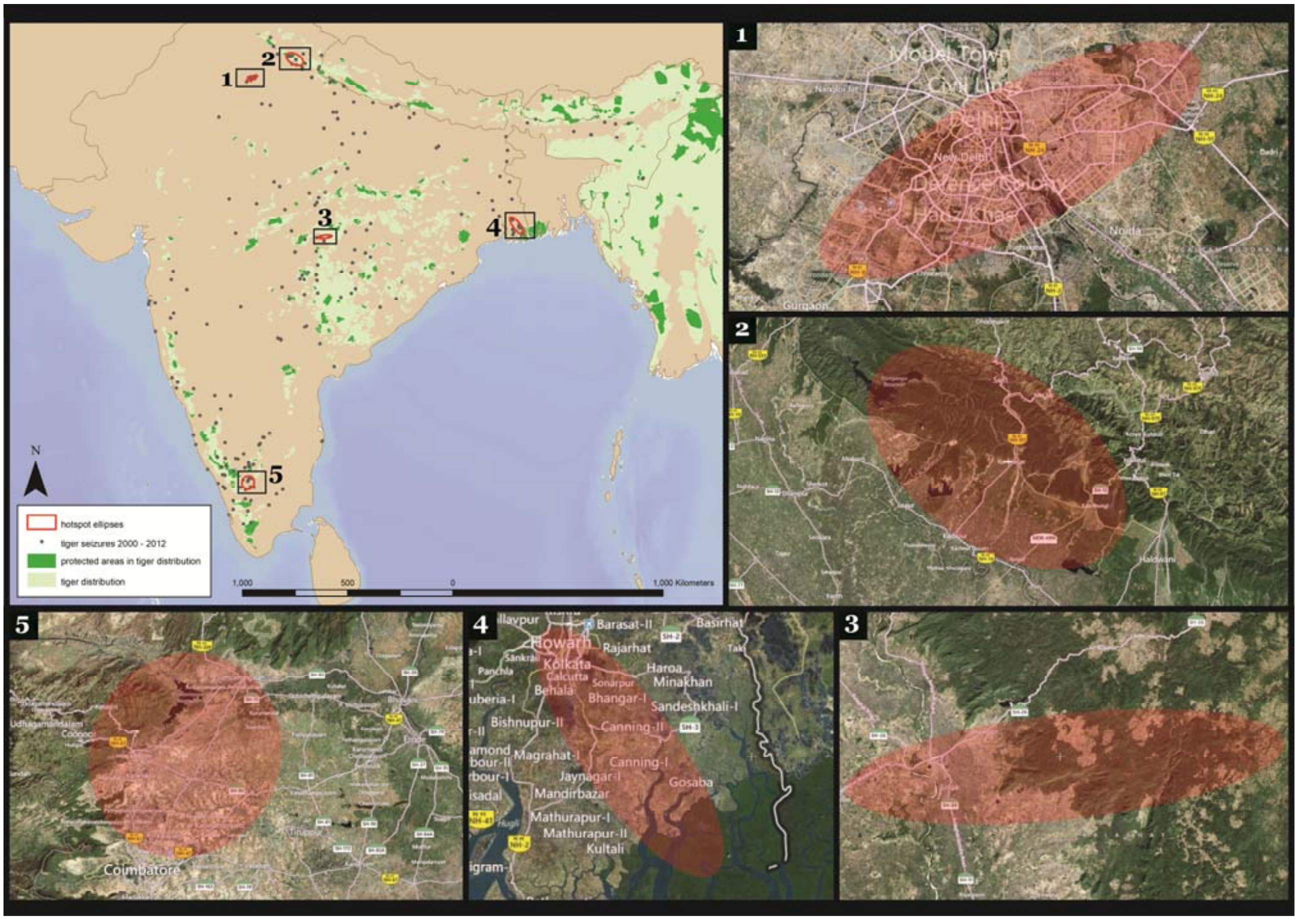
Sarah Stoner
TRAFFIC International
Kuala Lumpur, Malaysia

India is home to over half of the world's wild Tiger population and as a consequence records the greatest number of seizures globally. Since 2000, 336 seizures have been reported equating to an estimated 529 dead Tigers. Hotspot analysis of Tiger seizures has never been conducted in India and determining where clusters of activity exist is problematic.

Using the Crimestat nearest neighbour hierarchical clustering routine (*Nnh*), five significant clusters of seizures were identified. ArcGIS was used to map both the seizures and one standard deviational ellipses and were overlaid on tiger distribution and Protected Area* layers. Four of the ellipses were related to towns or cities which are also within close proximity of a Tiger reserve. Furthermore, transboundary trading of Tigers is prevalent but often securing agreement to combat trade at this level is challenging. Two clusters were also close to the borders of Nepal and Bangladesh. These findings will create leverage for law enforcement agencies to focus on the areas where seizures are most likely to occur to affect the greatest impact and will help create collaborative partnerships with neighbouring countries to tackle the issue at a regional level.

***A clearly defined geographical space, recognised, dedicated and managed, through legal or other effective means, to achieve the long-term conservation of nature with associated ecosystem services and cultural values. **SOURCE**: World Database Protected Area*

Figure 1: Tiger seizures in India (2000-2012, n=336


1
2
3
4
5
N
hotspot ellipses
tiger seizures 2000 - 2012
protected areas in tiger distribution
tiger distribution
1,000
500
0
1,000 Kilometers
Model Town
Civil Lines
Delhi
New Delhi
Defence Colony
Hauz Khas
Noida
Gurgaon
Haldwani
Udhagamandalam
Coonoor
Coimbatore
Barasat-II
Basirhat
Howrah
Rajarhat
Kolkata
Calcutta
Haroa
Minakhan
Behala
Sonarpur
Bhangar-I
Sandeshkhali-I
Bishnupur-II
Canning-II
Magrahat-I
Canning-I
Jaynagar-I
Gosaba
Mandirbazar
Mathurapur-I
Mathurapur-II
Kultali