Document Title:	CrimeStat II: Spatial Description, Part II
Author(s):	Ned Levine and Associates
Document No.:	195773
Date Received:	August 13, 2002
Award Number:	99-IJ-CX-0044

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

195773
Pt II

CrimeStat II

Part II: Spatial Description

Chapter 4
Spatial Distribution

In this chapter, the spatial distribution of crime incidents will be discussed. The statistics that are used in describing the spatial distribution of crime incidents will be explained and will be illustrated with examples from *CrimeStat®*. For the examples, crime incident data from Baltimore County and Baltimore City will be used. Figure 4.1 shows the user interface for the spatial distribution statistics in *CrimeStat*. For each of these, the statistics will first be presented followed by examples of their use in crime analysis.

Centrographic Statistics

The most basic type of descriptors for the spatial distribution of crime incidents are *centrographic statistics*. These are indices which estimate basic parameters about the distribution (Lefever, 1926; Furfey, 1927; Bachi, 1957; Neft, 1962, Hultquist, Brown and Holmes, 1971; Ebdon, 1988). They include:

1. Mean center
2. Median center
3. Center of minimum distance
4. Standard deviation of X and Y coordinates
5. Standard distance deviation
6. Standard deviational ellipse

They are called centrographic in that they are two dimensional correlates to the basic statistical moments of a single-variable distribution - mean, standard deviation, skewness, and kurtosis (see Bachi, 1957). They have been applied to crime analysis by Stephenson (1980) and, more recently, by Langworthy and Jefferis (1998).

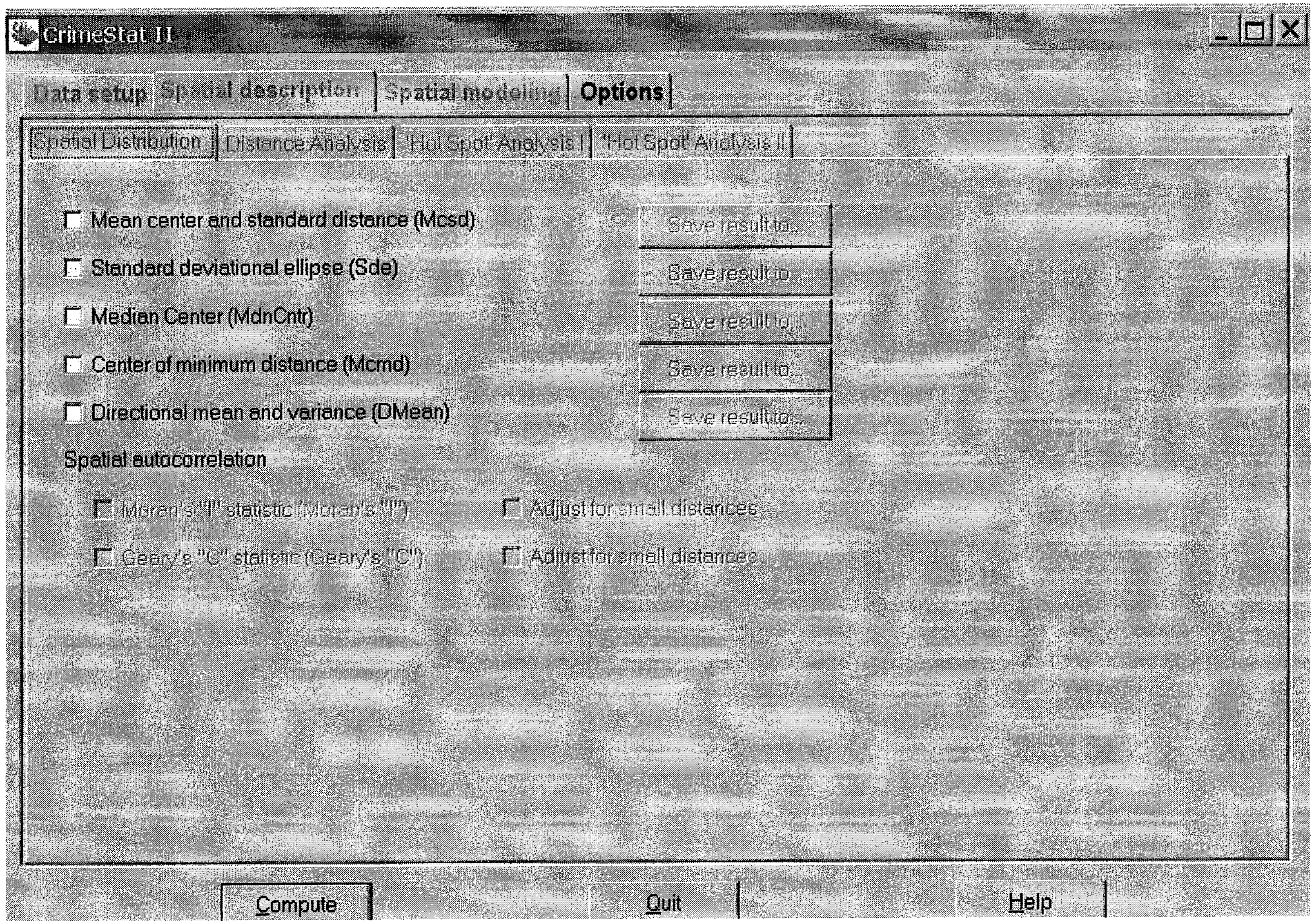
Because two dimensions adds complexity not seen in one dimension, these statistical moments have been modified to be appropriate. Figure 4.2 shows how the centrographic statistics are selected in *CrimeStat.*

Mean Center

The simplest descriptor of a distribution is the *mean center.* This is merely the mean of the X and Y coordinates. It is sometimes called a *center of gravity* in that it represents the point in a distribution where all other points are balanced if they existed on a plane and the mean center was a fulcrum (Ebdon, 1988; Burt and Barber, 1996).

For a single variable, the mean is the point at which the sum of all differences between the mean and all other points is zero. Unfortunately, for two variables, such as the location of crime incidents, the mean center is not necessarily the point at which the sum of all distances to all other points is minimized. That property is attributed to the center of minimum distance (see below). However, the mean center can be thought of as a

Figure 4.1: **Spatial Distribution Screen**



□ Mean center and standard distance (Mcsd)

□ Standard deviational ellipse (Sde)

□ Median Center (MdnCntr)

□ Center of minimum distance (Mcmd)

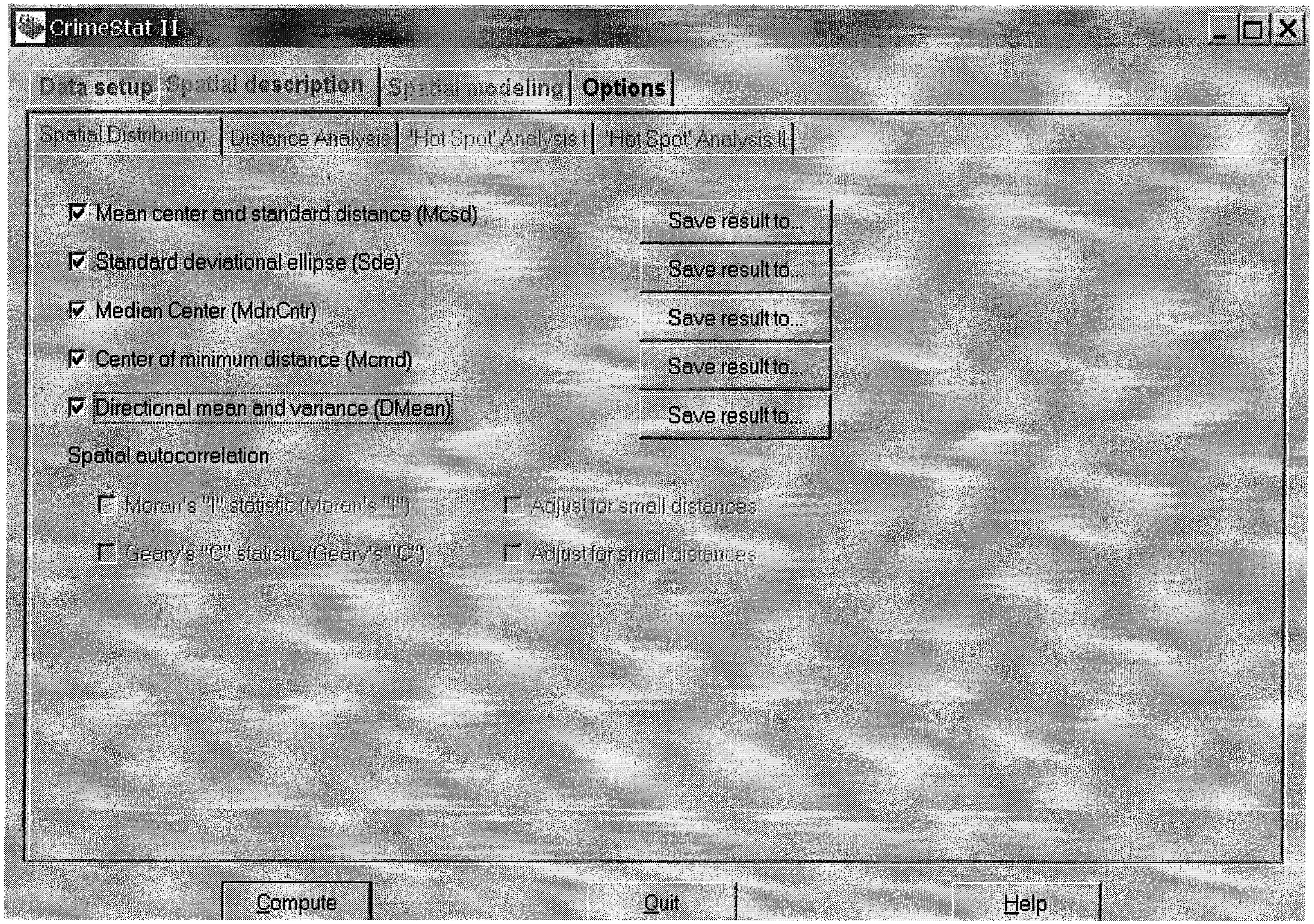
□ Directional mean and variance (DMean)

Spatial autocorrelation

□ Moran's "I" statistic (Moran's "I") □ Adjust for small distances

□ Geary's "C" statistic (Geary's "C") □ Adjust for small distances

Figure 4.2: **Selecting Centrographic Statistics**



☑ Mean center and standard distance (Mcsd)

☑ Standard deviational ellipse (Sde)

☑ Median Center (MdnCntr)

☑ Center of minimum distance (Mcmd)

☑ Directional mean and variance (DMean)

Spatial autocorrelation

☐ Moran's "I" statistic (Moran's "I") ☐ Adjust for small distances

☐ Geary's "C" statistic (Geary's "C") ☐ Adjust for small distances

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

point where both the sum of all differences between the mean X coordinate and all other X coordinates is zero and the sum of all differences between the mean Y coordinate and all other Y coordinates is zero.

The formula for the mean center is:

$$\bar{X} = \sum_{i=1}^{N} \frac{X_i}{N} \qquad \bar{Y} = \sum_{i=1}^{N} \frac{Y_i}{N} \qquad (4.1)$$

where X_i and Y_i are the coordinates of individual locations and N is the total number of points.

To take a simple example, the mean center for burglaries in Baltimore County has spherical coordinates of longitude -76.608482, latitude 39.348368 and for robberies longitude -76.620838, latitude 39.334816. Figure 4.3 illustrates these two mean centers.

Weighted Mean Center

A weighted mean center can be produced by weighting each coordinate by another variable, W_i. For example, if the coordinates are the centroids of census tracts, then the weight of each centroid could be the population within the census tract. Formula 4.1 is extended slightly to include a weight.

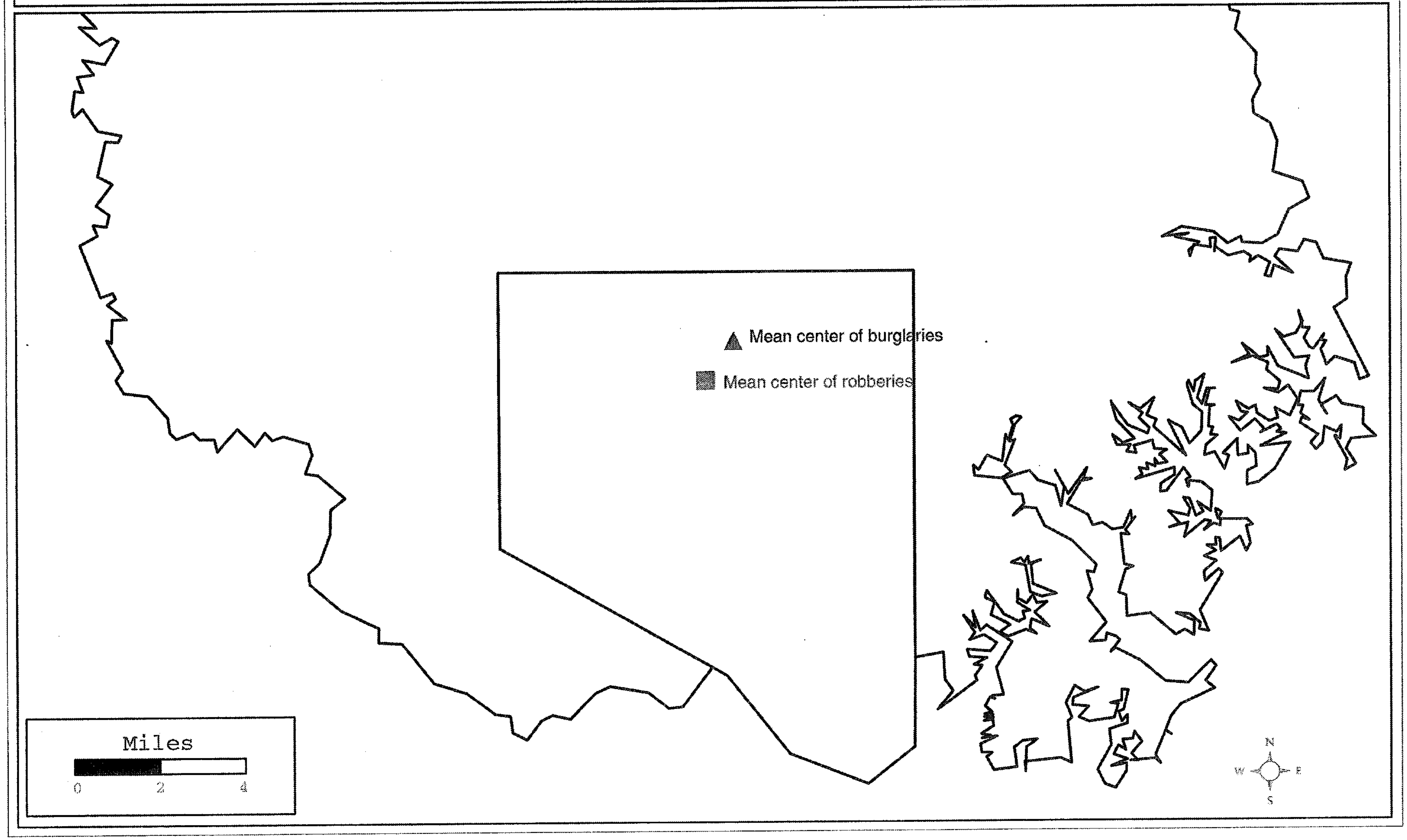
$$\bar{X} = \sum_{i=1}^{N} \frac{W_i X_i}{N} \qquad \bar{Y} = \sum_{i=1}^{N} \frac{W_i Y_i}{N} \qquad (4.2)$$

The advantage of a weighted mean center is that points associated with areas can have the characteristics of the areas included. For example, if the coordinates are the centroids of census tracts, then the weight of each centroid could be the population within the census tract. This will produce a different center of gravity than, say, the unweighted center of all census tracts. *CrimeStat* allows the mean to be weighted by either the weighting variable or by the intensity variable. Users should be careful, however, not to weight the mean with both the weighting and intensity variable unless there is an explicit distinction being made between weights and intensities.

To take an example, in the six jurisdictions making up the metropolitan Baltimore area (Baltimore City, and Baltimore, Carroll, Harford, Howard and Anne Arundel counties), the mean center of all census block groups is longitude -76.619121, latitude 39.304344. This would be an *unweighted* mean center of the block groups. On the other hand, the mean center of the 1990 population for the Baltimore metropolitan area had coordinates of longitude -76.625186 and latitude 39.304186, a position slightly southwest of the unweighted mean center. Weighting the block groups by median household income

Figure 4.3: Burglary and Robbery in Baltimore County

Comparison of Mean Centers



produces a mean center which is still more southwest. Figure 4.4 illustrates these three mean centers.

Weighted mean centers can be useful because they describe spatial differentiation in the metropolitan area and factors that may correlate with crime distributions. Another example is the weighted mean centers of different ethnic groups in the Baltimore metropolitan area (figure 4.5). The mean center of the White population is almost identical to the unweighted mean center. On the other hand, the mean center of the African-American/Black population is southwest of this and the mean center of the Hispanic/Latino population is considerably south of that for the White population. In other words, different ethnic groups tend to live in different parts of the Baltimore metropolitan area. Whether this has any impact on crime distributions is an empirical question. As we will see, there is not a simple spatial correlation between these weighted mean centers and particular crime distributions.

When the *Mcsd* box is checked, *CrimeStat* will run the routine. *CrimeStat* has a status bar that indicates how much of the routine has been run (Figure 4.6).[1] The results of these statistics are shown in the *Mcsd* output table (figure 4.7).

Median Center

The median center is the intersection between the median of the X coordinate and the median of the Y coordinate. The concept is simple. However, it is not strictly a median. For a single variable, such as median household income, the median is that point at which 50% of the cases fall below and 50% fall above. On a two dimensional plane, however, there is not a single median because the location of a median is defined by the way that the axes are drawn. For example, in figure 4.8, there are eight incident points shown. Four lines have been drawn which divide these eight points into two groups of four each. However, the four lines do not identify an exact location for a median. Instead, there is an area of non-uniqueness in which any part of it could be considered the 'median center'. This violates one of the basic properties of a statistic is that it be a unique value.

Nevertheless, as long as the axes are not rotated, the median center can be a useful statistic. The *CrimeStat* routine outputs three statistics:

1. The sample size
2. The median of X
3. The median of Y

The tabular output can be printed and the median center can be output as a graphical object to ArcView 'shp', MapInfo 'mif' or Atlas*GIS 'bna' files. A root name should be provided. The median center is output as a point (MdnCntr<root name>).

Figure 4.4: Center of Baltimore Metropolitan Population

Mean Center of Block Groups Weighted By Selected Variables

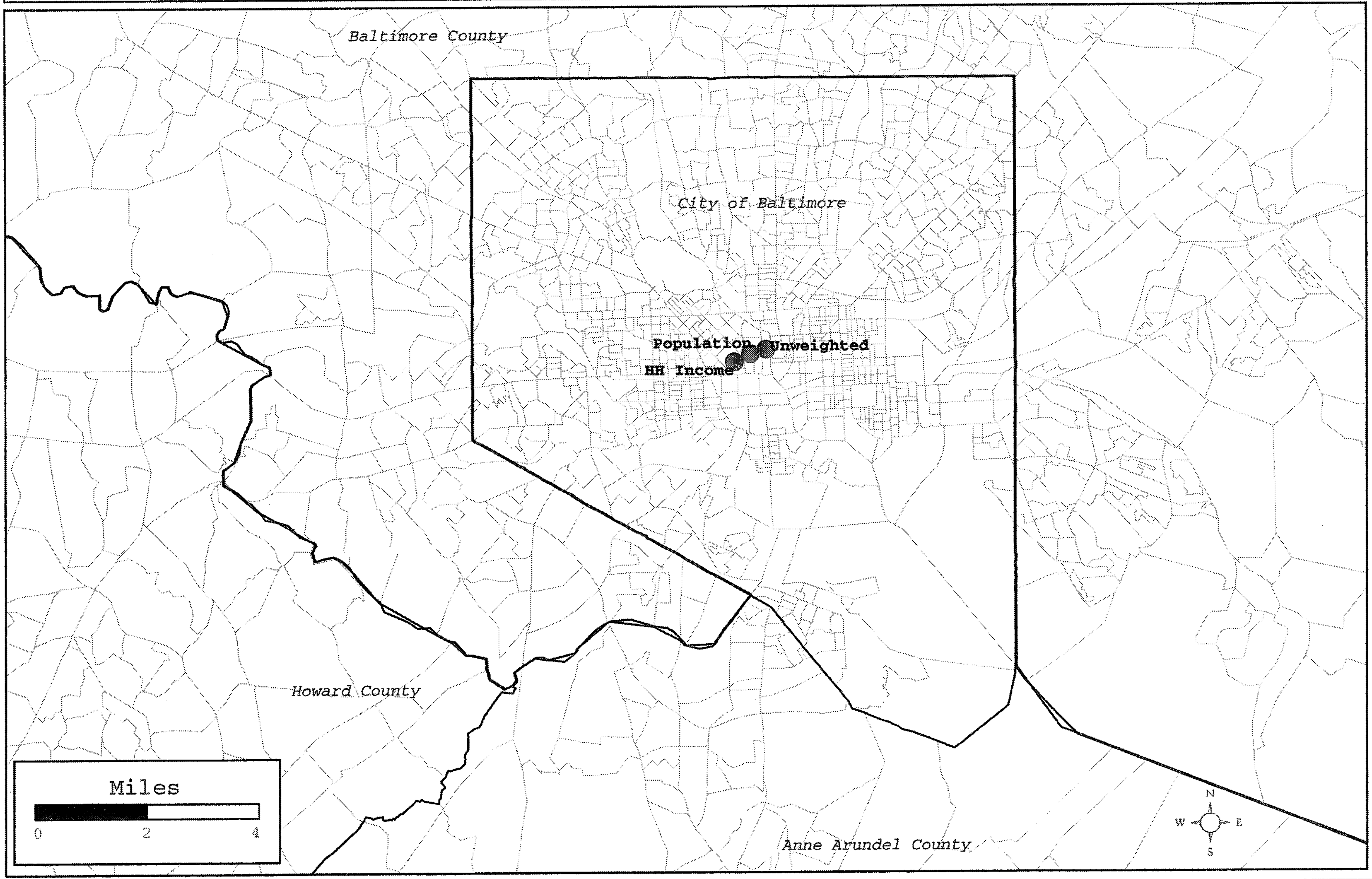


Figure 4.5: Center of Baltimore Metropolitan Population

Mean Center of Block Groups Weighted By Selected Variables

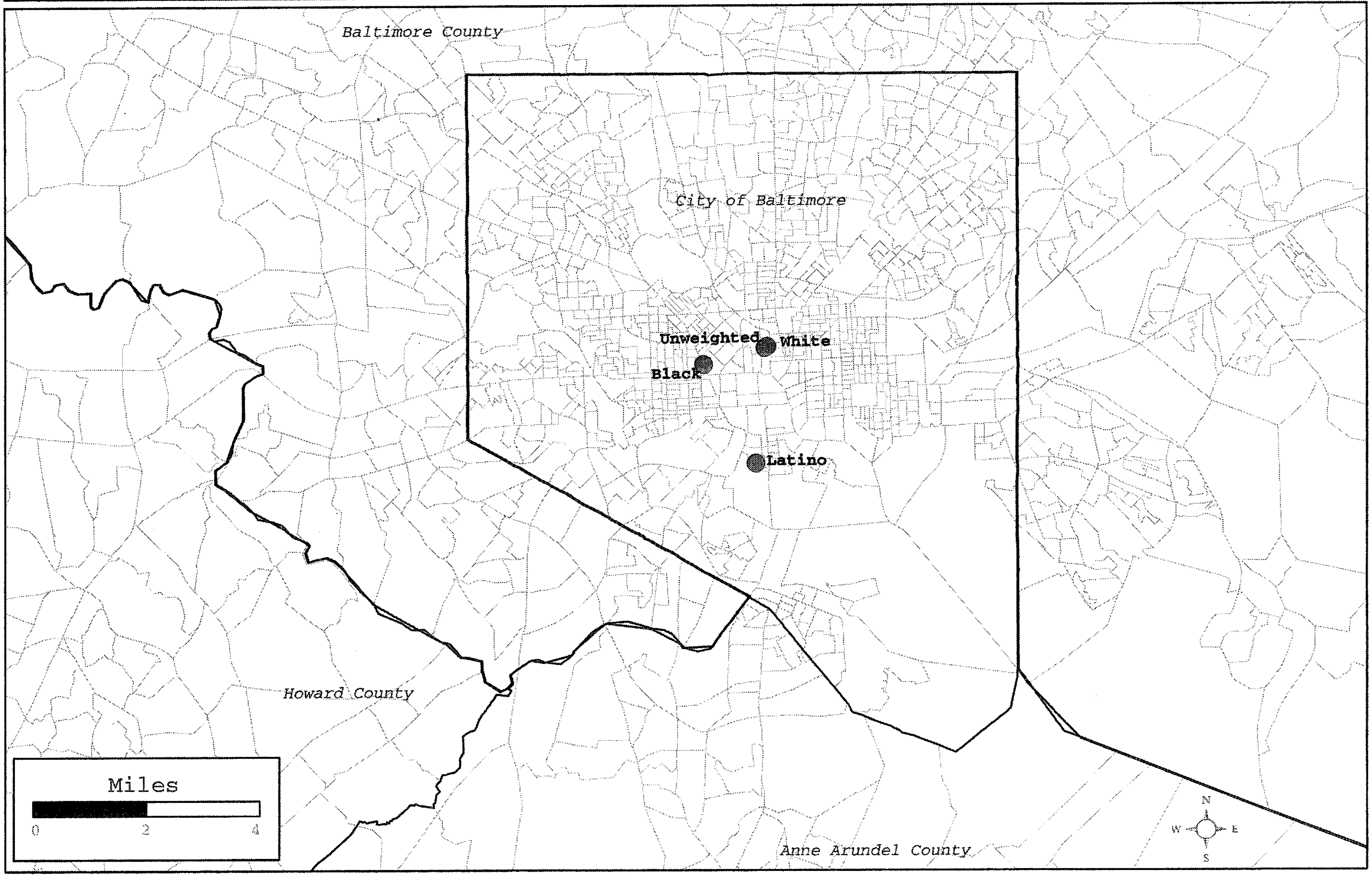
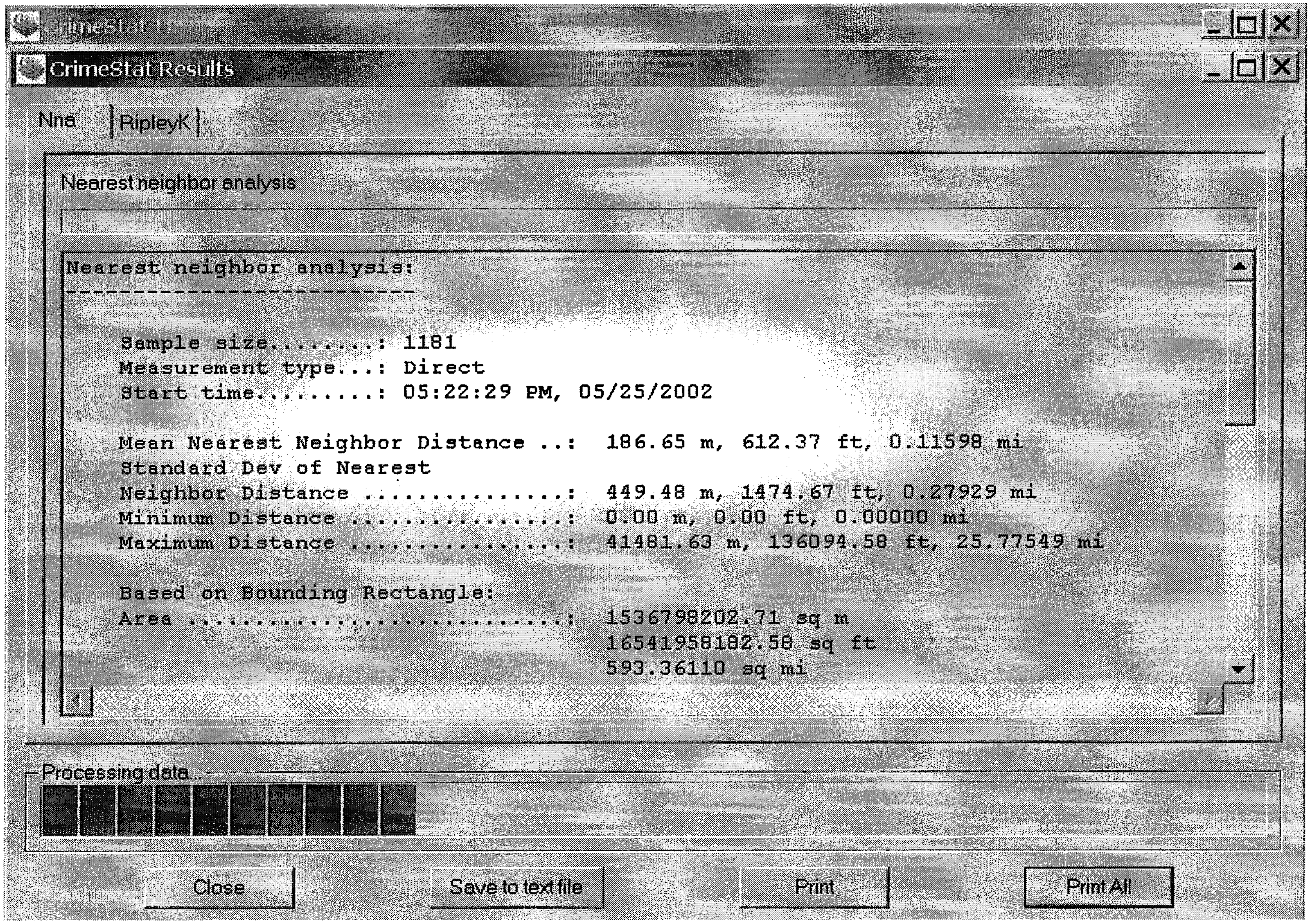
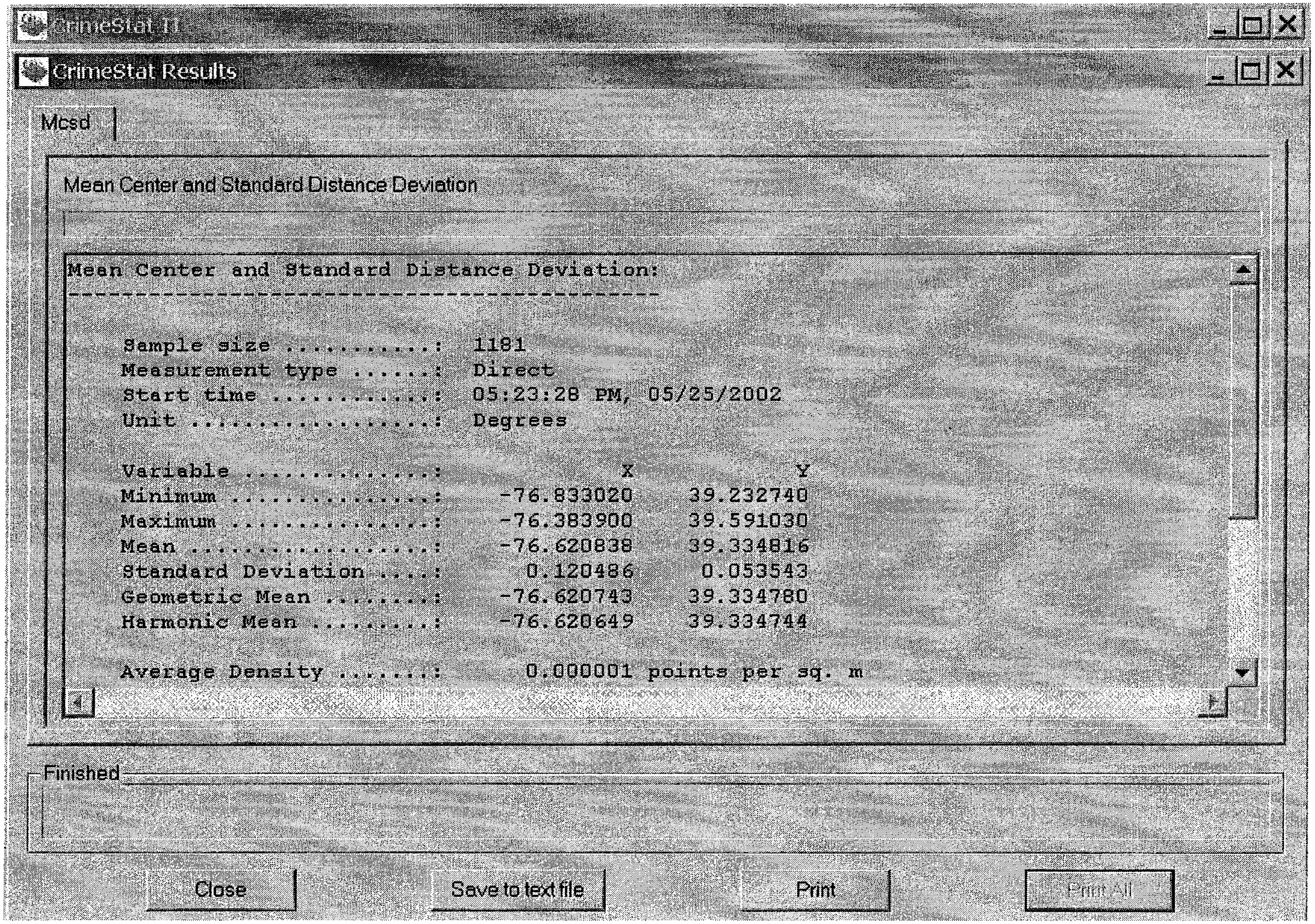


Figure 4.6: ***CrimeStat* Calculating A Routine**



```
Nearest neighbor analysis:
-------------------------

    Sample size........: 1181
    Measurement type...: Direct
    Start time.........: 05:22:29 PM, 05/25/2002

    Mean Nearest Neighbor Distance ..:  186.65 m, 612.37 ft, 0.11598 mi
    Standard Dev of Nearest
    Neighbor Distance ...............:  449.48 m, 1474.67 ft, 0.27929 mi
    Minimum Distance ................:  0.00 m, 0.00 ft, 0.00000 mi
    Maximum Distance ................:  41481.63 m, 136094.58 ft, 25.77549 mi

    Based on Bounding Rectangle:
    Area ............................:  1536798202.71 sq m
                                        16541958182.58 sq ft
                                        593.36110 sq mi
```

Figure 4.7: **Mean Center and Standard Distance Deviation Output**

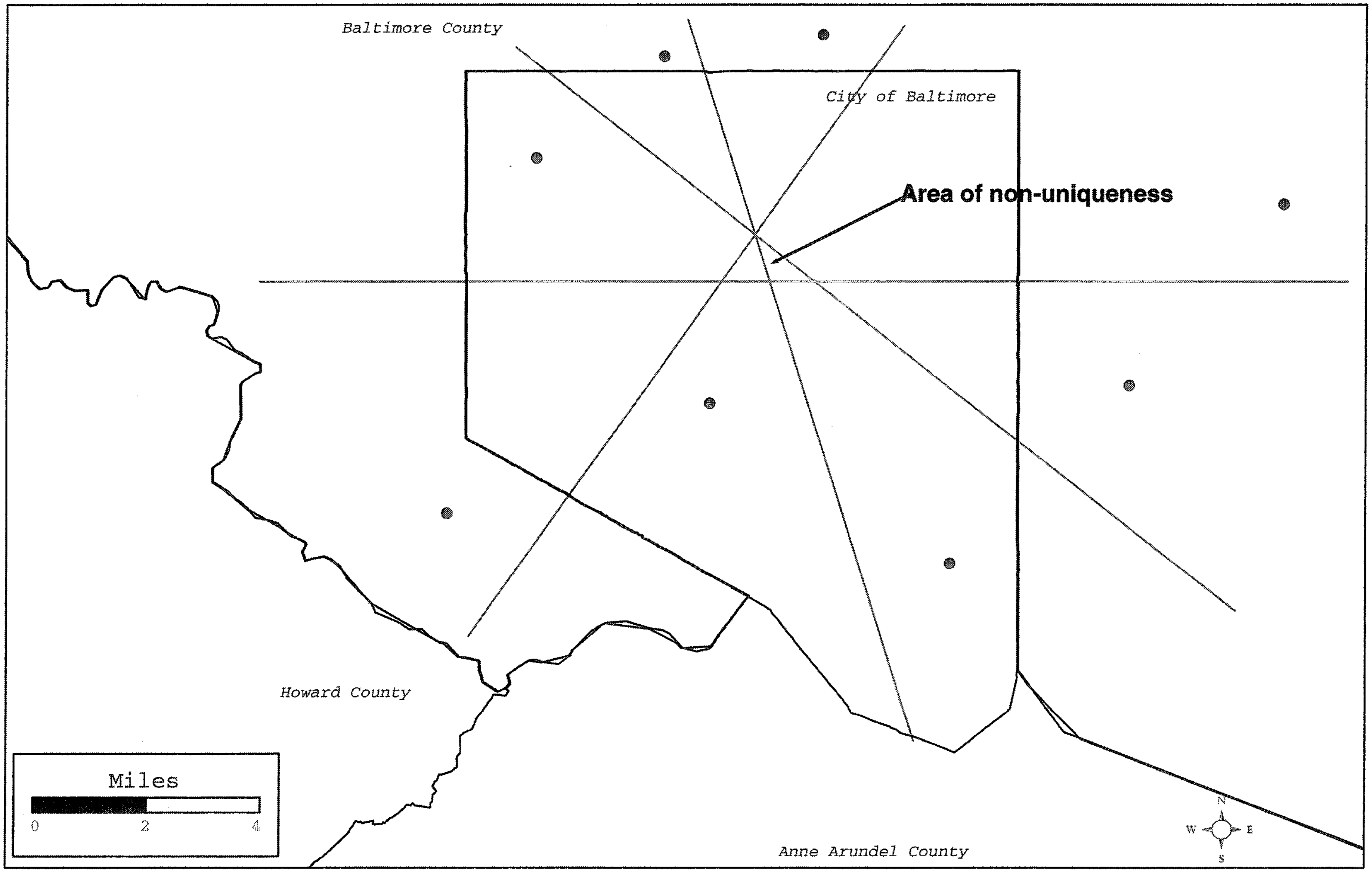


```
Mean Center and Standard Distance Deviation:
-------------------------------------------


    Sample size ...........:  1181
    Measurement type ......:  Direct
    Start time ............:  05:23:28 PM, 05/25/2002
    Unit ..................:  Degrees

    Variable ..............:             X            Y
    Minimum ...............:    -76.833020    39.232740
    Maximum ...............:    -76.383900    39.591030
    Mean ..................:    -76.620838    39.334816
    Standard Deviation ....:      0.120486     0.053543
    Geometric Mean ........:    -76.620743    39.334780
    Harmonic Mean .........:    -76.620649    39.334744

    Average Density .......:      0.000001 points per sq. m
```

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Figure 4.8: Non-Uniqueness of a Median Center

Lines Splitting Incident Locations Into Two Halves



Center of Minimum Distance

Another centrographic statistic is the *center of minimum distance*. Unfortunately, this statistic is sometimes also called the *median center*, which can make it confusing since the above statistic has the same name. Nevertheless, unlike the median center above, the center of minimum distance is a unique statistic in that it defines the point at which the sum of the distance to all other points is the smallest (Burt and Barber, 1996). It is defined as:

$$\text{Center of Minimum Distance} = C = \sum_{i=1}^{N} d_{ic} \text{ is a minimum} \qquad (4.3)$$

where d_{ic} is the distance between a single point, i, and C, the center of minimum distance (with an X and Y coordinate). Unfortunately, there is not a formula that can calculate this location.

Instead, an iterative algorithm is used which approximates this location (Kuhn and Kuenne, 1962; Burt and Barber, 1996). Depending on whether the coordinates are spherical or projected, *CrimeStat* will calculate distance as either Great Circle (spherical) or Euclidean (projected), as discussed in the previous chapter.[2] The results are shown in the *Mcmd* output table (figure 4.9).

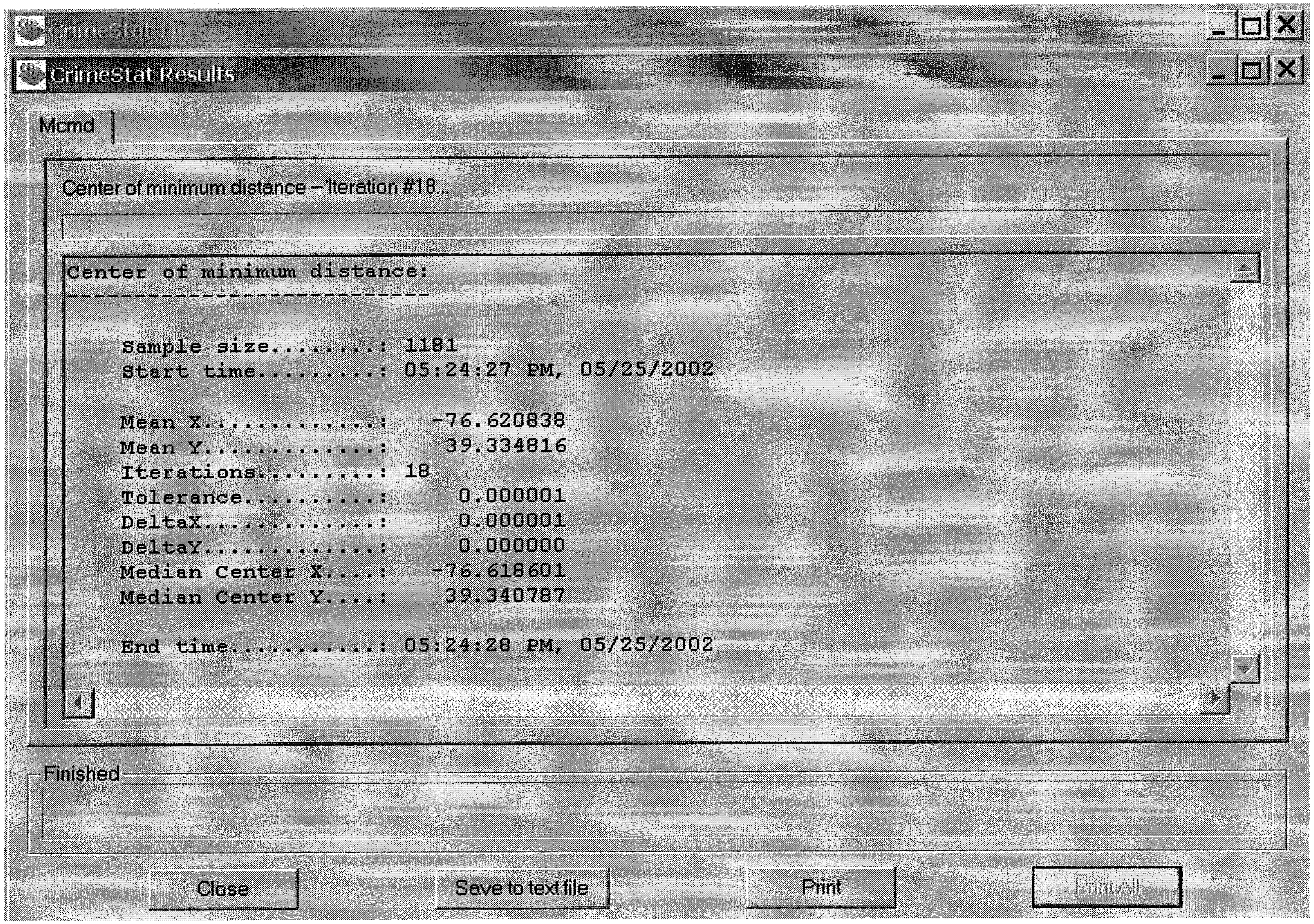
The importance of the center of minimum distance is that it is a location where distance to all the defining incidents is the smallest. Since *CrimeStat* only measures distances as either direct or indirect, actual travel time is not being calculated. But in many jurisdictions, the minimum distance to all points is a good approximation to the point where travel distances are minimized. For example, in a police precinct, a patrol car could be stationed at the center of minimum distance to allow it to respond quickly to calls for service.

For example, figure 4.10 maps the center of minimum distance for 1996 auto thefts in both Baltimore City and Baltimore County and compares this to both the mean center and the median center statistic. As seen, both the center of minimum distance and the median center are south of the mean center, indicating that there are slightly more incidents in the southern part of the metropolitan area than in the northern part. However, the difference in these three statistics is very small, especially the median center and the center of minimum distance.

Standard Deviation of the X and Y Coordinates

In addition to the mean center and center of minimum distance, *CrimeStat* will calculate various measures of spatial distribution, which describe the dispersion, orientation, and shape of the distribution of a variable (Hammond and McCullogh 1978; Ebdon 1988). The simplest of these is the raw standard deviations of the X and Y

Figure 4.9: **Center of Minimum Distance Output**



Center of minimum distance – Iteration #18...

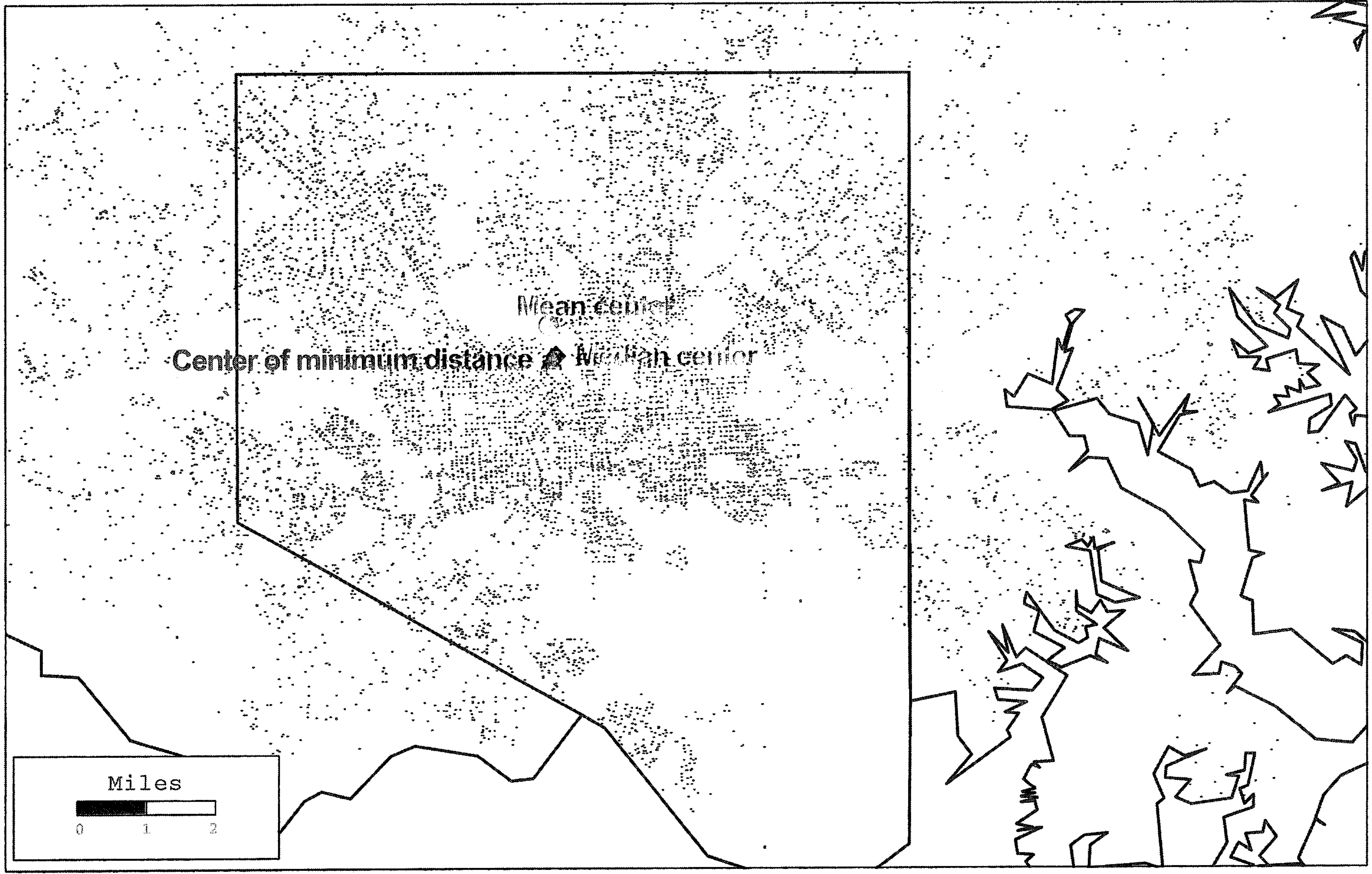
```
Center of minimum distance:
---------------------------

    Sample size.........: 1181
    Start time..........: 05:24:27 PM, 05/25/2002

    Mean X..............:   -76.620838
    Mean Y..............:    39.334816
    Iterations..........: 18
    Tolerance...........:     0.000001
    DeltaX..............:     0.000001
    DeltaY..............:     0.000000
    Median Center X.....:   -76.618601
    Median Center Y.....:    39.340787

    End time............: 05:24:28 PM, 05/25/2002
```

Finished

Figure 4.10:

1996 Metropolitan Baltimore Vehicle Thefts

Mean Center and Center of Minimum Distance for 1996 Auto Thefts



coordinates, respectively. The formulas used are the standard ones found in most elementary statistics books:

$$S_X = \mathrm{SQRT}\left[\sum_{i=1}^{N} \frac{(X_i - \bar{X})^2}{N-1} \right] \tag{4.4}$$

$$S_Y = \mathrm{SQRT}\left[\sum_{i=1}^{N} \frac{(Y_i - \bar{Y})^2}{N-1} \right] . \tag{4.5}$$

where X_i and Y_i are the X and Y coordinates for individual points, $\bar{X}$ and $\bar{Y}$ are the mean X and mean Y, and N is the total number of points. Note that 1 is subtracted from the number of points to produce an unbiased estimate of the standard deviation.

The standard deviations of the X and Y coordinates indicate the degree of dispersion. Figure 4.11 shows the standard deviation of the coordinates for auto thefts and represents this as a rectangle. As seen, the distribution of auto thefts spreads more in an east-west direction than in a north-south direction.

Standard Distance Deviation

While the standard deviation of the X and Y coordinates provides some information about the dispersion of the incidents, there are two problems with it. First, it does not provide a single summary statistic of the dispersion in the incident locations and is actually two separate statistics (i.e., dispersion in X and dispersion in Y). Second, it provides measurements in the units of the coordinate system. Thus, if spherical coordinates are being used, then the units will be decimal degrees.

A measure which overcomes these problems is the *standard distance deviation* or *standard distance*, for short. This is the standard deviation of the *distance* of each point from the mean center and is expressed in measurement units (feet, meters, miles). It is the two-dimensional equivalent of a standard deviation.

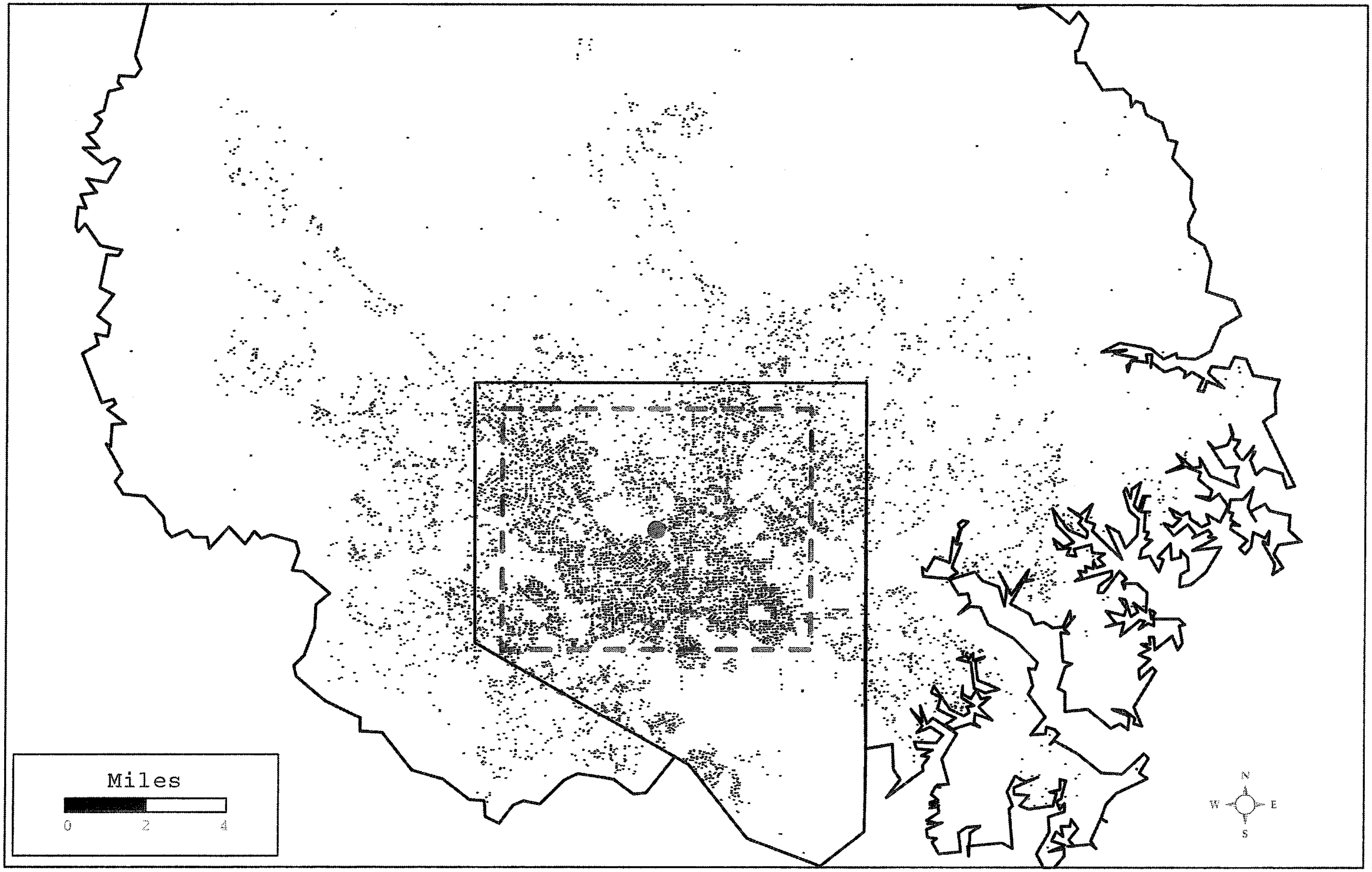
The formula for it is

$$S_{XY} = \mathrm{Sqrt}\left[\sum_{i=1}^{N} \left\{ \frac{(d_{iMC})^2}{N-2} \right\} \right] \tag{4.6}$$

Figure 4.11:

1996 Metropolitan Baltimore Auto Thefts

Mean Center and Standard Deviations of X and Y Coordinates



where d_{iMC} is the distance between each point, i, and the mean center and N is the total number of points. Note that 2 is subtracted from the number of points to produce an unbiased estimate of standard distance since there are two constants from which this distance is measured (mean of X, mean of Y).[3]

The standard distance can be represented as a single vector rather than two vectors as with the standard deviation of the X and Y coordinates. Figure 4.12 shows the mean center and standard distance deviation of both robberies and burglaries for 1996 in Baltimore County represented as circles. It is clear that the spatial distributions of these two types of crime vary with robberies being slightly more concentrated.

Standard Deviational Ellipse

The standard distance deviation is a good single measure of the dispersion of the incidents around the mean center. However, with two dimensions, distributions are frequently skewed in one direction or another (a condition called *anisotropy*). Instead, there is another statistic which gives dispersion in two dimensions, the *standard deviational ellipse* or *ellipse*, for short (Ebdon, 1988; Cromley, 1992).

The standard deviational ellipse is derived from the bivariate distribution (Furfey, 1927; Neft, 1962; Bachhi, 1957) and is defined by

$$\text{Bivariate Distribution} = \text{SQRT} \frac{[\sigma^2_x + \sigma^2_y]}{2} \qquad (4.7)$$

The two standard deviations, in the X and Y directions, are orthogonal to each other and define an ellipse. Ebdon (1988) rotates the X and Y axis so that the sum of squares of distances between points and axes are minimized. By convention, it is shown as an ellipse.

Aside from the mean X and mean Y, the formulas for these statistics are as follows:

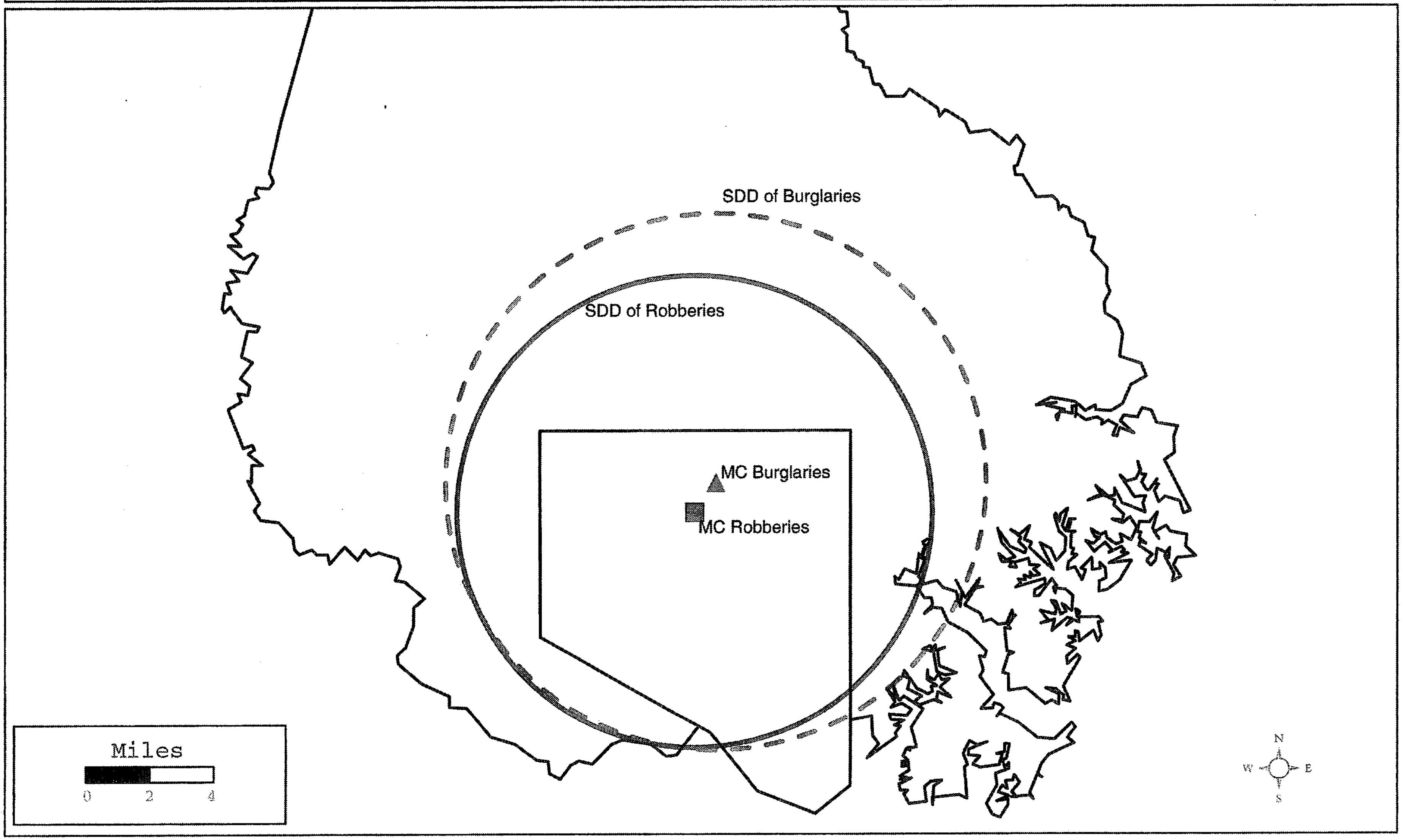
1. The Y-axis is rotated *clockwise* through an angle, θ, where

$$\theta = \text{ARCTAN } \{(\Sigma(X_i-\bar{X})^2 - \Sigma(Y_i-\bar{Y})^2) + [(\Sigma(X_i-\bar{X})^2 - \Sigma(Y_i-\bar{Y})^2)^2 + 4(\Sigma(X_i-\bar{X})(Y_i-\bar{Y}))^2]^{1/2}\}/(2\Sigma(X_i-\bar{X})(Y_i-\bar{Y})) \qquad (4.8)$$

where all summations are for i=1 to N (Ebdon, 1988).

Figure 4.12: 1996 Baltimore County Burglaries and Robberies

Comparison of Mean Centers and Standard Distance Deviations



2. Two standard deviations are calculated, one along the transposed X-axis and one along the transposed Y-axis.

$$S_x = SQRT(2)\{\sum_{i=1}^{N}[(X_i - \bar{X})Cos\theta - (Y_i - \bar{Y})Sin\theta]^2/(N-2)\}^{½} \qquad (4.9)$$

$$S_y = SQRT(2)\{\sum_{i=1}^{N}[(X_i - \bar{X})Sin\theta - (Y_i - \bar{Y})Cos\theta]^2/(N-2)\}^{½} \qquad (4.10)$$

where N is the number of points. Note, again, that 2 is subtracted from the number of points in both denominators to produce an unbiased estimate of the standard deviational ellipse since there are two constants from which the distance along each axis is measured (mean of X, mean of Y).[4]

3. The X-axis and Y-axis of the ellipse are defined by

$$Length_x = 2S_x \qquad (4.11)$$

$$Length_y = 2S_y \qquad (4.12)$$

4. The area of the ellipse is

$$A = \pi S_x S_y \qquad (4.13)$$

Figure 4.13 shows the output of the ellipse routine and figure 4.14 maps the standard deviational ellipse of auto thefts in Baltimore City and Baltimore County for 1996.

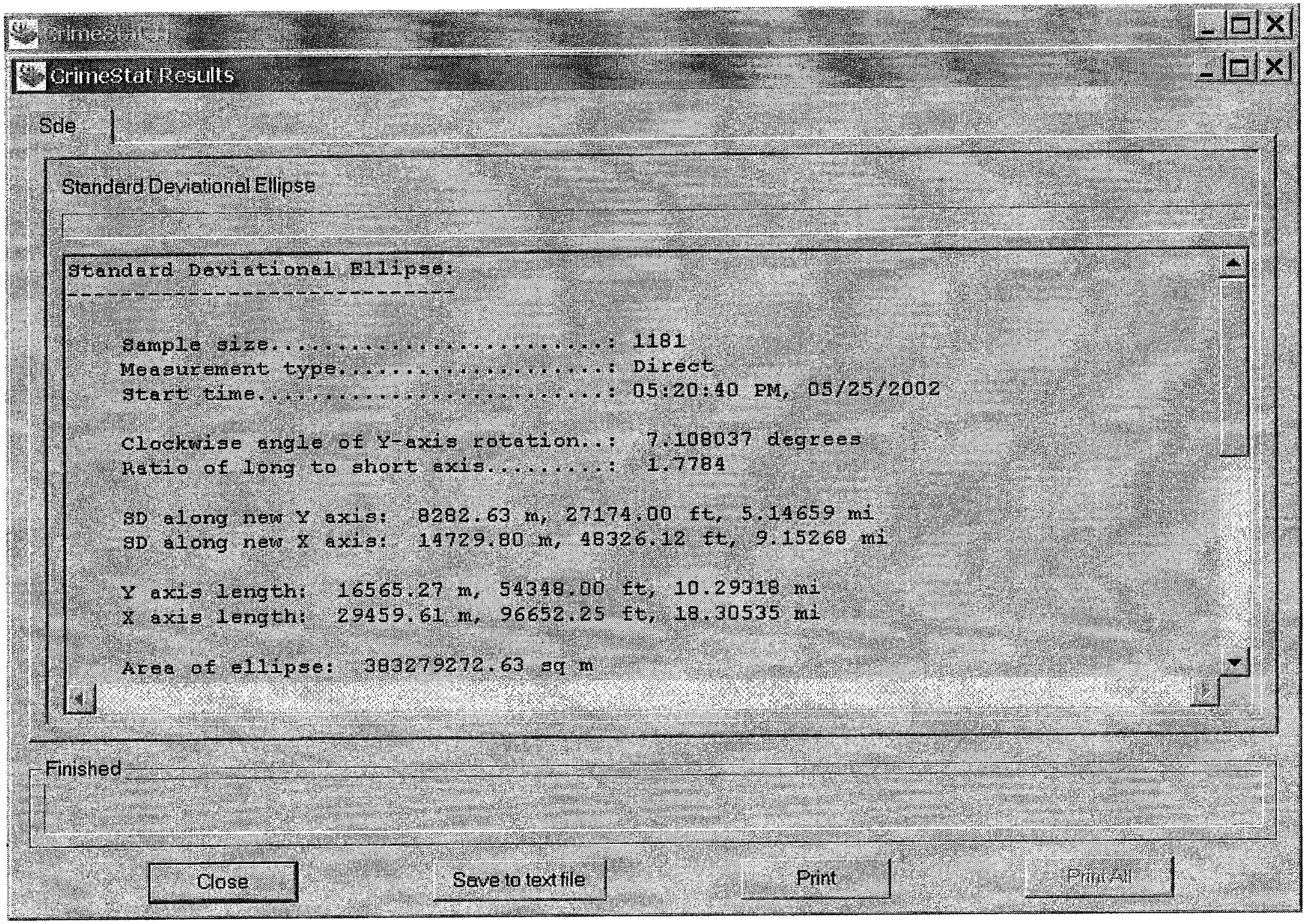
Geometric Mean

The mean center routine (Mcsd) includes two additional means. First, there is the geometric mean, which is a mean associated with the mean of the logarithms. It is defined as:

$$\text{Geometric Mean of X} = GM(X) = \prod_{i=1}^{N} (X_i)^{1/N} \qquad (4.14)$$

$$\text{Geometric Mean of Y} = GM(Y) = \prod_{i=1}^{N} (Y_i)^{1/N} \qquad (4.15)$$

Figure 4.13: **Standard Deviational Ellipse Output**

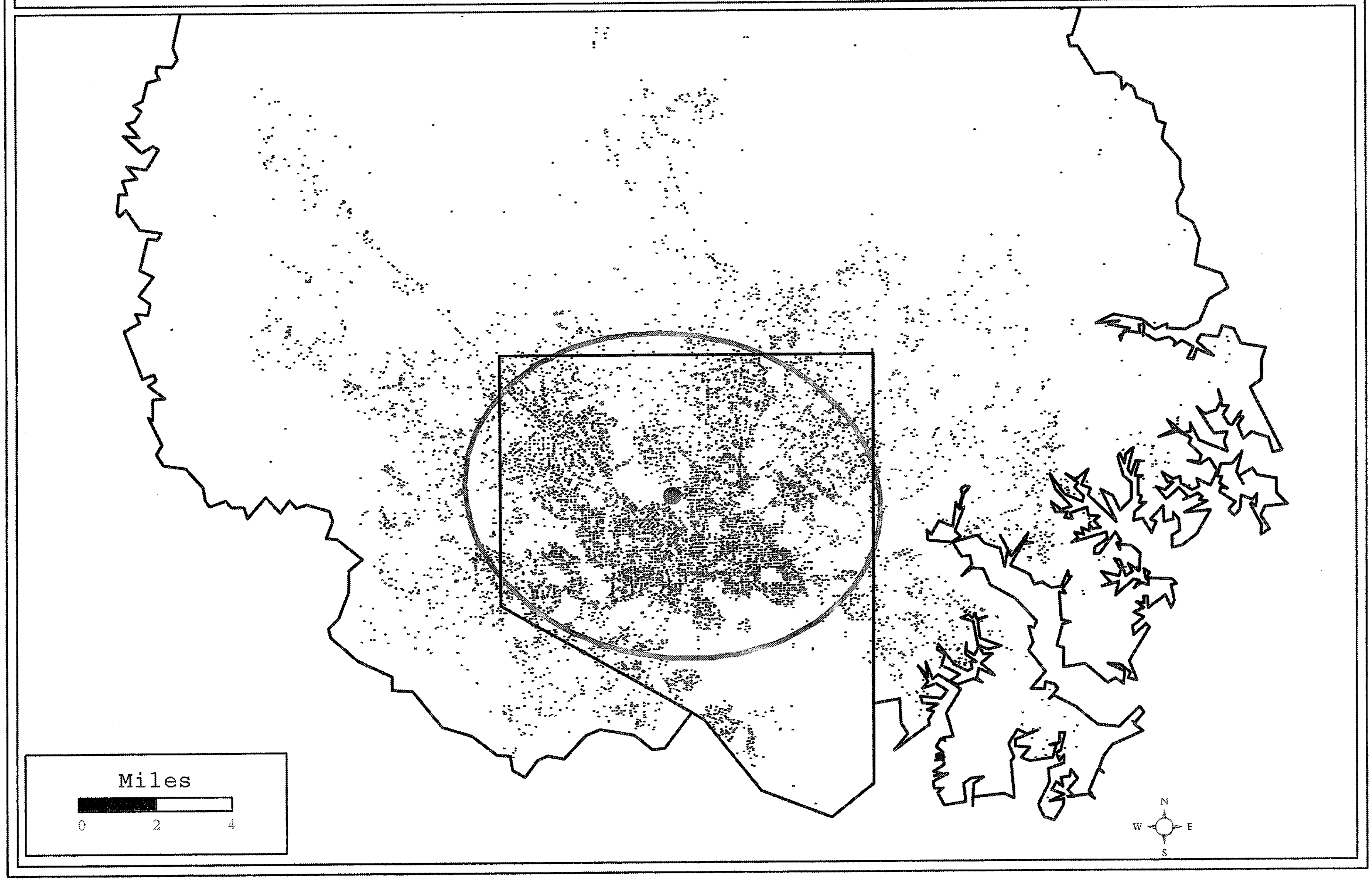


```
Standard Deviational Ellipse:
-----------------------------

    Sample size........................: 1181
    Measurement type...................: Direct
    Start time.........................: 05:20:40 PM, 05/25/2002

    Clockwise angle of Y-axis rotation..:  7.108037 degrees
    Ratio of long to short axis.........:  1.7784

    SD along new Y axis:  8282.63 m, 27174.00 ft, 5.14659 mi
    SD along new X axis:  14729.80 m, 48326.12 ft, 9.15268 mi

    Y axis length:  16565.27 m, 54348.00 ft, 10.29318 mi
    X axis length:  29459.61 m, 96652.25 ft, 18.30535 mi

    Area of ellipse:  383279272.63 sq m
```

Figure 4.14:

1996 Metropolitan Baltimore Auto Thefts

Mean Center and Standard Deviational Ellipse



where Π is the product term of each point value, i (i.e., the values of X or Y are multiplied times each other), and N is the sample size (Everitt, 1995). The equation can be evaluated by logarithms.

$$Ln[GM(X)] = \frac{1}{N} [Ln(X_1) + Ln(X_2) + ... + Ln(X_N)] = \frac{1}{N} \Sigma Ln(X_i) \qquad (4.16)$$

$$Ln[GM(Y)] = \frac{1}{N} [Ln(Y_1) + Ln(Y_2) + ... + Ln(Y_N)] = \frac{1}{N} \Sigma Ln(Y_i) \qquad (4.17)$$

$$GM(X) = e^{Ln(GM(X)} \qquad (4.18)$$

$$GM(Y) = e^{Ln(GM(Y)} \qquad (4.19)$$

The geometric mean is the anti-log of the mean of the logarithms. Because it first converts all X and Y coordinates into logarithms, it has the effect of discounting extreme values. The geometric mean is output as part of the Mcsd routine and has a 'Gm' prefix before the user defined name.

Harmonic Mean

The harmonic mean is also a mean which discounts extreme values, but is calculated differently. It is defined as

$$\text{Harmonic mean of X} = HM(X) = \frac{N}{\Sigma (1/X_i)} \qquad (4.20)$$

$$\text{Harmonic mean of Y} = HM(Y) \quad \frac{N}{\Sigma (1/Y_i)} \qquad (4.21)$$

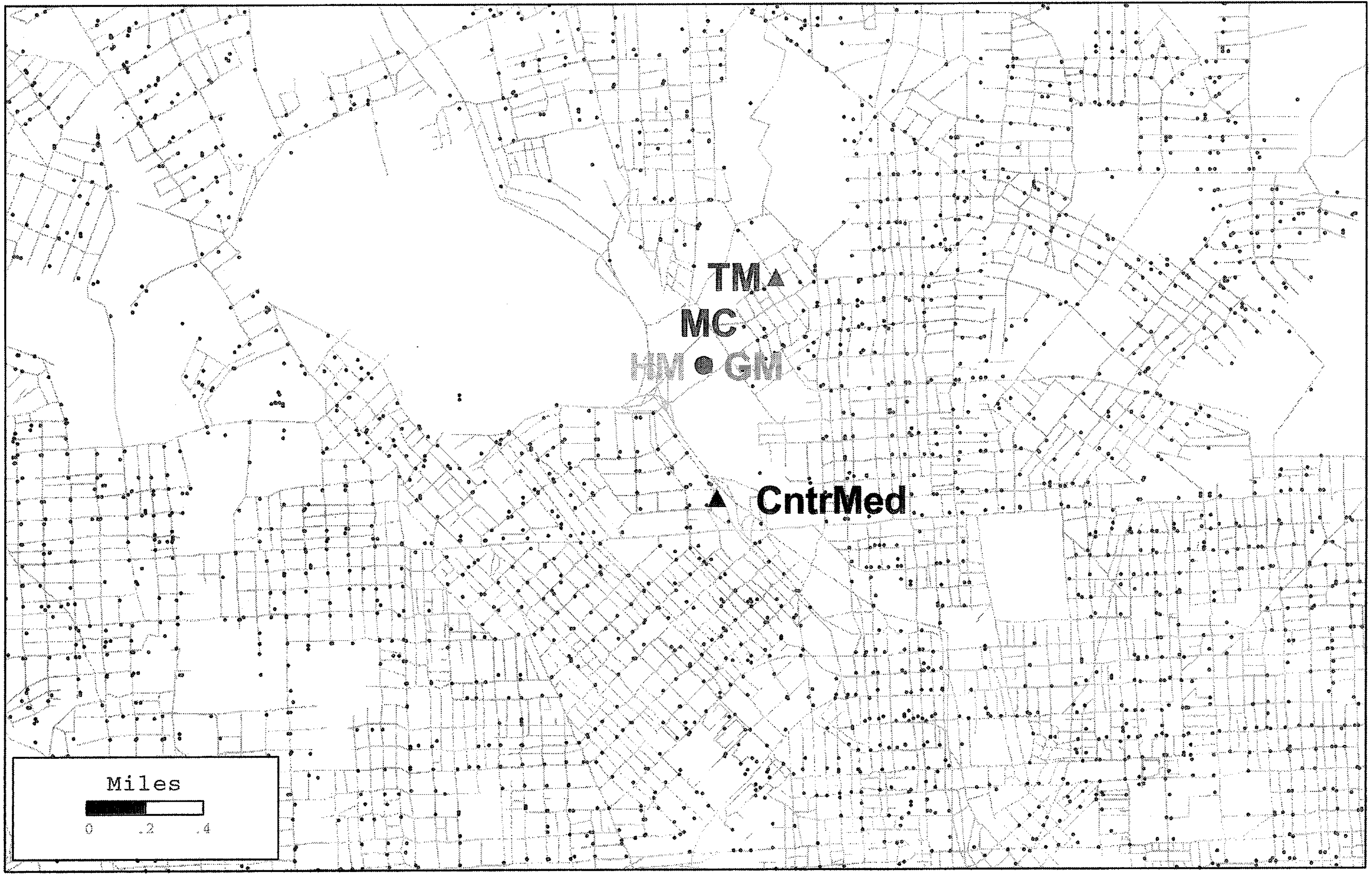
In other words, the harmonic mean of X and Y respectively is the inverse of the mean of the inverse of X and Y respectively (i.e., take the inverse; take the mean of the inverse; and invert the mean of the inverse). The harmonic mean is output as part of the Mcsd routine and has a 'Hm' prefix before the user defined name.

The geometric and harmonic means are discounted means that 'hug' the center of the distribution. They differ from the mean center when there is a very skewed distribution. To contrast the different means, figure 4.15 below shows five different means for Baltimore County motor vehicle thefts:

Figure 4.15:

Five Mean Centers for 1996 Baltimore Vehicle Thefts

Five Different Means Compared



1. Mean center;
2. Center of minimum distance;
3. Geometric mean;
4. Harmonic mean; and
5. Triangulated mean (discussed below)

In the example, the mean center, geometric mean, and harmonic mean fall almost on top of each other; however, they will not always be so. The center of minimum distance approximates the geographical center of the distribution. The triangulated mean is defined by the angularity and distance from the lower-left and upper-right corners of the data set (see below).

Centrographic descriptors can be very powerful tools for examining spatial patterns. They are a first step in any spatial analysis, but an important one. The above example illustrates how they can be a basis for decision-making, even with small samples. A couple of other examples can be illustrated.

Average Density

The average density is the number of incidents divided by the area. It is a measure of the average number of events per unit of area; it is sometimes called the *intensity*. If the area is defined on the measurement parameters page, the routine uses that value; otherwise, it takes the rectangular area defined by the minimum and maximum X and Y values (the bounding rectangle).

Output Files

Calculating the Statistics

Once the statistics have been selected, the user clicks on *Compute* to run the routine. The results are shown in a results table.

Tabular Output

For each of these statistics, *CrimeStat* produces tabular output. In *CrimeStat*, all tables are labeled by symbols, for example Mcsd for the mean center and standard distance deviation or Mcmd for the center of minimum distance. All tables present the sample size.

Graphical Objects

The six centrographic statistics can be output as graphical objects. The mean center and center of minimum distance are output as single points. The standard deviation of the X and Y coordinates is output as a rectangle. The standard distance deviation is output as a circle and the standard deviational ellipse is output as an ellipse.

CrimeStat currently supports graphical outputs to *ArcView* '.shp' files, to *MapInfo* '.mif' and to *Atlas*GIS* '.bna' files. Before running the calculation, the user should select the desired output files and specify a root name (e.g., Precinct1Burglaries). Figure 4.16 shows a dialog box for selecting for the GIS program output. For *MapInfo* output only, the user has to also indicate the name of the projection, the projection number and the datum number. These can be found in the *MapInfo* users guide. By default, *CrimeStat* will use the standard parameters for a spherical coordinate system (Earth projection, projection number 1, and datum number 33). If a user requires a different coordinate system, the appropriate values should be typed into the space. Figure 4.17 shows the selection of the *MapInfo* coordinate parameters.

If requested, the output files are saved in the specified directory under the specified (root) name. For each statistic, *CrimeStat* will add prefix letters to the root name.

MC<*root*> for the mean center
MdnCntr<root> for the median center
Mcmd<*root*> for center of minimum distance
XYD<*root*> for the standard deviation of the X and Y coordinates
SDD<*root*> for the standard distance deviation
SDE<*root*> for the standard deviational ellipse.

The '.shp' files can be read directly into *ArcView* as themes. The '.mif' and '.bna' files have to be imported into *MapInfo* and *Atlas*GIS,* respectively.[5]

Statistical Testing

While the current version of *CrimeStat* does not conduct statistical tests that compare two distributions, it is possible to conduct such tests. Appendix B presents a discussion of the statistical tests that can be used. Instead, the discussion here will focus on using the outputs of the routines without formal testing.

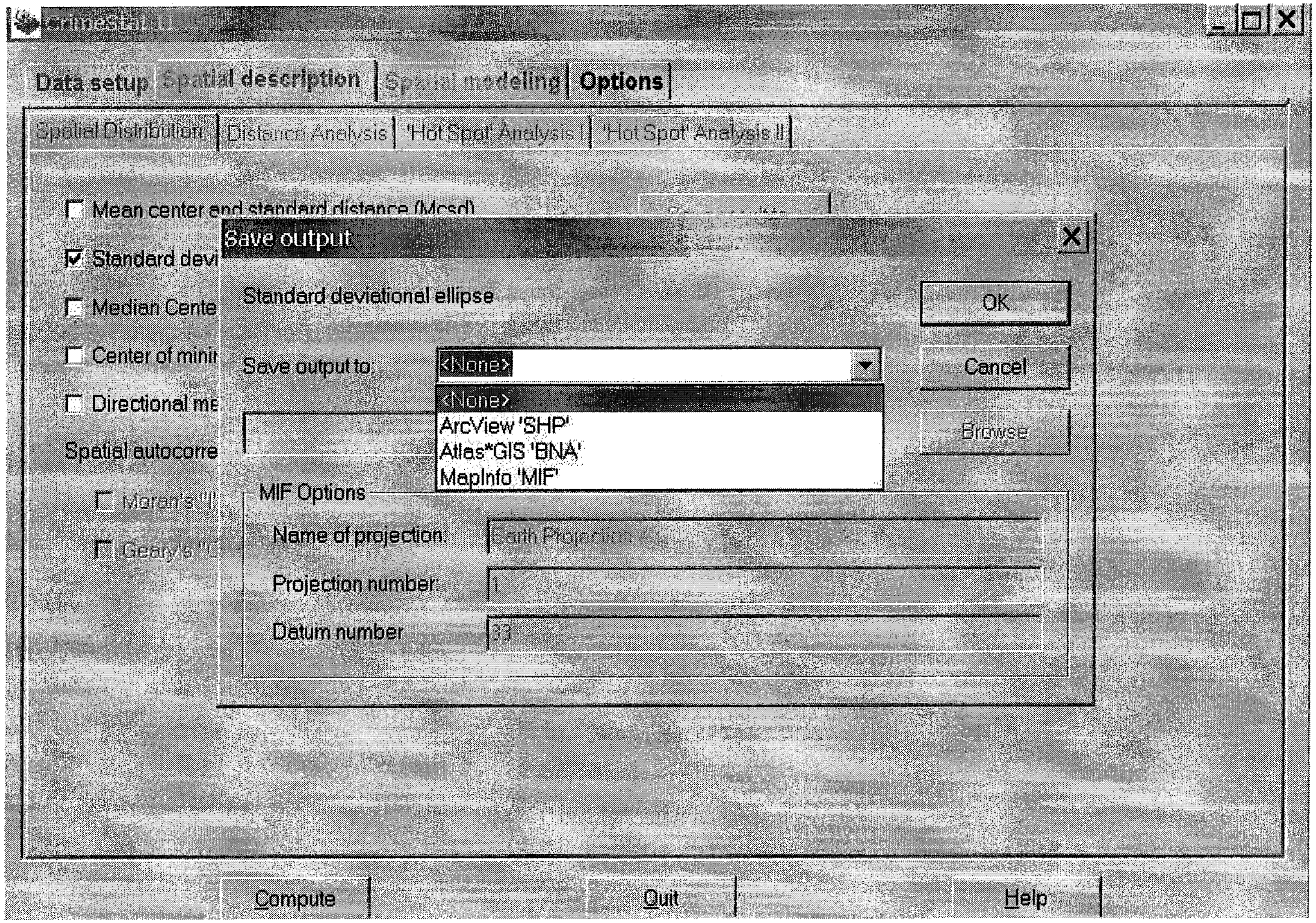
Decision-making Without Formal Tests

Formal significance testing has the advantage of providing a consistent inference about whether the difference in two distributions is likely or unlikely to be due to chance. Almost all formal tests compare the distribution of a statistic with that of a random distribution. However, police departments frequently have to make decisions based on small samples, in which case the formal tests are less useful than they would with larger samples. Still, the centrographic statistics calculated in *CrimeStat* can be useful and can help a police department make decision even in the absence of formal tests.

Example 1: June and July auto thefts in Precinct 11

We want to illustrate the use of these statistics to make decisions with two examples. The first is a comparison of crimes in small geographical areas. In most metropolitan areas, most analysts will concentrate on particular sub-areas of the

Figure 4.16: Outputting Objects to A GIS Program



□ Mean center and standard distance (Mcsd)

☑ Standard devi

□ Median Cente

□ Center of minii

□ Directional me

Spatial autocorre

□ Moran's "I

□ Geary's "C

Save output

Standard deviational ellipse

Save output to: <None>

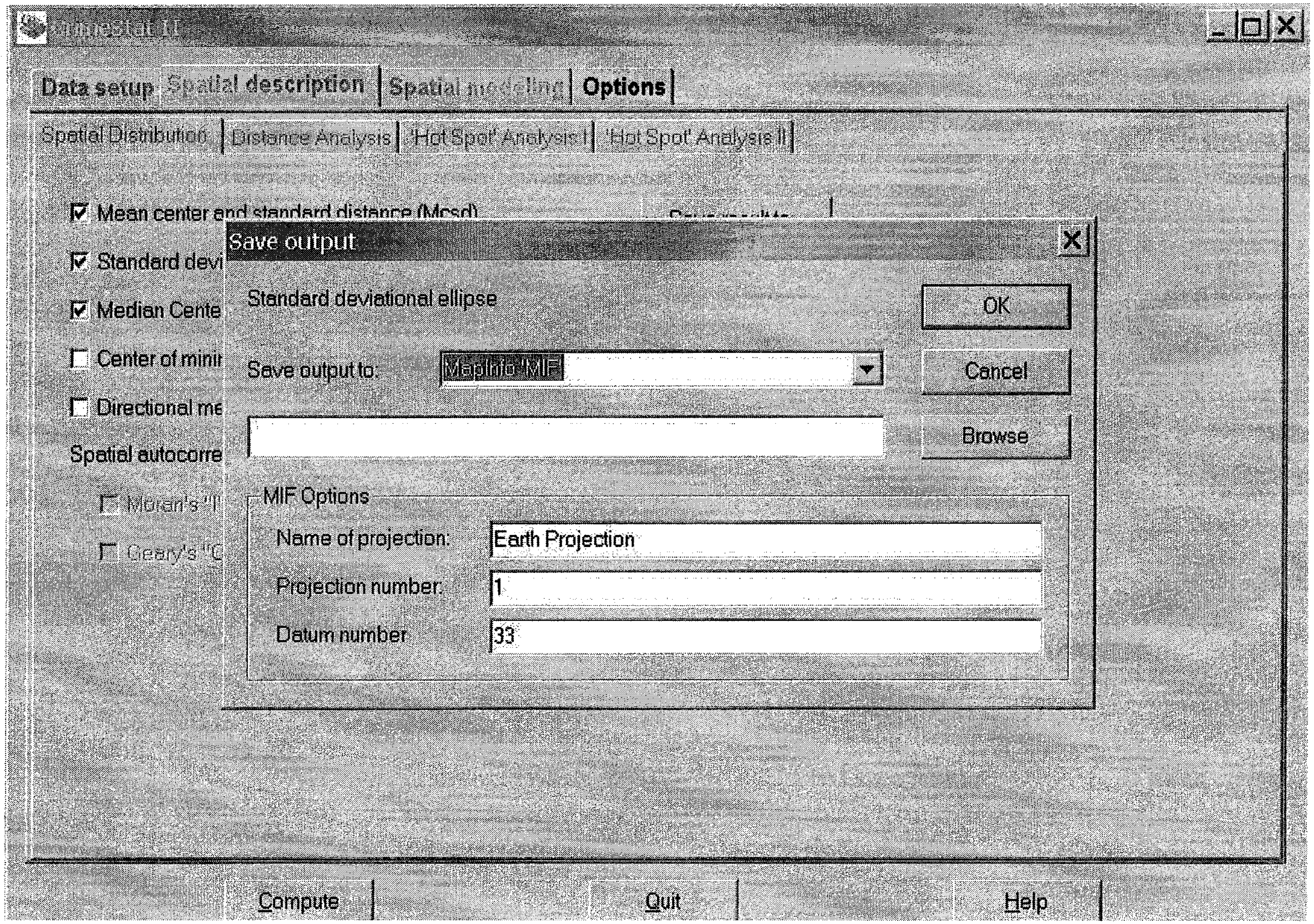
MIF Options

Name of projection: Earth Projection

Projection number: 1

Datum number: 33

Figure 4.17: *MapInfo* Output Options



☑ Mean center and standard distance (Mcsd)

☑ Standard devi

☑ Median Cente

□ Center of mini

□ Directional me

Save output

Standard deviational ellipse

Save output to: MapInfo 'MIF'

MIF Options

Name of projection: Earth Projection

Projection number: 1

Datum number: 33

jurisdiction, rather than on the jurisdiction itself. In Baltimore County, for instance, analysis is done both for the jurisdiction as a whole as well as by individual precincts. Below in Figure 4.18 are the standard deviational ellipses for 1996 auto thefts for June and July in Precinct 11 of Baltimore County. As can be seen, there was a spatial shift that occurred between June and July of that year, the result most probably of increased vacation travel to the Chesapeake Bay. While the comparison is very simple, involving looking at the graphical object created by *CrimeStat,* such a month to month comparison can be useful for police departments because it points to a shift in incident patterns, allowing the police department to reorient their patrol units.

Example 2: Serial burglaries in Baltimore City and Baltimore County

The second example illustrates a rash of burglaries that occurred on both sides of the border of Baltimore City and Baltimore County. On one hand there were ten residential burglaries that occurred on the western edge of the City/County border within a short time period of each other and, on the other hand, there were 13 commercial burglaries that occurred in the central part of the metropolitan areas. Both police departments suspected that these two sets were the work of a serial burglar (or group of burglars). What they were not sure about was whether the two sets of burglaries were done by the same individuals or by different individuals.

The number of incidents involved are too small for significance testing; only one of the parameters tested was significant and that could easily be due to chance. However, the police do have to make a guess about the possible perpetrator even with limited information. Let's use *CrimeStat* to try and make a decision about the distributions.

Figure 4.19 illustrates these distributions. The thirteen commercial burglaries are shown as squares while the ten residential burglaries are shown as triangles. Figure 4.20 plots the mean centers of the two distributions. They are close to each other, but not identical. An initial hunch would suggest that the robberies are committed by two perpetrators (or groups of perpetrators), but the mean centers are not different enough to truly confirm this expectation. Similarly, figure 4.21 plots the center of minimum distance. Again, there is a difference in the distribution, but it is not great enough to truly rule out the single perpetrator theory.

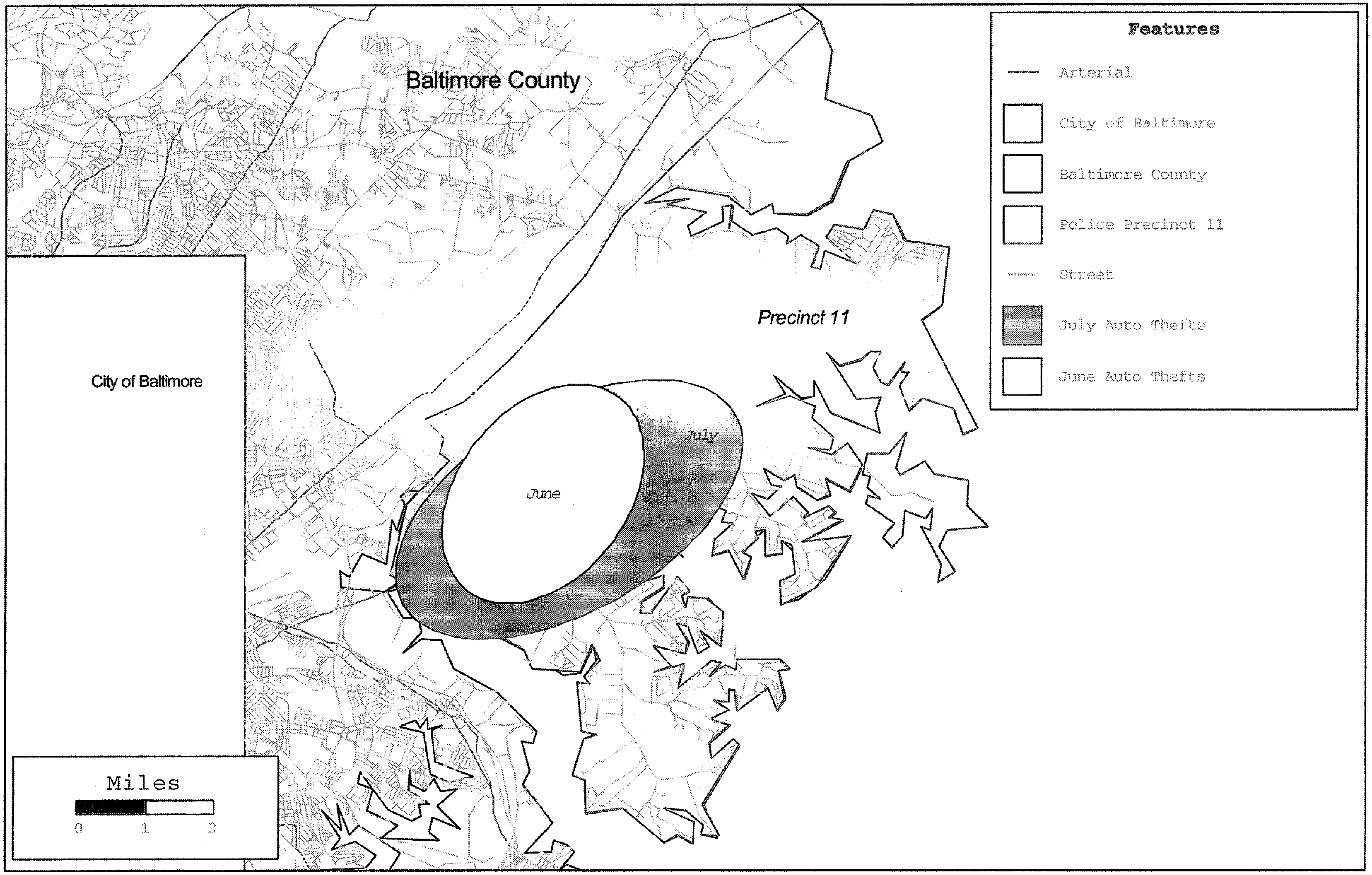
Figure 4.22 plots the raw standard deviations, expressed as a rectangle by *CrimeStat*. The dispersion of incidents overlaps to a sizeable extent and the area defined by the rectangle is approximately the same. In other words, the search area of the perpetrator or perpetrators is approximately the same. This might argue for a single perpetrator, rather than two. Figure 4.23 shows the standard distance deviation of the two sets of incidents. Again, there is sizeable overlap and the search radiuses are approximately the same.

Only with the standard deviational ellipse, however, is there a fundamental difference between the two distributions (figure 4.24). The pattern of commercial robberies is falling along a northeast-southwest orientation while that for residential robberies along

Figure 4.18:

Vehicle Theft Change in Precinct 11

Standard Deviational Ellipses for June and July 1996



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Figure 4.19:

Identifying Serial Burglars

Incident Distribution of Two Serial Offenders

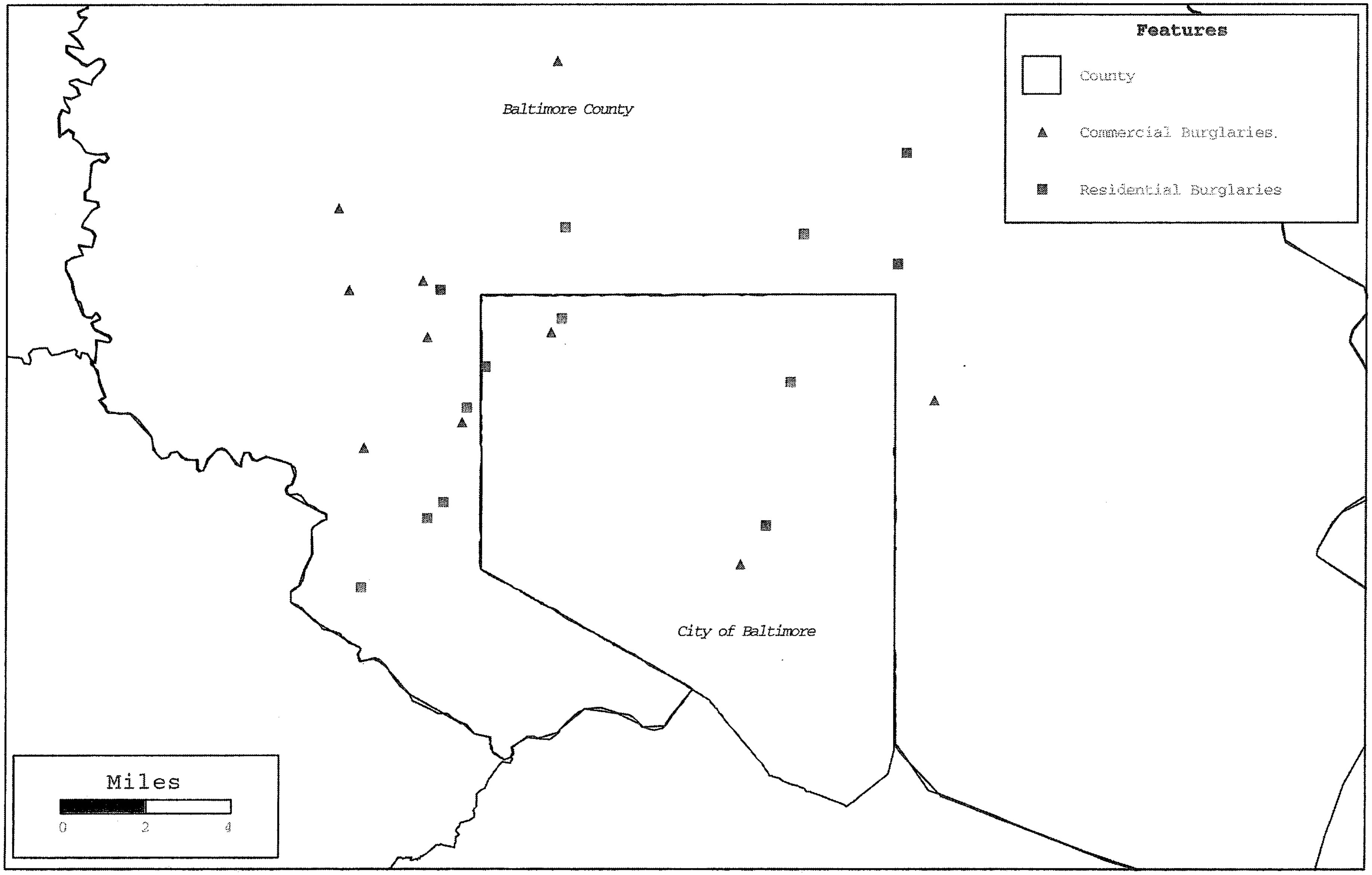


Figure 4.20:

Identifying Serial Burglars

Mean Centers of Incidents for Two Serial Offenders

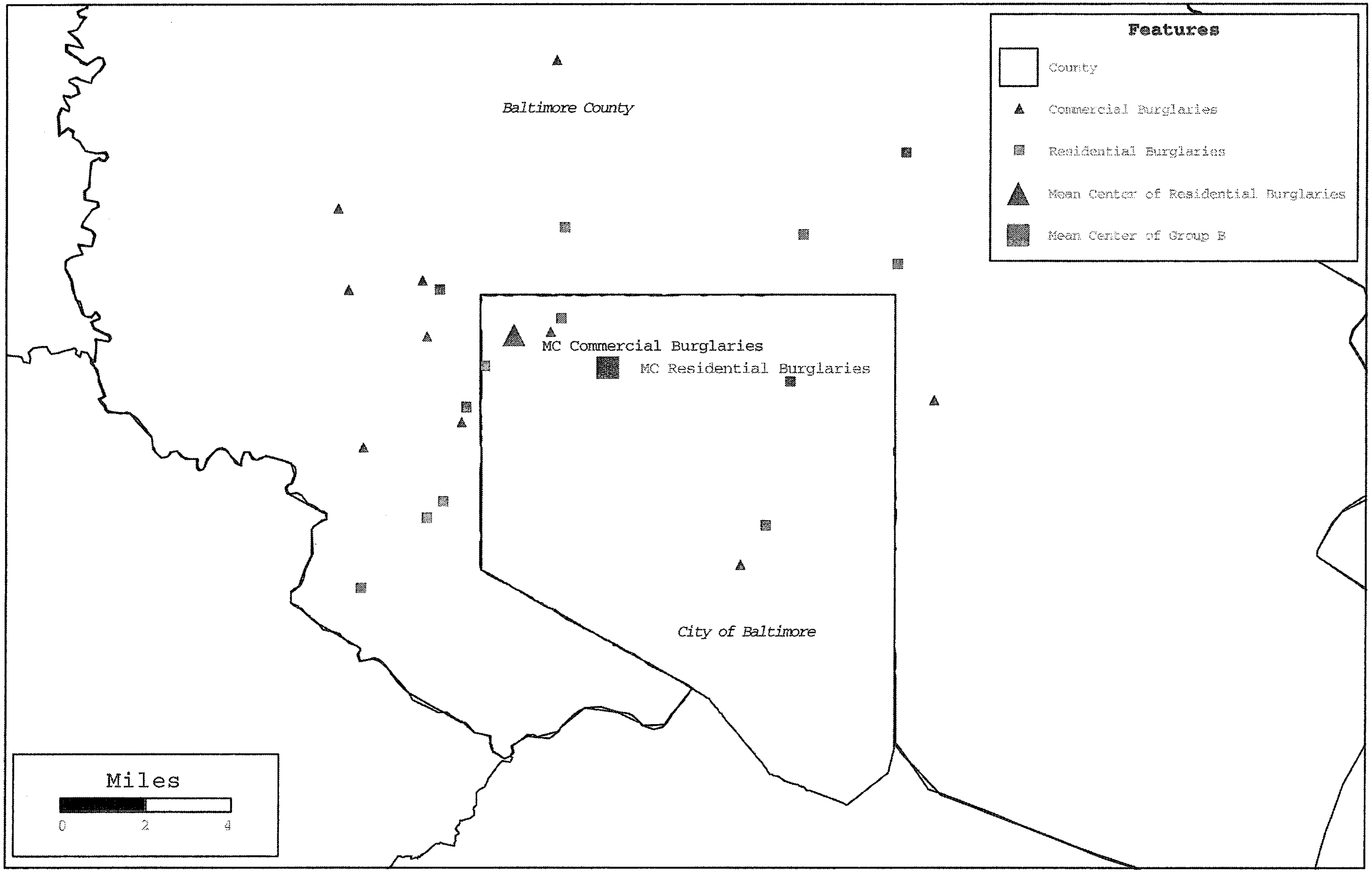


Figure 4.21:

Identifying Serial Burglars

Center of Minimum Distances for Incidents for Two Serial Offenders

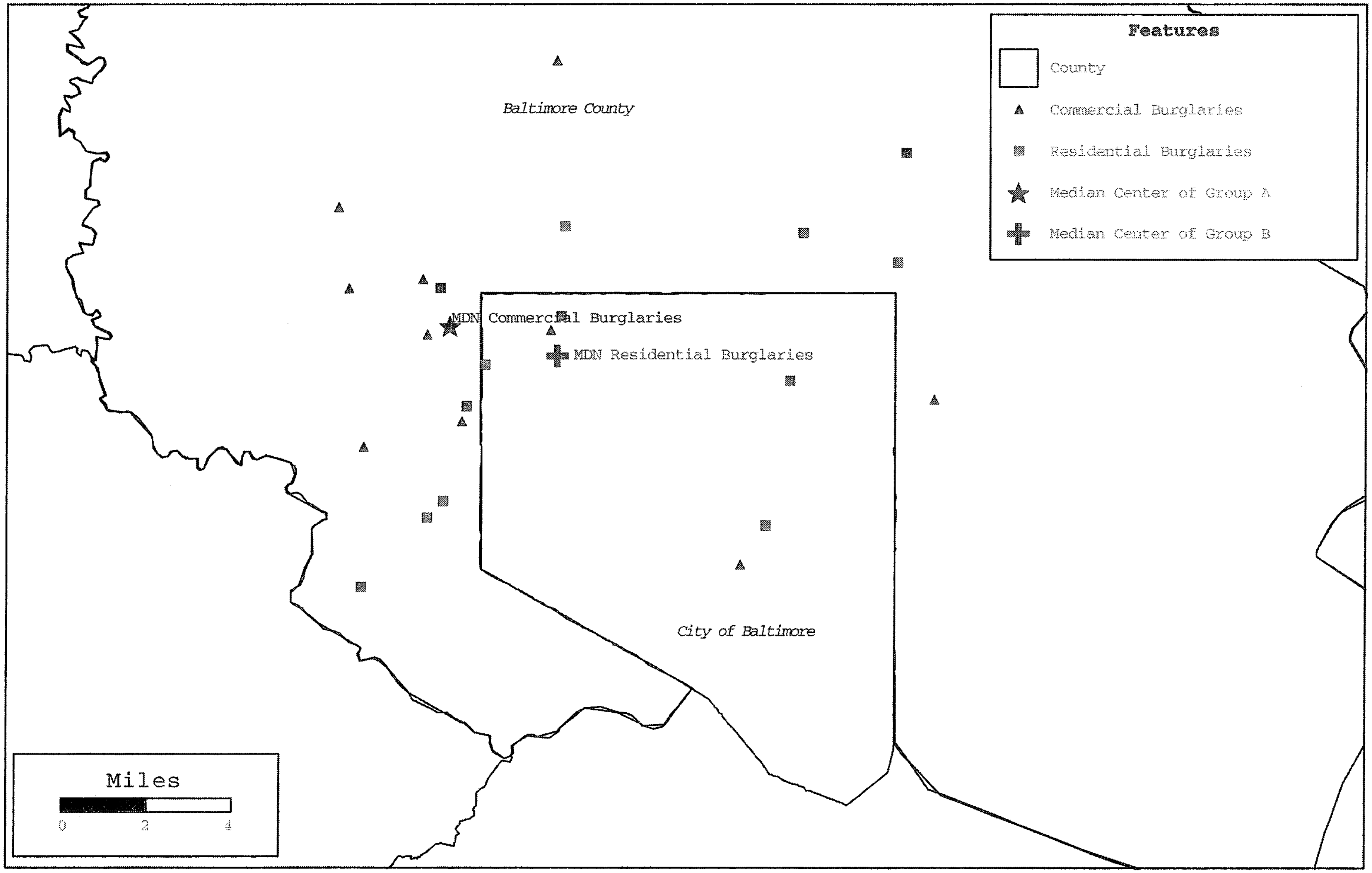


Figure 4.22:

Identifying Serial Burglars

Standard Deviations of Incidents for Two Serial Offenders

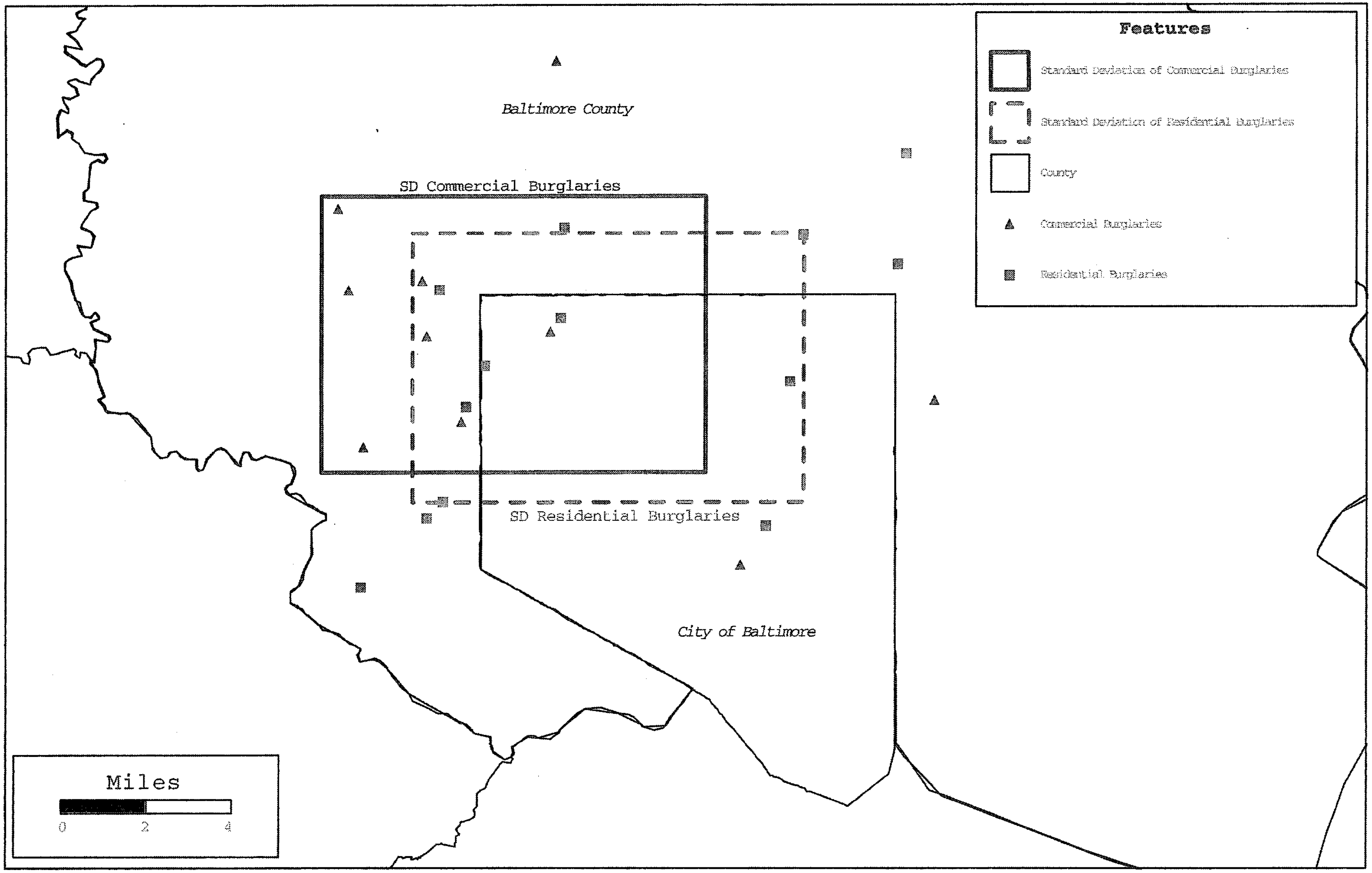


Figure 4.23:

Identifying Serial Burglars

Standard Distance Deviation of Incidents for Two Serial Offenders

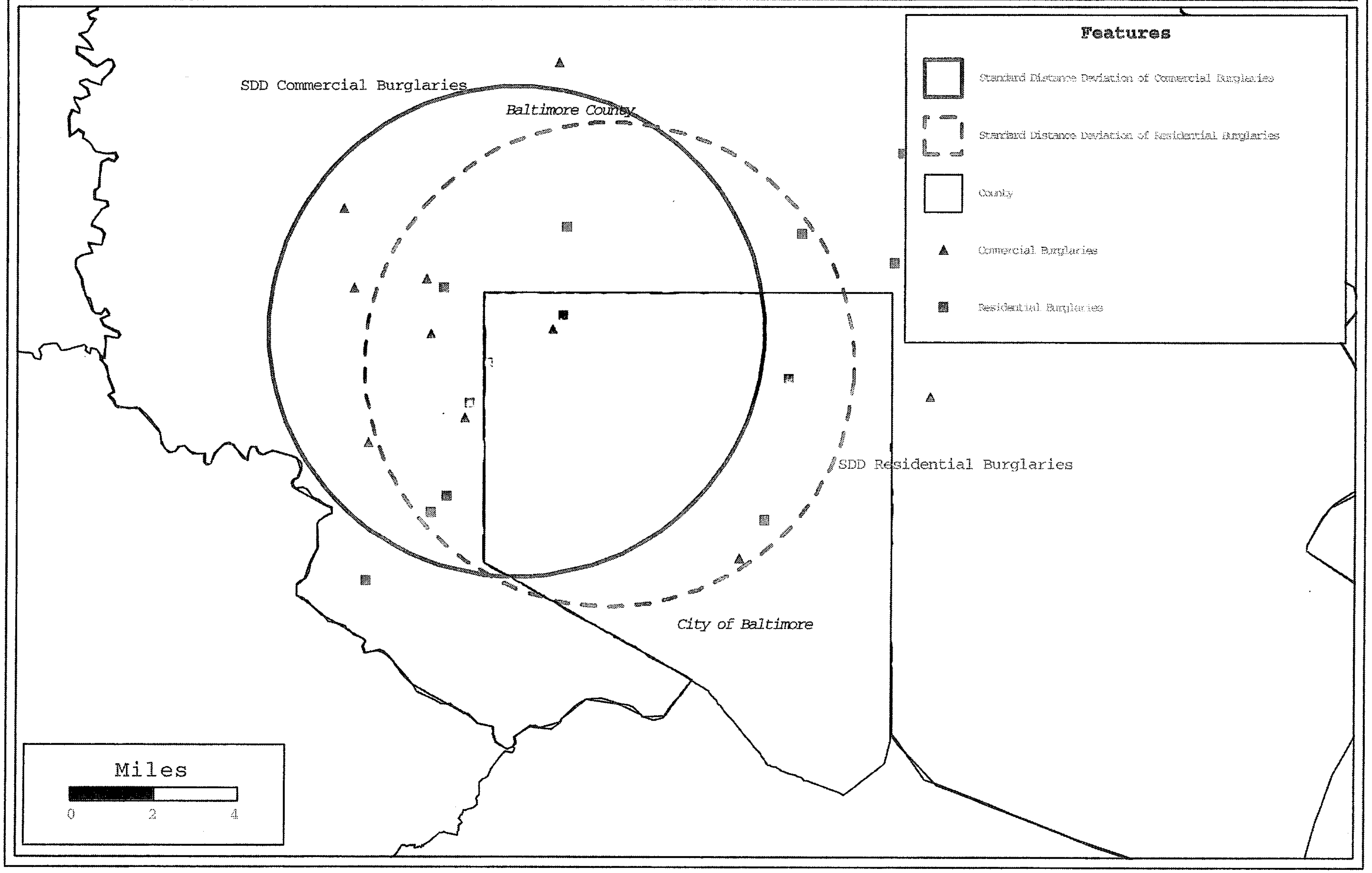
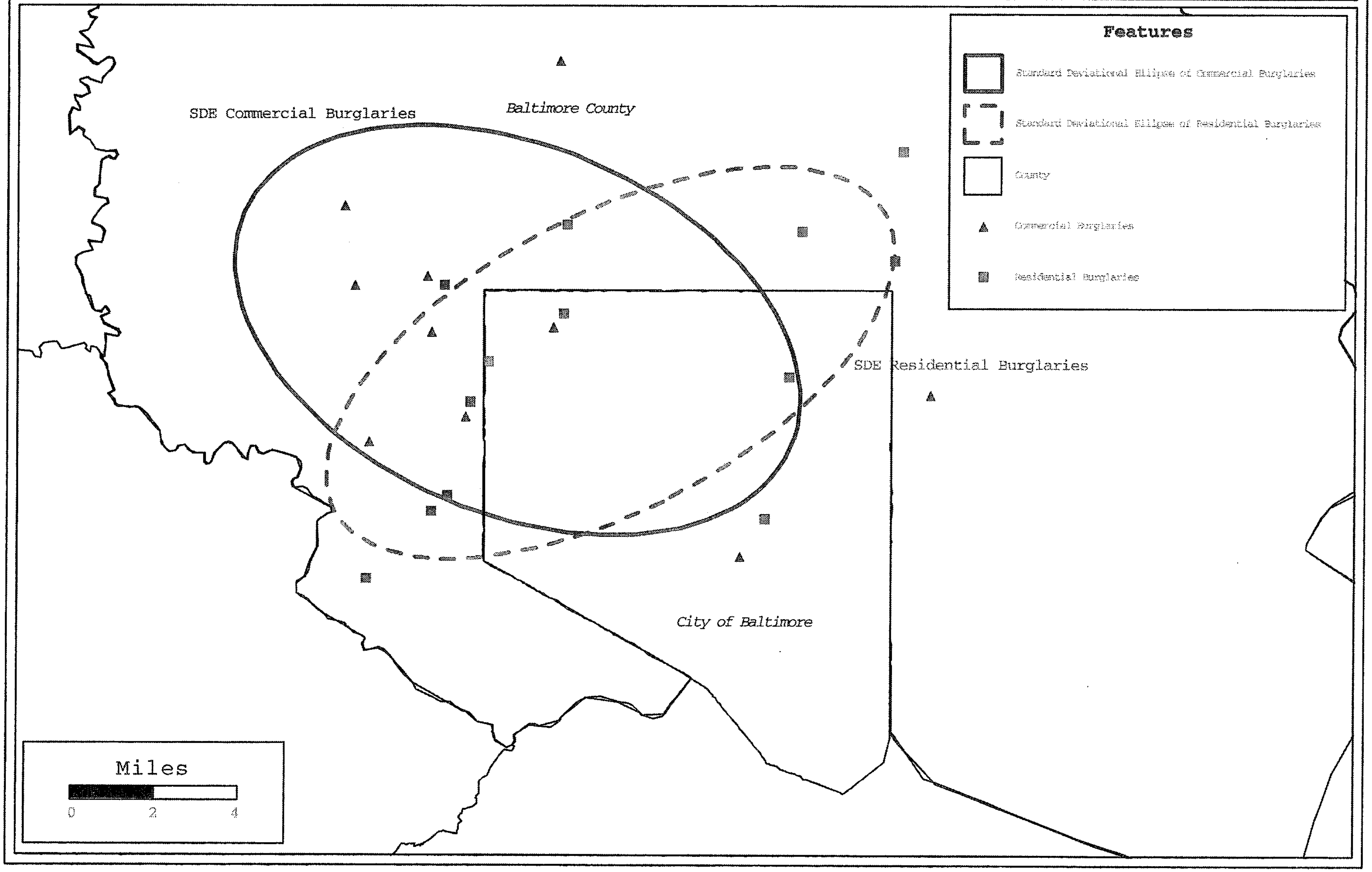

SDD Commercial Burglaries
Baltimore County
SDD Residential Burglaries
City of Baltimore
Features
Standard Distance Deviation of Commercial Burglaries
Standard Distance Deviation of Residential Burglaries
County
Commercial Burglaries
Residential Burglaries
Miles
0
2
4

Figure 4.24:

Identifying Serial Burglars

Standard Deviational Ellipse of Incidents for Two Serial Offenders


SDE Commercial Burglaries
Baltimore County
SDE Residential Burglaries
City of Baltimore
Features
Standard Deviational Ellipse of Commercial Burglaries
Standard Deviational Ellipse of Residential Burglaries
County
Commercial Burglaries
Residential Burglaries
Miles
0
2
4

a northwest-southeast axis. In other words, when the orientation of the incidents is examined, as defined by the standard deviational ellipse, there are two completely opposite patterns. Unless this difference can be explained by an obvious factor (e.g., the distribution of commercial establishments), it is probable that the two sets of robberies were committed by two different perpetrators (or groups of perpetrators).

Directional Mean and Variance

Centrographic statistics utilize the coordinates of a point, defined as an X and Y value on either a spherical or projected/Cartesian coordinate system. There is another type of metric that can be used for identifying incident locations, namely a *polar coordinate* system. A *vector* is a line with direction and length. In this system, there is a reference vector (usually 0^0 due North) and all locations are defined by angular deviations from this reference vector. By convention, angles are defined as deviations from 0^0, clockwise through 360^0. Note the measurement scale is a circle which returns back on itself (i.e. 0^0 is also 360^0). Point locations can be represented as vectors on a polar coordinate system.

With such a system, ordinary statistics cannot be used. For example, if there are five points which on the northern side of the polar coordinate system and are defined by their angular deviations as 0^0, 10^0, 15^0, 345^0, and 350^0 from the reference vector (moving clockwise from due North), the statistical mean will produce an erroneous estimate of 144^0. This vector would be southeast and will lie in an opposite direction from the distribution of points.

Instead, statistics have to be calculated by trigonometric functions. The input for such a system is a set of vectors, defined as angular deviations from the reference vector and a distance vector. Both the angle and the distance vector are defined with respect to an origin. The routine can calculate angles directly or can convert all X and Y coordinates into angles with a bearing from an origin. For reading angles directly, the input is a set of vectors, defined as angular deviations from the reference vector. *CrimeStat* calculates the mean direction and the circular variance of a series of points defined by their angles. On the primary file screen, the user must select Direction (angles) as the coordinate system.

If the angles are to be calculated from X/Y coordinates, the user must define an origin location. On the reference file page, the user can select among three origin points:

1. The lower-left corner of the data set (the minimum X and Y values). This is the default setting.

2. The upper-right corner of the data set (the maximum X and Y values); and

3. A user-defined point.

Users should be careful about choosing a particular location for an origin, either lower-left, upper-right or user-defined. If there is a point at that origin, *CrimeStat* will drop that case since any calculations for a point with zero distance are indeterminate.

Users should check that there is no point at the desired origin. If there is, then the origin should be adjusted slightly so that no point falls at that location (e.g., taking slightly smaller X and Y values for the lower-left corner or slightly larger X and Y values for the upper right corner).

The routine converts all X and Y points into an angular deviation from true North relative to the specified origin and a distance from the origin. The bearing is calculated with different formulae depending on the quadrant that the point falls within.

First Quadrant

With the lower-left corner as the origin, all angles are in the first quadrant. The clockwise angle, θ_i is calculated by

$$\theta_i = \quad \text{Arctan}\left[\frac{\text{Abs}(X_i - X_O)}{\text{Abs}(Y_i - Y_O)}\right] \tag{4.22}$$

where X_i is the X-value of the point, Y_i is the Y-value of the point, X_O is the X-value of the origin, and Y_O is the Y-value of the origin.

The angle, θ_i, is in radians and can be converted to polar coordinate degrees using:

$$\theta_i \text{ (degrees)} = \theta_i \text{ (radians)} * 180/\pi \tag{4.23}$$

Third Quadrant

With the upper-right corner as the origin, all angles are in the third quadrant. The clockwise angle, θ_i, is calculated by

$$\theta_i = \quad \pi + \quad \text{Arctan}\left[\frac{\text{Abs}(X_i - X_O)}{\text{Abs}(Y_i - Y_O)}\right] \tag{4.24}$$

where the angle, θ_i, is again in radians. Since there are 2π radians in a circle, π radians is 180^0. Again, the angle in radians can be converted into degrees with formula 4.23 above.

Second and Fourth Quadrants

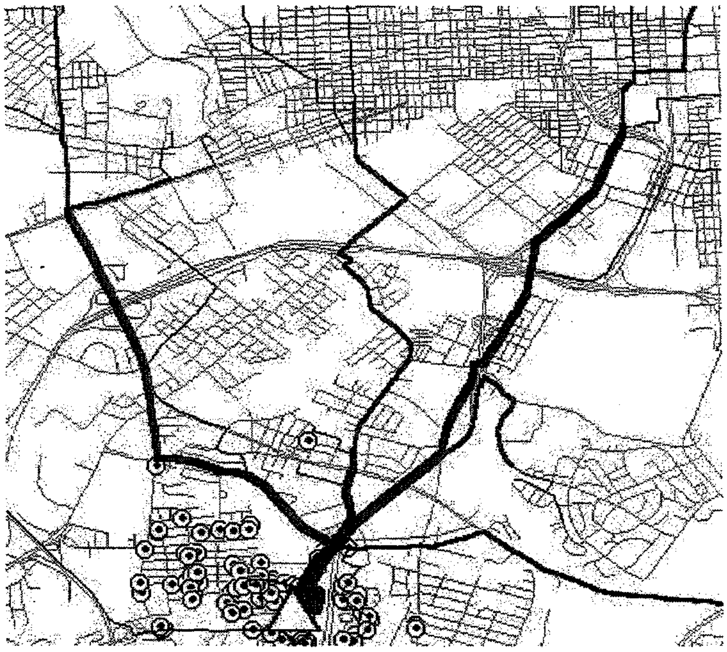
When the origin is user-defined, each point must be evaluated as to which quadrant it is in. The second and fourth quadrants define the clockwise angle, θ_i, differently

Using Spatial Measures of Central Tendency with Network Analyst to Identify Routes Used by Motor Vehicle Thieves

Philip R. Canter
Baltimore County Police Department
Towson, Maryland

Motor vehicle thefts have been steadily declining countywide over the last 5 years, but one police precinct in southwest Baltimore County was experiencing significant increases over several months. Cases were concentrated in several communities, but directed deployment and saturated patrols had minimal impact. In addition to increasing patrols in target communities, the precinct commander was interested in deploying police on roads possibly used by motor vehicle thieves. Police analysts had addresses for theft and recovery locations; it was a matter of using the existing highway network to connect the two locations.

To avoid analyzing dozens of paired locations, analysts decided to set up a database using one location representing the origin of motor vehicle thefts for a particular community. The origin was computed using *CrimeStat*'s median center for motor vehicle theft locations reported for a particular community. The median center is the position of minimum average travel and is less affected by extreme locations compared to the arithmetic mean center. The database consisted of the median center paired with a recovery location. Using Network Analyst, a least-effort route was computed for cases reported by community. A count was assigned to each link along a roadway identified by Network Analyst. Analysts used the count to thematically weight links in ArcView. The precinct commander deployed resources along these routes with orders to stop suspicious vehicles. This operation resulted in 27 arrests, and a reduction in motor vehicle thefts.

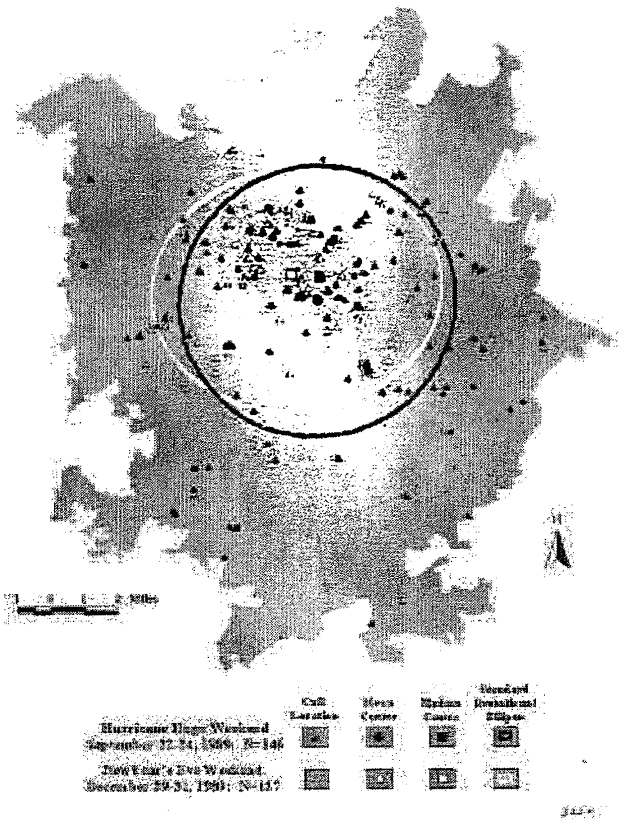


Distance Analysis
Man With A Gun Calls For Service
Charlotte, N.C., 1989

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

Hurricane Hugo arrived on Friday, September 22, 1989 in Charlotte, North Carolina. That weekend experienced the highest counts of *Man With A Gun* calls for service for the year. The locations of the calls during the Hugo Weekend are compared with the following New Year's Eve weekend.

CrimeStat was used to compare the two weekends. Compared to the New Year's Eve weekend: 1) Hugo's mean and median centers are more easterly; 2) Hugo's ellipse is larger and more circular; and 3) Hugo's ellipse shifts more to the east and southeast. The abrupt spatial change of *Man With A Gun* calls during a natural disaster might indicate more instances of defensive gun use for protection of property.



$$\theta_i = 0.5\pi + \text{Arctan}\left[\frac{\text{Abs}(Y_i - Y_O)}{\text{Abs}(X_i - X_O)}\right] \qquad (4.25)$$

$$\theta_i = 1.5\pi + \text{Arctan}\left[\frac{\text{Abs}(Y_i - Y_O)}{\text{Abs}(X_i - X_O)}\right] \qquad (4.26)$$

Once all X/Y coordinates are converted into angles, the mean angle is calculated.

Mean Angle

With either angular input or conversion from X/Y coordinates, the *Mean Angle* is the resultant of all individual vectors (i.e., points defined by their angles from the reference vector). It is an angle that summarizes the mean direction. Graphically, a *resultant* is the sum of all vectors and can be shown by laying each vector end to end. Statistically, it is defined as

$$\text{Mean angle} = \bar{\theta} = \text{Abs}\left\{\text{Arctan}\left[\frac{\Sigma\, d_i \sin\theta_i}{\Sigma\, d_i \cos\theta_i}\right]\right\} \qquad (4.27)$$

where the summation of sines and cosines is over the total number of points, i, defined by their angles, θ_i. Each angle, θ_i, can be weighted by the length of the vector, d_i. In an unweighted angle, d_i is assumed to be of equal length, 1. The absolute value of the ratio of the sum of the weighted sines to the sum of the weighted cosines is taken. All angles are in radians. In determining the mean angle, the quadrant of the resultant must be identified:

1. If $\Sigma \sin\theta_i>0$ and $\Sigma \cos\theta_i>0$, then $\bar{\theta}$ can be used directly as the mean angle

2. If $\Sigma \sin\theta_i>0$ and $\Sigma \cos\theta_i<0$, then the mean angle is $\pi/2 + \bar{\theta}$.

3. If $\Sigma \sin\theta_i<0$ and $\Sigma \cos\theta_i<0$, then the mean angle is $\pi + \bar{\theta}$.

4. If $\Sigma \sin\theta_i<0$ and $\Sigma \cos\theta_i>0$, then the mean angle is $1.5\pi + \bar{\theta}$.

Formulas 4.22, 4.24, 4.25 and 4.26 above are then used to convert the directional mean back to an X/Y coordinate, depending on which coordinate it falls within.

Circular Variance

The dispersion (or variance) of the angles are also defined by trigonometric functions. The unstandardized variance, R, is sometimes called the *sample resultant length* since it is the resultant of all vectors (angles).

$$R = \text{SQRT} [(\Sigma d_i \sin \theta_i)^2 + (\Sigma d_i \cos \theta_i)^2] \qquad (4.28)$$

where d_i is the length of vector, i, with an angle (bearing) for the vector of θ_i. For the unweighted sample resultant, d_i is 1.

Because R increases with sample size, it is standardized by dividing by N to produce a *mean resultant length*.

$$\bar{R} = \frac{R}{N} \qquad (4.29)$$

where N is the number points (sample size).

Finally, the average distance from the origin, D, is calculated and the *circular variance* is calculated by

$$\text{Circular variance} = \frac{1}{D} \{D - \frac{R}{N}\} = (D - \bar{R})/D = 1 - \frac{\bar{R}}{D} \qquad (4.30)$$

This is the standardized variance which varies from 0 (no variability) to 1 (maximum variability). The details of the derivations can be found in Burt and Barber (1996) and Gaile and Barber (1980).

Mean Distance

The mean distance, $\bar{d}$, is calculated directly from the X and Y coordinates. It is identified in relation to the defined origin.

Directional Mean

The directional mean is calculated as the intersection of the mean angle and the mean distance. It is not a unique position since distance and angularity are independent dimensions. Thus, the directional mean calculated using the minimum X and minimum Y location as the reference origin (the 'lower left corner') will yield a different location from the directional mean calculated using the maximum X and maximum Y location as the origin (the 'upper right corner'). There is a weighted and unweighted directional mean.

Though *CrimeStat* calculates the location, users should be aware of the non-uniqueness of the location. The unweighted directional mean can be output with a 'Dm' prefix. The weighted directional mean is not output.

Triangulated Mean

The triangulated mean is defined as the intersection of the two vectors, one from the lower-left corner of the study area (the minimum X and Y values) and the other from the upper-right corner of the study area (the maximum X and Y values). It is calculated by estimating mean angles from each origin (lower left and upper right corners), translating these into equations, and finding the point at which these equations intersect (by setting the two functions equal to each other).

Directional Mean Output

The directional mean routine outputs nine statistics:

1. The sample size;
2. The unweighted mean angle;
3. The weighted mean angle;
4. The unweighted circular variance;
5. The weighted circular variance;
6. The mean distance;
7. The intersection of the mean angle and the mean distance;
8. The X and Y coordinates for the triangulated mean; and
9. The X and Y coordinates for the weighted triangulated mean.

The directional mean and triangulated mean can be saved as an *ArcView* 'shp', *MapInfo* 'mif', or *Atlas*GIS* 'bna' file. The unweighted directional mean - the intersection of the mean angle and the mean distance is output with the prefix 'Dm' while the unweighted triangulated mean location is output with a 'Tm' prefix. The weighted triangulated mean is output with a 'TmWt' prefix. The directional mean can be saved as an *ArcView* 'shp', *MapInfo* 'mif', or *Atlas*GIS* 'bna' file. The letters 'Dm' are prefixed to the user defined file name. See the example below.

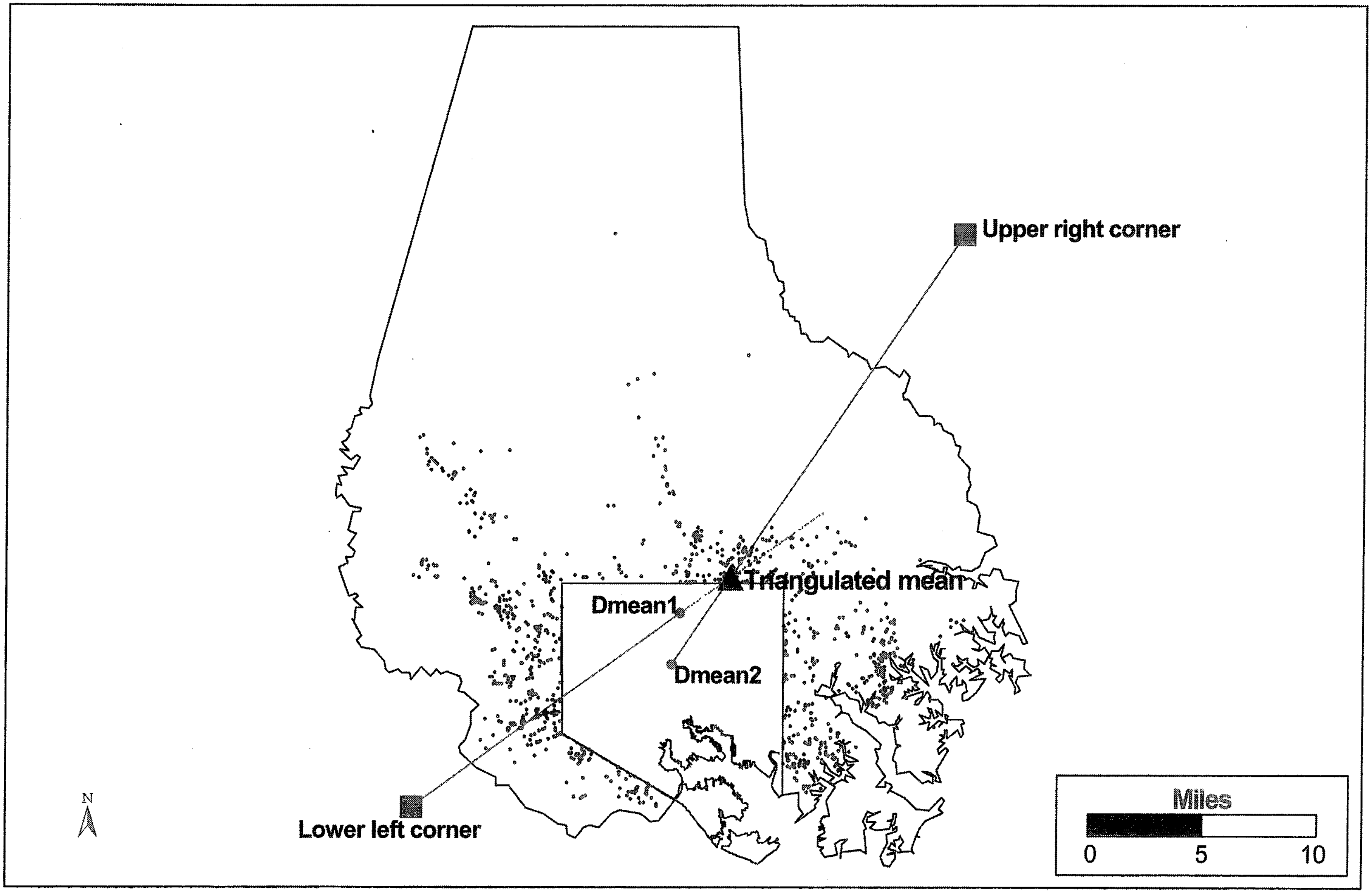
Figure 4.25 shows the unweighted triangular mean for 1996 Baltimore County robberies and compares it to the two directional means calculated using the lower-left corner (Dmean1) and the upper-right corner (Dmean2) respectively as origins. As can be seen, the two directional means fall at different locations. Lines have been drawn from each origin point to their respective directional means and are extended until they intersect. As seen, the triangulated mean falls at the location where the two vectors (i.e., mean angles) intersect.

Because the triangulated mean is calculated with vector geometry, it will not necessarily capture the central tendency of a distribution. Asymmetrical distributions can cause it to be placed in peripheral locations. On the other hand, if the distribution is

Figure 4.25:

Triangulated Mean for Baltimore County Robberies

Defined by the Intersection of Two Mean Angles



relatively balanced in each direction, it can capture the center of orientation perhaps better than other means, as figure 4.25 shows.

Appendix B includes a discussion of how to formally tests the mean direction between two different distributions.

Spatial Autocorrelation

The concept of *spatial autocorrelation* is one of the most important in spatial statistics. Spatial *independence* is an arrangement of incident locations such that there are no spatial relationships between any of the incidents. The intuitive concept is that the location of an incident (e.g., a street robbery, a burglary) is unrelated to the location of any other incident. The opposite condition - spatial autocorrelation, is an arrangement of incident locations where the location of points are related to each other, that is they are not statistically independent of one another. In other words, spatial autocorrelation is a spatial arrangement where spatial independence has been violated.

When events or people or facilities are clustered together, we refer to this arrangement as *positive* spatial autocorrelation. Conversely, an arrangement where people, events or facilities are dispersed is referred to as *negative* spatial autocorrelation; it is a rarer arrangement, but does exist (Levine, 1999).

Many, if not most, social phenomena are spatially autocorrelated. In any large metropolitan area, most social characteristics and indicators, such as the number of persons, income levels, ethnicity, education, employment, and the location of facilities are not spatially independent, but tend to be concentrated.

There are practical consequences. Police and crime analysts know from experience that incidents frequently cluster together in what are called 'hot spots'. This non-random arrangement allows police to target certain areas or zones where there are high concentrations as well as prioritize areas by the intensity of incidents. Many of the incidents are committed by the same individuals. For example, if a particular neighborhood had a concentration of street robberies over a time period (e.g., a year), many of these robberies will have been committed by the same perpetrators. Statistical dependence between events often has common causes.

Statistically, however, non-spatial independence suggests many statistical tools and inferences are inappropriate. For example, the use of correlation coefficients or Ordinary Least Squares regression (OLS) to predict a consequence (e.g., the correlates or predictors of burglaries) assumes that the observations have been selected randomly. If the observations, however, are spatially clustered in some way, the estimates obtained from the correlation coefficient or OLS estimator will be biased and overly precise. They will be biased because the areas with higher concentration of events will have a greater impact on the model estimate and they will overestimate precision because, since events tend to be concentrated, there are actually fewer number of independent observations than are being assumed. This concept of spatial autocorrelation underlies almost all the spatial statistics tools which are included in *CrimeStat*. We will return to the concept in each of the next three chapters because the concept is implicit in all the tools that will be discussed.

Indices of Spatial Autocorrelation

There are a number of formal statistics which attempt to measure spatial autocorrelation. This include simple indices, such as the Moran's I" or Geary's C statistic; derivatives indices, such as Ripley's K statistic (Ripley, 1976) or the application of Moran's I to individual zones (Anselin, 1995); and multivariate indices, such as the use of a spatial autocorrelation parameter in a bivariate regression model (Cliff and Ord, 1973; Griffith, 1987) or the use of a spatially-lagged dependent variable in a multiple variable regression model (Anselin, 1992). The simple indices attempt to identify whether spatial autocorrelation exists for a single variable, while the more complicated indices attempt to estimate the effect of spatial autocorrelation on other variables.

CrimeStat includes two simple indices: Moran's I statistic and Geary's C statistic. They are very similar indices and are often used in conjunction. The Moran statistic is slightly more robust than the Geary, but the Geary is often used as well.

Moran's I Statistic

Moran's I statistic (Moran, 1950) is one of the oldest indicators of spatial autocorrelation. It is applied to zones or points which have continuous variables associated with them (intensities). For any continuous variable, X_i, a mean can be calculated and the deviation of any one observation from that mean can also be calculated. The statistic then compares the value of the variable at any one location with the value at all other locations (Ebdon, 1985; Griffith, 1987; Anselin, 1992). Formally, it is defined as

$$I = \frac{N \ \Sigma_i \Sigma_j W_{ij} (X_i - \bar{X})(X_j - \bar{X})}{(\Sigma_i \Sigma_j W_{ij}) \Sigma_i (X_i - \bar{X})^2} \qquad (4.31)$$

where N is the number of cases, X_i is the variable value at a particular location, i, X_j is the variable value at another location (where $i \neq j$), $\underline{X}$ is the mean of the variable and W_{ij} is a weight applied to the comparison between location i and location j.

In Moran's initial formulation, the weight variable, W_{ij}, is a contiguity matrix. If zone j is adjacent to zone i, the interaction receives a weight of 1. Otherwise, the interaction receives a weight of 0. Cliff and Ord (1973) generalized these definitions to include any type of weight. In more current use, W_{ij} is a distance-based weight which is the inverse distance between locations i and j ($1/d_{ij}$). *CrimeStat* uses this interpretation. Essentially, it is a *weighted* Moran's I where the weight is an inverse distance.

The weighted Moran's I is similar to a correlation coefficient in that it compares the sum of the cross-products of values at different locations, two at a time weighted by the inverse of the distance between the locations, with the variance of the variable. Like the correlation coefficient, it varies between -1.0 and + 1.0. When nearby points have similar values, the cross-product is high. Conversely, when nearby points have dissimilar values, the cross-product is low. Consequently, an I value which is high indicates more spatial autocorrelation than an I which is low.

However, unlike the correlation coefficient, the theoretical value of the index does not equal 0 for lack of spatial dependence, but instead a number which is negative but very close to 0.

$$E(I) = -\frac{1}{N-1} \qquad (4.32)$$

Values of I above the theoretical mean, E(I), indicate positive spatial autocorrelation while values of I below the theoretical mean indicate negative spatial autocorrelation.

Adjustment for small distances

CrimeStat calculates the weighted Moran's I formula using equation 4.31. However, there is one problem with this formula that can lead to unreliable results. The distance weights between two locations, W_{ij}, is defined as the reciprocal of the distance between the two points:

$$W_{ij} = \frac{1}{d_{ij}} \qquad (4.33)$$

Unfortunately, as d_{ij} becomes small, then W_{ij} becomes very large, approaching infinity as the distance between the points approaches 0. If the two zones were next to each other, which would be true for two adjacent blocks for example, then the pair of observations would have a very high weight, sufficient to distort the I value for the entire sample. Further, there is a scale problem which alters the value of the weight. If the zones are police precincts, for example, then the minimum distance between precincts will be a lot larger than the minimum distance between a smaller type of geographical unit, such as blocks. We need to take into account these different scales.

CrimeStat includes an adjustment for small distances so that the maximum weight can never be greater than 1.0. The adjustment scales distances to one mile, which is a typical distance unit in the measurement of crime incidents. When the small distance adjustment is turned on, the minimal distance is automatically scaled to be one mile. The formula used is

$$W_{ij} = \frac{\text{one mile}}{\text{one mile} + d_{ij}} \qquad (4.34)$$

in the units are specified. For example, if the distance units, d_{ij}, are calculated as feet, then

$$W_{ij} = \frac{5{,}280}{5{,}280 + d_{ij}}$$

where 5,280 is the number of feet in a mile. This has the effect of insuring that the weight of a particular pair of point locations will not have an undue influence on the overall statistic. The traditional measure of I is the default condition in *CrimeStat* (figure 4.26), but the user can turn on the small distance adjustment.

Testing the significance of the weighted Moran's I

The empirical distribution can be compared with the theoretical distribution by dividing by an estimate of the theoretical standard deviation

$$Z(I) = \frac{I - E(I)}{S_{E(I)f}} \qquad (4.35)$$

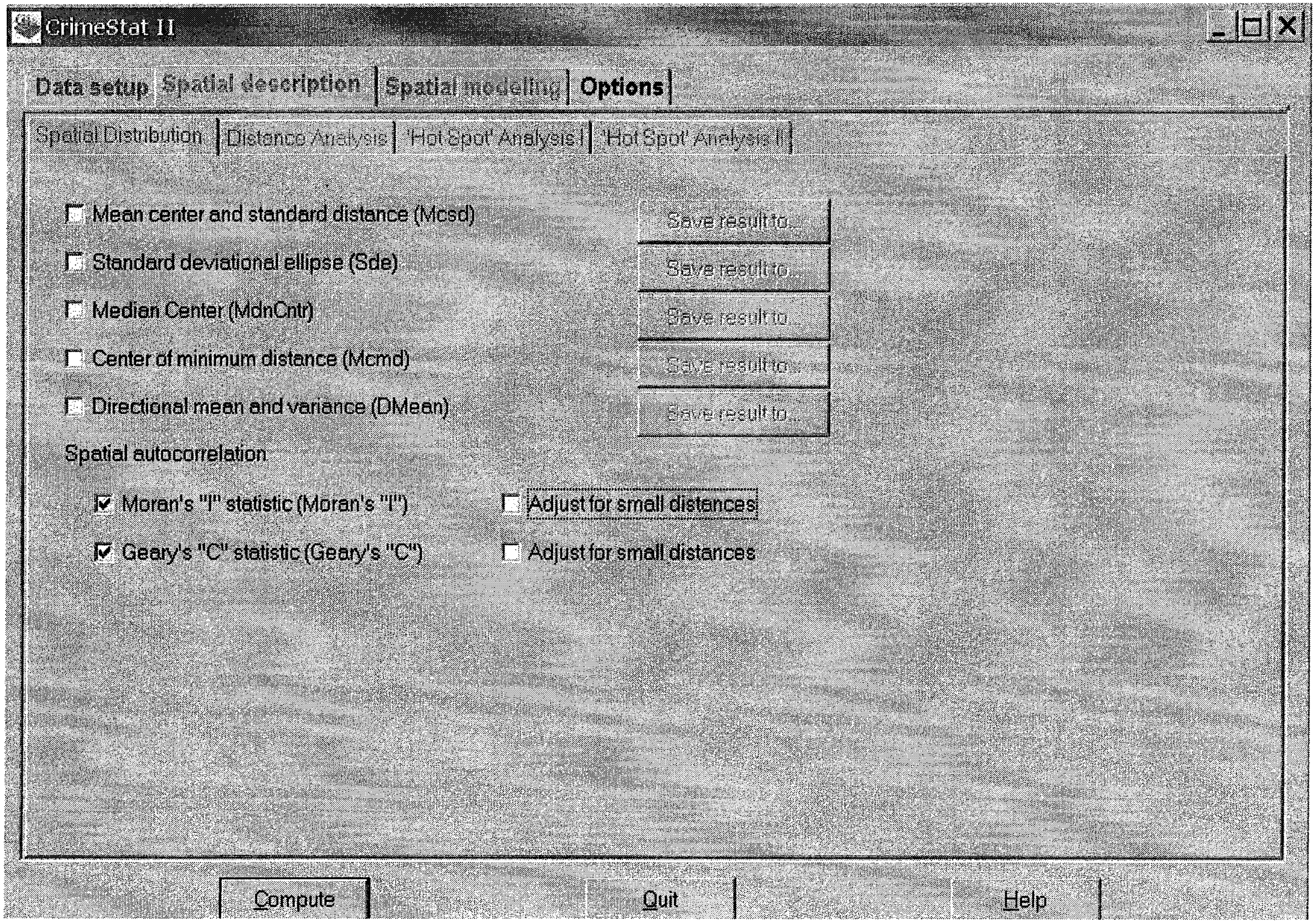
where I is the empirical value calculated from a sample, E(I) is the theoretical mean of a random distribution and $S_{E(I)}$ is the theoretical standard deviation of E(I).

There are several interpretations of the theoretical standard deviation which affect the particular statistic used for the denominator as well as the interpretation of the significance of the statistic (Anselin, 1992). The most common assumption is to assume that the standardized variable, Z(I), has a sampling distribution which follows a standard normal distribution, that is with a mean of 0 and a variance of 1. This is called the *normality* assumption.[6] A second interpretation assumes that each observed value could have occurred at any location, that is the location of the values and their spatial arrangement is assumed to be unrelated. This is called the *randomization* assumption and has a slightly different formula for the theoretical standard deviation of I.[7] *CrimeStat* outputs the Z-values for both the normality and randomization assumptions (figure 4.27).

Example 3: Testing auto thefts with the weighted Moran's I

To illustrate the use of Moran's I with point locations requires data to have intensity values associated with each point. Since most crime incidents are represented as a single point, they do not naturally have associated intensities. It is necessary, therefore, to adapt crime data to fit the form required by Moran's I. One way to do this is assign crime incidents to geographical areas and count the number of incidents per area.

Figure 4.26: **Selecting Spatial Autocorrelation Statistics**



□ Mean center and standard distance (Mcsd)

□ Standard deviational ellipse (Sde)

□ Median Center (MdnCntr)

□ Center of minimum distance (Mcmd)

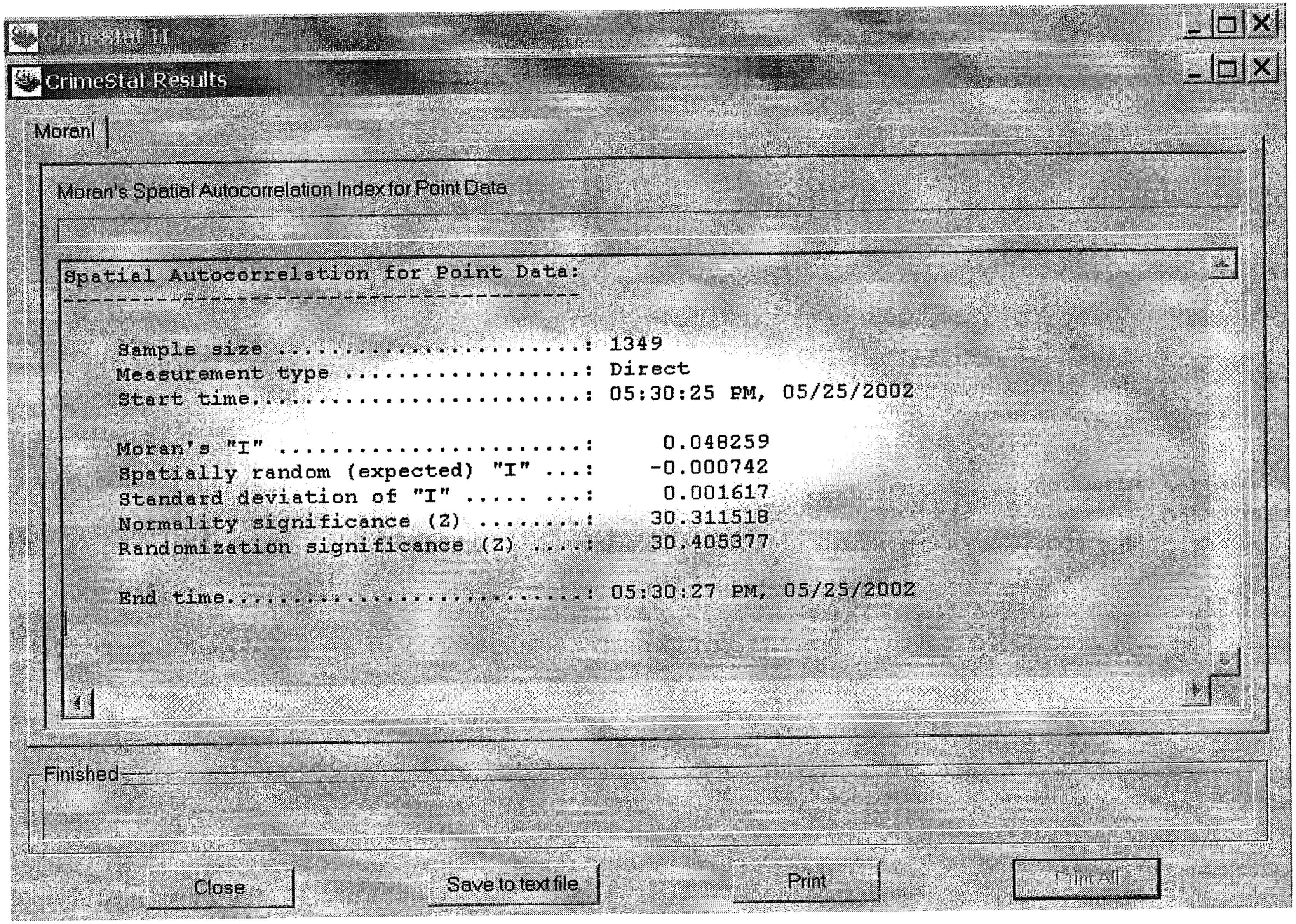
□ Directional mean and variance (DMean)

Spatial autocorrelation

☑ Moran's "I" statistic (Moran's "I") □ Adjust for small distances

☑ Geary's "C" statistic (Geary's "C") □ Adjust for small distances

Figure 4.27: Moran's I Statistic Output



```
Spatial Autocorrelation for Point Data:
--------------------------------------


    Sample size ......................: 1349
    Measurement type .................: Direct
    Start time........................: 05:30:25 PM, 05/25/2002

    Moran's "I" ......................:     0.048259
    Spatially random (expected) "I" ...:    -0.000742
    Standard deviation of "I" ..... ...:     0.001617
    Normality significance (Z) ........:    30.311518
    Randomization significance (Z) ....:    30.405377

    End time..........................: 05:30:27 PM, 05/25/2002
```

Figure 4.28 shows 1996 motor vehicle thefts in both Baltimore County and Baltimore City by individual blocks. With a GIS program, 14,853 vehicle theft locations were overlaid on top of a map of 13,101 census blocks and the number of motor vehicle thefts within each block were counted and then assigned to the block as a variable. The numbers varied from 0 incidents (for 7,675 blocks) up to 46 incidents (for 1 block). The map shows the plot of the number of auto thefts per block.

Clearly, aggregating incident locations to zones, such as blocks, eliminates some information since all incidents within a block are assigned to a single location (the centroid of the block). The use of Moran's I, however, requires the data to be in this format. Using data in this form, Moran's I was calculated using the small distance adjustment because many blocks are very close together. *CrimeStat* calculated I as 0.012464 and the theoretical value of I as -0.000076. The test of significance using the normality assumption gave a Z-value of 125.13, a highly significant value. Below are the calculations.

$$Z(I) = \frac{I - E(I)}{S_{E(I)}} = \frac{0.012464 - (-0.00076)}{0.000100} = 125.13 \ (p \leq .001)$$

In other words, motor thefts are highly and positively spatially autocorrelated. Blocks with many incidents tend to be located close to blocks which also have many incidents and, conversely, blocks with few or no incidents tend to be located close to blocks which also have few or no incidents.

How does this compare with other distributions? Finding positive spatial autocorrelation for auto thefts is not surprising given that there is such a high concentration of population (and, hence, motor vehicles) towards the metropolitan center. For comparison, we ran Moran's I for the population of the blocks (Figure 4.29).[8] With these data, Moran's I for population was 0.001659 with a Z-value of 17.32; the theoretical I is the same since the same number of blocks is being used for the statistic (n=13,101).

Comparing the I value for motor vehicle thefts (0.012464) with that of population (0.00166) suggests that motor vehicle thefts are slightly more concentrated than would be expected on the basis of the population distribution. We can set up an approximate test of this hypothesis. The joint sampling distribution for two variables, such as motor vehicle thefts and population, is not known. However, if we assume that the standard error of the distribution follows a spatially random distribution under the assumption of normality, then equation 4.35 can be applied:

$$Z(I) = \frac{I_{MV} - I_P}{S_{E(I)}} = \frac{0.012464 - 0.001659}{0.000100} = 108.05 \ (p \leq .001)$$

where I_{MV} is the I value for motor vehicle thefts, I_P is the I value for population, and $S_{E(I)}$ is the standard deviation of I under the assumption of normality. The high Z-value suggests

Figure 4.28:

1996 Baltimore Metropolitan Auto Thefts

Number of Auto Thefts Per Block

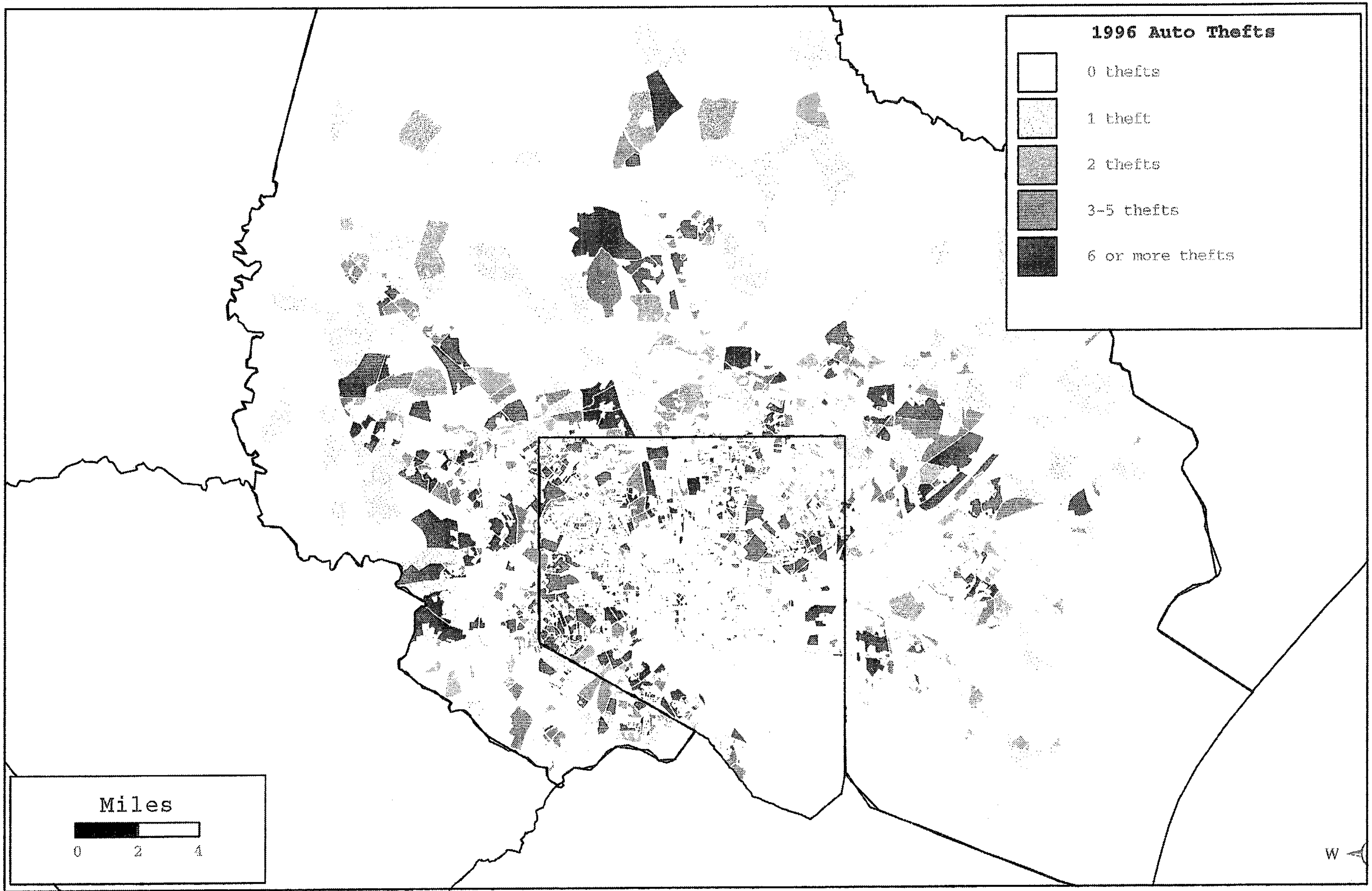
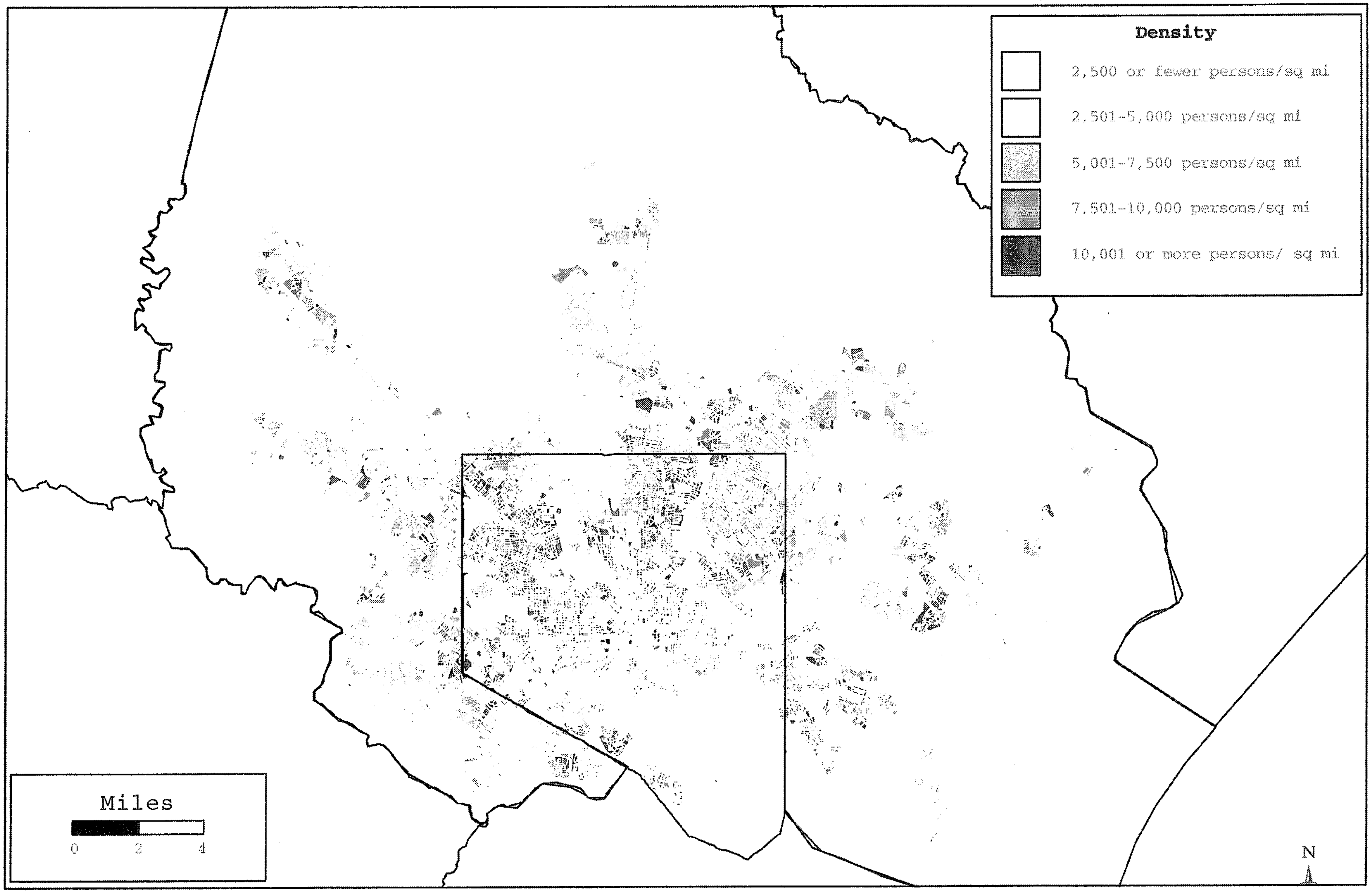


Figure 4.29:

1990 Baltimore Population Density

Number of Persons Per Square Mile by Block


Density
2,500 or fewer persons/sq mi
2,501-5,000 persons/sq mi
5,001-7,500 persons/sq mi
7,501-10,000 persons/sq mi
10,001 or more persons/ sq mi
Miles
0
2
4
N

that motor vehicle thefts are much more clustered than the clustering of population. To put it another way, they are more clustered than would be expected from the population distribution. As mentioned, this is an approximate test since the joint distribution of I for two empirical distributions of I is not known.

Geary's C Statistic

Geary's C statistic is similar to Moran's I (Geary, 1954). In this case, however, the interaction is not the cross-product of the deviations from the mean, but the deviations in intensities of each observation location with one another. It is defined as

$$C = \frac{(N -1) [\Sigma_i \Sigma_j W_{ij} (X_i - X_j)^2]}{2(\Sigma_i \Sigma_j W_{ij}) \Sigma_i (X_i - \bar{X})^2} \qquad (4.36)$$

The values of C typically vary between 0 and 2 although 2 is not a strict upper limit (Griffith, 1987). The theoretical value of C is 1; that is, if values of any one zone are spatially unrelated to any other zone, then the expected value of C would be 1. Values less than 1 (i.e., between 0 and 1) typically indicate positive spatial autocorrelation while values greater than 1 indicate negative spatial autocorrelation. Thus, this index is inversely related to Moran's I. It will not provide identical inference because it emphasizes the differences in values between pairs of observations comparisons rather than the covariation between the pairs (i.e., product of the deviations from the mean). The Moran coefficient gives a more global indicator whereas the Geary coefficient is more sensitive to differences in small neighborhoods.

Adjustment for small distances

Like Moran's I, the weights are defined as the inverse of the distance between the paired points:

$$W_{ij} = \frac{1}{d_{ij}} \qquad \begin{matrix}(4.33)\\ \text{repeat}\end{matrix}$$

However, the. weights will tend to increase substantially as the distance between points decreases. Consequently, a small distance adjustment is allowed which ensures that no weight is greater than 1.0. The adjustment scales the distances to one mile

$$W_{ij} = \frac{\text{one mile}}{\text{one mile} + d_{ij}} \qquad \begin{matrix}(4.34)\\ \text{repeat}\end{matrix}$$

in the units are specified. This is the default condition although the user can calculate all weights as the reciprocal distance by turning off the small distance adjustment.

Global Moran's I and Small Distance Adjustment: Spatial Pattern of Crime in Tokyo

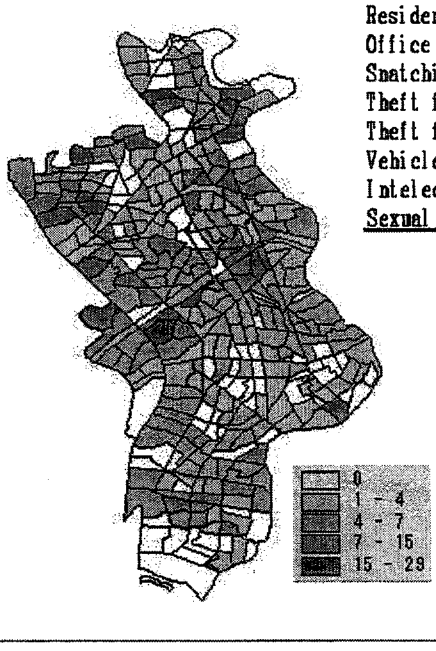
Takahito Shimada
National Research Institute of Police Science
National Police Agency, Chiba, Japan

Crimestat calculates spatial autocorrelation indicators such as Moran's I and Geary's C. These indicators can be used to compare the spatial patterns among crime types. Moran's I is calculated based on the spatial weight matrix where the weight is the inverse of the distance between two points. There is a problem that could occur for incident locations in that the weight could become very large as the distance between points become closer. In *Crimestat*, the small distance adjustment is available to solve this problem. The adjustment produces a maximum weight of 1 when the distance between points is 0.

The number of reported crimes in Tokyo increased from 1996 to 2000 although the city is generally very safe. For this analysis, 68,400 cases reported in the eastern parts of Tokyo were aggregated by census tracts (N=350). Then *Crimestat* calculated Moran's I for each crime type with and without the small distance adjustment.

The "I" value for most crime types, including burglary, theft, purse snatching, showed significantly positive autocorrelation. The results with and without the small distance adjustment were generally very close. The Pearson's correlation between the original and adjusted Moran's I is .98. Among 10 crime types, relatively strong spatial patterns were detected for car theft, sexual assaults, and residential burglary.

Spatial Patterns of Residential Burglary:
Moran's I = 0.023. z=7.58



Calculated Moran's I by Crime Types

	Original			Adjustment		
Crime Type	Moran's I	z		Moran's I	z	
Felonious Offense	0.018	4.09	**	0.003	0.96	
Violent Offense	0.030	6.27	**	0.007	3.03	**
Residential Burglary	0.055	11.21	**	0.023	7.58	**
Office Burglary	0.028	5.93	**	0.012	4.34	**
Snatching	0.031	6.48	**	0.006	2.45	*
Theft from Vender	0.030	6.38	**	0.012	4.28	**
Theft from Cars	0.081	16.08	**	0.044	13.75	**
Vehicle Theft	0.047	9.65	**	0.018	6.14	*
Intelectual Offense	0.023	4.99	**	0.003	1.79	
Sexual Assault	0.080	16.00	**	0.045	14.04	**

**: p<.01 *: p<.05

Preliminary Statistical Tests for Hotspots: Examples from London, England

Spencer Chainey
InfoTech Enterprises Europe
London, England

Preliminary statistical tests for clustering and dispersion can provide insight into what types of patterns will be expected when the crime data is mapped. Global tests can confirm whether there is statistical evidence of clusters (i.e. hotspots) in crime data which can be mapped, rather than mapping data as a first step and struggling to accurately identify hotspots when none actually exist.

Using *CrimeStat,* four statistical tests were compared for robbery, residential burglary and vehicle crime data for the London Borough of Croydon, England. For the incident data, the standard distance deviation and nearest neighbor index were used. For crime incidents aggregated to Census block areas, Moran's I and Geary's C spatial autocorrelation indices were compared. The crime data is for the period June 1999 – May 2000.

Crime type	Number of crime records	Standard distance	NN Index	z-score (test statistic)	***Evidence of Clustering?***
Robbery	1132	3119.5 m	0.47	-34.2	***Yes***
Residential burglary	3104	3664.6 m	0.46	-57.5	***Yes***
Vehicle crime	9314	3706.2 m	0.26	-137.0	***Yes***

Crime type	**Moran's I**	**Geary's C**
All crime	0.0067	1.14
Robbery	0.0078	1.15
Residential burglary	0.014	0.99
Vehicle crime	0.0082	1.08

With the point statistics, all three crime types show evidence of clustering. Vehicle crime shows the more dispered pattern suggesting that whilst hotspots do exist, they may be more spread out over the Croydon area than that of the other two crime types. For the two spatial autocorrelation measures, there are differences in the sensitivities of the two tests. For example, for robbery, there is evidence of global positive spatial autocorrelation (i.e. evidence that, overall, Census blocks that are close together have similar values than those that are further apart). On the other hand, the Geary coefficient suggests that, at a smaller neighbourhood level, areas with a high number of robberies are surrounded by areas with a low number of robberies.

Testing the significance of Geary's C

The empirical C distribution can be compared with the theoretical distribution by dividing by an estimate of the theoretical standard deviation

$$Z(C) = \frac{C - E(C)}{S_{E(C)}} \tag{4.37}$$

where C is the empirical value calculated from a sample, E(C) is the theoretical mean of a random distribution and $S_{E(C)}$ is the theoretical standard deviation of E(C). The usual test for C is to assume that the sample Z follows a standard normal distribution with mean of 0 and variance of 1 (normality assumption). *CrimeStat* only calculates the normality assumption though it is possible to calculate the standard error under a randomization assumption (Ripley, 1981).[9] Figure 4.30 illustrates the output.

Example 4: Testing auto thefts with Geary's C

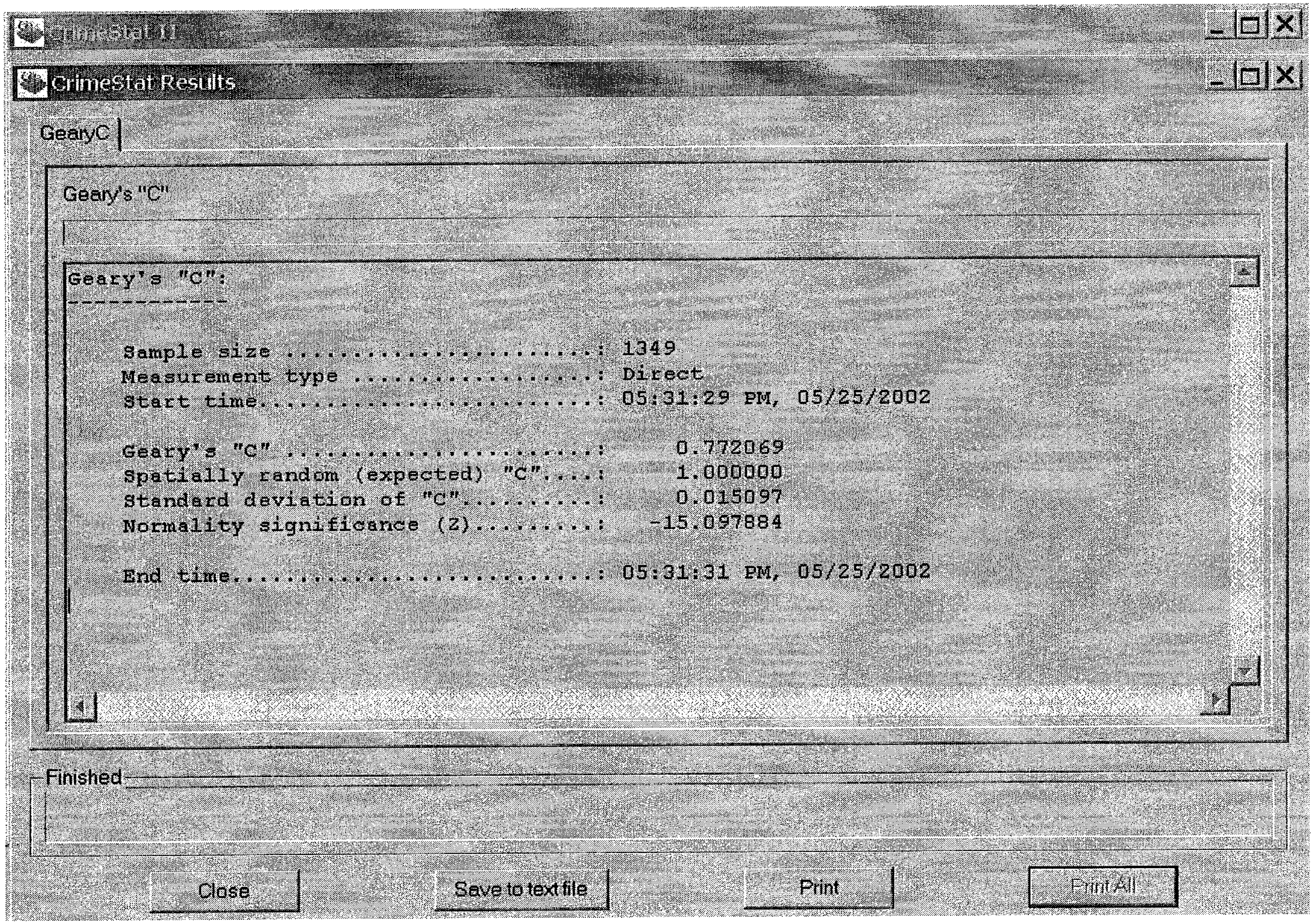
Using the same data on auto thefts for Baltimore County and Baltimore City, the C value for auto thefts was 1.0355 with a Z-value of 10.68 ($p \leq .001$) while that for population was 0.924811 with a Z-value of 122.61 ($p \leq .001$). The C value of motor vehicle thefts is greater than the theoretical C of 1 and suggests *negative* spatial autocorrelation, rather than positive spatial autocorrelation. That is, the index suggests that blocks with a high number of auto thefts are adjacent to blocks with a low number of auto thefts or with low population density. The C value of population, on the other hand, is below the theoretical C of 1 and points to positive spatial autocorrelation. Thus, Geary's C provides a different inference from Moran's I regarding the spatial distribution of the blocks.

In the example above, Moran's I indicated positive spatial autocorrelation for both auto thefts and population density. An inspection of figure 4.28 above show however, that there are little 'peaks' and 'valleys' among the blocks. Several blocks have a high number of auto thefts, but are surrounded by blocks with a low number of auto thefts.

In other words, the Moran coefficient has indicated that there is more positive spatial autocorrelation for motor vehicle thefts among the 13,101 blocks while the Geary coefficient has emphasized the irregular patterning among the blocks. The Geary index is more sensitive to local clustering (second-order effects) than the Moran index, which is better seen as measuring first-order spatial autocorrelation. This illustrates how these indices have to be used with care and cannot be generalized by themselves. Each of them emphasizes slightly different information regarding spatial autocorrelation, yet neither is sufficient by itself. They should be used as part of a larger analysis of spatial patterning.[10]

The next chapter will examine tools for measuring *second-order* effects using properties of the distances between incident locations.

Figure 4.30: Geary's C Statistic Output



```
Geary's "C":
------------

    Sample size .......................: 1349
    Measurement type ..................: Direct
    Start time.........................: 05:31:29 PM, 05/25/2002

    Geary's "C" .......................:     0.772069
    Spatially random (expected) "C"....:     1.000000
    Standard deviation of "C"..........:     0.015097
    Normality significance (Z).........:   -15.097884

    End time...........................: 05:31:31 PM, 05/25/2002
```

Endnotes for Chapter 4

1. Hint. There are 40 bars indicated in the status bar while a routine is running. For long runs, users can estimate the calculation time by timing how long it takes for two bars to be displayed and then multiply by 20.

2. *CrimeStat's* implementation of the Kuhn and Kuenne algorithm is as follows (from Burt and Barber, 1996, 112-113):

   A. Let t be the number of the iteration. For the first iteration only (i.e., t=1) the weighted mean center is taken as the initial estimate of the median location, X_t and Y_t.

   B. Calculate the distance from each point, i, to the current estimate of the median location, d_{ict}, where i is a single point and ct is the current estimate of the median location during iteration t.

      a. If the coordinates are spherical, then Great Circle distances are used.

      b. If the coordinates are projected, then Euclidean distances are used.

   C. Weight each case by a weight, W_i, and calculate

   $$K_{it} \quad = \quad W_i \, e^{-d(ict)}$$

   where **e** is the base of the natural logarithm(2.7183..) and $d_{(ict)}$ is an alternative way to write d_{ict}.

      a. If no weights are defined in the primary file, W_i is assumed to be 1.

      b. If weights are defined in the primary file, W_i takes their values.

   Note that as the distance, d_{ict}, approaches 0, then $e^{-d(ict)}$ becomes 1.

   D. Calculate a new estimate of the center of minimum distance from

   $$X^{t+1} \quad = \quad \frac{\Sigma K_{it} X_i}{\Sigma K_{it}} \quad \text{for } i=1...n$$

   $$Y^{t+1} \quad = \quad \frac{\Sigma K_{it} Y_i}{\Sigma K_{it}} \quad \text{for } i=1...n$$

where X_i and Y_i are the coordinates of point i (either lat/lon for spherical or feet or meters for projected).

E. Check to see how much change has occurred since the last iteration

$$ABS|\ X^{t+1} - X^t\ |\ \leq 0.000001$$

$$ABS|\ Y^{t+1} - Y^t\ |\ \leq 0.000001$$

a. If either the X or Y coordinates have changed by greater than 0.000001 between iterations, substitute X^{t+1} for X^t and Y^{t+1} for Y^t and repeat steps <u>B</u> through <u>D</u>.

b. If *both* the change in X and the change in Y is less than or equal to 0.000001, then the estimated X_t and Y_t coordinates are taken as the center of median distance.

3. With a weight for an observation, w_i, the squared distance is weighted and the formula becomes

$$S_{XY} \quad = \quad SQRT \ \frac{\sum w_i(d_{iMC})^2}{(\sum w_i) - 2}$$

Both summations are over all points, N.

4. Formulas for the new axes provided by Ebdon (1988) and Cromley (1992) yield standard deviational ellipses that are too small, for two different reasons. First, they produce transformed axes that arc too small. If the distribution of points is random and even in all directions, ideally the standard deviational ellipse should be equal to the standard distance deviation, since $S_x = S_y$. The formula used here has this property. Since the formula for the standard distance deviation is (4.6):

$$SDD = SQRT[\ \frac{\Sigma(X_i - \bar{X})^2 + \Sigma(Y_i - \bar{Y})^2}{N-2}\ ]$$

If $S_x = S_y$, then $\Sigma(X_i - \bar{X})^2 = \Sigma(Y_i - \bar{Y})^2$, therefore

$$SDD = SQRT[2 * \frac{\Sigma(X_i - \bar{X})^2}{N-2}]$$

Similarly, the formula for the transformed axes are (4.9, 4.10):

$$S_x = \text{SQRT}[\,2*\frac{\Sigma\{(X_i - \bar{X})\,\text{Cos}\theta\ -\ \Sigma(Y_i - \bar{Y})\,\text{Sin}\theta\,\}^2}{N-2}\,]$$

$$S_y = \text{SQRT}[\,2*\frac{\Sigma\{(X_i - \bar{X})\,\text{Sin}\theta\ -\ \Sigma(Y_i - \bar{Y})\,\text{Cos}\theta\,\}^2}{N-2}\,]$$

However, if $S_x = S_y$, then $\theta = 0$, $\text{Cos}0 = 1$, $\text{Sin}0 = 0$ and, therefore,

$$S_x = S_y = \quad \text{SQRT}[\,2*\ \frac{\Sigma(X_i - \bar{X})}{N-2}\,]$$

which is the same as for the standard distance deviation (SDD) under the same conditions. The formulas used by Ebdon (1988) and Cromley (1992) produce axes which are SQRT(2) times too small.

The second problem with the Ebdon and Cromley formulas is that they do not correct for degrees of freedom and, hence, produce too small a standard deviational ellipse. Since there are two constants in each equation, MeanX and MeanY, then there are only N-2 degrees of freedom. The cumulative effect of using transformed axes that are too small and not correcting for degrees of freedom yields a much smaller ellipse than that used here.

5. In *MapInfo*, the command is *T*able *I*mport <*Mapinfo interchange file*>. With *Atlas*GIS*, the command is *F*ile *O*pen <*boundary (*.bna) file*>. With the DOS version of *Atlas*GIS*, the *Atlas Import-Export* program has to be used to convert the '.bna' output file to an *Atlas*GIS* '.agf' file.

6. The theoretical standard deviation of I under the assumption of normality is (from Ebdon, 1985):

$$S_{E(I)} \quad = \quad \text{SQRT}[\ \frac{N^2\,\Sigma_{ij}\,w_{ij}^2 + 3(\Sigma_{ij}\,w_{ij})^2\ - N\,\Sigma_i\,(\Sigma_j\,w_{ij})^2}{(N^2 - 1)\,(\Sigma_{ij}\,w_{ij})^2}\,]$$

7. The formula for the theoretical standard deviation of I under the randomization assumption is (from Ebdon, 1985):

$$S_{E(I)} = \text{SQRT}[\frac{N\{(N^2{+}3{-}3N)\Sigma_{ij}\,w_{ij}^2{+}3(\Sigma_{ij}\,w_{ij})^2{-}N\Sigma_i(\Sigma_j\,w_{ij})^2\}{-}k((N^2{-}N)\Sigma_{ij}\,w_{ij}^2{+}6(\Sigma_{ij}\,w_{ij})^2{-}2N(\Sigma_i(\Sigma_j\,w_{ij})^2)}{(N-1)(N-2)(N-3)(\Sigma_{ij}\,w_{ij})^2}]$$

8. We could have compared Moran's I for auto thefts with that of population, rather than population density. However, since the areas of blocks tend to get larger the farther the distance from the metropolitan center, the effect of testing only population is partly being minimized by the changing sizes of the blocks. Consequently, population density was used to provide a more accurate measure of population concentration. In any case, Moran's I for population is also highly significant: I = 0.00166 (Z=17.32).

9. The theoretical standard deviation for C under the normality assumption is (from Ripley, 1981):

$$S_{E(I)} = \text{SQRT}\left[\frac{(2 \Sigma_{ij} w_{ij}^2 + \Sigma_i (\Sigma_j w_{ij})^2)(N-1) - 4(\Sigma_{ij} w_{ij})^2}{2(N+1)(\Sigma_{ij} w_{ij})^2} \right]$$

10. Anselin (1992) points out that the results of the two indices are determined to a large extent by the type of weighting used. In the original formulation, where adjacent weights of 1 and 0 are used, the two indices are linearly related, though moving in opposite directions (Griffith, 1987). Thus, only adjacent zones have any impact on the index. With inverse distance weights, however, zones farther removed can influence the overall index so it is possible to have a situation whereby adjacent zones have similar values (hence, are positively autocorrelated) whereas zones farther away could have dissimilar values (hence, are negatively autocorrelated).

Chapter 5
Distance Analysis

In this chapter, tools that identify characteristics of the distances between points will be described. The previous chapter provided tools for describing the general spatial distribution of crime incidents or *first-order* properties of the incident distribution (Bailey and Gattrell, 1995). First-order properties are global because they represent the dominant pattern of distribution - where it is centered, how far it spreads out, and whether there is any orientation or direction to its dispersion. *Second-order* (or *local*) properties, on the other hand, refer to sub-regional patterns or 'neighborhood' patterns within the overall distribution. If there are distinct 'hot spots' where many crime incidents cluster together, their distribution is spatially related not so much to the overall global pattern as to something unique in the sub-region or neighborhood. Thus, second-order characteristics tell something about particular environments that may concentrate crime incidents. Figure 5.1 shows the distance analysis screen and the distance statistics that are calculated by *CrimeStat*.

Nearest Neighbor Index (Nna)

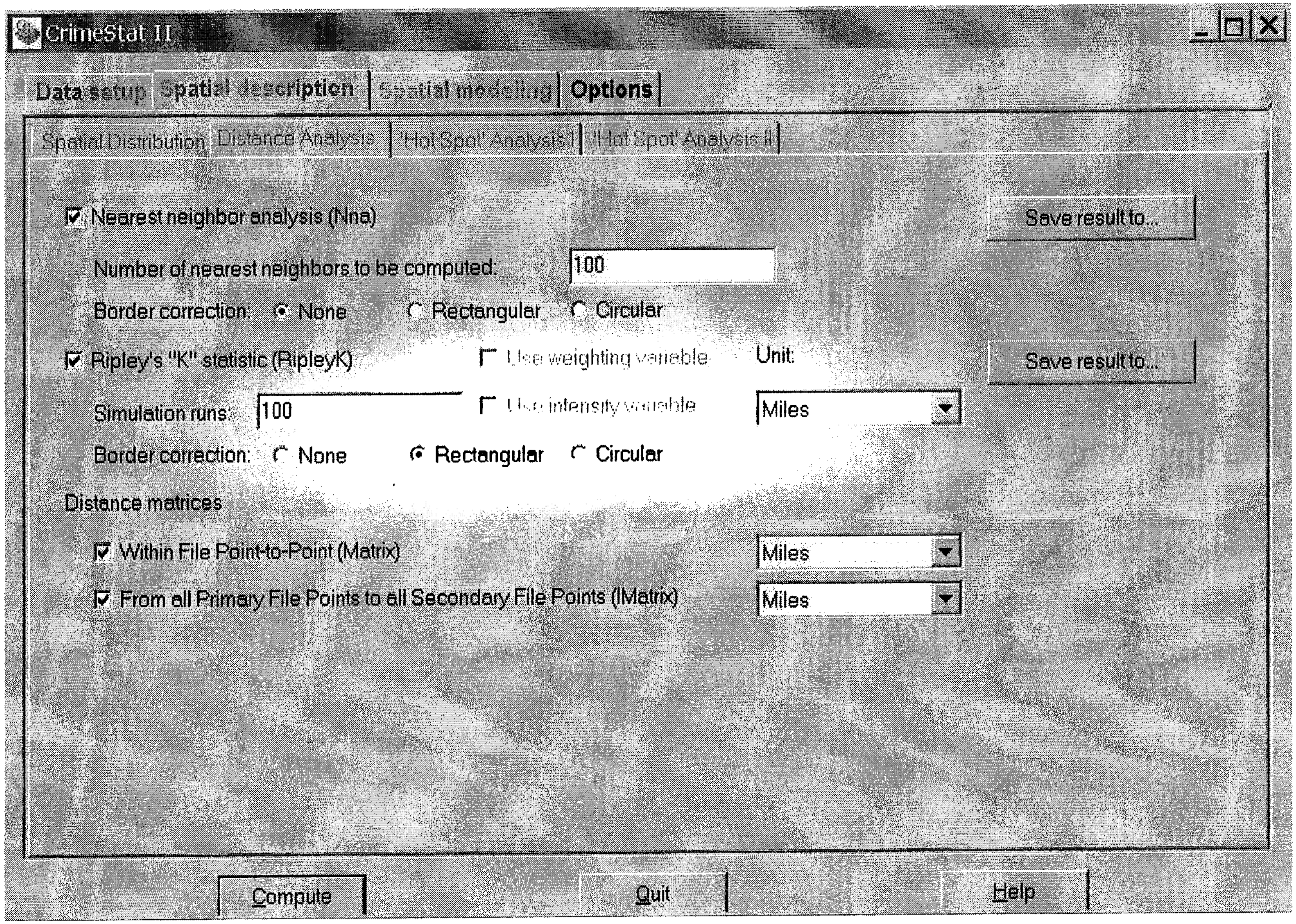
One of the oldest distance statistics is the *nearest neighbor index*. It is particularly useful because it is a simple tool to understand and to calculate. It was developed by two botanists in the 1950s (Clark and Evans, 1954), primarily for field work, but it has been used in many different fields for a wide variety of problems (Cressie, 1991). It has also become the basis of many other types of distance statistics, some of which are implemented in *CrimeStat*.

The nearest neighbor index compares the distances between nearest points and distances that would be expected on the basis of chance. It is an index that is the ratio of two summary measures. First, there is the *nearest neighbor distance*. For each point (or incident location) in turn, the distance to the closest other point (nearest neighbor) is calculated and averaged over all points.

$$\text{Nearest Neighbor Distance} = d(NN) = \sum_{i=1}^{N}\left[\frac{\text{Min}(d_{ij})}{N}\right] \qquad (5.1)$$

where $\text{Min}(d_{ij})$ is the distance between each point and its nearest neighbor and N is the number of points in the distribution. Thus, in *CrimeStat*, the distance from a single point to every other point is calculated and the smallest distance (the minimum) is selected. Then, the next point is taken and the distance to all other points (including the first point measured) is calculated with the nearest being selected and added to the first minimum distance. This process is repeated until all points have had their nearest neighbor selected. The total sum of the minimum distances is then divided by N, the sample size, to produce an average minimum distance.

Figure 5.1: **Distance Analysis Screen**



☑ Nearest neighbor analysis (Nna)

Number of nearest neighbors to be computed: 100

Border correction: ● None ○ Rectangular ○ Circular

☑ Ripley's "K" statistic (RipleyK)

□ Use weighting variable

□ Use intensity variable

Unit: Miles

Simulation runs: 100

Border correction: ○ None ● Rectangular ○ Circular

Distance matrices

☑ Within File Point-to-Point (Matrix) — Miles

☑ From all Primary File Points to all Secondary File Points (IMatrix) — Miles

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

The second summary measure is the expected nearest neighbor distance if the distribution of points is completely spatially random. This is the *mean random distance* (or the mean random nearest neighbor distance). It is defined as

$$\text{Mean Random Distance} = d(ran) = 0.5 \text{ SQRT } \left[\frac{A}{N} \right] \qquad (5.2)$$

where A is the area of the region and N is the number of incidents. Since A is defined by the square of the unit of measurement (e.g., square mile, square meters, etc.), it yields a random distance measure in the same units (i.e., miles, meters, etc.).[1] If defined on the measurement parameters page by the user, *CrimeStat* will use the specified area in calculating the mean random distance. If no area measurement is provided, *CrimeStat* will take the rectangle defined by the minimum and maximum X and Y points.

The nearest neighbor index is the ratio of the observed nearest neighbor distance to the mean random distance

$$\text{Nearest Neighbor Index} = NNI = \frac{d(NN)}{d(ran)} \qquad (5.3)$$

Thus, the index compares the average distance from the closest neighbor to each point with a distance that would be expected on the basis of chance. If the observed average distance is about the same as the mean random distance, then the ratio will be about 1.0. On the other hand, if the observed average distance is smaller than the mean random distance, that is, points are actually closer together than would be expected on the basis of chance, then the nearest neighbor index will be less than 1.0. This is evidence for clustering. Conversely, if the observed average distance is greater than the mean random distance, then the index will be greater than 1.0. This would be evidence for dispersion, that points are more widely dispersed than would be expected on the basis of chance.

Testing the Significance of the Nearest Neighbor Index

Some differences from 1.0 in the nearest neighbor index would be expected by chance. Clark and Evans (1954) proposed a Z-test to indicate whether the observed average nearest neighbor distance was significantly different from the mean random distance (Hammond and McCullagh, 1978; Ripley, 1981). The test is between the observed nearest neighbor distance and that expected from a random distribution and is given by

$$Z = \frac{d(NN) - d(ran)}{SE_{d(ran)}} \qquad (5.4)$$

where the standard error of the mean random distance is approximately given by:

$$SE_{d(ran)} \approx SQRT\left[\frac{(4-\pi)\,A}{4\pi\,N^2}\right] \approx \frac{0.26136}{SQRT[\,N^2/A\,]} \qquad (5.5)$$

with A being the area of region and N the number of points. There have been other suggested tests for the nearest neighbor distance as well as corrections for edge effects (see below). However, equations 5.4 and 5.5 are used most frequently to test the average nearest neighbor distance. See Cressie (1991) for details of other tests.

Calculating the statistics

Once nearest neighbor analysis has been selected, the user clicks on *Compute* to run the routine. The program outputs 10 statistics:

1. The sample size
2. The mean nearest neighbor distance
3. The standard deviation of the nearest neighbor distance
4. The minimum distance
5. The maximum distance
6. The mean random distance for both the bounding rectangle and the user input area, if provided
7. The mean dispersed distance for both the bounding rectangle and the user input area, if provided
8. The nearest neighbor index for both the bounding rectangle and the user input area, if provided
9. The standard error of the nearest neighbor index for both the maximum bounding rectangle and the user input area, if provided
10. A significance test of the nearest neighbor index (Z-test)

In addition, the output can be saved to a '.dbf' file, which can then be imported into spreadsheet or graphics programs.

Example 1: The nearest neighbor index for street robberies

In 1996, there were 1181 street robberies in Baltimore County. The area of the County is about 607 square miles and is specified on the measurement parameters page. *CrimeStat* returns the statistics shown in Table 5.1 with the NNA routine.

CrimeStat does not provide the significance level of the test, but only the Z-value. However, the significance level of the Z-value can be found in any table of standard normal deviants. In this case, a Z-value of -44.4672 is highly significant ($p \leq .001$). In other words, the distribution of the nearest neighbors of street robberies in Baltimore County is significantly smaller than the expected distribution of nearest neighbors.

Table 5.1
Nearest Neighbor Statistics for
1996 Street Robberies in Baltimore County
N=1181

Mean nearest neighbor distance:	0.11598 mi
Mean random distance based on user input area:	0.35837 mi
Nearest neighbor index:	0.3236
Standard error:	0.00545 mi
Test Statistic (Z):	-44.4672

It should be noted that the significance test for the nearest neighbor index is not a test for complete spatial randomness, for which it is sometimes mistaken. It is only a test whether the average nearest neighbor distance is significantly different than what would be expected on the basis of chance. In other words, it is a test of *first-order* nearest neighbor randomness.[2] There are also second-order, third-order, and so forth distributions that may or may not be significantly different from their corresponding orders under complete spatial randomness. A complete test would have to test for all those effects, what are called *K-order* effects.

Example 2: The nearest neighbor index for residential burglaries

The nearest neighbor index and test can be very useful for understanding the degree of clustering of crime incidents in spite of its limitations. For example, in Baltimore County, the distribution of 6051 residential burglaries in 1996 yields the following nearest neighbor statistics (Table 5.2):

Table 5.2
Nearest Neighbor Statistics for
1996 Residential Burglaries in Baltimore County
N=6051

Mean nearest neighbor distance:	0.07134 mi
Mean random distance based on user input area:	0.16761 mi
Nearest neighbor index:	0.4256
Standard error:	0.00113 mi
Test Statistic (Z):	-85.4750

The distribution of residential burglaries is also highly significant. Now, suppose we want to compare the distribution of street robberies (table 5.1) with that residential burglaries (table 5.2). The significance test is not very useful for the comparison because the sample sizes are so large (1181 v. 6051); the much higher Z-value for residential burglaries indicates primarily that there was a larger sample size to test it. However, comparing the relative nearest neighbor indices can be meaningful.

$$\text{Relative Nearest Neighbor Comparison} = \frac{NNI(A)}{NNI(B)} \qquad (5.6)$$

where NNI(A) is the nearest neighbor index for one group (A) and NNI(B) is the nearest neighbor index for another group (B). Thus, comparing street robberies with residential burglaries, we have

$$\frac{NNI\ (A)}{NNI\ (B)} = \frac{NNI\ (\text{robberies})}{NNI\ (\text{burglaries})} = \frac{0.3057}{0.4256} = 0.7182$$

In other words, the distribution of street robberies relative to an expected random distribution appears to be more concentrated than that of burglaries relative to an expected random distribution. There is no simple significance test of this comparison since the standard error of the joint distributions is not known. But the relative index suggests that robberies are more concentrated than burglaries and, hence, are more likely to have 'hot spot' or 'hot zones' where they are particularly concentrated. This index, of course, does not prove that there are 'hot spots', but only points us towards the higher concentration of robberies relative to burglaries. In the previous chapter, it was shown that robberies had a smaller dispersion than burglaries. Here, however, the analysis is taken a step further to suggest that robberies are more concentrated than burglaries.

K-Order Nearest Neighbors

As mentioned above, the nearest neighbor index is only an indicator of first-order spatial randomness. It compares the average distance for the nearest neighbor to an expected random distance. But what about the second nearest neighbor? Or the third nearest neighbor? Or the K^{th} nearest neighbor? *CrimeStat* constructs K-order nearest neighbor indices. On the distance analysis page, the user can specify the number of nearest neighbor indices to be calculated.

The K-order nearest neighbor routine returns four columns:

1. The order, starting from 1
2. The mean nearest neighbor distance for each order (in meters)
3. The expected nearest neighbor distance for each order (in meters)
4. The nearest neighbor index for each order

For each order, *CrimeStat* calculates the K^{th} nearest neighbor distance for each observation and then takes the average. The expected nearest neighbor distance for each order is calculated by:

$$\text{Mean Random Distance to K}^{th}\text{ nearest neighbor} = d(K_{ran}) = \frac{K\,(2K)!}{(2^K K!)^2 \text{ SQRT } [N/A]} \qquad (5.7)$$

where K is the order and ! is the factorial operation (e.g., 4! = 4 x 3 x 2 x 1; Thompson, 1956). The K^{th} nearest neighbor index is the ratio of the observed K^{th} nearest neighbor distance to the K^{th} mean random distance. There is not a good significance test for the K^{th} nearest neighbor index due to the non-independence of the different orders, though there have been attempts (see examples in Getis and Boots, 1978; Aplin, 1983). Consequently, *CrimeStat* does not provide a test of significance.

There are no restrictions on the number of nearest neighbors that can be calculated. However, since the average distance increases with higher-order nearest neighbors, the potential for bias from edge effects will also increase. It is suggested that not more than 100 nearest neighbors be calculated.[3]

Nevertheless, the K-order nearest neighbor distance and index can be useful for understanding the overall spatial distributions. Figure 5.2 compares the K-order nearest neighbor index for street robberies with that of residential burglaries. The output was saved as a ‘.dbf’and was then imported into a graphics program. The graph shows the nearest neighbor indices for both robberies and burglaries up to the 50th order (i.e., the 50th nearest neighbor). The nearest neighbor index is scaled from 0 (extreme clustering) up to 1 (extreme dispersion). Since a nearest neighbor index of 1 is expected under randomness, the thin straight line at 1.0 indicates the expected K-order index. As can be seen, both street robberies and residential burglaries are much more concentrated than K-order spatial randomness. Further, robberies are more concentrated than even burglaries for each of the 50 nearest neighbors. Thus, the graph reinforces the analysis above that robberies are more concentrated than burglaries, and both are more concentrated than a random distribution.

In other words, even though there is not a good significance test for the K-order nearest neighbor index, a graph of the K-order indices (or the K-order distances) can give a picture of how clustered the distribution is as well as allow comparisons in clustering between the different types of crimes (or the same crime at two different time periods).

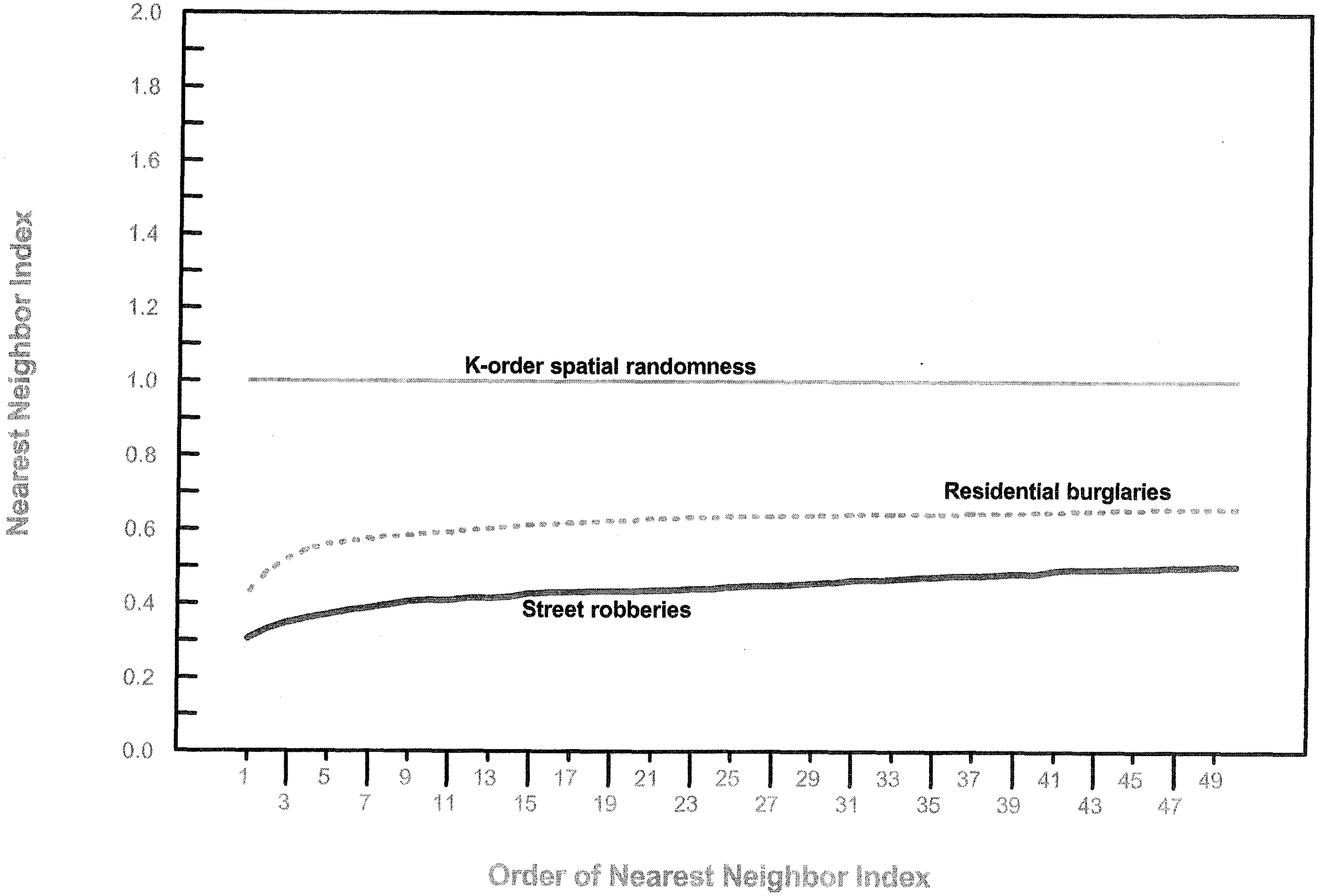
Edge Effects

It should be noted that there are potential edge effects that can bias the nearest neighbor index. An incident occurring near the border of the study area may actually have its nearest neighbor on the other side of the border. However, since there are usually no data on the distribution of incidents outside the study area, the program selects another point within the study area as the nearest neighbor of the border point. Thus, there is the potential for exaggerating the nearest neighbor distance, that is, the observed nearest neighbor distance is probably greater than what it should be and, therefore, there is an *overestimation* of the nearest neighbor distance. In other words, the incidents are probably more clustered than what has been measured (see Cressie, 1991 for details).

Figure 5.2

K-Order Nearest Neighbor Indices

1996 Street Robberies and Residential Burglaries

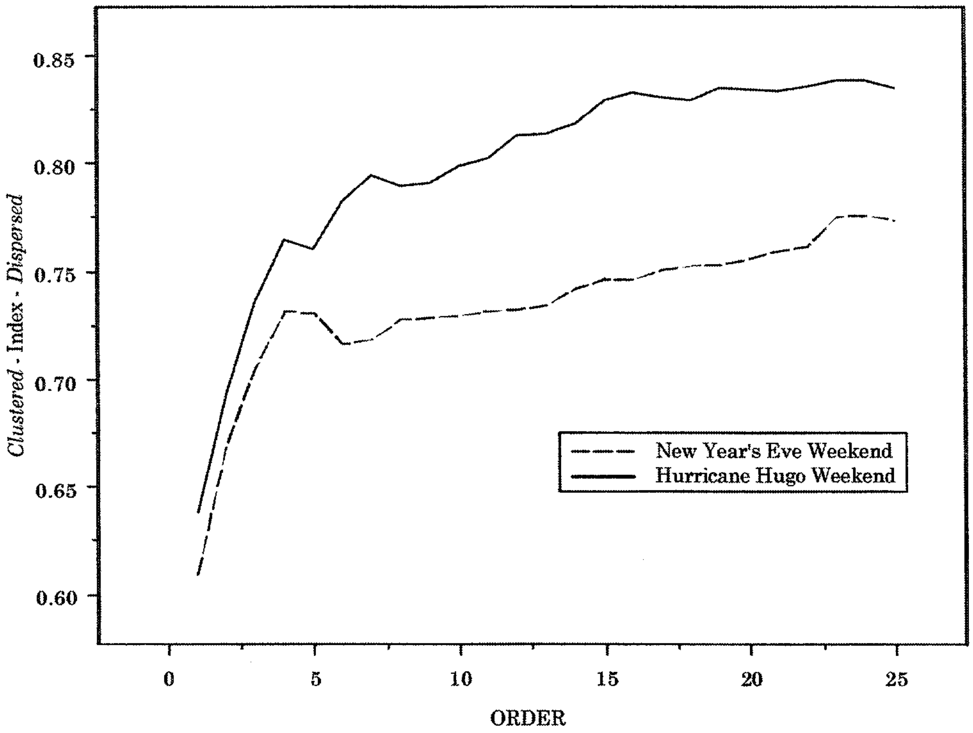


Nearest Neighbor Analysis
Man With A Gun Calls
Charlotte, N.C.: 1989

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

A comparison was made of *Man with a Gun* calls for the weekend in which Hurricane Hugo hit the North Carolina coast (September 22 – 24) with the following New Year's Eve weekend (December 29-31, 1989). There were 146 *Man with a Gun* calls during the Hurricane Hugo weekend compared to 137 calls for New Year's Eve.

Nearest Neighbor Index of Man With A Gun Calls



The Nearest Neighbor Index in *CrimeStat* was used to compare the distributions. From the onset, the Hurricane Hugo *Man With a Gun* locations are more dispersed than New Year's Eve. After the 5th nearest neighbor (Order 5) the differences become more pronounced

Nearest Neighbor Edge Corrections

The default condition is no edge correction. However, one way that the measured distance to the nearest neighbor can be corrected for possible edge effects is to assume for each observed point that there is another point just outside the border at the closest distance. If the distance from a point to the border is shorter than to its measured nearest neighbor, then the nearer theoretical point is taken as a proxy for the nearest neighbor. Thus, with each point in the data set, the observed nearest neighbor distance is compared to the distance, the measured distance is kept. This correction has the effect of reducing the average neighbor distance. Since it assumes that there is always another point at the border, it probably *underestimates* the true nearest neighbor distance. The true value is probably somewhere in between the measured and the assumed nearest neighbor distance.

CrimeStat has two different edge corrections. Because *CrimeStat* is not a GIS package, it cannot locate the actual border of a study area. One would need a topological GIS package in which the distance from each point to the nearest boundary is calculated. Instead, there are two different geometric models that can be applied. The first assumes that the study area is a rectangle while the second assumes that the study area is a circle. Depending on the shape of the actual study area, one or either of these models may be appropriate.

Rectangular study area

In the rectangular adjustment, the area of the study area, A, is first calculated, either from the user input on the measurement parameters tab or from the maximum bounding rectangle defined by the minimum and maximum X/Y values (see chapter 3). If the user provides an estimate of the area, the rectangle is proportionately re-scaled so that the area of the rectangle equals A. Second, for each point, the distance to the nearest other point is calculated. This is the observed nearest neighbor distance for point i.

Third, the minimum distance to the nearest edge of the rectangle is calculated and is compared to the observed nearest neighbor distance for point i. If the observed nearest neighbor distance for point i is equal to or less than the distance to the nearest border, it is retained. On the other hand, if the observed nearest neighbor distance for point i is greater than the distance to the nearest border, the distance to the border is used as a proxy for the nearest neighbor distance of point i.

Circular study area

In the circular adjustment, first, the area of the study area is calculated, either from the user input on the measurement parameters tab (see chapter 3) or from the maximum bounding rectangle defined by the minimum and maximum X/Y values. If the user has specified a study area on the measurement parameters page, then that value is taken for A and the radius of the circle is calculated by

$$R = \text{SQRT}\left[A / \pi\right] \qquad (5.8)$$

If the user has not specified a study area on the measurement parameters page, then A is calculated from the minimum and maximum X and Y coordinates (the bounding rectangle) and the radius of the circle is calculated with equation 5.8.

Second, for each point, the distance to the nearest other point is calculated. This is the observed nearest neighbor distance for point i. Third, for each point, i, the distance from that point to the mean center is calculated, R_i. Fourth, the minimum distance to the nearest edge of the circle is calculated using

$$R_{iC} = R - R_i \qquad (5.9)$$

Fifth, for each point, i, the observed minimum distance is compared to the nearest edge of the circle, R_{iC}. If the observed nearest neighbor distance for point i is equal to or less than the distance to the nearest edge, it is retained. On the other hand, if the observed nearest neighbor distance for point i is greater than the distance to the nearest edge, the distance to the border is used as a proxy for the true nearest neighbor distance of point i.

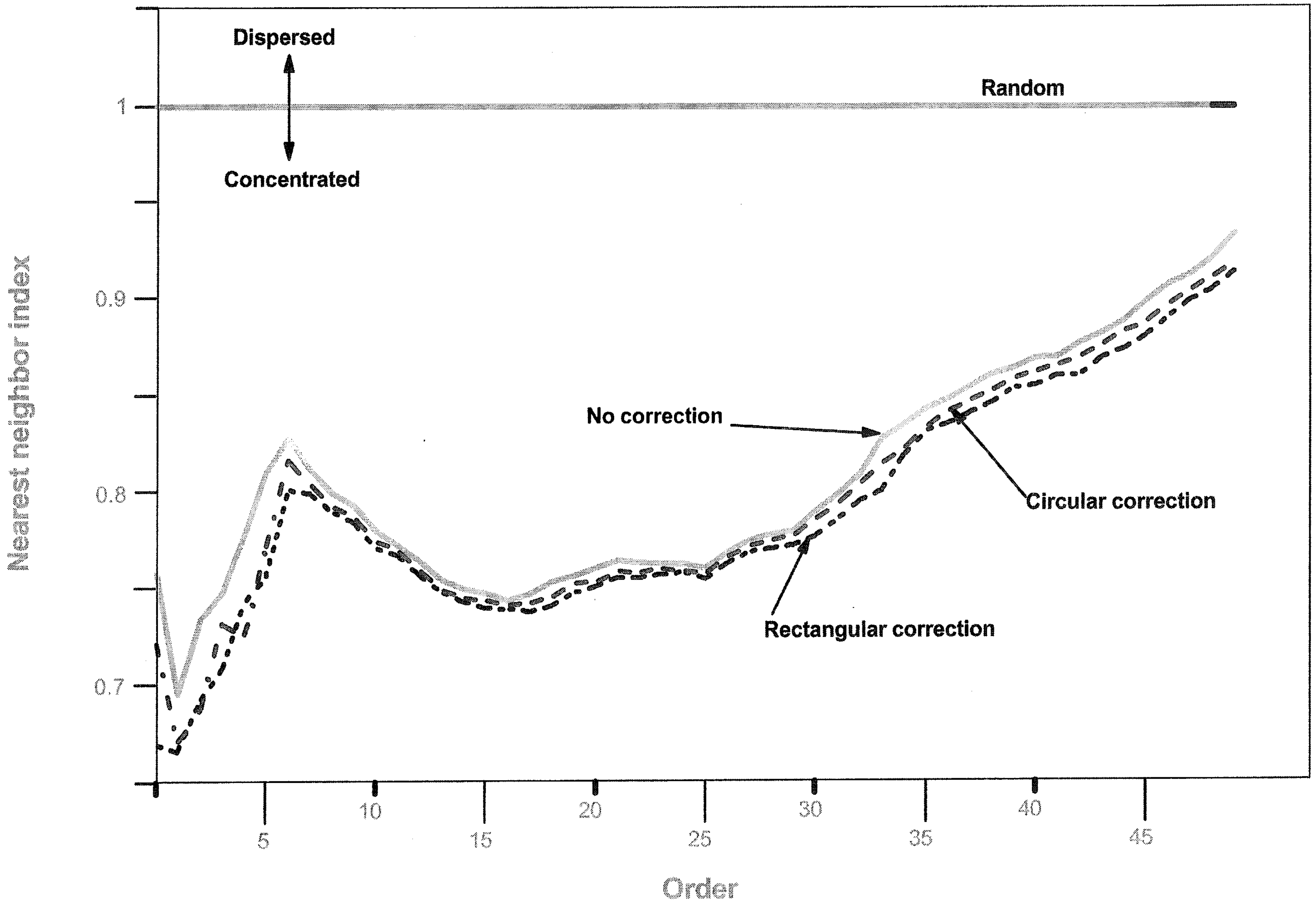
For either correction

The average nearest neighbor distance is calculated and compared to the theoretical average nearest neighbor distance under random conditions. The indices and tests are as before (see chapter 4). Figure 5.3 below shows a graph of the K-order nearest neighbor index for the 50 nearest neighbors for 1996 motor vehicle thefts in police Precinct 11 of Baltimore County. The uncorrected nearest neighbor indices are compared with those corrected by a rectangle and a circle. As can be seen, both corrections are very similar to the uncorrected. However, they both show greater concentrations than the uncorrected index. The rectangular correction shows greater concentration than the circular because it is less compact (i.e., the average distance from the center of the geometric object to the border is slightly larger). In general, the rectangle will lead to more correction than the circle since it substitutes a greater nearest neighbor distance, on average, for a point nearer the border than to its measured nearest neighbor.

The user has to decide whether either of these corrections are meaningful or not. Depending on the shape of the study area, either correction may or may not be appropriate. If the study area is relatively rectangular, then the rectangular model may provide a good approximation. Similarly, if the study area is compact (circular), then the circular model may provide a good approximation. On the other hand, if the study area is of irregular shape, then either of these corrections may produce more distortion than the raw nearest neighbor index. One has to use these corrections with judgement. Also, in some cases, it may not make any sense to correct the measured nearest neighbor distances. In Honolulu, for example, one would not correct the measured nearest neighbor distances because there are no incidents outside the island's boundary.

Correction of Nearest Neighbor Indices

Motor Vehicle Thefts in Precinct 11



Linear Nearest Neighbor Index (Lnna)

The *linear nearest neighbor index* is a variation on the nearest neighbor routine, but one applied to a street network. All distances along this network are assumed to travel along a grid, hence indirect distances are used. Whereas the nearest neighbor routine calculates the distance between each point and its nearest neighbor using direct distances, the linear nearest neighbor routine uses indirect ('Manhattan') distances (see chapter 3). Similarly, whereas the nearest neighbor routine calculates the expected distance between neighbors in a random distribution of N points using the geographical area of the study region, the linear nearest neighbor routine uses the total length of the street network.

The theory of linear nearest neighbors comes from Hammond and McCullagh (1978). The observed linear nearest neighbor distance, Ld(NN), is calculated by *CrimeStat* as the average of indirect distances between each point and its nearest neighbor. The expected linear nearest neighbor distance is given by

$$Ld(ran) = 0.5 \left[\frac{L}{N - 1} \right] \qquad (5.10)$$

where L is the total length of street network and N is the sample size (Hammond and McCullagh, 1978, 279). Consequently, the linear nearest neighbor index is defined as

$$\text{Linear Nearest Neighbor Index} = LNNI = \frac{Ld(NN)}{Ld(ran)} \qquad (5.11)$$

Testing the Significance of the Linear Nearest Neighbor Index

Since the theoretical standard error for the random linear nearest neighbor distance is not known, the author has constructed an approximate standard deviation for the observed linear nearest neighbor distance:

$$S_{Ld(NN)} \approx SQRT \left[\frac{\Sigma (Min(d_{ij}) - Ld(NN))^2}{N - 1} \right] \qquad (5.12)$$

where $Min(d_{ij})$ is the nearest neighbor distance for point i and Ld(NN) is the average linear nearest neighbor distance. This is the standard deviation of the linear nearest neighbor distances. The standard error is calculated by

$$SE_{Ld(NN)} = \frac{S_{Ld(NN)}}{SQRT[N]} \qquad (5.13)$$

An approximate significance test can be obtained by

$$t = \frac{Ld(NN) - Ld(ran)}{SE_{Ld(NN)}} \qquad (5.14)$$

where Ld(NN) is the average linear nearest neighbor distance, Ld(ran) is the expected linear nearest neighbor distance (equation 5.10), and $SE_{Ld(NN)}$ is the approximate standard error of the linear nearest neighbor distance (equation 5.13). Since the empirical standard deviation of the linear nearest neighbor is being used instead of a theoretical value, the test is a *t-test* rather than a Z-test.

Calculating the statistics

On the measurements parameters page, there are two parameters that are input, the geographical area of the study region and the length of street network. At the bottom of the page, the user must select which type of distance measurement to use, direct or indirect. If the measurement type is direct, then the nearest neighbor routine returns the standard nearest neighbor analysis (sometimes called *areal* nearest neighbor). On the other hand, if the measurement type is indirect, then the routine returns the linear nearest neighbor analysis. To calculate the linear nearest neighbor index, therefore, distance measurement must be specified as indirect and the length of the street network must be defined.

Once nearest neighbor analysis has been selected, the user clicks on *Compute* to run the routine. The *Lnna* routine outputs 9 statistics:

1. The sample size
2. The mean linear nearest neighbor distance
3. The minimum linear distance between nearest neighbors
4. The maximum linear distance between nearest neighbors
5. The mean linear random distance
6. The linear nearest neighbor index
7. The standard deviation of the linear nearest neighbor distance
8. The standard error of the linear nearest neighbor distance
9. A significance test of the nearest neighbor index (t-test)

Example 3: Auto thefts along two highways

The linear nearest neighbor index is useful for analyzing the distribution of crime incidents along particular streets. For example, in Baltimore County, state highway 26 in the western part and state highway 150 in the eastern part have high concentrations of motor vehicle thefts (figure 5.4). In 1996, there were 87 vehicle thefts on highway 26 and 47 on highway 150. A GIS can be used with the linear nearest neighbor index to indicate whether these incidents are greater than what would be expected on the basis of chance.

Figure 5.4:

1996 Auto Thefts in Baltimore County

Incident Distribution on State Highways 26 and 150

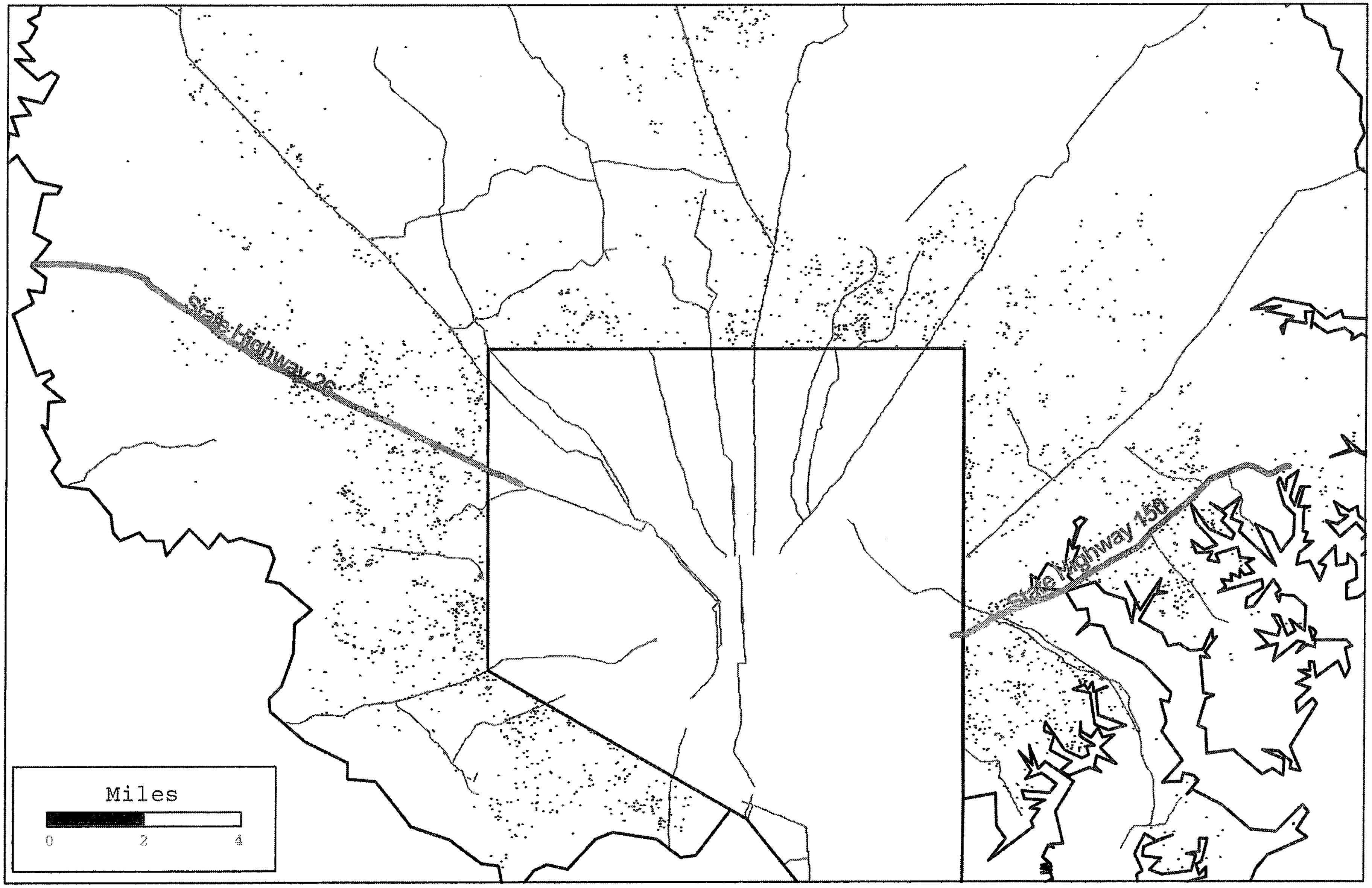


Table 5.3 presents the data. Using the GIS, we estimate that there are 3,333.54 miles of roadway segments; this number was estimated by adding up the total length of the street network in the GIS. Of all the road segments in Baltimore County, there are 241.04 miles of major arterial roads of which state highway 26 has a total length of 10.42 miles and state highway 150 has a total road length of 7.79 miles.

In 1996, there were 3,774 motor vehicle thefts in the county. If these thefts were distributed randomly, then the random expected distance between incidents would be 0.44 miles (equation 5.10). Using this estimate, table 5.3 shows the number of incidents that would be expected on each of the two state highways if the distribution were random and the ratio of the actual number of motor vehicle thefts to the expected number. As can be seen, the distribution of motor vehicle thefts is not random. On all major arterial roads, there are 2.2 times as many thefts as would be expected by a random spatial distribution. In fact, in 1996, of 28,551 road segments in Baltimore County, only 7791 (27%) had one or more motor vehicle thefts occur on them; most of these are major roads. Further, on highway 26 there were 7.4 times as much and on highway 150 there were 5.3 times as much as would be expected if the distribution was random. Clearly, these two highways had more than their share of auto thefts in 1996.

But what about the distribution of the incidents *along* each of these highways? If there were any pattern, for example, most of the incidents clustering on the western edge or in the center, then police could use that information to more efficiently deploy vehicles to respond quickly to events. On the other hand, if the distribution along these highways were no different than a random distribution, then police vehicles must be positioned in the middle, since that would minimize the distance to all occurring incidents.

Unfortunately, the results appear to be close to a random distribution. *CrimeStat* calculates that for highway 26, the average linear nearest neighbor distance is 0.05 miles which is close to the average random linear nearest neighbor distance (0.06 miles). The ratio - the linear nearest neighbor index, is 0.96 with a t-value of -0.16, which is not significantly different from chance. Similarly, for highway 150, the average linear nearest neighbor distance is 0.079 miles which, again, is almost identical to the average random linear nearest neighbor distance (0.084 miles); the nearest neighbor index is 0.94 and the t-value is -0.41 (not significant). In short, even though there was a higher concentration of vehicle thefts on these two state highways than would be expected on the basis of chance, the distribution *along* each highway is not very different than what would be expected on the basis of chance.[4]

K-Order Linear Nearest Neighbors

There is also a K-order linear nearest neighbor analysis, as with the areal nearest neighbors. The user can specify how many additional nearest neighbors are to be calculated. The linear K-order nearest neighbor routine returns four columns:

1. The order, starting from 1
2. The mean linear nearest neighbor distance for each order (in meters)

Table 5.3

Comparison of 1996 Baltimore County Auto Thefts for Different Types of Roads
(N = 3774 Incidents)

Length of Road Segments:

Highway 26	10.42 mi
Highway 150	7.79 mi
All Major Arterials	241.04 mi
All Roads	3333.54 mi

Random Expected Distance Between Incidents = 0.44 miles

Proportional To Network				**Proportional to Same Road**		
Where Incidents Occurred	**Number of Incidents**	**Expected Number *If* Random**	***"Relative to Random"* Ratio of Frequency**	**Average Linear Nearest Neighbor Distance**	**Average Random Linear Nearest Neighbor Distance**	***"Relative to Itself"* Linear Nearest Neighbor Index**
Highway 26	87	11.8	7.4	0.05 mi	0.06	0.96
Highway 150	47	8.8	5.3	0.08 mi	0.08	0.94
All Major Arterials	607	272.8	2.2	0.13 mi	0.20	0.64 (p≤.001)
All Roads	3774	3774.0	1.0	0.09 mi	044	0.21 (p≤.001)

3. The expected linear nearest neighbor distance for each order (in meters)
4. The linear nearest neighbor index for each order

Since the expected linear nearest neighbor distance has not been worked out for orders higher than one, the calculation produced here is a rough approximation. It applies equation 5.10 only adjusting for the decreasing sample size, N_k, which occurs as degrees of freedom are lost for each successive order. In this sense, the index is really the k-order linear nearest neighbor distance relative to the expected linear neighbor distance for the first order. It is not a strict nearest neighbor index for orders above one.

Nevertheless, like the areal k-order nearest neighbor index, the k-order linear nearest neighbor index can provide insights into the distribution of the points, even if the first-order is random. Figure 5.5 shows a graph of 50 linear nearest neighbors for 1996 residential burglaries and street robberies for Baltimore County. As with the areal k-order nearest neighbors (see figure 5.3) both burglaries and robberies show evidence of clustering. For both, the first nearest neighbors are closer together than a random distribution. Similarly, over the 50 orders, street robberies are more clustered than burglaries. However, measuring distance on a grid shows that for burglaries, there is only a small amount of clustering. After the fourth order neighbor, the distribution for burglaries is more dispersed than a random distribution. An interpretation of this is that there are small number of burglaries which are clustered, but the clusters are relatively dispersed. Street robberies, on the other hand, are highly clustered, up to over 30 nearest neighbors.

The linear k-order nearest neighbor distribution gives a slightly different perspective on the distribution than the areal. For one thing, the index is slightly biased as the denominator - the K-order expected linear neighbor distance, is only approximated. For another thing, the index measures distance *as if* the street follow a true grid, oriented in an east-west and north-south direction. In this sense, it may be unrealistic for many places, especially if streets traverse in diagonal patterns; in these cases, the use of indirect distance measurement will produce greater distances than what actually occur on the network. Still, the linear nearest neighbor index is an attempt to approximate travel along the street network. To the extent that a particular jurisdiction's street pattern fall in this manner, it can provide useful information.

Ripley's K Statistic

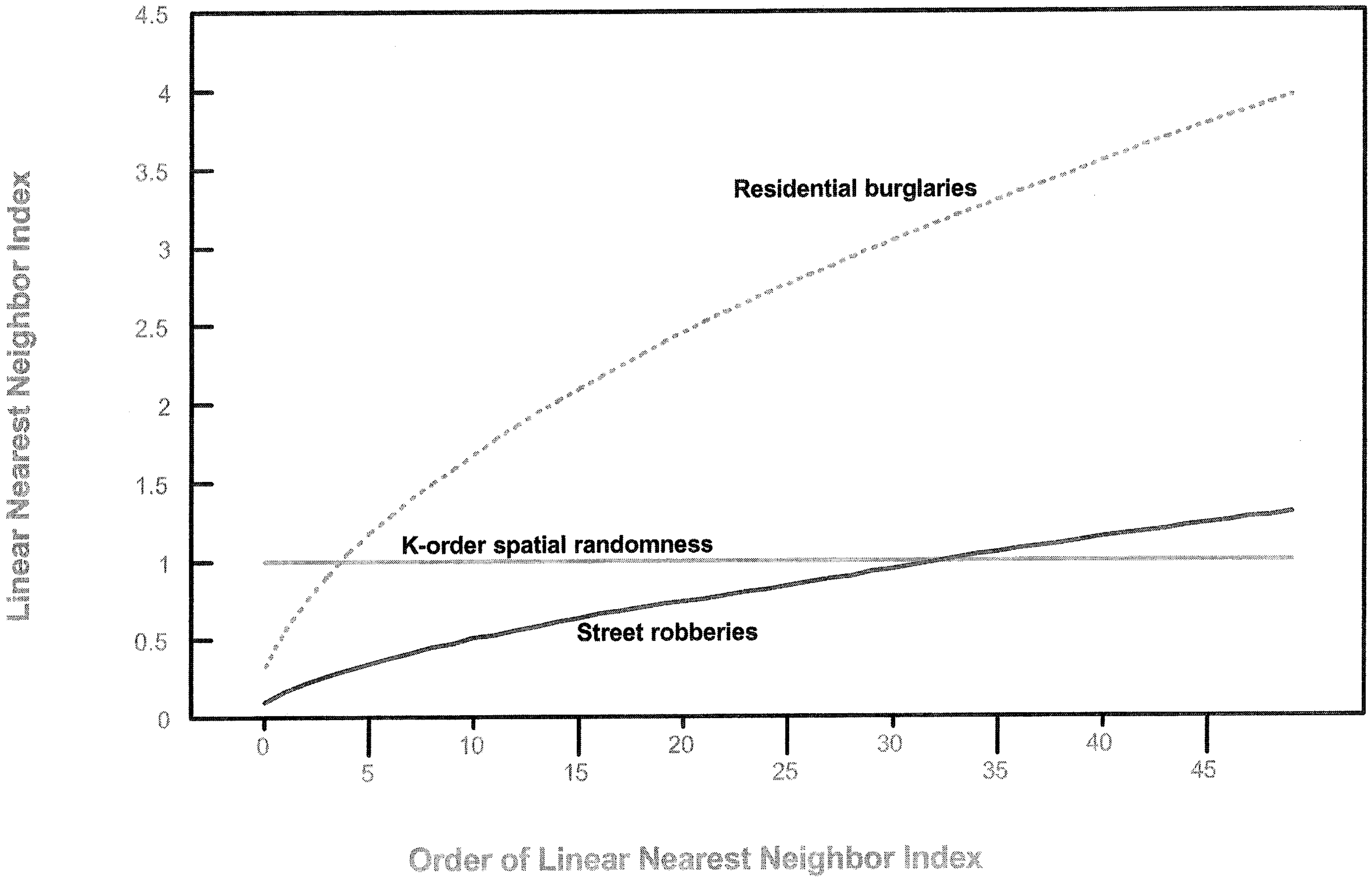
Ripley's K statistic is an index of non-randomness for different scale values (Ripley, 1976; Ripley, 1981; Bailey and Gattrell, 1995; Venables and Ripley, 1997) . In this sense, it is a 'super-order' nearest neighbor statistic, providing a test of randomness for every distance from the smallest up to the size of the study area. It is sometimes called the *reduced second moment measure*, implying that it is designed to measure second-order trends (i.e., local clustering as opposed to a general pattern over the region). However, it is also subject to first-order effects so that it is not strictly a second-order measure.

Consider a *spatially random* distribution of N points. If circles of radius, d_s, are drawn around each point, where s is the order of radii from the smallest to the largest, and the

Figure 5.5

K-Order Linear Nearest Neighbor Indices

1996 Street Robberies and Residential Burglaries



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

number of other points that are found within the circle are counted and then summed over all points (allowing for duplication), then the expected number of points within that radius are

$$E(\#\text{ of points within distance } d_i) = \frac{N}{A} K(d_s) \qquad (5.15)$$

where N is the sample size, A is the total study area, and $K(d_s)$ is the area of a circle defined by radius, d_s. For example, if the area defined by a particular radius is one-fourth the total study area and *if* there is a spatially random distribution, on average approximately one-fourth of the cases will fall within any one circle (plus or minus a sampling error). More formally, with *complete spatial randomness* (csr), the expected number of points within distance, d_s, is

$$E(\#\text{ under csr}) = \frac{N}{A} \pi d_s^2 \qquad (5.16)$$

On the other hand, if the average number of points found within a circle for a particular radius placed over each point, in turn, is greater than that found in equation 5.16, this points to clustering, that is points are, on average, closer than would be expected on the basis of chance for that radius. Conversely, if the average number of points found within a circle for a particular radius placed over each point, in turn, is less than that found in equation 5.16, this points to dispersion; that is points are, on average, farther apart than would be expected on the basis of chance for that radius. By counting the number of total numbers within a particular radius and comparing it to the number expected on the basis of complete spatial randomness, the statistic is an indicator of non-randomness.

In this sense, the K statistic is similar to the nearest neighbor distance in that it provides information about the average distance between points. However, it is more comprehensive than the nearest neighbor statistic for two reasons. First, it applies to all orders cumulatively, not just a single order. Second, it applies to all distances up to the limit of the study area because the count is conducted over successively increasing radii.

Under unconstrained conditions, K is defined as

$$K(d_s) = \frac{A}{N^2} \sum_i \sum_j I\,(d_{ij}) \qquad (5.17)$$

where $I\,(d_{ij})$ is the number of other points, j, found within distance, d_s, summed over all points, i. That is, a circle of radius, d_s, is placed over each point, i. Then, the number of other points, ij, are counted. The circle is moved to the next i and the process is repeated. Thus, the double summation points to the count of all j's for each i, over all i's. After this process is completed, the radius of the circle is increased, and the entire process is repeated. Typically, the radii of circles are increased in small increments so that there are 50-100 intervals by which the statistic can be counted. In *CrimeStat*, 100 intervals (radii) are used, based on

$$d_s = \frac{R}{100} \qquad (5.18)$$

where R is the radius of a circle for whose area is equal to the study area (i.e., the area entered on the measurement parameters page).

One can graph $K(d_s)$ against the distance, d_s, to reveal whether there is any clustering at certain distances or any dispersion at others (if there is clustering at some scales, then there must be dispersion at others). Such a plot is non-linear, however, typically increasing exponentially (Kaluzny et al, 1998. Consequently, $K(d_s)$ is transformed into a square root function, $L(d_s)$, to make it more linear. $L(d_s)$ is defined as:

$$L(d_s) = \text{SQRT}\left[\frac{K(d_s)}{\pi}\right] - d_s \qquad (5.19)$$

That is, $K(d_s)$ is divided by π and then the square root is taken. Then the distance interval (the particular radius), d_s, is subtracted from this.[5] In practice, only the L statistic is used even though the name of the statistic K is based on the K derivation. Figure 5.6 shows a graph of L against distance for 1996 robberies in Baltimore County. As can be seen, L increases up to a distance of about 3 miles whereupon it decreases again.

Comparison to A Spatially Random Distribution

To understand whether an observed K distribution is different from chance, one typically uses a random distribution. Because the sampling distribution of $L(d_s)$ is not known, a simulation can be conducted by randomly assigning points to the study area. Because any one simulation might produce a clustered or dispersed pattern strictly by chance, the simulation is repeated many times, typically 100 or more. Then, for each random simulation, the L statistic is calculated for each distance interval. Finally, after all simulations have been conducted, the highest and lowest L-values are taken for each distance interval. This is called an *envelope*. Thus, by comparing the distribution of L to the random envelope, one can assess whether the particular observed pattern is likely to be different from chance.[6]

Specifying simulations

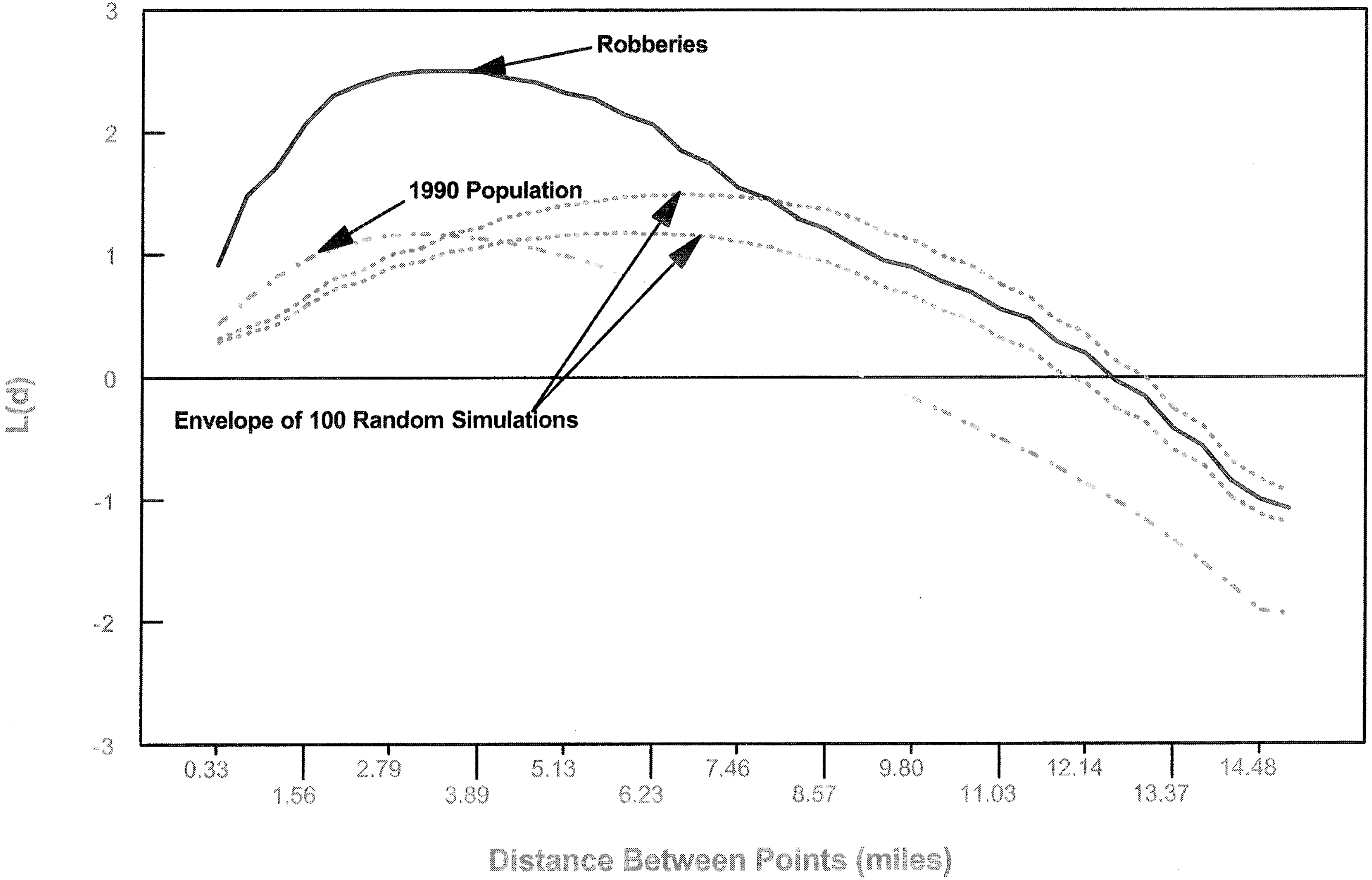
Because simulations can take a long time, particularly if the data sets are large, the default number of simulations is 0. However, a user can conduct simulations by writing a positive number (e.g., 10, 100, 300). If simulations are selected, *CrimeStat* will conduct the number of simulations specified by the user and will calculate the upper and lower limits for each distance interval, as well as the 0.5%, 2.5%, 5%, 95%, 97.5% and 99% intervals; these latter statistics only make sense if many simulation runs are conducted (e.g. 1000).

Figure 5.6:

K Statistic For 1996 Robberies

Compared to Random and Population Distributions

L(d) = Sqrt[K(d)/pi] - d



The way *CrimeStat* conducts the simulation is as follows. It takes the maximum bounding rectangle of the distribution, that is the rectangle formed by the maximum and minimum X and Y coordinates respectively and re-scales this (up or down) until the rectangle has an area equal to the study area (defined on the measurement parameters page). It then assigns N points, where N is the same number of points as in the incident distribution, using a uniform random number generator to this rectangle and calculates the L statistic. It then repeats the experiment for the number of specified simulations, and calculates the above statistics. For example, with 1181 robberies for 1996, the Ripley's K function calculates the empirical L statistics for 100 distance intervals and compares this to a simulation of 1181 points randomly distributed over a rectangle k times, where k is a user-defined number.

In practice, the simulation test also has biases associated with edges. Unlike the theoretical L under uniform conditions of complete spatial randomness (i.e., stretching in all directions well beyond the study area) where L is a straight horizontal line, the simulated L also declines with increasing distance separation between points. This is a function of the same type of edge bias. Consequently, it is possible to compare the empirical L with the random L for even longer distance separations since both have edge biases. There are some subtle differences between the two, however, so some care should be used. The empirical L is obtained from the points within the study area, the geography of which is usually irregular. The random L, however, is calculated from a rectangle. Thus, the differences in the shape comparisons may account for some variations.

Comparison to Baseline Populations

For most social distributions, such as crime incidents, randomness is not a very meaningful baseline. Most social characteristics are non-random. Consequently, to find that the amount of clustering that is occurring is greater than what would be expected on the basis of chance is not very useful for crime analysts. However, it is possible to compare the distribution of L for crime incidents with the distribution of L for various baseline characteristics, for example, for the population distribution or the distribution of employment. In almost all metropolitan areas, population is more concentrated towards the center than at the periphery; the drop-off in population density is very sharp as was shown in the last chapter. All other things being equal, one would expect more incidents towards the metropolitan center than at the periphery; consequently, the average distance between incidents will be shorter in the center than farther out. This is nothing more than a consequence of the distribution of people. However, to say something about concentrations of incidents above-and-beyond that expected by population requires us to examine the pattern of population as well as of crime incidents.

CrimeStat allows the use of intensity and weighting variables in the calculation of the K statistic. The user must define an intensity or a weight (or both in special circumstances) on the primary file page. The K routine will then use the intensity (or weight) in the calculation of L. In Figure 5.6 above, there is an envelope produced from 100 random simulations as well as the L distribution from the 1990 population; the latter variable was obtained by taking the centroid of census block groups from the 1990 census and using population as the intensity variable. As can be seen, the amount of clustering for robberies is much greater than both the random envelope as well as the distribution of population. In other words, robberies are more clustered together than even what would be expected on the basis of the population distribution and this holds for distances up to about 7 miles, whereupon the distribution of robberies is indistinguishable from a random distribution. For comparison, figure 5.7 below shows the distribution of 1996 burglaries, again compared to a random envelope and the distribution of population. We find that burglaries are more clustered than even population, but less so than for robberies; the L value is higher for robberies than for burglaries for near distances. Thus, the distribution of L confirms the result that burglaries tend to be spread over a much larger geographical area in smaller clusters than street robberies, which tend to be more concentrated in large clusters. In terms of looking for 'hot spots', one would expect to find more with robberies than with burglaries.

Edge Corrections for Ripley's K

The L statistic is prone to edge effects just like the nearest neighbor statistic. That is, for points located near the boundary of the study area, the number enumerated by any circle for those points will, all other things being equal, necessarily be less than points in the center of the study area because points outside the boundary are not counted. Further, the greater the distance between points that are being tested (i.e., the greater the radius of the circle placed over each point), the greater the bias. Thus, a plot of L against distance will show a declining curve as distance increases as figures 5.6 and 5.7 show.

There are various adjustments to the function to help correct the bias. One is a 'guard rail' within the study area so that points outside the guard rail, but inside the study area can only be counted for points inside the guard rail, but cannot be used for enumerating other points within a circle placed over them (that is, they can only be j's and not i's, to use the language of equation 5.17). Such an operation, however, requires manually constructing these guard rails and enumerating whether each point can be both an enumerator and a recipient or a recipient only. For complex boundaries, such as are found in most police departments, this type of operation is extremely tedious and difficult.[7]

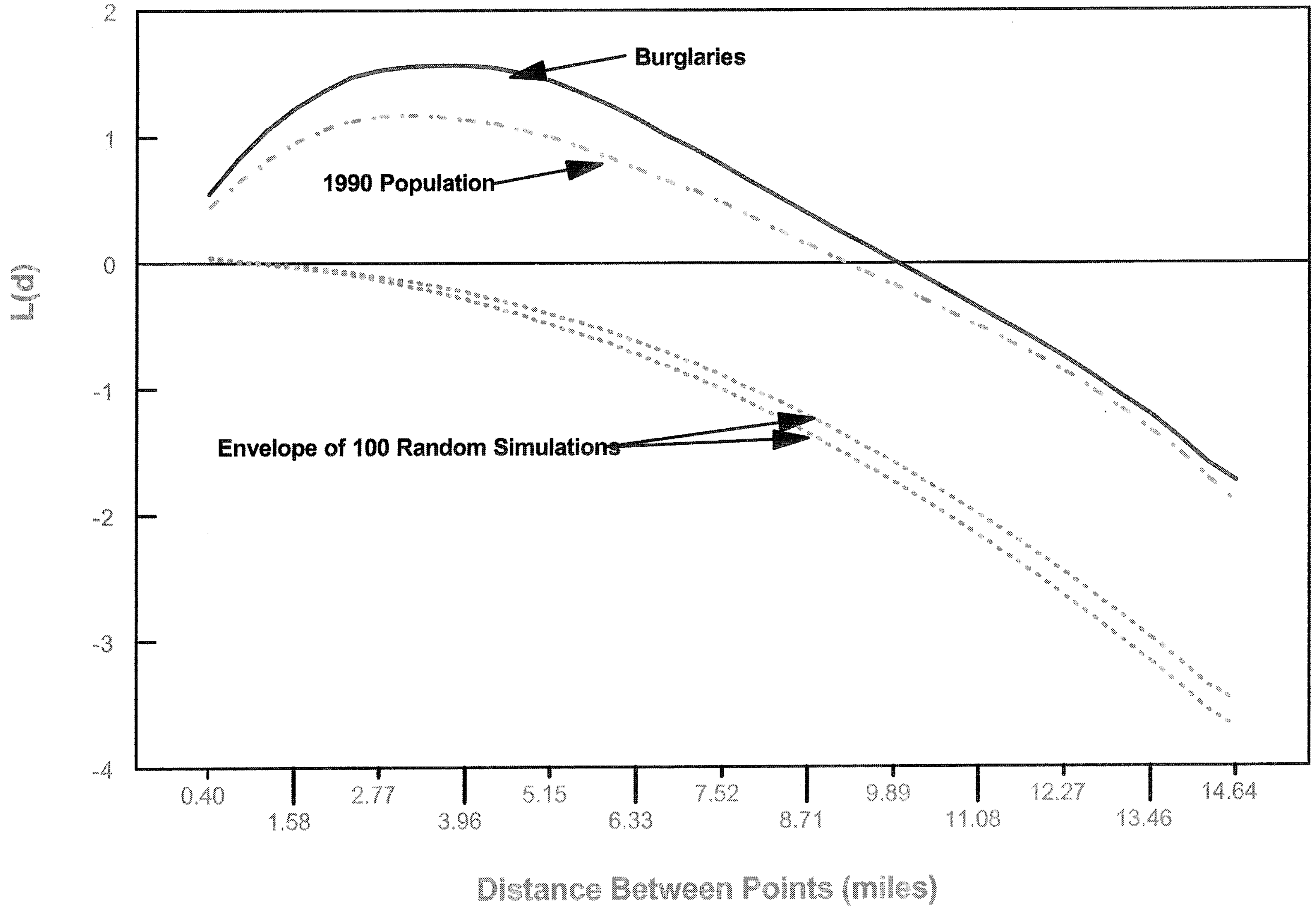
Similarly, Ripley has proposed a simple weighting to account for the proportion of the circle placed over each point that is within the study area (Venables and Ripley, 1997). Thus, equation 5.17 is re-written as:

$$K(d_s) = \frac{A}{N^2} \sum_i \sum_j W_{ij}^{-1} I(d_{ij}) \qquad (5.20)$$

where W_{ij}^{-1} is the inverse of the proportion of a circle of radius, d_s, placed over each point which is within the total study area. Thus, if a point is near the study area border, it will receive a greater weight because a smaller proportion of the circle placed over it will be within the study area.

K Statistic For 1996 Burglaries
Compared to Random and Population Distributions

L(d) = Sqrt[K(d)/pi] - d



Using this latter concept, two edge corrections for Ripley's K statistic are provided, also following rectangular and circular models. The logic is slightly different than with the edge corrections for the nearest neighbor index. The Ripley's K routine places a search circle of radius, R_j, over each point and the number of other points within the circle is counted. The circle is moved to the next point and the cumulative count continued. After all points are visited by the circle and a cumulative count enumerated, the count is transformed into K and then L (see chapter 4). The process then continues with a slightly larger radius, $R_j + i$, where i is the bin width.

Ripley's K has the same potential edge problem as the nearest neighbor index. For points located near the border of the study area, the cumulative count will frequently be smaller than points more central because there are no measured points that fall within the circle beyond the border. Thus, they underestimate the number of points found within a certain distance. Ripley (1976) suggested that each point be weighted by the inverse of the proportion of the search circle within the study area.

Define this as an edge weight, E,

$$E = 1/p \qquad (5.21)$$

where p is the proportion of the search circle within the study area. If the entire search circle is within the study area, then E = 1/1 = 1. If the point is on the border of the study area, then for the rectangle only half the radius of the search circle is within the study area and E= 1/0.5 = 2; for the circle, it is slightly less than half. In between are various values of E (i.e., E varies between 1 and 2).

The following is an approximation of the intermediate weights (between 1 and approximately 2) using either a rectangular or circular correction.

Rectangular correction

In the rectangular correction for Ripley's K, the search circle radius, R_j, is compared to the edge of an assumed rectangle with area, A, centered at the mean center. First, the area to be analyzed is defined. If the user has specified a study area on the measurement parameters page, then that value for A is taken. The maximum bounding rectangle is taken (i.e., rectangle defined by the minimum and maximum X/Y values) and proportionately re-scaled so that the area of the rectangle is equal to A. If the user does not specify an area on the measurement parameters page, then the maximum bounding rectangle is taken for A.

Second, for each point, the minimum distance to the nearest edge of this rectangle is calculated in both the horizontal and vertical directions, $d(minR_X)$ and $d(minR_Y)$. Third, each of the minimum distances are compared to the search circle radius, R_j:

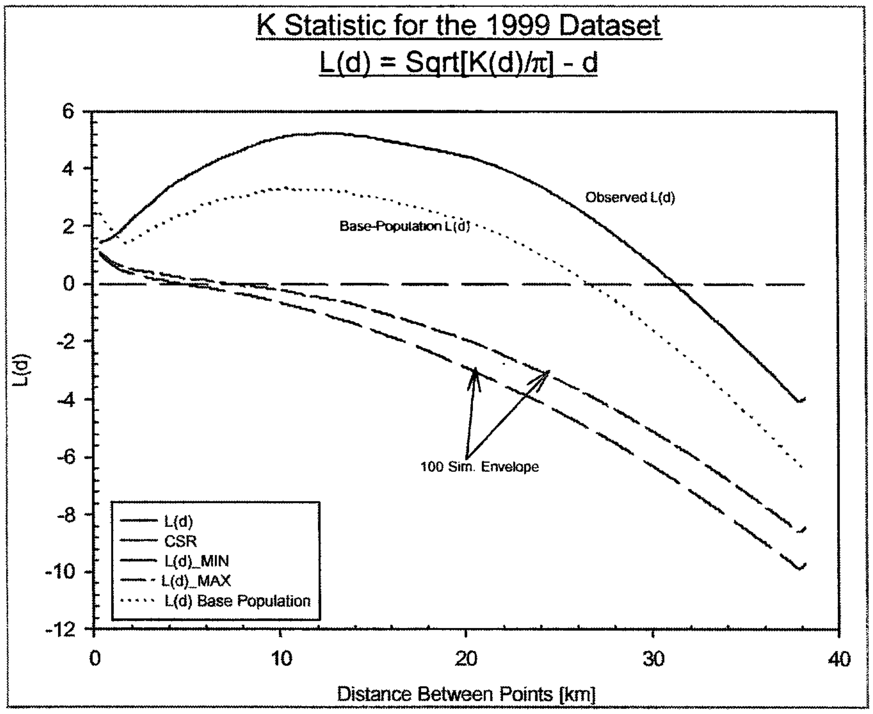
1. If neither the minimum distance in the X-direction - $d(minR_X)$, nor the minimum distance in the Y-direction - $d(minR_Y)$, are less than the search circle radius, R_j, then the circle falls entirely within the rectangle and E = 1;

K-Function Analysis to Determine Clustering in the *Police Confrontations* Dataset in Buenos Aires Province, Argentina: 1999

Gastón Pezzuchi, Crime Analyst
Buenos Aires Province Police Force
Buenos Aires, Argentina

Sometimes crime analysts tend to produce beautiful hot spot maps without any formal evidence that clustering is indeed present in the data. One excellent and powerful tool that *CrimeStat* provides is the computation of the K function, which summarizes spatial dependence over a wide range of scales, and uses the information of all events.

We computed the K function using 1999 police confrontations data (mostly shootings) within our study area[1] and ran 100 Monte Carlo simulations in order to test for spatial randomness[2] (see figure below); the K function showed clustering up to about 30 Km. Yet, spatial randomness is not a particularly meaningful hypothesis to test considering that the "population at risk" are highly clustered. Hence we used police deployment data as a base population and calculated the K function for that data set. As can seen, the amount of clustering for the confrontation dataset is much greater than both the random envelope as well as the distribution of police officers.



[1] A years worth dataset of events occurring within a 9,500 km2 area around the Federal Capital (29 counties).

[2] Remember that Pr(L(d) > Lmax) = Pr(L(d) < Lmin) = 1 / (m + 1) where m is the number of independent simulations,

2. If either the minimum distance in the X-direction - d(minR_X), or the minimum distance in the Y-direction - d(minR_Y), but NOT BOTH, are less than the search circle radius, R_j, then part of the search circle falls outside the rectangle and an adjustment is necessary. An approximate adjustment is made than is inversely proportional to the area of the search circle within the rectangle. The values of E will vary between 1 and 2 since up to one-half of the search circle could fall outside the rectangle;

3. If both the minimum distance in the X-direction - d(minR_X), and the minimum distance in the Y-direction - d(minR_Y), are less than the search circle radius, R_j, then a greater adjustment is required since E could vary between 1 and 4 since up to three-fourth of the search circle could fall outside the rectangle.

Circular correction

In the circular correction for Ripley's K, the search circle radius, R_j, is compared to the edge of an assumed circle with area, A, centered at the mean center. First, the area to be analyzed is defined. If the user ha specified a study area on the measurement parameters page, then that value for a is taken. The radius of the circle, R_j, is calculated by equation 5.8 above. If the user has not specified a study area on the measurement parameters page, then A is calculated from the maximum bounding rectangle and the radius of the circle is calculated by equation 5.8 above.

Second, for each point, the distance from that point to the mean center, R_j, is calculated. The nearest distance from the point to the circle's edge is given by

$$R_{jC} = R - R_j \tag{5.22}$$

Third, the search circle radius, R_j, is compared to the nearest edge of the circle, R_{iC}:

1. If the search area radius, R_j, is less than or equal to R_{jC}, then the entire search circle falls within the model circle and E=1.

2. If the search area radius, R_j, is greater than R_{jC}, then an adjustment is made for the approximate proportion of the search circle within the model circle with E varying between 1 and 2.2.

For either correction

During the calculation of Ripley's K, each point is multiplied by E (aside from W or I) and the K and L statistics are calculated as before (see chapter 5). The simulation of random point distributions is treated in an analogous way. Figure 5.8 below shows a Ripley's K distribution for 1996 Baltimore County burglaries, with and without edge corrections. As can be seen, the uncorrected L distribution (the transformation of K) decreases and falls below the theoretical random count (L=0) after about 8 miles whereas neither the L distribution with the rectangular correction nor the L distribution with the circular distribution do so. As expected, the rectangular distribution produces the most concentration.

Distance Matrices

CrimeStat has the capability for outputting distance matrices. There are two types of matrices that can be output. First, the distance between every point in the primary file and every other point can be calculated in miles, nautical miles, feet, kilometers or meters. This is called the *within file point-to-point matrix* (Matrix). Second, if there is also a secondary file, *CrimeStat* can calculate the distance from every point in the primary file to every point in the secondary file, again in miles, nautical miles, feet, kilometers or meters. This is called the *From all primary file points to all secondary file points matrix* (Imatrix).

Both types of matrices can be displayed or saved to a text file for import into another program. Each matrix defines incidents by the order in which they occur in the files (i.e., Record number 1 is listed as '1'; record number 2 is listed '2'; and so forth). Only a subset of each matrix is displayed on the results tab. However, there are horizontal and vertical slider bars that allow the user to scroll through the matrix. The user should move the vertical slide bar first to an approximate proportion of the matrix and click the *Go* button. The matrix will scroll through the rows of the matrix to a place which represents that proportion indicated in the slide bar. The user can then scroll across the rows with the upper slide bar.

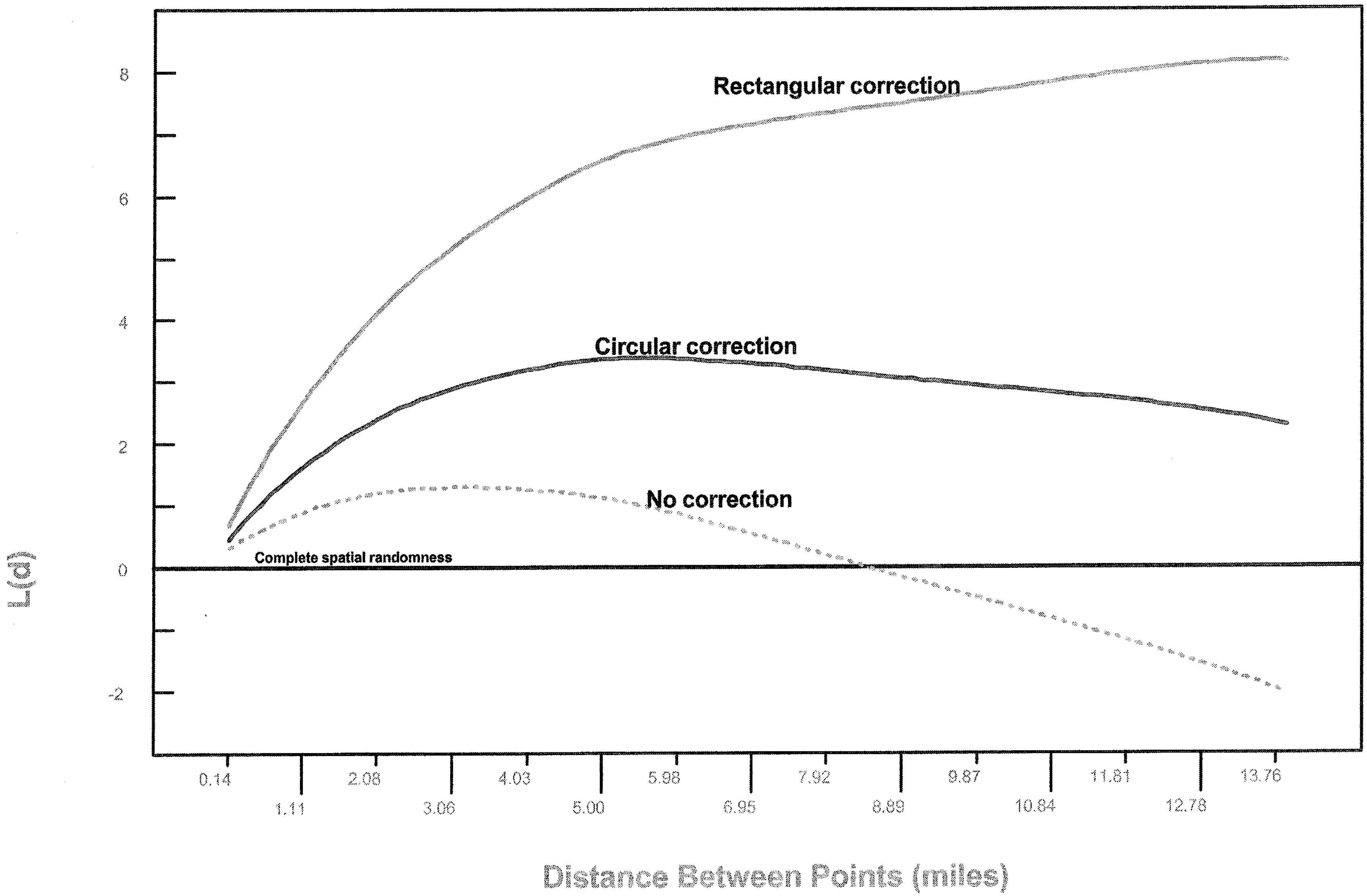
The matrices can be used for various purposes. The *within file point-to-point matrix* can be used to examine distances between particular incidents. The *saved '.txt' matrix* can also be imported into a network program for estimating transportation routes. The *primary-to-secondary file matrix* can be used in optimization routines, for example in trying to assess optimal allocation of police cars in order to minimize response time in a police district.

The next chapter will discuss how to identify 'hot spots' with *CrimeStat*.

Figure 5.8:

"K" Statistic For 1996 Burglaries With Different Types of Corrections

$L(d) = Sqrt[K(d)/pi] - d$



Endnotes for Chapter 5

1. There is also a mean random distance for a dispersed pattern, called the *mean dispersed distance* (Ebdon, 1988). It is defined as

$$d(dis) = \frac{SQRT[2]}{3^{1/4}\ SQRT[\ N/A\ ]}$$

A nearest neighbor index can be set up comparing the observed mean neighbor distance with that expected for a dispersed pattern. *CrimeStat* only provides the traditional nearest neighbor index, but it does output the mean dispersed distance.

2. Unfortunately, the term *order* when used in the context of nearest neighbor analysis has a slightly different meaning than when used as *first-order* compared to *second-order* statistics. In the nearest neighbor context, *order* really means *neighbor* whereas in the type of statistics context, *order* means the scale of the statistics, global or local. The use of the terms is historical.

3. There is not a hard-and-fast rule about how many K-order nearest neighbor distances may be calculated. Cressie (1991, p. 613) shows that error increases with increasing order and the degree of divergence from an edge-corrected measure increases over time. In a test case of 584 point locations, he shows that even after only 25 nearest neighbors, the uncorrected measure yields opposite conclusions about clustering from the corrected measures. So, as a rough approximation, orders no greater than 2.5% of the cases should be calculated.

4. Because *CrimeStat* uses indirect distance for the linear nearest neighbor index (i.e. measurement only in an horizontal or vertical direction), there is a slight distortion that can occur if the incidents are distributed in a diagonal manner, such as with State Highways 26 and 150 in Figure 5.4. The distortion is very small, however. For example, with the incidents along State Highway 26, after rotating the incident points so that they fell approximately in a horizontal orientation, the observed average linear nearest neighbor distance decreased slightly from 0.05843 miles to 0.05061 miles and the linear nearest neighbor index became 0.8354 (t=-.91; not significant). In other words, the effects of the diagonal distribution lengthened the estimate for the average linear nearest neighbor distance by about 41 feet compared to the actual distances between incidents. For a small sample size, this could be relevant, but for a larger sample it generally will be a small distortion. However, if a more precise measure is required, then the user should rotation the distribution so that the incidents have as closely as possible a horizontal or vertical orientation.

5. This form of the $L(d_s)$ is taken from Cressie (1991). In Ripley's original formulation (Ripley, 1976), distance is not subtracted from the square root function. The advantage of the Cressie formulation is that a complete random distribution will be a straight line that is parallel to the X-axis.

6. Note, that since there is not a formal test of significance, the comparison with an envelope produced from a number of simulations provides only approximate confidence about whether the distribution differs from chance or not. That is, one cannot say that the likelihood of obtaining this result by chance is less than 5%, for example.

7. The 'guard rail' concept, while frequently used, is poor methodology because it involves ignoring data near the boundary of a study area. That is, points within the guard rail are only allowed to be selected by other points and not, in turn, be allowed to select others. This has the effect of throwing out data that could be very important. It is analogous to the old, but fortunately now discarded, practice of throwing out 'outliers' in regression analysis because the outliers were somehow seen as 'not typical'. The guard rail concept is also poor policing practice since incidents occurring near a border may be very important to a police department and may require coordination with an adjacent jurisdiction. In short, use mathematical adjustments for edge corrections or, failing that, leave the data as it is.

Chapter 6
'Hot Spot' Analysis I

In this and the next chapter, we describe seven tools for identifying clusters of crime incidents. The discussion has been divided into two 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. The next chapter discusses STAC, the K-means algorithm, and Anselin's Local Moran statistics. However, the seven 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. Police have learned from experience that there are particular environments that attract drug trading and 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 and Green, 1995; Sherman, Gartin and Buerger, 1989; Maltz, Gordon, and 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 and Block, 1995; Levine, Wachs and 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. 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 and Weisburd, 1995). In short, the concept is very useful. Nevertheless, the concept is a perceptual construct. 'Hot spots' may not exist in reality, but could be areas where there is sufficient concentration of certain activities (in this case, crime incidents) such that they get labeled as being an area of high concentration. There is not a boundary 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, 1974). 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, 1974; Çan 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. *Hierarchical* techniques (Sneath, 1957; McQuitty, 1960; Sokal and Sneath, 1963; King, 1967; Sokal and 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 fails. Thus, there is a hierarchy of clusters that can be displayed with a dendogram (an inverted tree diagram).

   Figure 6.1 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 clusters which, in turn, are grouped into second-order clusters which, in turn, are grouped into third-order clusters 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 two hierarchical techniques: a *Nearest Neighbor Hierarchical Clustering* routine in this chapter and the *Spatial and Temporal Analysis of Crime* module (STAC) which will be discussed in chapter 7;

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 6.2 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;

Figure 6.1:

Hierarchical Clustering Technique

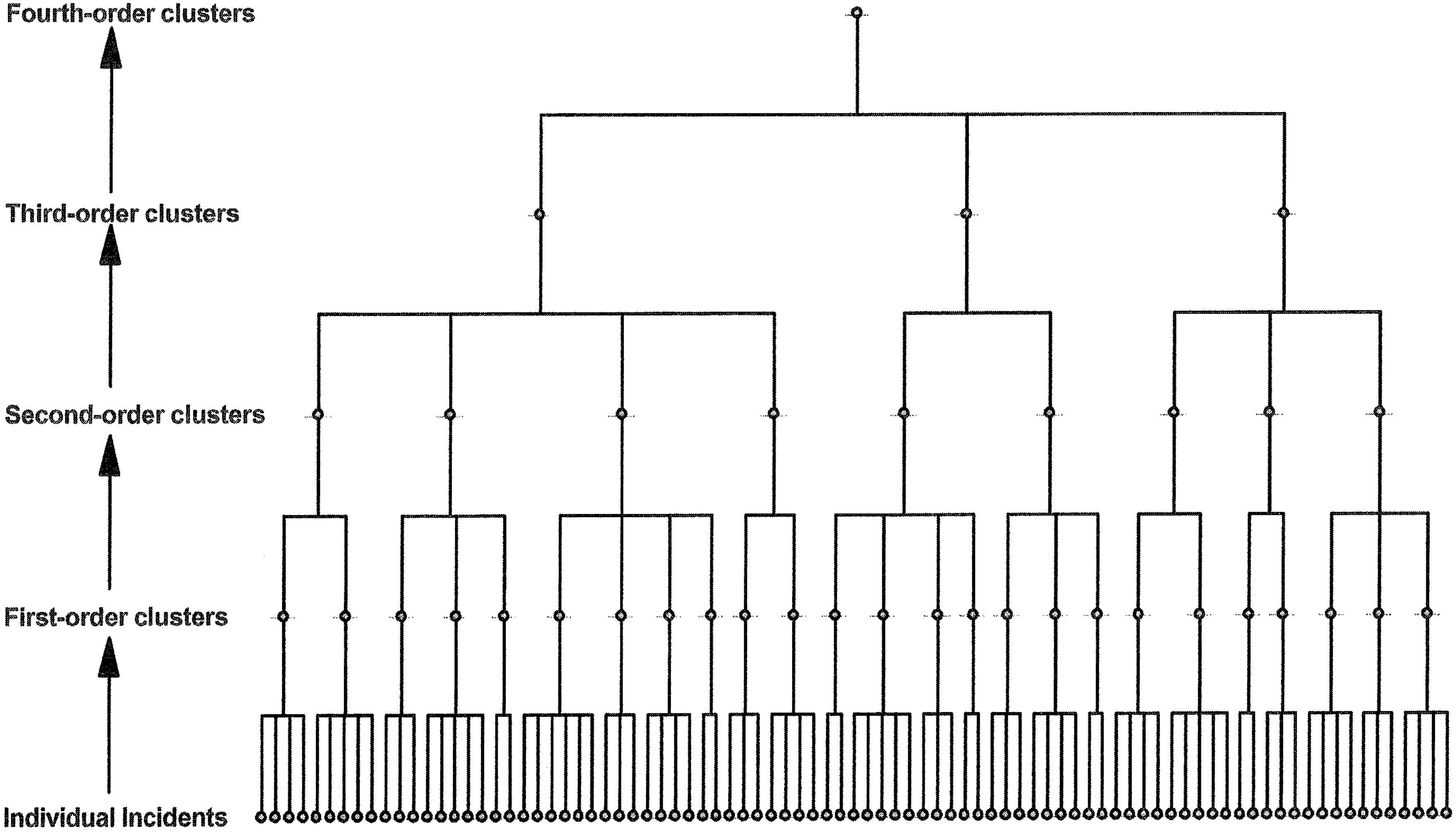
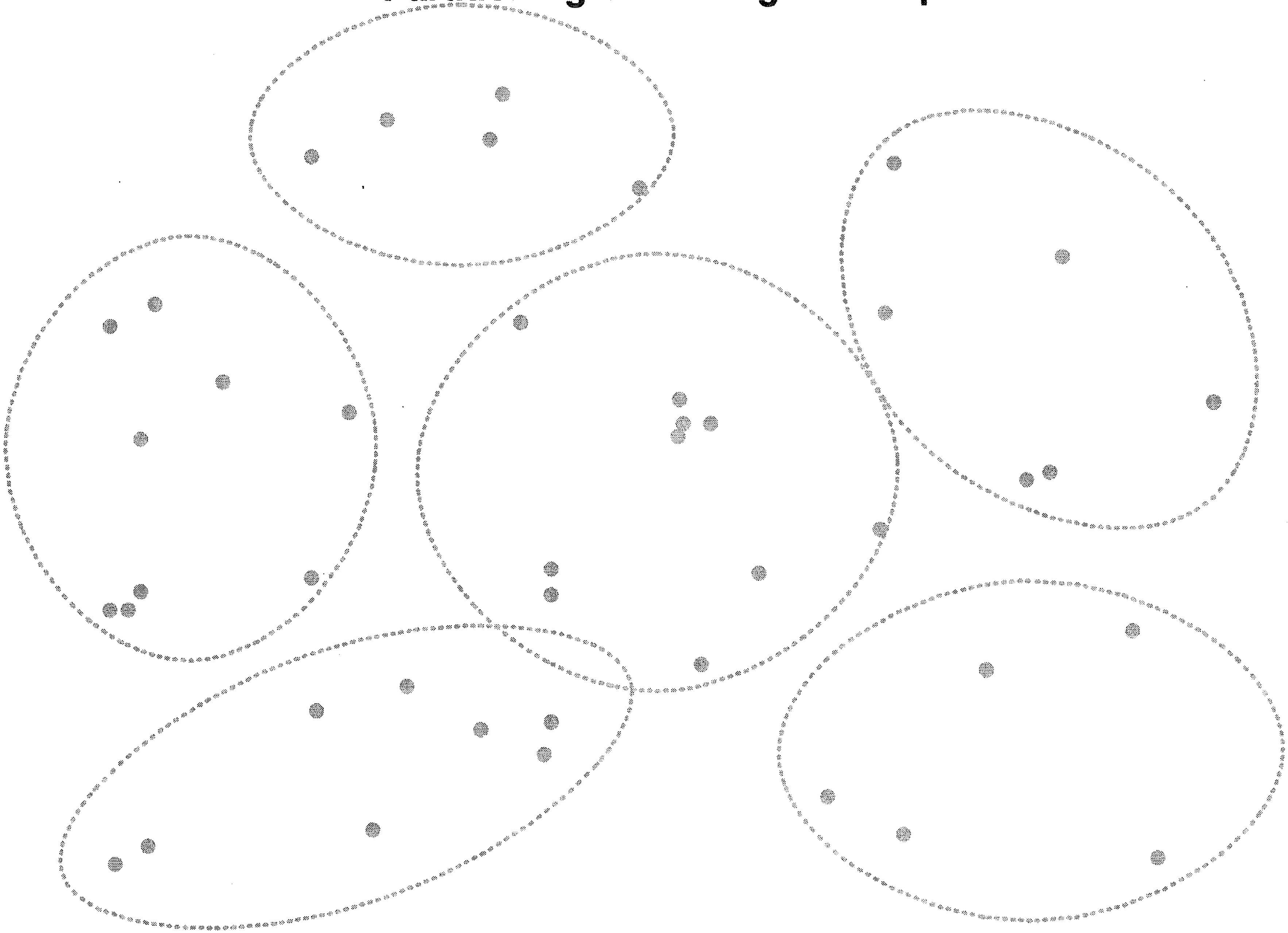


Figure 6.2:

Partitioning Clustering Technique



4. *Density* techniques identify clusters by searching for dense concentrations of incidents (Carmichael et al, 1968; Gitman and Levine, 1970; Cattell and Coulter, 1966; Wishart, 1969). *CrimeStat* has one type of density search algorithm using the *Single Kernel Density* method and will be presented in chapter 8;

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

6. *Risk-based* techniques identify clusters in relation to an underlying base 'at risk' variable, such as population, employment, or active targets (Jefferis, 1998; Kulldorff, 1997; Kulldorff and Nagarwalla, 1995). *CrimeStat* includes two risk-based technique - a *Risk-adjusted Nearest Neighbor Hierarchical Clustering* routine and a *Duel Kernel Density* method; and

7. *Miscellaneous* techniques are other methods that are less commonly used including techniques applied to zones, not incidents. *CrimeStat* includes *Anselin's Local Moran* technique for identifying neighborhood discrepancies (Anselin, 1995).

There are also hybrids between these methods. For example, *STAC* is primarily a partitioning method but with elements of hierarchical grouping (Block and Green, 1994).

Optimization Criteria

In addition to the different types of cluster analysis, there are different criteria that distinguish techniques applied to space. 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. Nevertheless, it should be clear that with the several types of cluster analysis and the many criteria that can be used for any particular technique, there is a large number of different cluster 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 which are unequivocally clear. In most cases analysts are not even sure why there are hot spots in the first place and, until that is solved, it would be unreasonable to expect a mathematical or statistical routine to solve that problem.

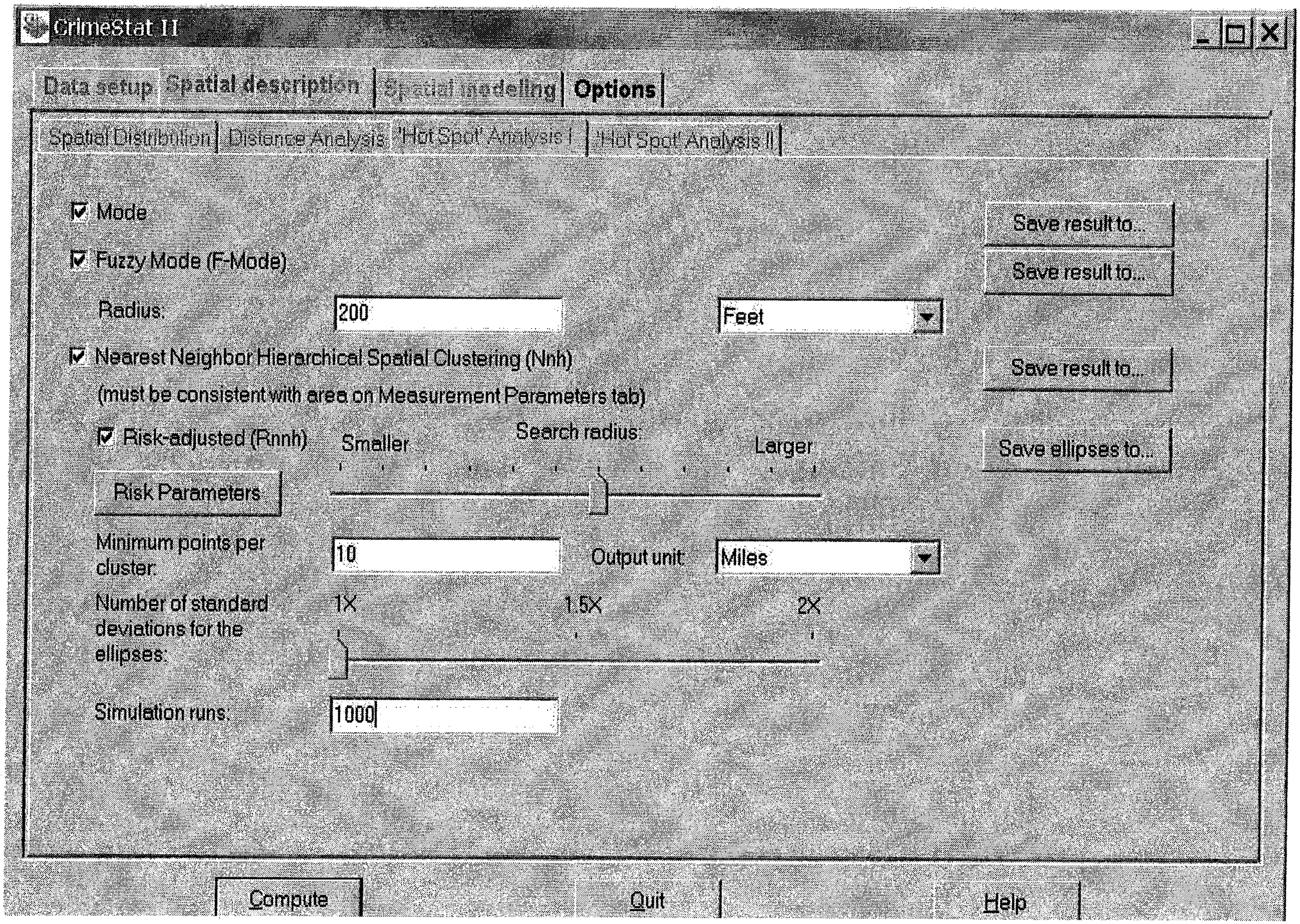
Cluster Routines in *CrimeStat*

Because of the variety of cluster techniques, *CrimeStat* includes seven 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 statistic

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 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. Figure 6.3 shows the 'Hot Spot' Analysis I page.

Figure 6.3: 'Hot Spot' Analysis I Screen



CrimeStat II

Data setup | Spatial description | Spatial modeling | Options

Spatial Distribution | Distance Analysis | 'Hot Spot' Analysis I | 'Hot Spot' Analysis II

☑ Mode — Save result to...

☑ Fuzzy Mode (F-Mode) — Save result to...

Radius: 200 — Feet

☑ Nearest Neighbor Hierarchical Spatial Clustering (Nnh) — Save result to...

(must be consistent with area on Measurement Parameters tab)

☑ Risk-adjusted (Rnnh) — Search radius: Smaller ... Larger — Save ellipses to...

Risk Parameters

Minimum points per cluster: 10 — Output unit: Miles

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

Simulation runs: 1000

Compute | Quit | Help

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

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 6.1 presents the formatted output for the ten most frequent locations for motor vehicle thefts in the Baltimore region in 1996 (the rest were ignored) and figure 6.4 maps the ten locations.[1] 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 number 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.

The mode is a very simple measure, but one that can be very useful. In the example, it's 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 managements 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.

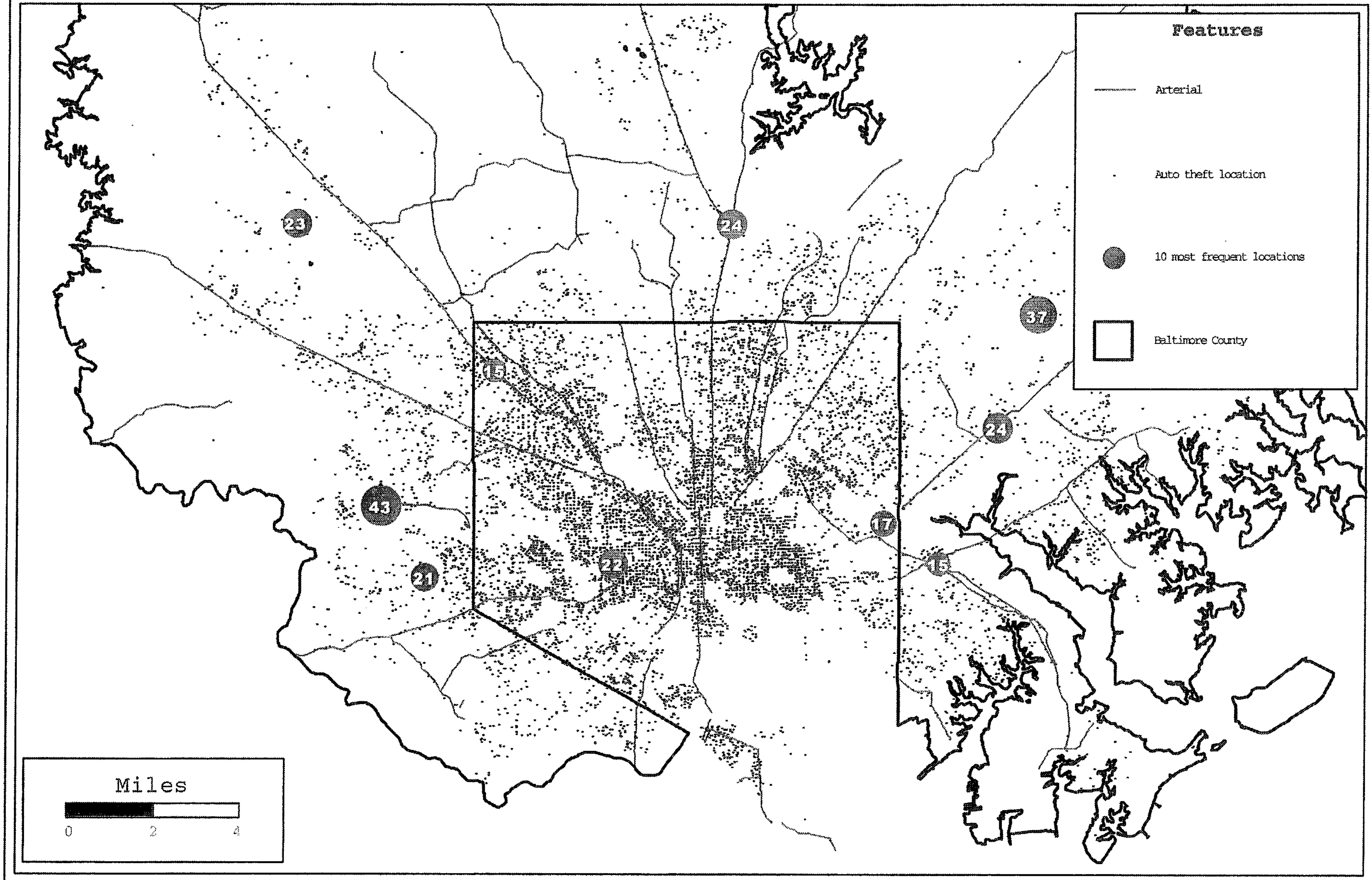
Fuzzy Mode

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

Figure 6.4:

Ten Most Frequent Locations for Motor Vehicle Theft

Symbol Area Proportional to Frequency



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Table 6.1

Mode Output for
Most Frequent Locations for Motor Vehicle Thefts
Baltimore: 1990

```
Mode:
-------

        Sample size............: 14853
        Measurement type.......: Direct
        Start time.............: 12:46:15 PM, 07/15/2001
        End time...............: 12:50:19 PM, 07/15/2001

        Displaying 45 results(s) starting from 1 (ONLY 10 SHOWN)

        Rank  Freq      X            Y
        ----  ----  ---------    --------
        1     43    -76.75070    39.31150
        2     37    -76.47100    39.37410
        3     24    -76.48800    39.33720
        4     24    -76.60150    39.40420
        5     23    -76.78770    39.40460
        6     22    -76.65170    39.29270
        7     21    -76.73190    39.28800
        8     17    -76.53630    39.30600
        9     15    -76.70260    39.35600
        10    15    -76.51280    39.29270
```

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 in different buildings with different addresses.

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 (150 feet) 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 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. 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 6.5 maps the top 13 locations for vehicle thefts identified by the fuzzy mode routine using a search radius of 100 yards. Thirteen locations are included because four were tied for number 10. The 13 locations are displayed by a magenta triangle and are compared to the 10 locations identified by the mode (blue circle). Three of the locations identified by the fuzzy mode routine are at the same approximate locations as that identified by the mode, but the remaining eight locations are clustered at a place not identified by the mode.

Figure 6.6 zooms in to display the eight clustered locations. This is a small regional mall within Baltimore city that has a subway station, a Maryland state motor vehicle administration office, and a parole/probation office. There are multiple parking lots located within the mall. Within this space, approximately 29 vehicle thefts occurred in 1996.

Figure 6.5:

Most Frequent Small Zones for Motor Vehicle Theft

Search Radius of 100 Yards

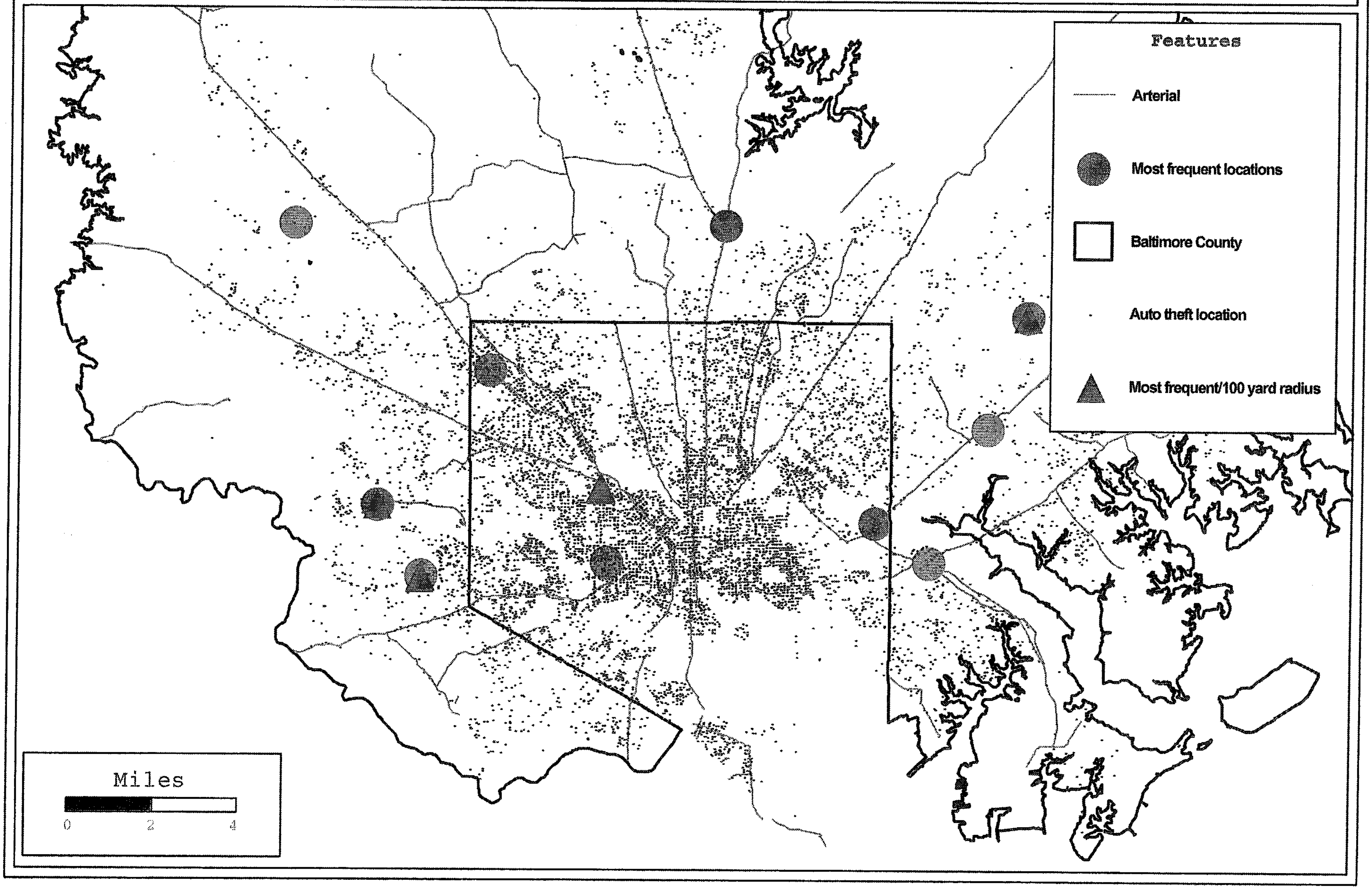
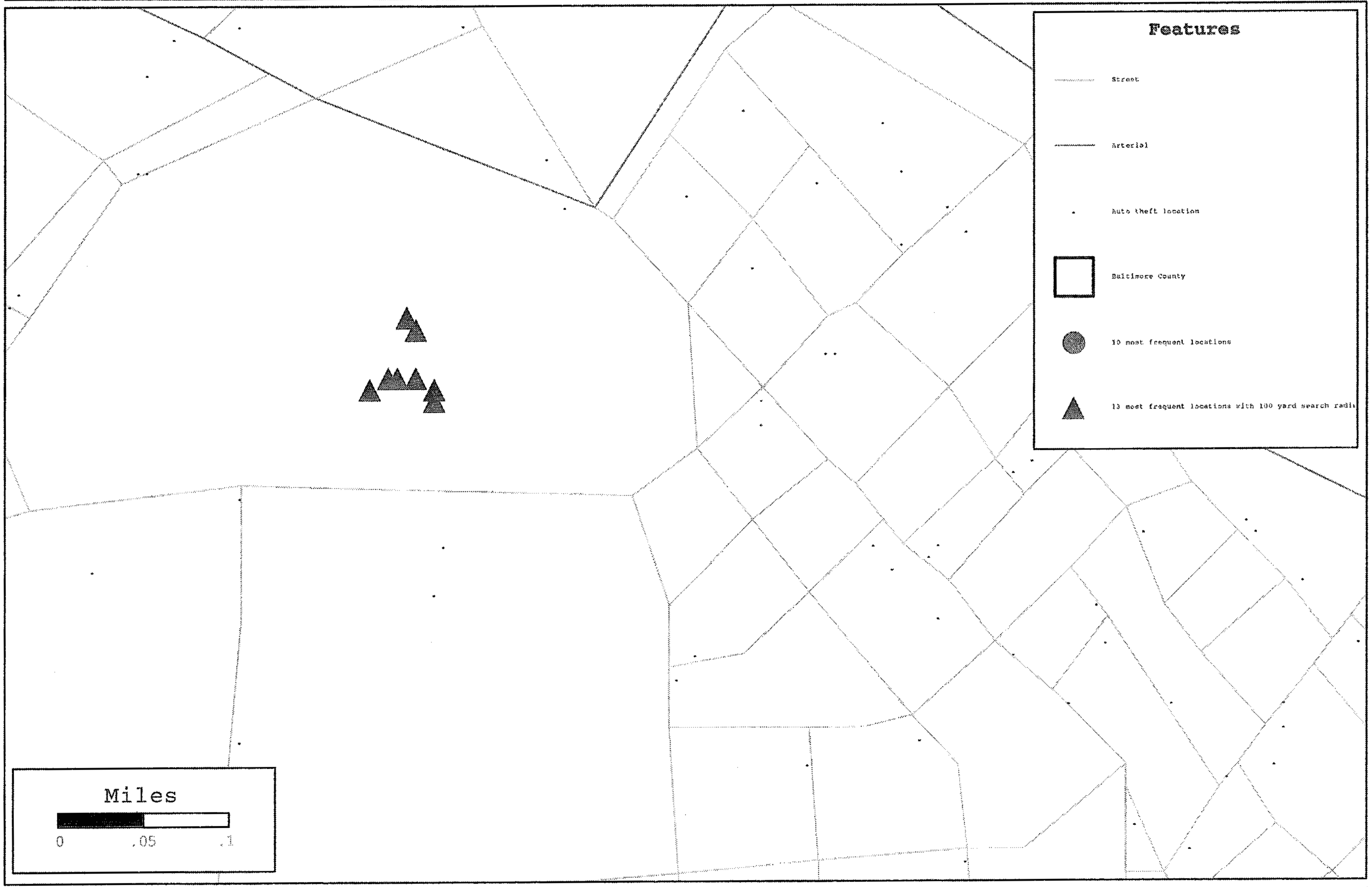


Figure 6.6:

Most Frequent Small Zones for Motor Vehicle Theft

Search Radius of 100 Yards


Features
Street
Arterial
Auto theft location
Baltimore County
10 most frequent locations
13 most frequent locations with 100 yard search radi
Miles
0
.05
.1

The fuzzy mode has identified a general location where there are multiple sub-locations in which vehicle thefts occur.

In other words, the fuzzy mode allows the identification of small hot spot areas, rather than exact locations. But, because all points within the user-defined search area are counted, points are counted multiple times. Thus, any one location may not have a sufficient number of incidents to be grouped in the 'top 10' by itself, but, because it is close to other locations that have incidents occurring, it may be elevated to the 'top 10' due to its adjacency to these other incident locations.

Still, the user must be careful in the analysis. 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).

Nearest Neighbor Hierarchical Clustering (Nnh)

The *nearest neighbor hierarchical clustering* (Nnh) 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 a criteria and proceeds to group the clusters together. The clustering is repeated until either all points are grouped into a single cluster or else the clustering criteria fails. Hierarchical clustering methods are among the oldest cluster routines (Everitt, 1974; King, 1967; Systat, 2000). 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 uses a nearest neighbor 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 no more clustering is possible, either all clusters converge into a single cluster or, more likely, the clustering criteria fail.

In order to conduct clustering, the user specifies two parameters:

1. First, for the threshold distance, 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]

2. Second, the minimum number of points that are required for each cluster. This criteria is used to reduce the number of very small clusters. The default is 10. By decreasing this number, more clusters are produced; conversely, by increasing this number, fewer clusters are produced.

Criteria 1: Nearest Neighbor Distance

The first criteria that is used for clustering points together is the confidence interval around the random expected nearest neighbor distance for first-order nearest neighbors.. This is controlled by a slide bar under the routine (see Figure 6.3). From chapter 5, the mean random distance was defined as

$$\text{Mean Random Distance} = d(ran) = 0.5 \text{ SQRT } \left[\frac{A}{N} \right] \qquad \begin{matrix}(5.2)\\ \text{repeat}\end{matrix}$$

where A is the area of the region and N is thc number of incidents. The confidence interval around that distance is defined as

$$\begin{aligned}\text{Confidence Interval for Mean Random Distance} &= \text{Mean Random Distance} \pm t^* SE_{d(ran)} \\ &= 0.5 \text{ SQRT } \left[\frac{A}{N} \right] \pm t \left[\frac{0.26136}{\text{SQRT}[N^2/A]} \right] \qquad (6.1)\end{aligned}$$

where A is the area of the region, N is the number of incidents, t is the t-value associated with a probability level in the Student's t-distribution.

The lower limit of this confidence interval is

$$\text{Lower Limit of Confidence Interval for Mean Random Distance} = 0.5 \text{ SQRT } \left[\frac{A}{N} \right] - t \left[\frac{0.26136}{\text{SQRT}[N^2/A]} \right] \qquad (6.2)$$

and the upper limit of this confidence interval is

$$\text{Upper Limit of Confidence Interval for Mean Random Distance} = 0.5\ \mathrm{SQRT}\left[\frac{A}{N}\right] + t\left[\frac{0.26136}{\mathrm{SQRT}\left[N^2/A\right]}\right] \qquad (6.3)$$

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 6.2):

Table 6.2

Approximate Probability Values Associated with Threshold Scale Bar

Scale Bar Position	**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 right
10	0.95	Third from right
11	0.99	Second from right
12	0.999	Far right point of slide bar

This is the *threshold distance* for the routine. Taking a broader conception of this, if there is a spatially random distribution, then for all distances between pairs of points, of which there are

$$\frac{N(N-1)}{2}$$

combinations, fewer than p% of the pairs will be shorter than this threshold distance.

Area must be defined correctly

Note: 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). 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.

Criteria 2: Minimum Number of Points

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.

The second criteria, therefore, is the minimum number of points that should be included in any cluster. The routine will only include points in the final clustering that are part of groups (or clusters) in which the minimum number is found.

First-order clustering

Using these criteria, *CrimeStat* constructs a first-order clustering of the points.[4] For each first-order cluster, the center of minimum distance is output as the cluster center, which can be saved as a '.dbf' file. To identify the approximate cluster location, a standard deviational ellipse is calculated for each cluster (see chapter 4 for definition). The user can choose between 1X (the default), 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 *ArcView* '.shp', *MapInfo* '.mif' or *Atlas*GIS* '.bna' formats.

In general, use a 1X standard deviational ellipse since 1.5X and 2X 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. 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.

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 are now treated as 'points' which themselves are clustered.[5] 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.

Guidelines for Selecting Parameters

In the Nnh routine, the user has to define three parameters - the likelihood (or p-value) for selecting a pair by chance (the threshold distance), the minimum number of points, and the number of standard deviations for the ellipses that are output. The p-value is selected with a likelihood slider bar (see figure 6.3). 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 6.3, 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, which are typically smaller (although not always).

If only pairs of points were being grouped, then the threshold distance would be critical. Thus, if the default $p \leq .5$ value is selected, 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 required, the likelihood of finding a cluster with the minimum number of points is much smaller. The higher the minimum number that is required, 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 distances may actually be appropriate; some of the ellipses seen in 6.7 below cover only a couple of street segments. These define micro-neighborhoods or almost pure hot spot locations. 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 covering 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 criteria, 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 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 criteria 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 output size of the clusters. For each cluster in turn, a standard deviational ellipse is calculated (see chapter 4). The user specifies the size of the ellipse in terms of standard deviations. The choices are 1X (the default), 1.5X and 2X 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, and two standard deviations will cover more than 99% of the cases, although the exact percentage will depend on the distribution.

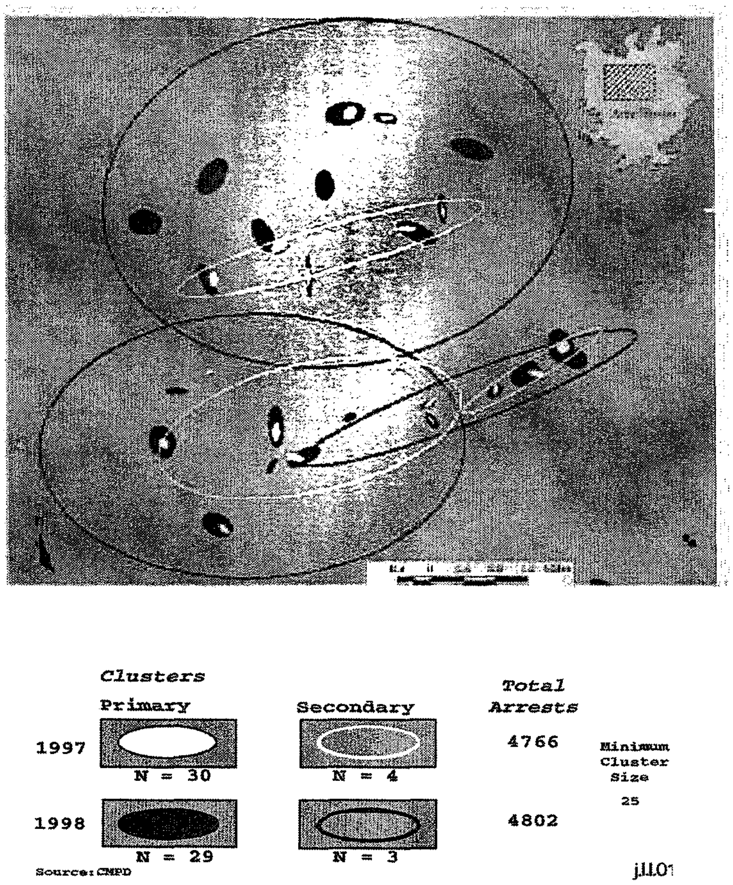
In general, use a one standard deviational ellipse since 1.5X and 2X standard deviations can create an exaggerated view of the underlying cluster. On the other hand, for a regional view, a one standard deviational ellipse may not be very visible. The user has to balance the need to accurately display the cluster compared to making it easier for a viewer to understand its location.

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.



Source: CMPD

jLL01

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, *CrimeStat* calculates 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. The size of the ellipses are 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 can be saved in *ArcView* '.shp', *MapInfo* '.mif' or *Atlas*GIS* '.bna' formats. Because there are also orders of clusters (i.e., first-order, second-order, etc.), there is a naming convention that distinguishes the order. 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

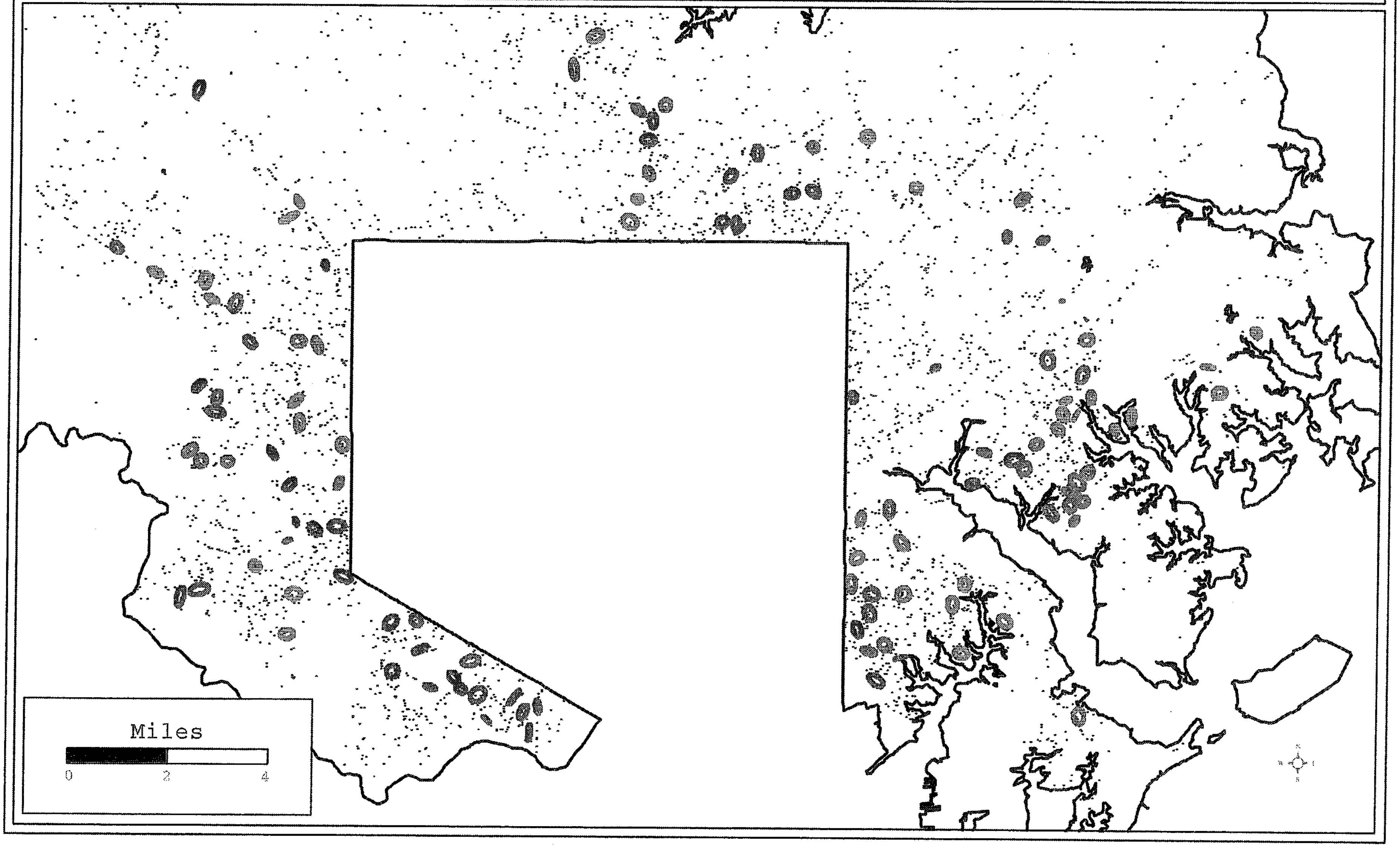
is the first ellipse within the second-order clusters for the file 'NightBurglaries'.

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.

Figure 6.7:

First-Order Baltimore County Burglary 'Hot Spots'

Using Nearest Neighbor Hierarchical Clustering Method



Example 1: Nearest neighbor hierarchical clustering of burglaries

The Nnh routine was applied to the Baltimore County 1996 burglary data (n=6,051 incidents). A default one-tailed probability level of .05 (or 5%) was selected and each cluster was required to contain a minimum of 10 points (the default). *CrimeStat* returned 122 first-order clusters, 15 second-order clusters and two third order clusters. Figure 6.7 shows the first-order clusters displayed as 1x 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 defined by the points in the cluster and includes approximately 50% of the points. Thus, the clusters actually extend a little beyond the ellipses.

Figure 6.8 shows the 20 second-order clusters (dashed lines) and the two third-order clusters (double lines). As seen, they cover much larger areas than the first-order clusters. Finally, figure 6.9 shows a part of east Baltimore County where there are 29 first-order clusters (solid line), five second-order clusters (dashed lines), and one third-order cluster (double line). The street network is presented to indicate the scale. Most first-order clusters cover an area the size of a small neighborhood while the second-order clusters cover larger neighborhoods.

Advantages of Hierarchical Clustering

There are four advantages to this 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 which generate crime incidents (Levine, Wachs and Shirazi, 1986; Maltz, Gordon and Friedman, 1989). The 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 p-value 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 prior to the analysis.

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 near other 'hot spots'. As we've seen from the maps of robbery, burglary and motor vehicle thefts in Baltimore County, there are large areas within the County that have a lot of incidents. Within these large areas, there are smaller hot spots and within some of those hot spots, there are even small 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.

Figure 6.8:

Second- and Third-Order Burglary 'Hot Spots'

Using Nearest Neighbor Hierarchical Clustering Method

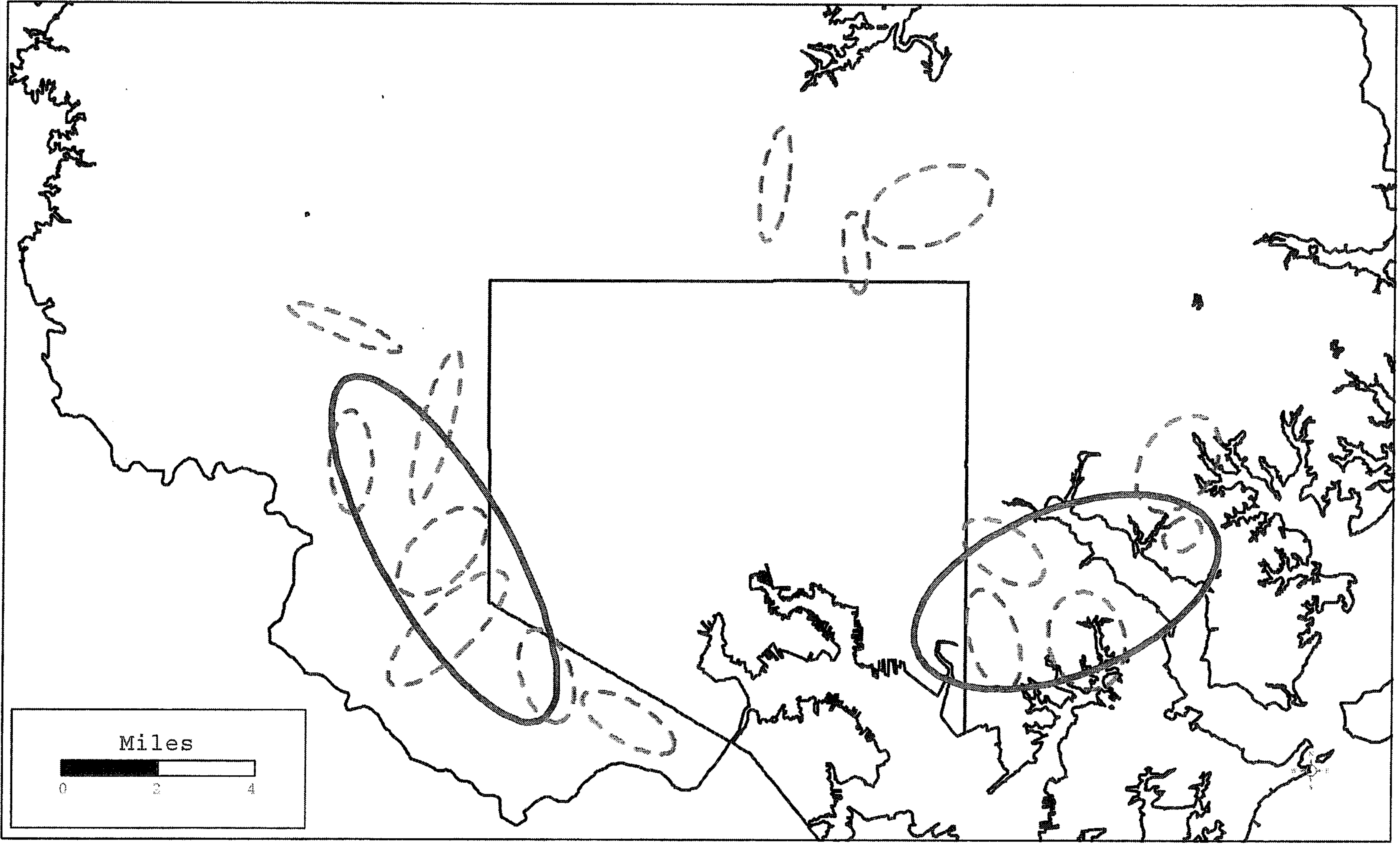
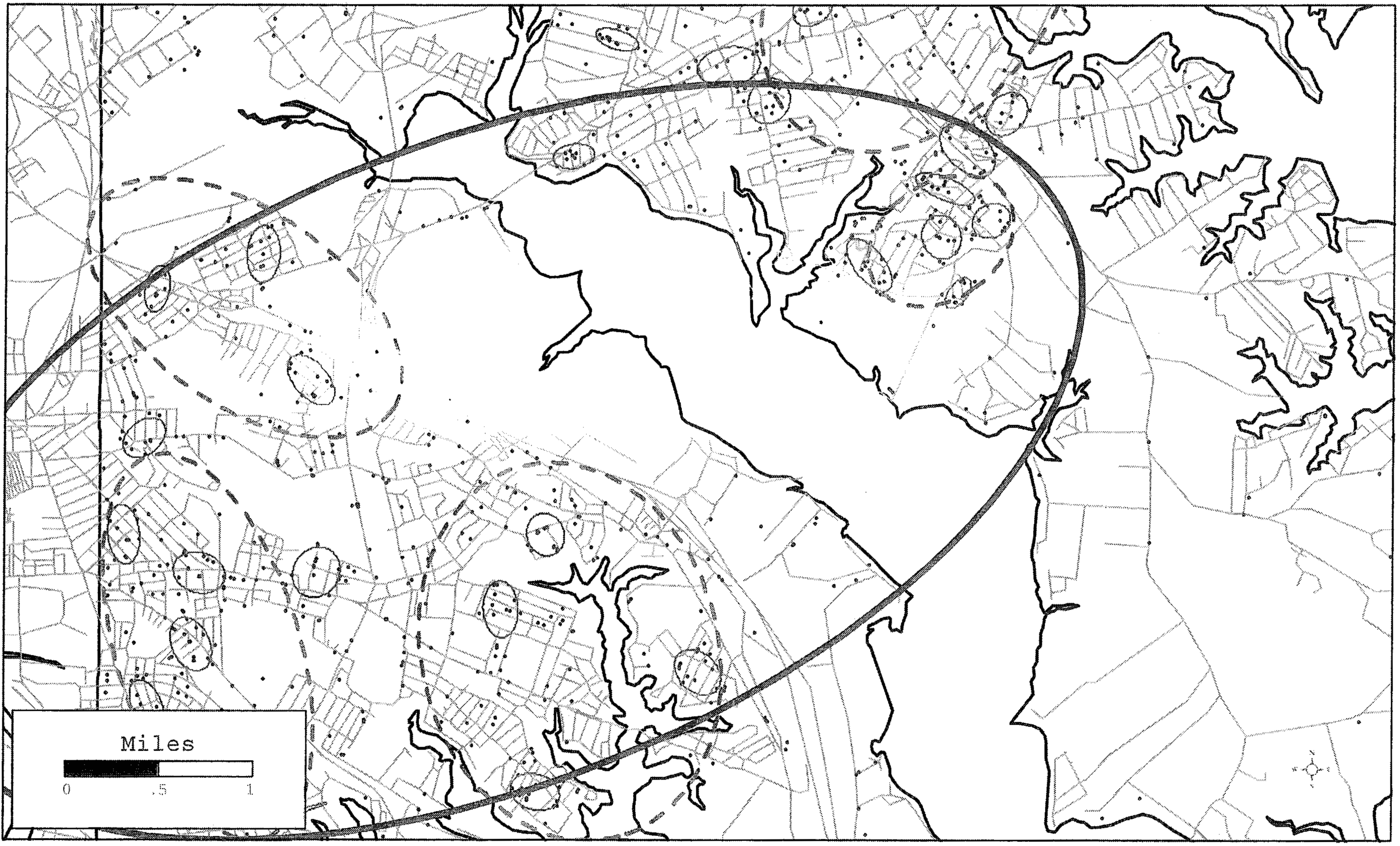


Figure 6.9:

First, Second- and Third-Order Burglary 'Hot Spots'

Using Nearest Neighbor Hierarchical Clustering Method



Fourth, each of the levels imply 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.

Simulating Statistical Significance

Testing the significance of clusters from the Nnh routine is difficult. Conceptually, 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. Further, the additional requirement is added that there be a 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 for the first-order Nnh clusters that has been run; 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 test 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 confidence intervals for the particular first-order Nnh.

The output includes five columns and twelve rows:

Columns:

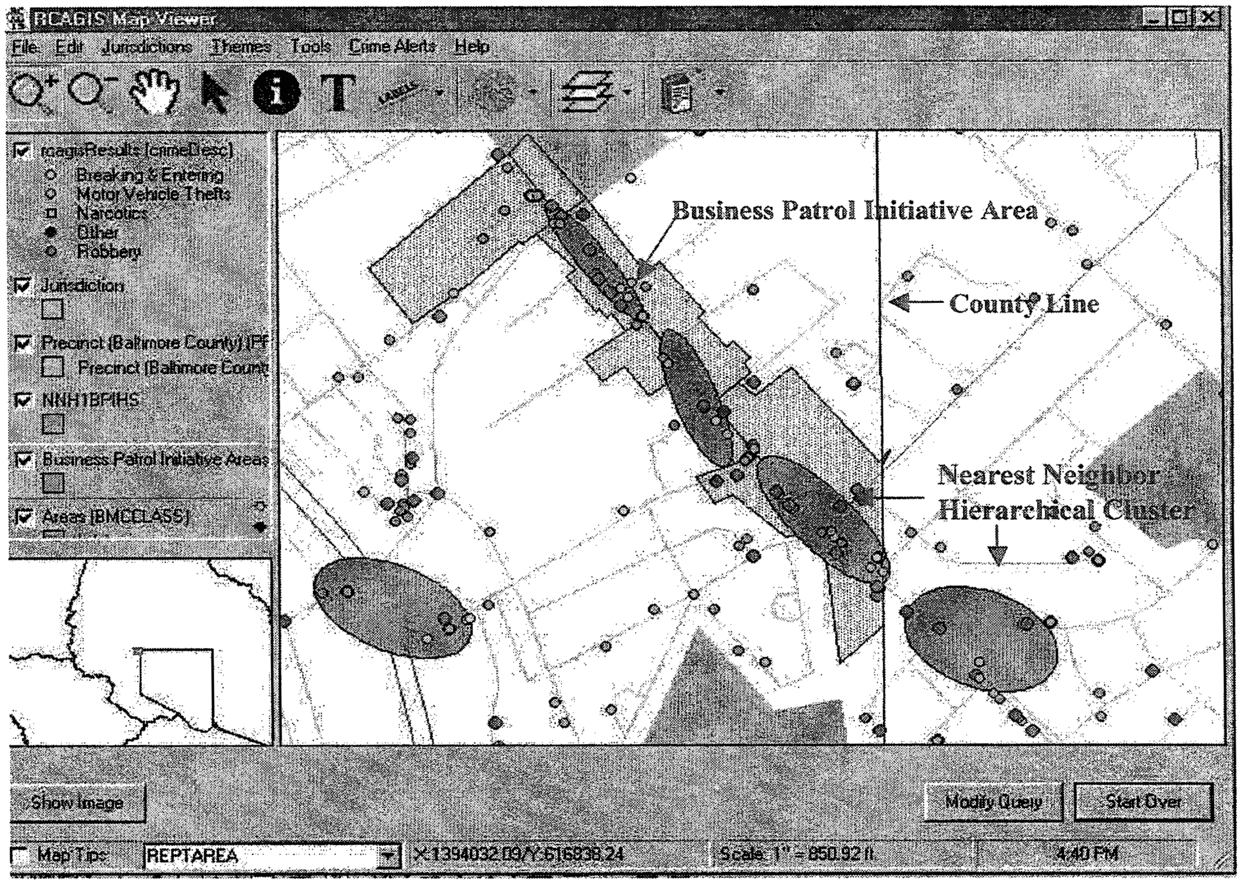
1. The percentile,
2. The number of first-order clusters found for that percentile,
3. The area of the cluster for that percentile,

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.



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 manner in which percentiles are calculated are 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 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.

Table 6.3 presents an example. An Nnh run was conducted on the Baltimore 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. For the actual data, the Nnh routine identified 69 first-order clusters and 7 second-order clusters. Table 6.3 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, let's look at each column 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

Table 6.3

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 ellipse(s) starting from 1

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      466.828013
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      223.684859
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
```

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.

Column 3 shows the areas of clusters that were found over the 1000 runs. 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 are 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 simulations, the numbers 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.

Thus, in conclusion, the simulation suggests that around 60% of the clusters are real with the other 40% being no different than might be expected on the basis of chance. There are far more clusters found in the actual Nnh than would be expected on the basis of chance and they are more compact than would be expected. On the other hand, only about half have densities that are higher than would be expected on the basis of chance.

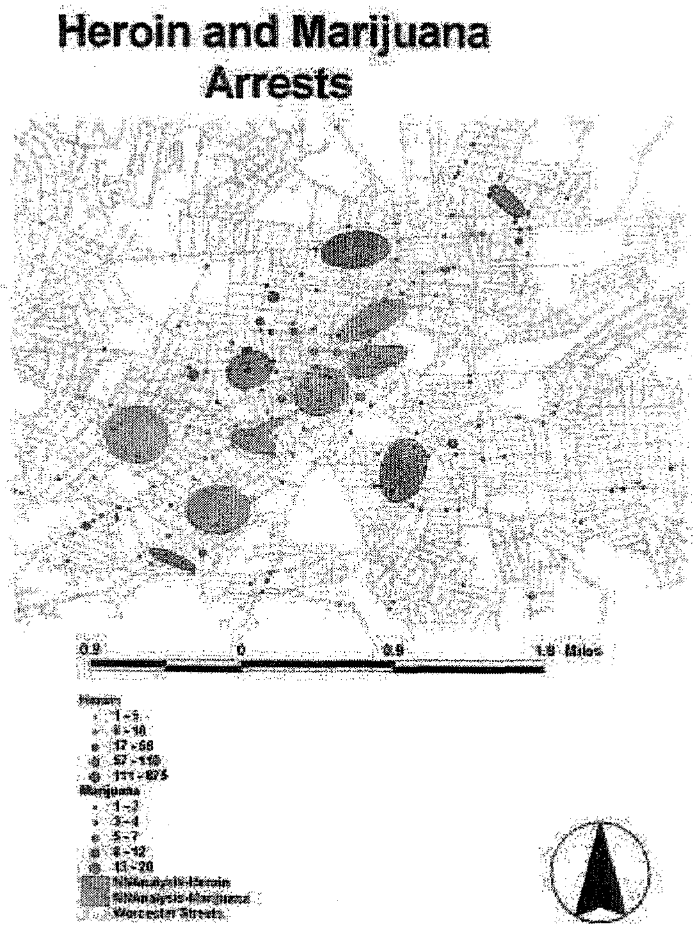
It should be clear that testing the significance of a cluster analysis is complex. In the example, some of the criteria chosen were definitely different than a chance distribution (as evidenced by the simulation) while other criteria were not very different. In this case, the user would be wise to re-run the Nnh and simulation under tighter conditions, either lowering the threshold distance or increasing the minimum number of points per cluster. With experimentation, it is frequently possible to obtain a solution in which all the criteria are greater than would be expected on the basis of chance.

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.

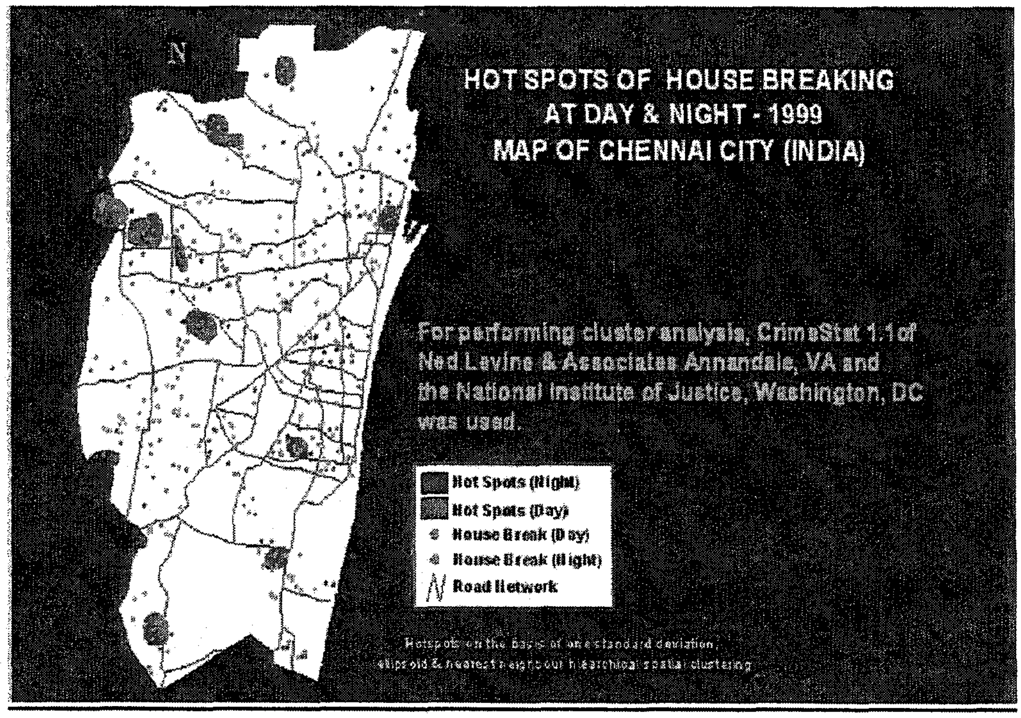


Use of *CrimeStat* in Crime Mapping in India: An Application for Chennai City Policing

Jaishankar Karuppannan
Department of Criminology
University of Madras
Chepauk, Chennai, Tamilnadu, India

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

When hotspot analysis was done to identify 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. This result complies with general perception that 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 & NIGHT - 1999
MAP OF CHENNAI CITY (INDIA)

For performing cluster analysis, CrimeStat 1.1of Ned Levine & Associates Annandale, VA and the National Institute of Justice, Washington, DC was used.

Hotspots on the basis of one standard deviation, ellipsoid & nearest neighbour hierarchical spatial clustering

Limitation to Hierarchical Clustering

At the same time, there are 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. Second, the size of the grouping area is dependent on the sample size since the confidence interval around the mean random distance is used as the criteria (see equation. 4.2). 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, the technique does not produce a consistent definition of a hot spot area.

Third, there is a certain 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 points are 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 don't 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 does produce a constant 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 theory or rationale behind the clusters. They are empirical derivatives of a procedures. Again, many clustering techniques are empirical groupings and also do not have any explanatory theory. However, 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 techniques, that are discussed in chapter 8.

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 point 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. In the next chapter, we will examine the distribution of population and how it affects incidence of crime over an entire metropolitan area. 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. The next chapter 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 8 be read before reading the rest of this section.

Steps in the Rnnh Routine

The Rnnh routine works as follows:

1. Both a primary and secondary file are required. The primary file are the basic 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 are 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., 2 miles). 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 8 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 6.2.

B. However, for each grid cell, a *unique threshold distance* is defined using formulas similar to 6.1 and 6.2. The difference is, however, that the formulas are applied to each grid cell with a unique distance for each grid cell (formulas 6.5-6.8):

$$\text{Mean Random Distance of Grid Cell } i = d(ran) = 0.5 \text{ SQRT } \left[\frac{A_i}{N_i} \right] \qquad (6.5)$$

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.

$$\text{Confidence Interval for Mean Random Distance of Grid Cell } i = \text{Mean Random Distance of grid cell } i \pm t^* SE_{d(ran)}$$

$$= 0.5 \text{ SQRT } \left[\frac{A_i}{N_i} \right] \pm t \left[\frac{0.26136}{\text{SQRT}[N_i^2 / A_i]} \right] \qquad (6.6)$$

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

$$\text{Lower Limit of Confidence Interval for Mean Random Distance of Grid Cell } i = 0.5 \text{ SQRT } \left[\frac{A_i}{N_i} \right] - t \left[\frac{0.26136}{\text{SQRT } [N_i^2 / A_i]} \right] \qquad (6.7)$$

and the upper limit of this confidence interval is

$$\text{Upper Limit of Confidence Interval for Mean Random Distance} = 0.5 \text{ SQRT } \left[\frac{A_i}{N_i} \right] + t \left[\frac{0.26136}{\text{SQRT } [N_i^2 / A_i]} \right] \qquad (6.8)$$

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. 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 are 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.

Area must be defined correctly

Note: 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 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

Another concern is that the bandwidth for the baseline variable be defined 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 produce a more stable average.

Example 2: Simulated Rnnh Clustering

To illustrate the logic of the Rnnh routine, a simulated example is presented. Twenty-seven points were assigned to three groups in the Baltimore metropolitan region (Figure 6.10). The 27 points were grouped in a similar pattern, but one was placed in the center of the metropolitan region (near downtown Baltimore) while the other two 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 27 points into three groups whereas the Rnnh routine should cluster only 18 of the points into two groups. The reason for the lack of a third group is that one would expect a high number of incidents in the center; consequently, it is not high relative to the underlying baseline population. Figures 6.11 and 6.12 show exactly this solution.

In other words, the Nnh routine clusters points together irrespective of the distribution of the baseline population whereas the Rnnh routine clusters points together relative to the baseline population.

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, the standard deviational ellipses of the clusters can be saved in *ArcView* '.shp', *MapInfo* '.mif' or *Atlas*GIS* '.bna' formats. The size of the ellipses are determined by the number of standard deviations to be calculated (see above). In general, use a 1X standard deviational ellipse since 1.5X or 2X standard deviations can create an exaggerated view of the underlying cluster. On the other hand, for a regional view, a one 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.

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

Figure 6.10:

Incidents in Relation to Population Density: Baltimore 1990

Incident Locations and Persons Per Square Mile

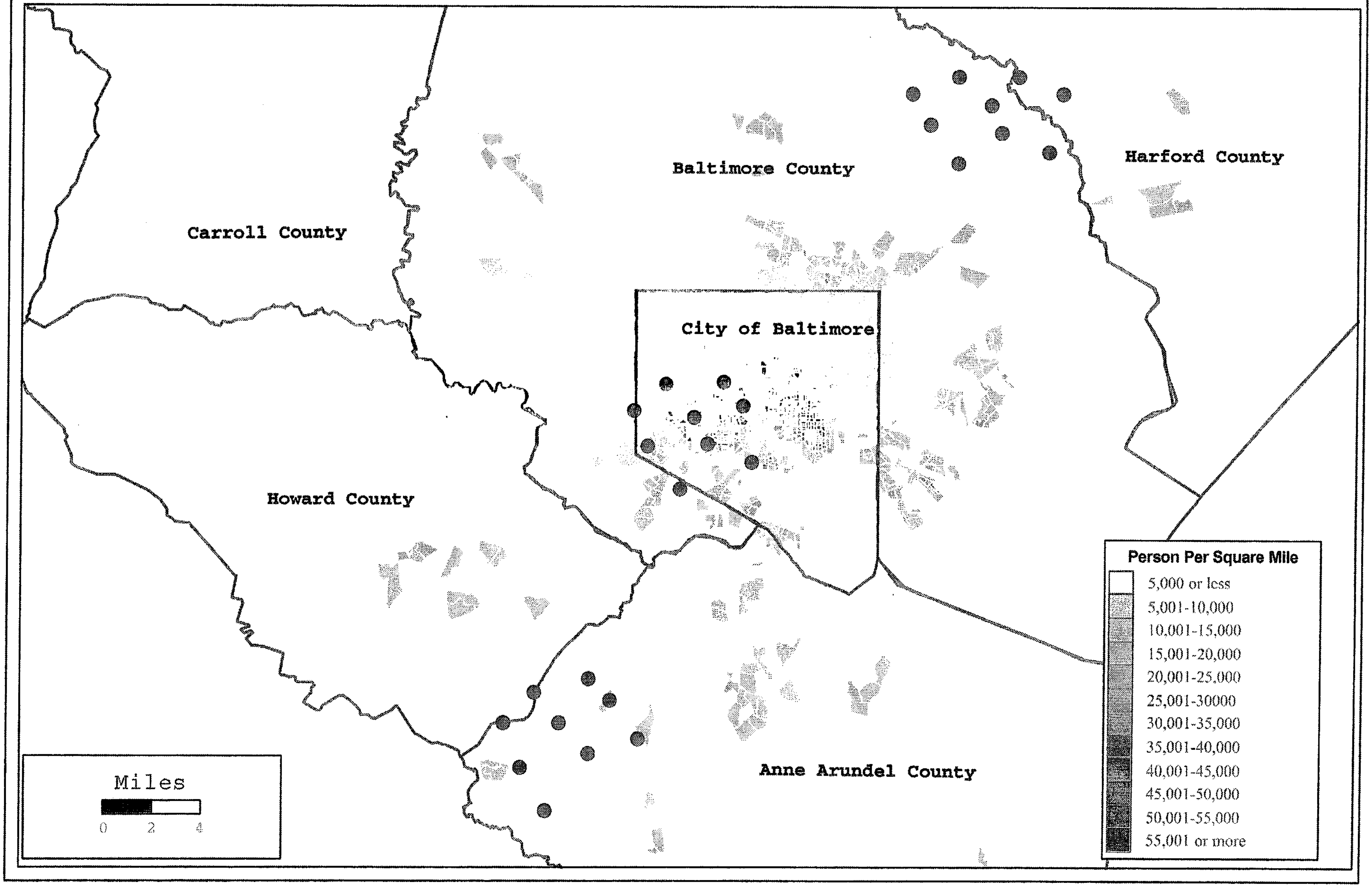


Figure 6.11:

Nearest Neighbor Clustering of Incidents

Nnh Clusters and Incident Locations

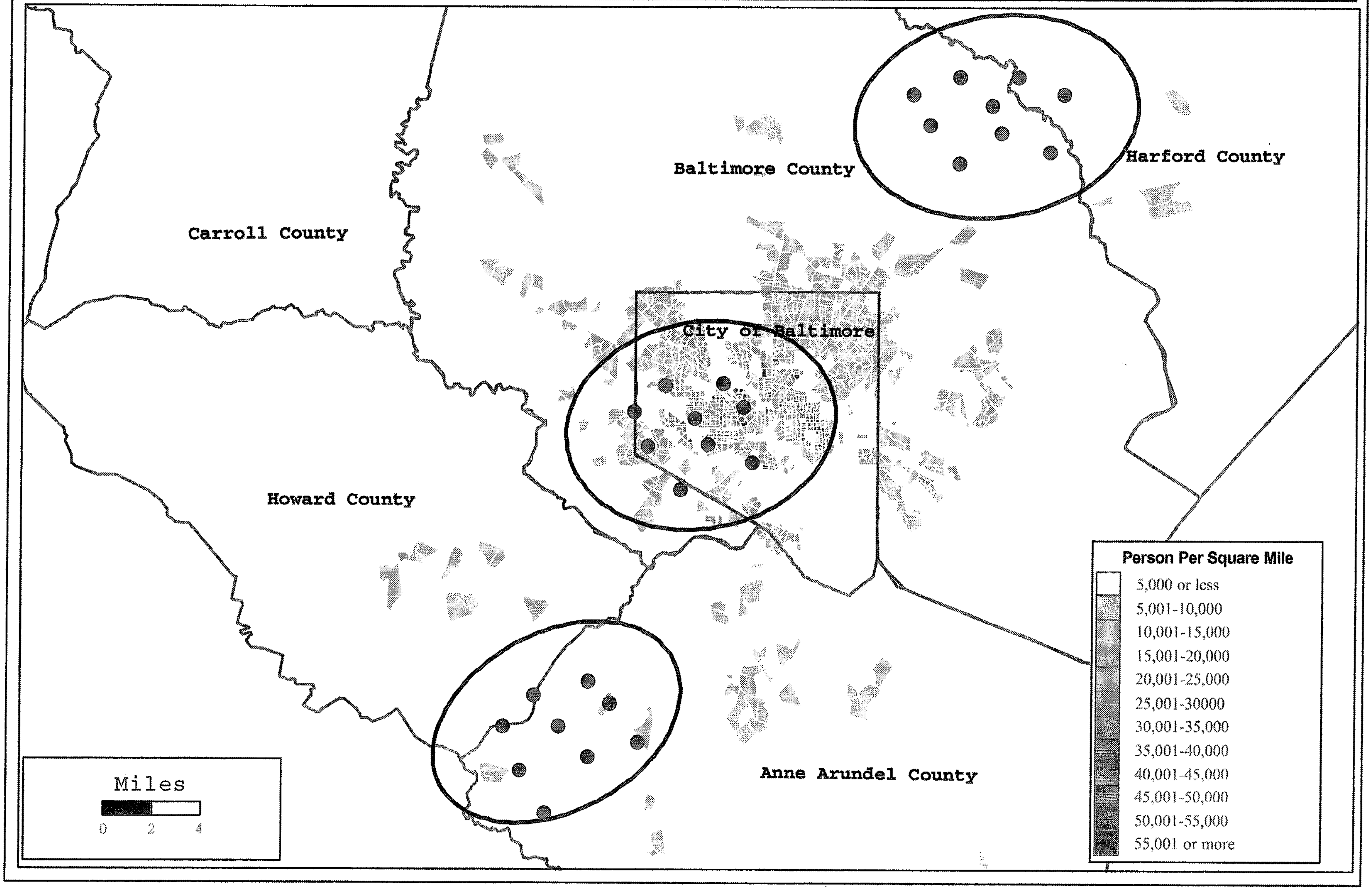
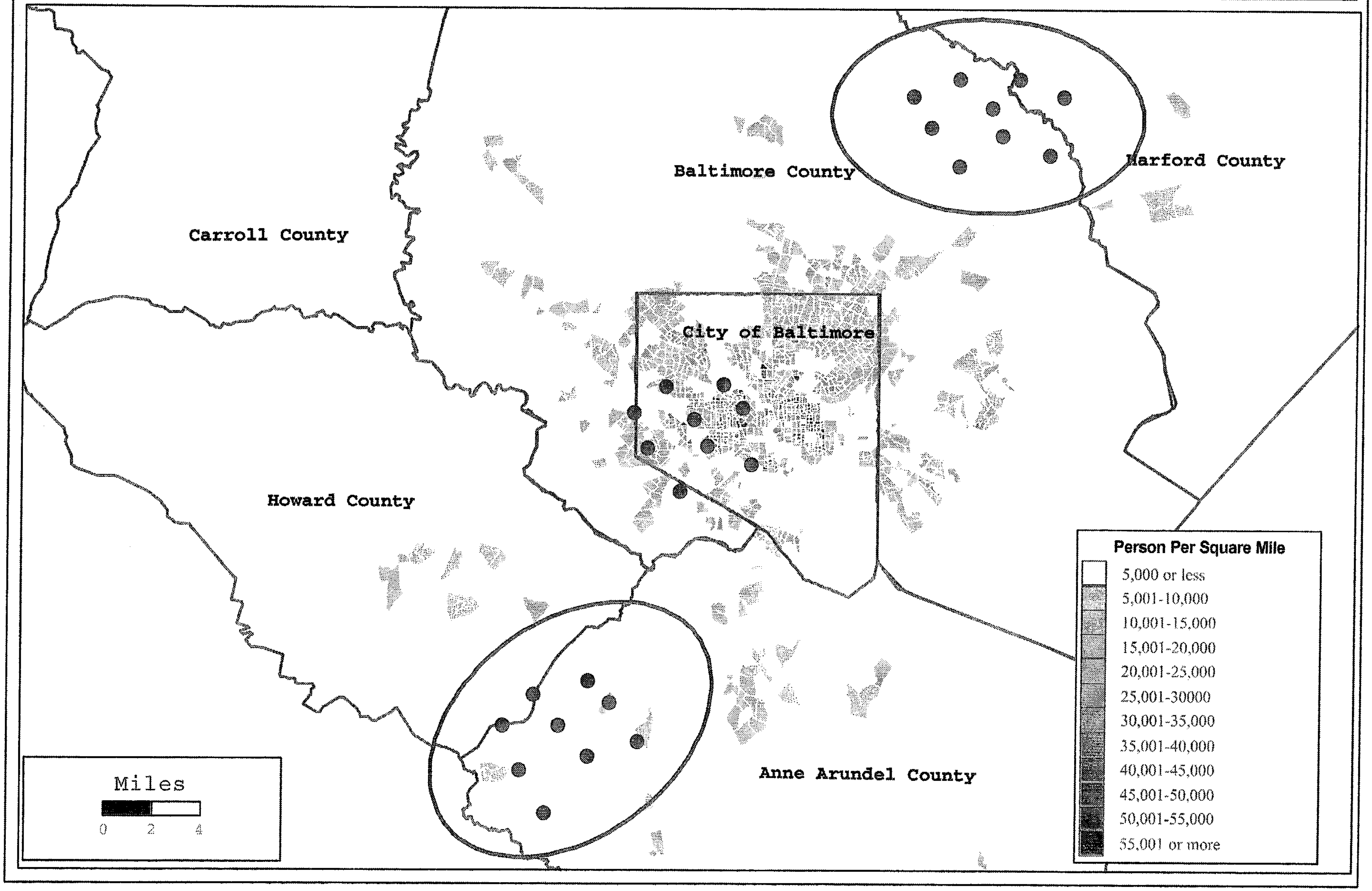


Figure 6.12:

Risk-Adjusted Nearest Neighbor Clustering of Incidents Relative to Population

Rnnh Clusters and Incident Locations



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

Rnnh1Ell10robbery

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

Rnnh2Ell1burglary

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

Example 3: Rnnh Clustering of Vehicle Thefts

A second example is the clustering of 1996 Baltimore vehicle thefts relative to the 1990 population of census block groups. The test is for clusters of vehicle thefts that are more concentrated than would be expected on the basis of the population distribution.[6] Using the default threshold probabilities and a minimum sample size per cluster of 25, the Rnnh routine identified five first-order and one second-order cluster (Figure 6.13). As seen, there are only five clusters, most of which are peripheral to the downtown area.

Compare this distribution with the results of the Nnh on the same data, using the same parameters (Figure 6.14). The Nnh found 28 first-order clusters and two second-order clusters. As expected, they are more concentrated in the center. Note that there are far fewer clusters identified in the Rnnh routine than in the Nnh. Many of the clusters in the Nnh routine are due to a higher concentration of population. Once this is normalized, one finds that there are only a few areas of very high risk for vehicle theft. In other words, the Rnnh routine identifies areas of high *risk* for vehicle theft whereas the Nnh routine identifies areas of high *volume* for vehicle theft.

Figure 6.13:

1996 Metropolitan Baltimore Vehicle Theft

Risk-Adjusted Nearest Neighbor Clusters

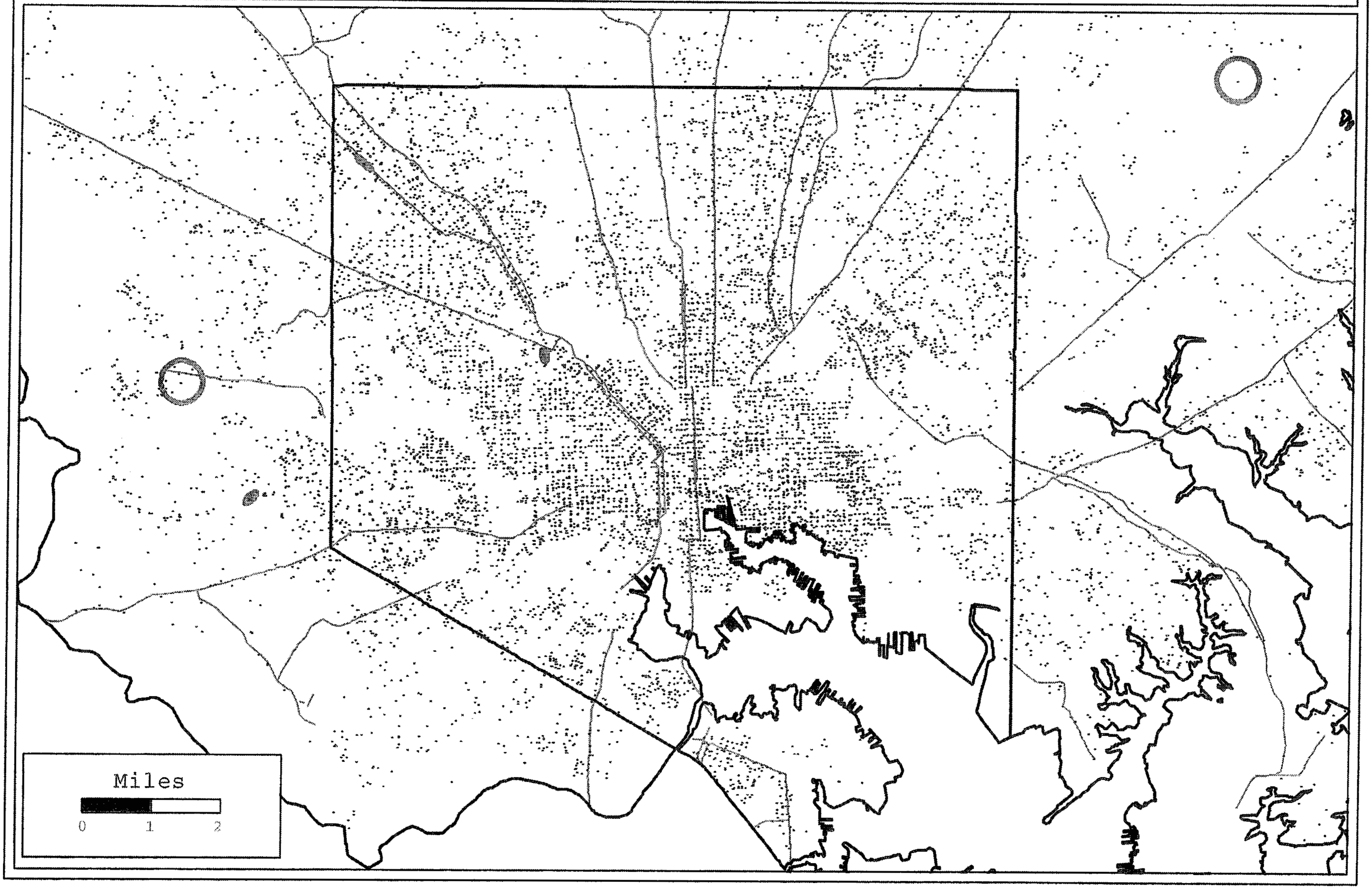


Figure 6.14:

1996 Metropolitan Baltimore Vehicle Theft

Nearest Neighbor Clusters

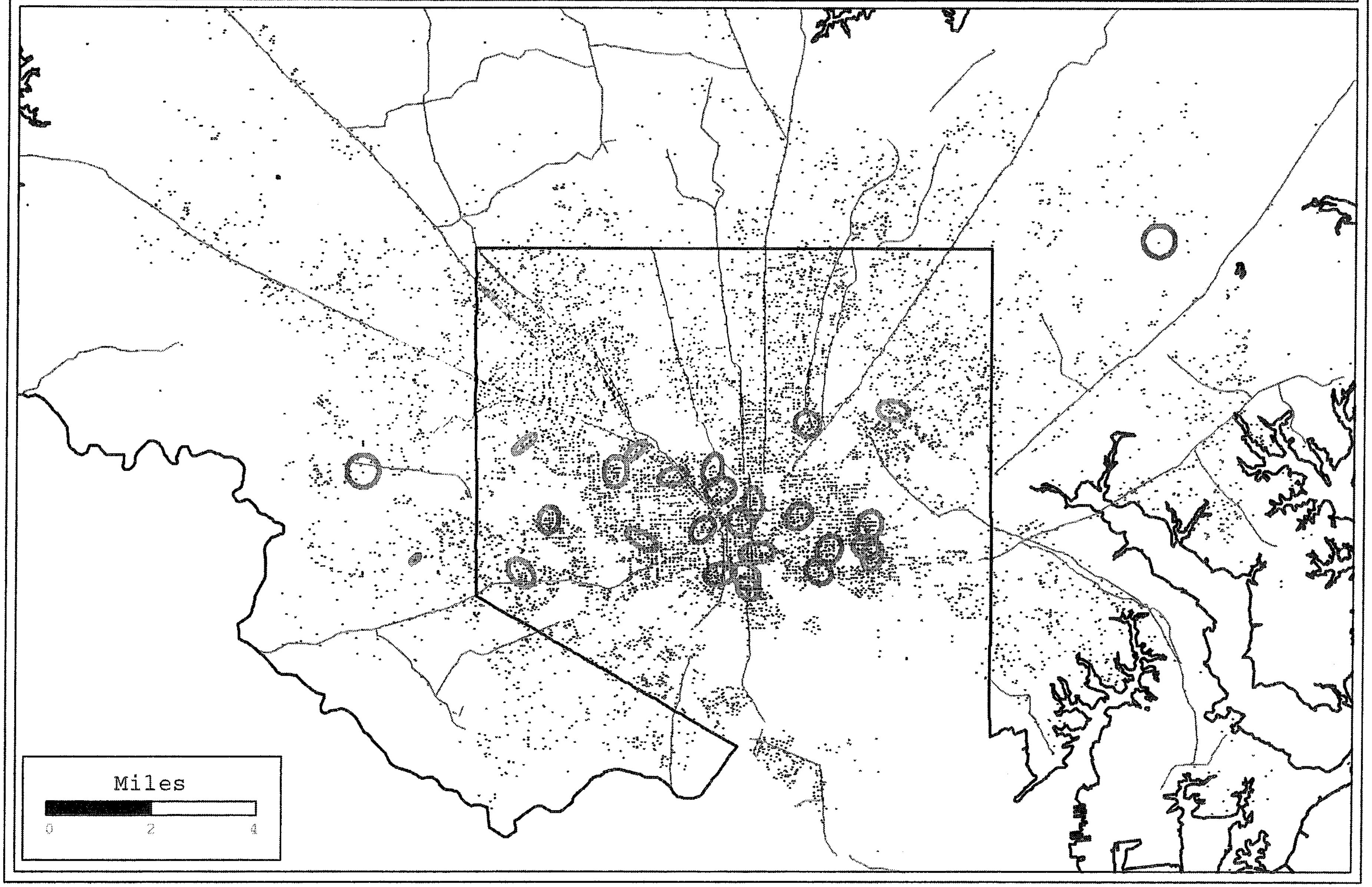


Table 6.4

Risk-adjusted Clustering of Baltimore County Robberies: 1996

```
Risk-Adjusted 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...............: 8
    Simulation runs..............: 1000

    Displaying 8 ellipse(s) starting from 1

Order Cluster  Mean X     Mean Y     Rotation  X-Axis    Y-Axis   Area     Points  Density
----- ------- --------   -------    --------- ------    ------   ----     ------  --------
  1      1   -76.44973  39.31523  73.89169  0.19429  0.09230  0.05634  31    550.251866
  1      2   -76.60194  39.40076   4.40641  0.12272  0.12929  0.04984  23    461.446220
  1      3   -76.78279  39.36184  62.61813  0.24605  0.15511  0.11990  26    216.852324
  1      4   -76.73157  39.34387   4.30498  0.08916  0.07321  0.02051  24   1170.341418
  1      5   -76.44539  39.30523  13.63299  0.19639  0.11154  0.06882  20    290.622622
  1      6   -76.75368  39.31132  89.56994  0.19748  0.22914  0.14216  22    154.753006
  1      7   -76.73132  39.28897  11.83419  0.09359  0.18312  0.05384  21    390.033756
  2      1   -76.74984  39.32650  66.40941  4.19556  1.63703 21.57723   4      0.185381

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

    Percentile Clusters            Area        Points           Density
    ---------- -------- --------------- ------------ ----------------
          min        1          1.67880           20         1.648432
          0.5        1          2.36257           20         1.874836
          1.0        1          2.51219           20         1.996056
          2.5        1          2.67031           20         2.208136
          5.0        1          2.98150           20         2.372246
         95.0        4         13.57660           50         7.365212
         97.5        4         13.95390           53         7.932653
         99.0        5         14.34076           56         8.643887
         99.5        5         14.60388           58         9.595312
          max        5         15.41259           67        11.913282
```

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 approximate confidence 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. Table 6.4 shows the output for 1996 Baltimore County robberies with the default search threshold and a minimum sample size of 20 incidents.

The results also show those obtained from 1000 Monte Carlo simulations. There were seven first-order clusters and one second-order cluster. Looking at the Monte Carlo simulations, the two most critical parameters are the number of first-order clusters found and the density of the clusters. In the simulation, the minimum number of clusters found under these conditions (i.e., with the default threshold distance and a minimum sample size of 20 incidents) was one while the maximum number was five. The 95th percentile was four incidents. Since the Rnnh routine produced seven first-order clusters, the routine has identified more clusters than would normally be expected on the basis of chance. Looking at the density estimates from the simulation, the maximum density was 11.913282 and the 95th percentile was 7.365212. Since all seven first-order clusters had densities higher than the 95th percentile, the density of these clusters is greater than what would normally be expected on the basis of chance. In other words, the routine has identified more clusters and higher density clusters than would be expected on the basis of chance.

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. 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 8). 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).

There are two types of 'fine tuning' that have to go on. First, the 'background' variation 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.

Second, there is 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.

Limitations of the Technique

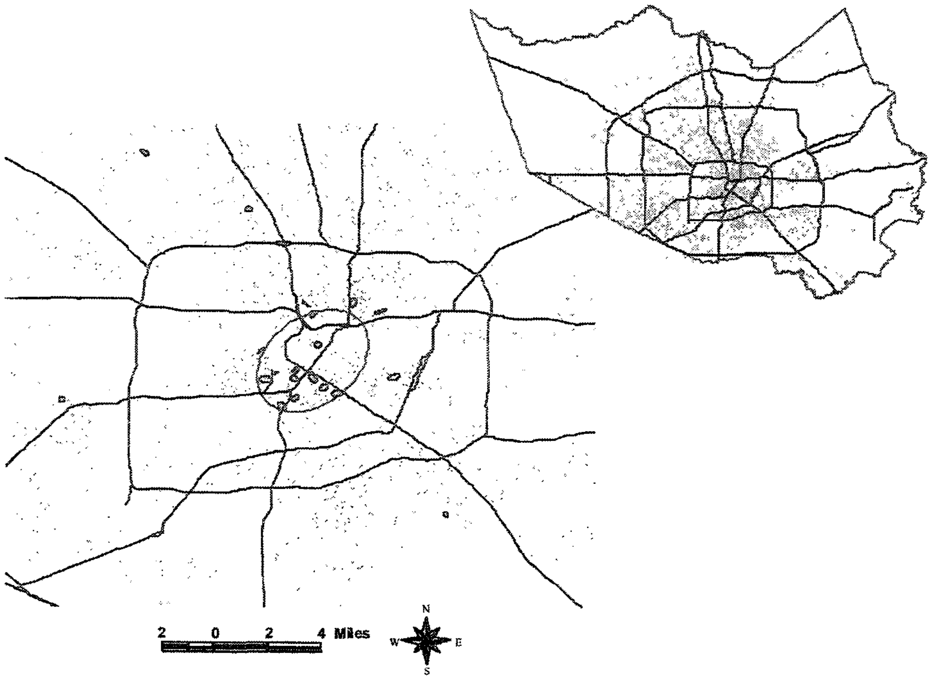
There are some technical limitations that the Rnnh routine shares with the Nnh routine. First, the method only clusters incidents (points); a weighting or intensity

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

Matthew L. Stone, MPH
Center for Health Policy Studies
University of Texas-Houston, School of Public Health-Houston, Texas

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.



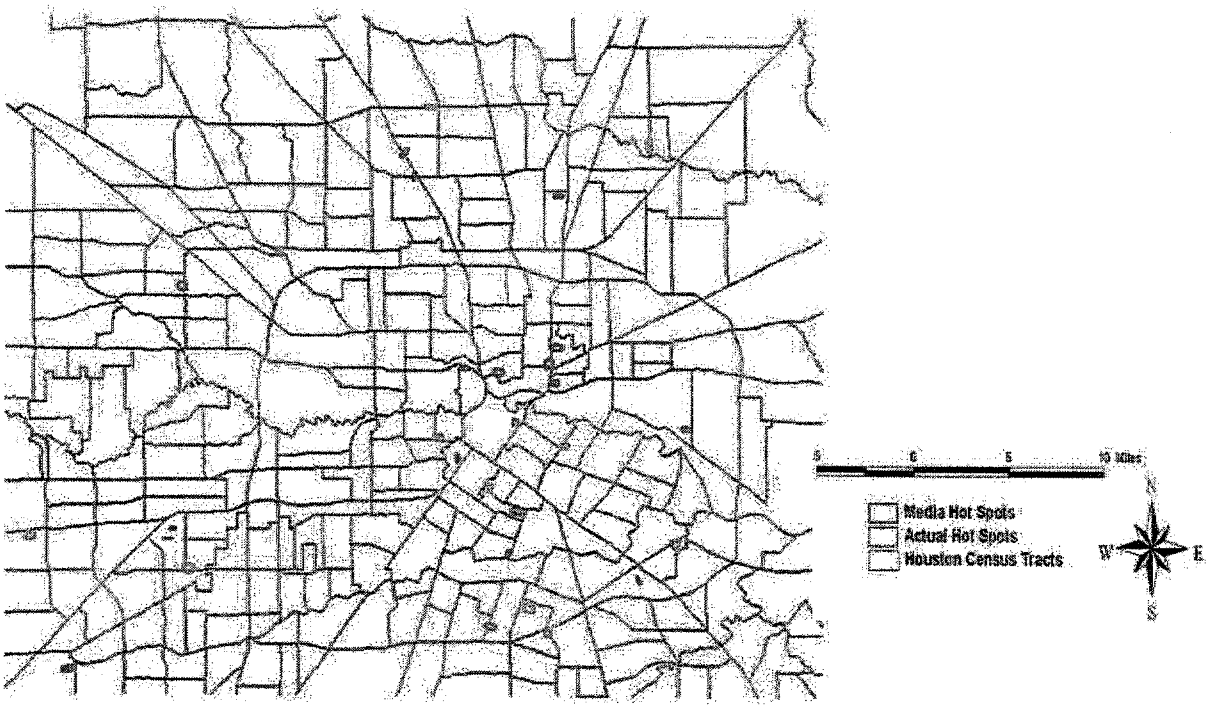
Using Risk Adjusted Nearest Neighbor Hierarchical Clustering to Compare Actual and Media Hotspots of Homicide

Derek J. Paulsen, Ph.D
Department of Political Science/Criminal Justice
Appalachian State University
Boone, NC

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 shows that media hot spots and actual hot spots do not coincide. Media coverage shows homicides to be concentrated in different areas than they are actually concentrated.

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



variable will have no effect. Second, the size of the grouping area is dependent on the sample size since the confidence interval around the mean random distance is used as the 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. It cannot work on calculated rates (e.g., incidents per capita by zones). For the latter, the user should look at techniques such as the SatScan method (Kulldorff, 1997).

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.

Endnotes for Chapter 6

1. The output in table 6.1 has been formatted. *CrimeStat* only outputs an Ascii file. In this case, the Ascii file was pasted into *Word Perfect*®, the word processing program used for this manual, and was then formatted so that the underscore was consistent with the title words and the columns lined up.

2. In the statistical literature, this type of statistic is 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).

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

4. The particular steps are as follows:

   A. 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.

   B. The mean expected random distance is calculated using formula 5.2 and the threshold distance (the confidence interval for the corresponding t) is calculated using formulas 6.2 and 6.3 depending on whether it is a lower or upper confidence interval. The particular interval is selected by user on the slide bar.

   C. All distance pairs smaller than the threshold distance are selected for clustering.

   D. For each incident point, the number of distances to other points that are smaller than the threshold distance are counted and placed in a *reduced matrix*. Any incident point which does not have another point within the threshold distance is not clustered. Any distance that is greater than the threshold distance is not considered for clustering.

   E. All points in the reduced matrix are sorted in descending order of the number of distances to other points shorter than the threshold distance, and the incident point with the largest number of below threshold distances is selected for the initial seed of the first cluster.

   F. All other incidents that are within the threshold distance of the initial seed point are selected for cluster 1.

G. The number of points within the cluster are 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.

H. For those clusters that are kept, the center of minimum distance is calculated for each to identify the cluster center.

I. The clustered points are removed from further clustering.

J. Of the remaining points, the incident point with the largest number of distances to other points shorter than the threshold distance is selected for the initial seed the second cluster.

K. All other points which are within the threshold distance of the first cluster seed point are selected for cluster 2.

L. The mean center of these selected points is calculated to identify the cluster center.

M. These points are removed from further clustering.

N. Steps J through M are repeated for all remaining points in the reduced matrix until no more points are remaining in the reduced matrix or until there are fewer than the specified minimum number of points for those remaining in the reduced matrix.

5. The steps are as follows:

A. 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

$$\text{Confidence Interval for Second-order Mean Random Distance} = 0.5\,\text{SQRT}\left[\frac{A}{M}\right] \pm t\left[\frac{0.26136}{\text{SQRT}\left[M^2/A\right]}\right] \qquad (6.1)$$

where A is the area of the region and M is the number of first-order clusters identified during first-order clustering (i.e., not N). 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.

B. 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.

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

D. If there are distances between first-order cluster centers that are smaller than the second-order threshold distance, then the steps specified in endnote 3 are repeated to produce second-order clusters. A minimum of four first-order clusters is required to allow a second- or higher-order cluster.

E. 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.

F. 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

6. It is not an exact risk test since we are comparing 1996 vehicle thefts with 1990 population. It is an approximate risk test.

Chapter 7
'Hot Spot' Analysis II

This chapter continues the discussion of hot spots. Three additional routines are discussed: STAC, K-means, and Anselin's Local Moran. Figure 7.1 displays the 'Hot Spot' Analysis II page. The first of these routines, the Spatial and Temporal Analysis of Crime (STAC), was developed by the Illinois Criminal Justice Information Authority. They have agreed to integrate STAC into *CrimeStat*. The second routine - K-means, is a partitioning technique. The third technique - Anselin's Local Moran, is a zonal hot spot method. We'll start first with STAC, and who better to explain it than the authors of the routine, Richard and Carolyn Block.

Spatial and Temporal Analysis of Crime (*STAC*)

by

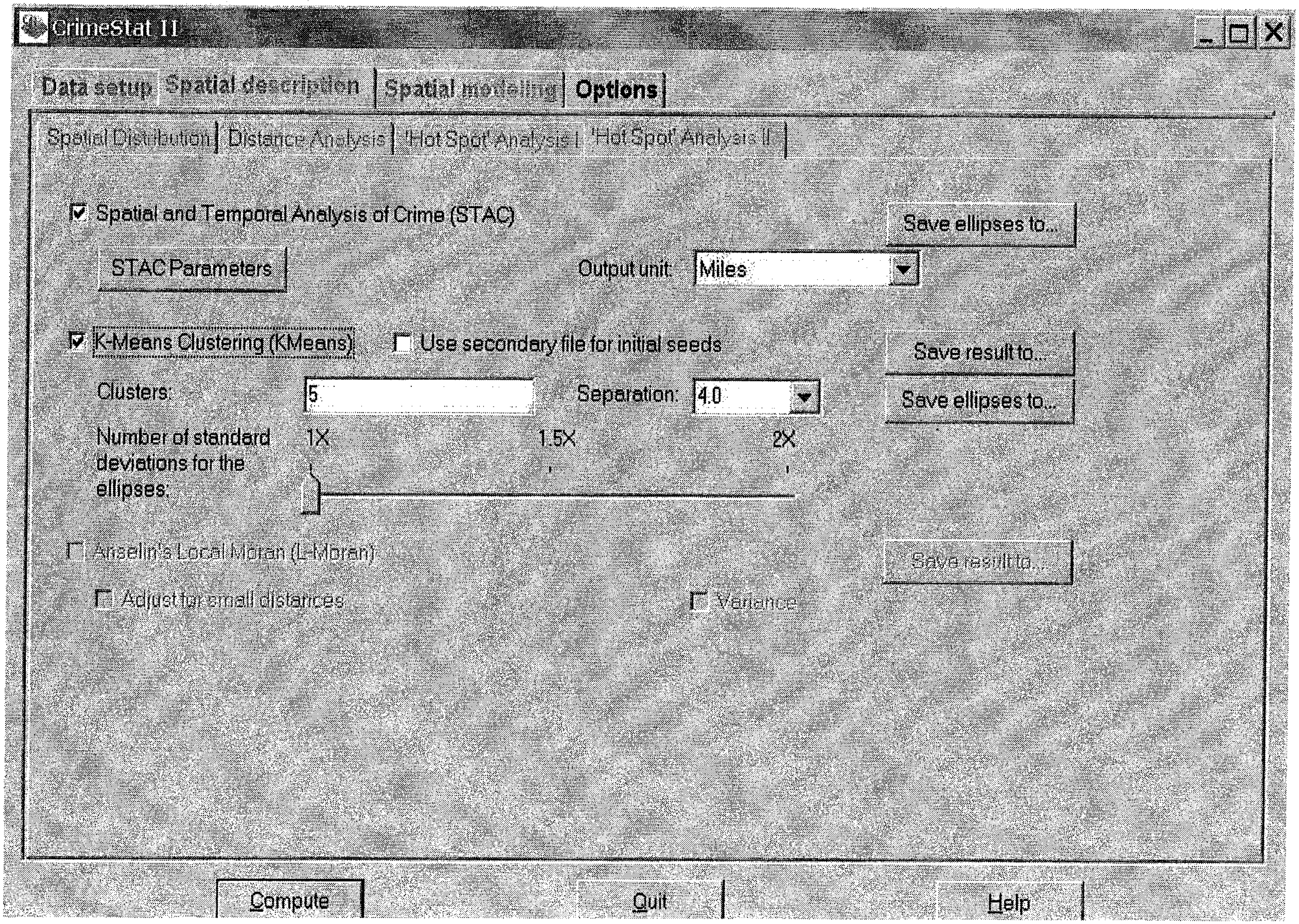
Richard Block	Carolyn Rebecca Block
Professor of Sociology	Senior Research Analyst
Criminal Justice	Illinois Criminal Justice Information Authority
Loyola University	Chicago, IL
Chicago, IL	

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

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

STAC is a scan-type clustering algorithm in which a circle is repeatedly laid over a grid and the number of points within the circle are counted (Openshaw, Charlton, Wymer and Craft, 1987; Openshaw, Craft, Charlton, and Birch, 1988; Turnbull, Iwano, Burnett, Howe, and Clark, 1990; Kuldorff, 1995). It, thus, shares with those other scan routines the property of multiple tests, but it differs in that the overlapping clusters are combined into larger cluster until there are no longer any overlapping circles. Thus, STAC clusters can be of differing sizes. The routine, therefore, combines some elements of partitioning clustering (the search circles) with hierarchical clustering (the aggregating of smaller clusters into larger clusters).

Figure 7.1: **'Hot Spot' Analysis II Screen**



☑ Spatial and Temporal Analysis of Crime (STAC)

Output unit: Miles

☑ K-Means Clustering (KMeans) □ Use secondary file for initial seeds

Clusters: 5

Separation: 4.0

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

□ Anselin's Local Moran (L-Moran)

□ Adjust for small distances □ Variance

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

The STAC Hot Spot Area routine in *CrimeStat* searches for and identifies the densest clusters of incidents based on the scatter of points on the map. The STAC Hot Spot Area routine creates areal units from point data. It identifies the major concentrations of points for a given distribution. It then represents each dense area by the best-fitting standard deviational ellipse (see chapter 4). The boundaries of the ellipses can easily be displayed as a mapped layer by standard GIS software.

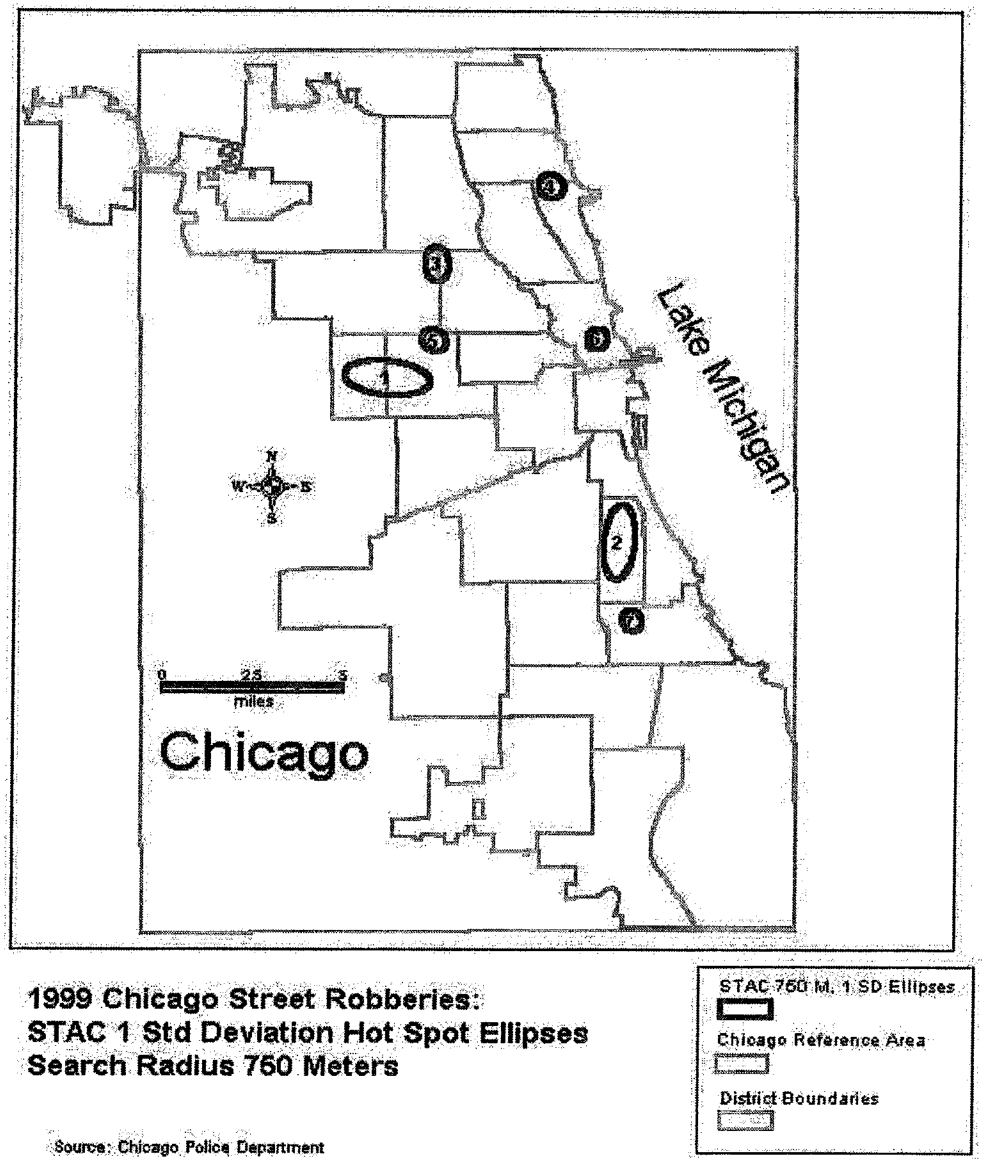
STAC is not constrained by artificial or political boundaries, such as police beats or census tracts. This is important, because clusters of events and places (such as drug markets, gang territories, high violence taverns, or graffiti) do not necessarily stop at the border of a police beat. Also, shading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

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

To take an example, Figure 7.2 shows the location of the seven densest Hot Spot Areas of street robbery in 1999 in Chicago. Four of the seven span the boundaries of police districts and two cover only a small part of a larger district. In a shaded area map, these dense clusters of robbery might be not easily identifiable. An area that is really dense might appear to be low-crime because it is divided by an arbitrary boundary. Using a shaded areal map aggregating the data within each district would give a general idea of the distribution of crime over the entire map, but it would not tell exactly where the clusters of crime are located.

For example, figure 7.3 zooms in on Hot Spot Area 4 (the northernmost Hot Spot Area in Figure 7.2). Hot Spot Area 4 covers parts of two districts (shown by a pink boundary line in figure 7.2) There are also four beats (shown by blue boundary lines). The shaded map indicates many incidents in beat 2311, but few in beats 2312, and 2313.[2] The incident distribution indicates that while few incidents occurred overall in 2312 and 2313, most of the incidents that did occur were near to beat 2311. Incidents in beat 2311 mainly occurred on its eastern boundary. Portions of the beat were relatively free from street robbery. The Hot Spot Area identifies this clustering that spans beats and districts. Hot Spot Areas that overlap beat and district boundaries might indicate to patrol officers in these neighboring areas that they should coordinate their efforts in combating crime.

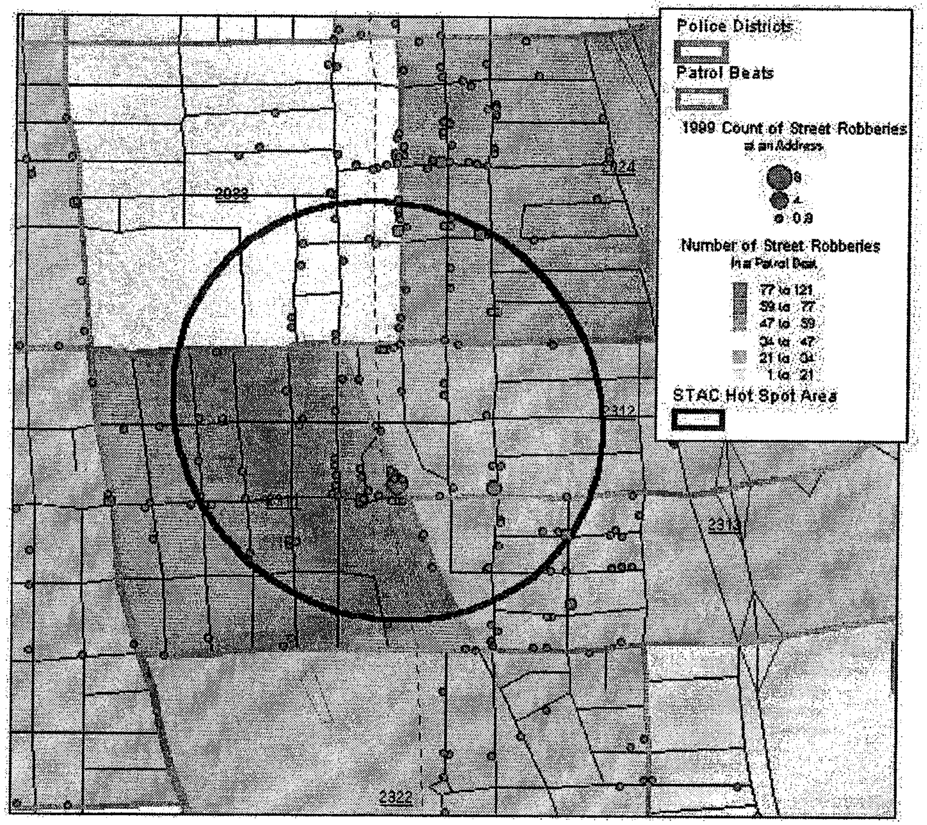
Figure 7.2: **STAC Hot Spots for 1999 Street Robberies**



1999 Chicago Street Robberies:
STAC 1 Std Deviation Hot Spot Ellipses
Search Radius 750 Meters

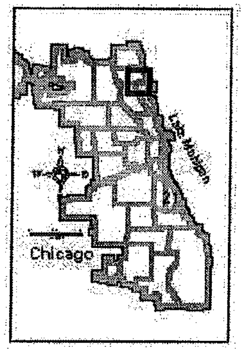
Source: Chicago Police Department

Figure 7.3: STAC 1999 Street Robbery Hot Spot Area 4



Location of 1999 Street Robberies
Chicago: Mid Northside

Source: Chicago Police Department



How STAC Identifies Hot Spot Areas

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

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

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

3. STAC counts the number of points falling within each circle, and ranks the circles in descending order.

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

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

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

7. Using the data points in each Hot Cluster, the program calculates the best-fitting standard deviational ellipse (see chapter 4). These are called *Hot Spot Areas*. Because the standard deviational ellipse is a statistical summary of the Hot Cluster points, it may not .contain every Hot Cluster point. It also may contain points that are not in the Hot Cluster.

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

Steps in Using *STAC*

STAC is available on the Hot Spot Analysis II tab under Spatial Description (see figure 7.1). A brief summary of the steps is as follows:

1. *STAC* requires a primary file and a reference file (see chapter 3). Optionally, STAC requires the reference file area (on the measurement parameters tab) if simulation runs are requested. Note: while *STAC* runs quite quickly, it runs more quickly with a Euclidean coordinate system such as UTM or State Plane. For example, an analysis of 13,000 street robberies in Chicago ran in less than two seconds on a 800 mhz PC with projected coordinates (Euclidean), while it took longer with spherical coordinates (latitude/longitude).

2. Define the reference file (see chapter 3). While *CrimeStat* does not include a data base manager or query system, a user can carry out analysis of different areas of a jurisdiction by using the boundaries of several reference areas. For example, define all of Chicago as a reference area and define each of the twenty-five police districts as additional reference areas. Hot Spot Areas can be identified for the city as a whole and for each district. In other words, the same incident file may be used for analysis of different map areas by using multiple reference files.

3. Define the search radius. Generally, a two-stage analysis is best. Start with a larger search radius and then analyze Hot Spot Areas with a smaller search radius. A search radius of more than one mile may not yield useful results in an area the size of Chicago (320 square miles).

4. Set the output units to miles or kilometers.

5. Specify the file output name for the ellipses.

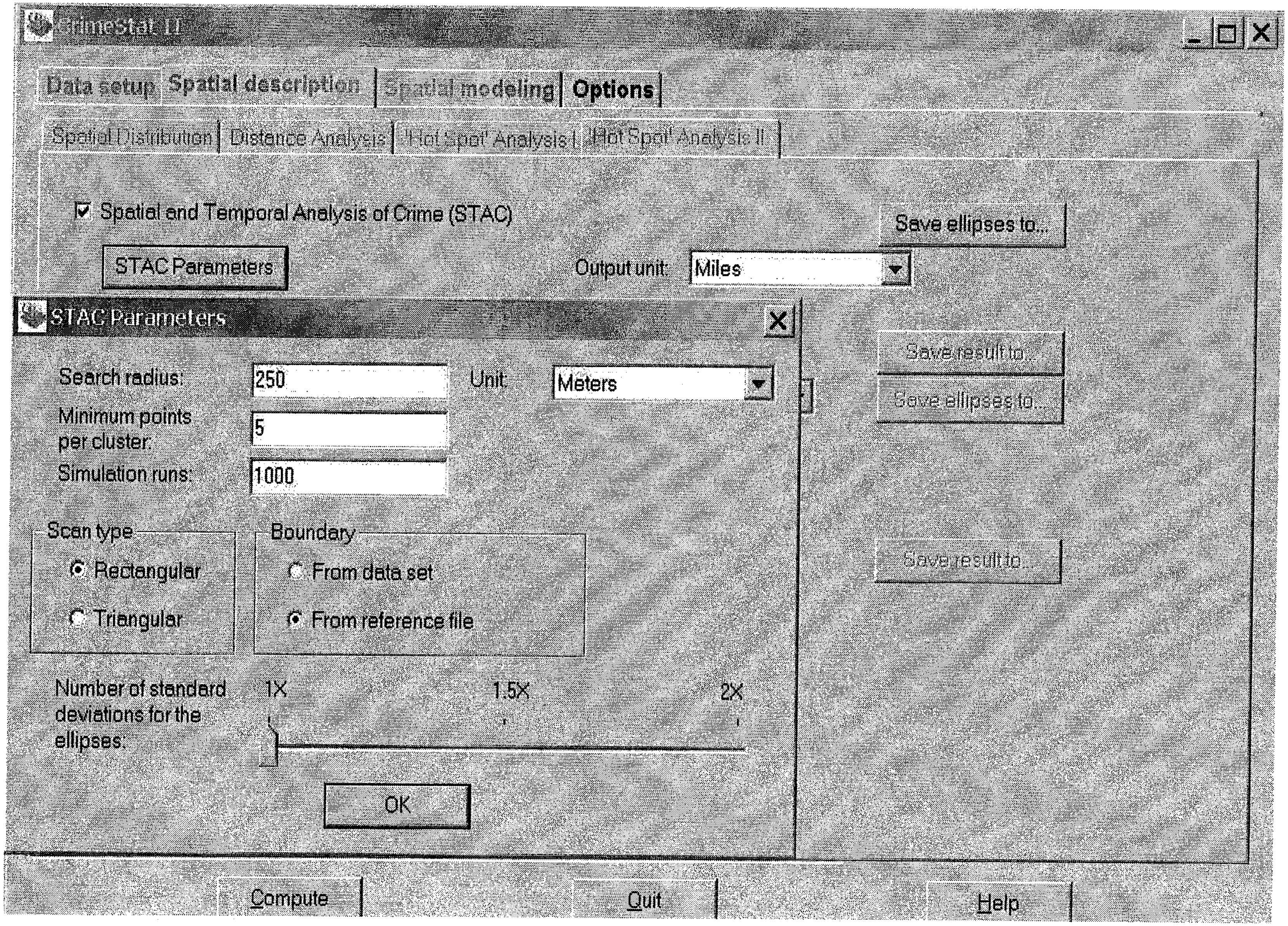
6. Click on the *STAC* parameters button.

The object of *STAC* is to identify hot spots and display them with ellipses. Its key function is visual. Save the ellipses in the form most appropriate for the system (e.g., *ArcView*, *Atlas*, *MapInfo*). Because the ellipses are generated as polygons, they can be used for selections, queries, or thematic maps in the GIS. In addition to the ellipses, a table is output with all the information on density and location for each ellipse. It can be saved to a 'dbf' file, which can then be read by any spreadsheet program. The ellipses are numbered in the sam eorder as the printed output.

STAC Parameters

The two most important parameters for running STAC are the boundary of the study area (reference area) and the search radius. A detailed discussion of the parameters follows. Figure 7.4 shows the *STAC* parameters screen.

Figure 7.4: STAC Parameters Setup



☑ Spatial and Temporal Analysis of Crime (STAC)

Output unit: Miles

STAC Parameters

Search radius: 250 Unit: Meters

Minimum points per cluster: 5

Simulation runs: 1000

Scan type
● Rectangular
○ Triangular

Boundary
○ From data set
● From reference file

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

This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Search Radius

1. The search radius is the key setting in *STAC*. In general, the larger the search radius, the more incidents that will be included in each Hot Cluster and the larger the ellipse that will be displayed. Smaller search radii generally result in more ellipses of a smaller size. A good strategy is to initially use a larger radius and then re-analyze areas that are 'hot' with a smaller radius. In Chicago, we have found that a 750 meter radius is appropriate for the city as a whole and a 200 meter search radius for one of the 25 districts. It will be necessary to experiment to determine an appropriate search radius.

Units

2. Specify the units for the search radius. The default is miles and the default search radius is 0.5 miles. Be careful about using larger search radii. In Chicago, a search radius larger than one mile generates ellipses that are too large to be of any tactical or planning use. Other good choices are 750 meters or 0.25 miles.

Minimum Points Per Cluster

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

Boundary

4. Select the reference file to be used for the analysis. The user can choose the boundary from the data set (i.e., the minimum and maximum X/Y values) or from the reference boundary. In our opinion, the choice of the reference boundary is best. If the data set is used to define the reference boundary, the smallest rectangle that encompasses all incident will be used.

Scan Type

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

Number of Standard Deviations for the Ellipses

6. Select the number of standard deviations for the ellipses. One (1X), 1.5X, and 2X standard deviations can be selected. One standard deviational ellipses should be sufficient for most analysis. While one standard deviational ellipses rarely overlap, 1.5X and 2X two standard deviational ellipses often

do. A larger ellipse will include more of the Hot Cluster points; a small ellipse will produce a more focused Hot Cluster identification. The user will have to work out a balance between defining a cluster precisely compared to making it so large as to be unclear where one starts and another ends.

Simulation Runs

7. Specify whether any simulation runs are to be made. To test the significance of *STAC* clusters, it is necessary to run a Monte Carlo simulation (Dwass, 1957; Barnard, 1963). *CrimeStat* includes a Monte Carlo simulation routine that produces approximate confidence intervals for the particular STAC model that has been run. The difference between the density of incidents in *STAC* ellipses in a spatially random data set and the *STAC* ellipses in the actual data set is a test of the strength of the clustering detected by *STAC*. Essentially, the Monte Carlo simulation assigns N cases randomly to a rectangle with the same area as the defined study area as specified on the Measurement Parameters tab and evaluates the number of clusters according to the defined parameters (i.e., search radius). It repeats this test 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 confidence intervals for the particular number of clusters and density of clusters. The default is zero simulation runs because the simulation run option usually increases the calculation time considerably. If a simulation run is selected, the user should identify the area of the study region on the Measurement Parameters tab. It is better to use the jurisidictional area rather than the reference area if the jurisdiction is irregularly shaped.

Output

Ellipses

The *STAC* ellipse files (see above) can easily be incorporated into a GIS system. *ArcView* shape files can be opened as themes. STAC ellipse files also can be added as a *MapInfo* layer using the Universal Translator Tool. *MapInfo* Mif/Mid files must be imported using the command table >import. Both *MapInfo* and *ArcView* files are polygons and can be used for queries, thematics, and selections.

Printed Output

Table 7.1 shows the printed output. Note that the printed output does not include the file name. Be sure to record the file name and the reference file (if any that is used).

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

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

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

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

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

6. Search Radius is the units specified in the setup. In Figure 7.2 above, this is meters.

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

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

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

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

    A. Cluster an indcntification number of each ellipse. This corresponds to their order in a table view in ArcView, or Browser in MapInfo.

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

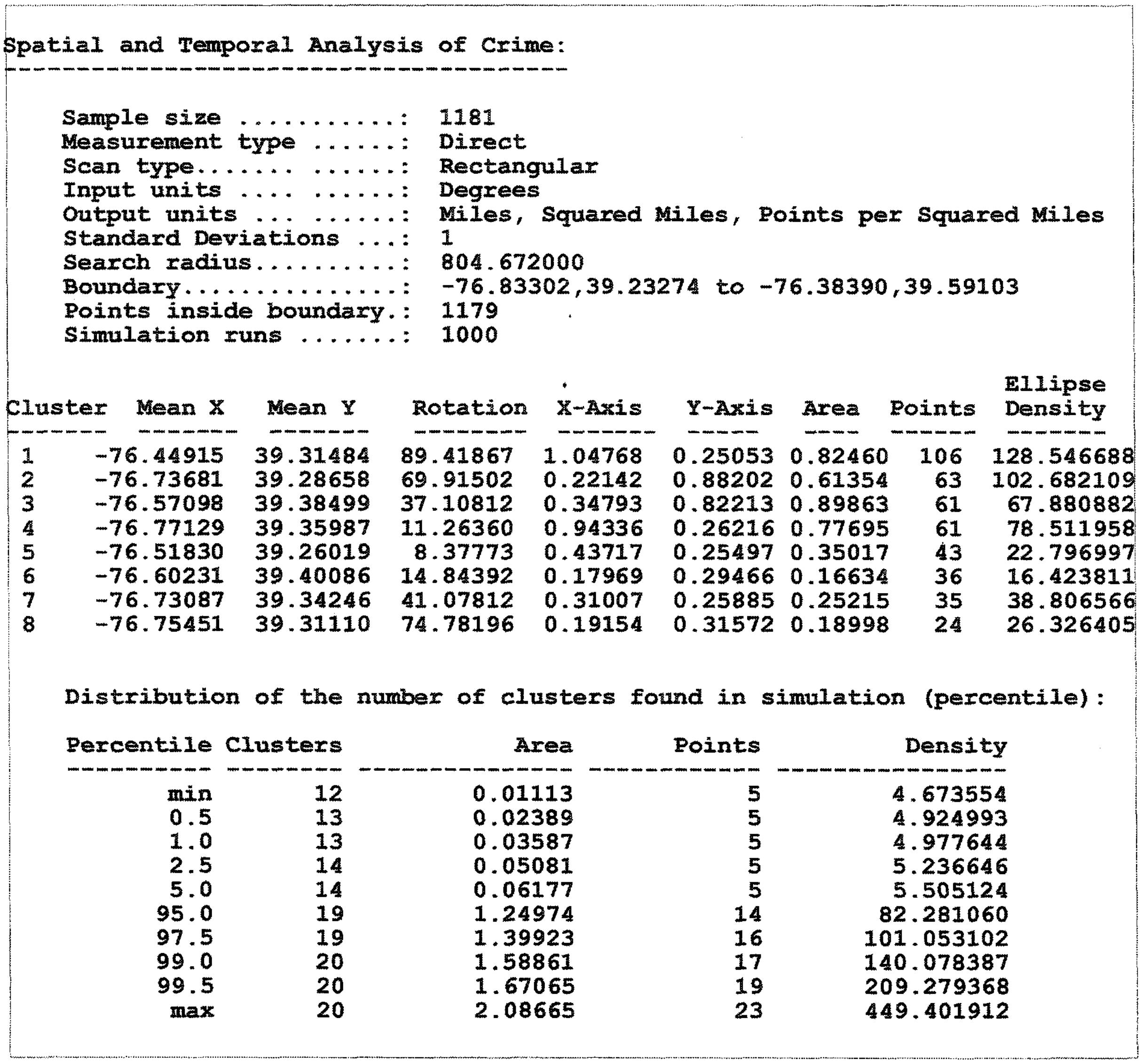
    C. Rotation- the degrees the ellipse is rotated (0 is horizontal; 90 is vertical).

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

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

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

Table 7.1
Printed Output for STAC



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

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

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

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

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

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

The best way to print or save *CrimeStat* printed output is to place the cursor inside the output window and *Select all,* then copy and paste the selection into a word processing document in landscape mode.

Make sure to adequately annotate the file, especially the type of incidents, the reference boundary, and the name of the output file. This can be very important for future reference.

For Old *STAC* Users

In general, *STAC* has retained all the functionality and speed of previous versions. The ellipses will look somewhat different than previous versions, because a more widely accepted method for calculating standard deviational ellipses has been used. *STAC for DOS* used a 1x standard deviation ellipse. Analysts who want results similar to *STAC* for DOS should set standard deviations to 1.

The *CrimeStat* version of *STAC* has the following improvements over STAC for DOS:

1. *STAC* no longer requires the use of a special ASCII data file. The data file can be any of those available in *CrimeStat.*

2. Any projection can be used, including latitude/longitude. Files are not converted into a Euclidean projection.

3. We have not found a limit on the number of points that can be analyzed with the *CrimeStat* version of STAC. Therefore, a small radius can now be used over large areas.

4. *STAC* can generate Shape files for ArcView or Mif/Mid files for MapInfo. Both are polygons-not points.

5. It is easier for the user to specify the number of standard deviations for an ellipse (1X, 1.5X, or 2X).

6. The user can run *STAC* on a spatially random data set to get an estimate of the degree of clustering detected by *STAC* in the incident data.

7. The study area boundary (reference file) can be generated from the data set (we would suggest not doing this).

Example 1: A STAC Analysis of 1999 Chicago Street Robberies

STAC Hot Spot Areas were calculated for all street (or sidewalk or alley) robberies occurring in Chicago in 1999 (n=13,009).[3] There were 13,007 within the search boundary. The search radius was set for 750 meters (approximately ½ mile), and the ellipses were set to one standard deviation. Ten was the minimum number of incidents per cluster.

In figure 7.2 (shown earlier), STAC detected seven ellipses. The areas of the seven ellipses ranged from 5 square kilometers to 0.7 square kilometers, and the number of incidents in an ellipse ranged from 760 to 153. The smallest ellipse (number 7 in figure 7.2) was the densest, 222 robberies per square kilometer. Of the 13,007 incidents, 2,375 were in a cluster. Therefore, 18 percent of all of Chicago's street robberies in 1999 occurred in 6% of its 233 square mile area.

To map the results, the ellipse boundaries were imported into *MapInfo* as a mif/mid file and overlaid on a map of police districts. The large blue rectangle in figure 7.2 designates the search boundary (reference file). O'Hare Airport was excluded because exact geo-coding is not possible for the few street robberies that occurred there. At a city-wide scale, the map is interesting, but is mainly useful for confirming what is already known. Ellipse 1, on the west side, has had a high level of violence for many years. Ellipses 2 and 6 are centered on areas where high rise public housing projects are gradually being abandoned. Overall, these ellipses are not very useful for tactical purposes. However, they point out that four Hot Spot Areas cross District boundaries, and that the large number of street robberies in these areas might be lost in separate district reports.

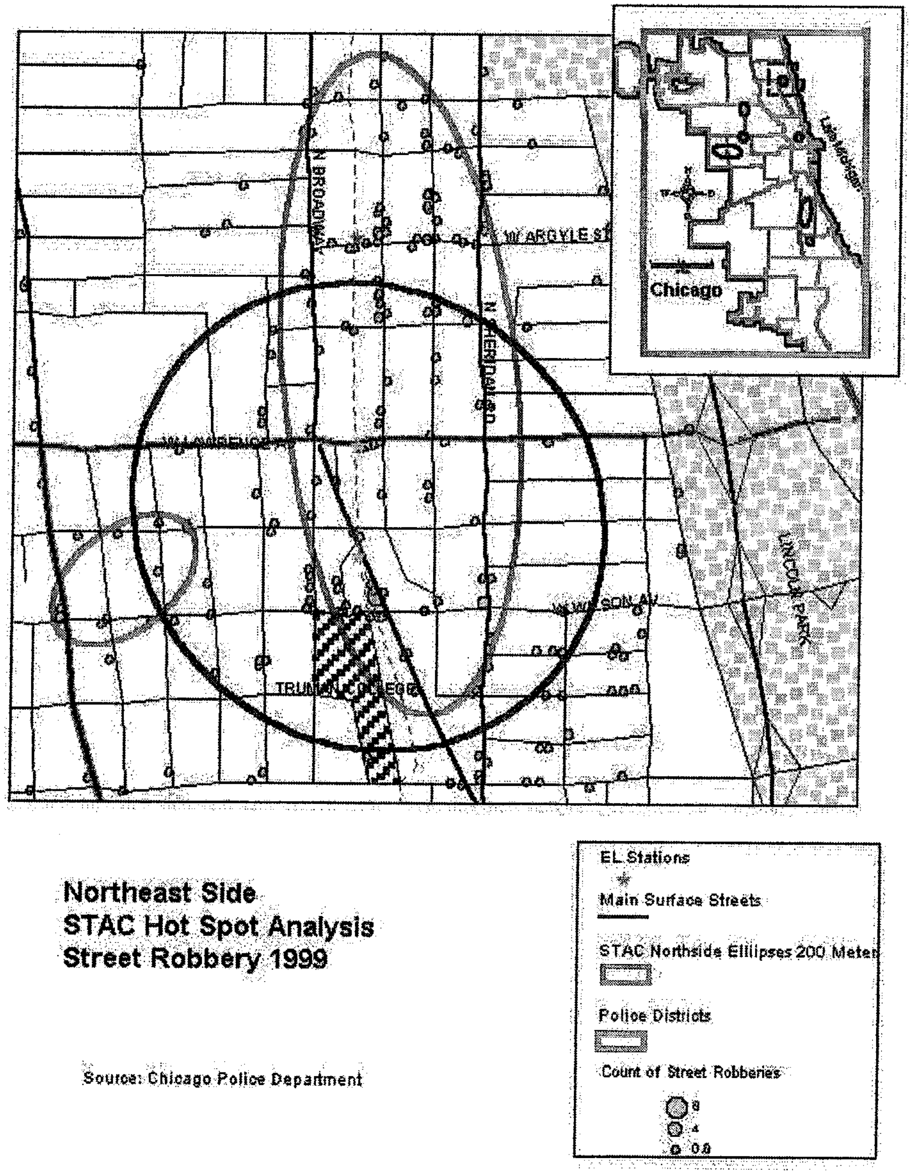
A Neighborhood STAC Analysis

The presence of Ellipse 4 (the northernmost ellipse in figure 7.2) might be unexpected to many Chicagoans. The mid-Northside, near the Lake Michigan, is generally considered to be a relatively affluent and safe neighborhood. However, the neighborhood around Ellipse 4 has had a high level of crime for many years. It was an entertainment center in the Roaring Twenties, and several institutions of that era remain. Today it is an area with multiple, often conflicting, uses. A more detailed analysis of the neighborhood with the help of STAC may point to specific areas that need increased patrol or prevention activities.

The second step of STAC analysis was to define a focused search boundary area around Ellipse 4. This was done easily by creating a new map layer in MapInfo and drawing a rectangle around the desired study area. Clicking on the study area gave the required *CrimeStat* reference boundary maximum and minimum coordinates. Using this more focused boundary, STAC was run a second time with a 200 meter search radius and the same file of 13,009 cases. The search boundary (reference file) now contained 442 incidents. STAC detected three ellipses that contained 231 incidents. The STAC ellipses were then imported into *MapInfo* and mapped (Figure 7.5).

As the area covered by a map grows smaller, detailed information about crime patterns and the community can be added. In this map, the STAC ellipses were overlain with the address locations of incidents (sized according to the number occurring at each location) and streets.[4] Much of the area is relatively crime-free. The most frequent locations for street robbery do not coincide with main streets. Street robbery incidents tend to cluster near rapid transit stations and the blocks immediately surrounding them. For example, Argyle Street, between Broadway and Sheridan, is the site of "New China Town." It is an an area with a number of street robberies and is a destination area for "Northsiders" who want an inexpensive Chinese or Vietnamese meal.

Figure 7.5: **STAC Hot Spots for Northeast Side Street Robberies**



Source: Chicago Police Department

There is a particularly risky area in the neighborhood of Broadway and Wilson adjacent to Truman Community College. In a previous analysis of the Bronx, Fordham University was shown to be a similar attractor for robbery incidents. Colleges supply good targets for street robbery. Also, authority for security is split between the college and the city police. The area around Broadway and Wilson has been risky for many years. Ninety years ago, it was the northern terminus of rapid transit, and the site of several very inexpensive hotels, two of which still exist. Today the area has several pawn shops and currency exchanges. There is an ATM located in the EL station. The area looks dangerous and dirty. Finally, the area has many blind corners and alleys that could serve as sites for robbery; this is unusual for Chicago. The census block that includes the northwest corner of Broadway and Wilson ranked fifth among Chicago's 21,000 census blocks in number of street robberies in 1999.

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

Advantages of STAC

STAC has a number of advantages as a clustering algorithm:

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

2. The STAC user controls the approximate size of the ellipses (search radius), the minimum number of points per ellipse, and the study area. These features allow for a broad search for Hot Spot Areas over an entire city and a second search concentrating on a smaller area and deriving focused Hot Spot Areas for local tactical use.

3. STAC and Heirarchical Clustering are complimentary. Heirarchical Clustering first derives small ellipses and then aggregates to larger ones. The recommended STAC procedure is to first derive large scale ellipses and then to analyze these for tactical use.

4. The visual display of STAC ellipses is quite intuitive.

5. Hot spots need not be limited to a single kind of crime, place or even. For example, ellipses of drug crime can be overlain on those for burglary. Some

causal factors are also analyzable with STAC ellipses. For example, ellipses of street robbery can be compared to those for liquor licenses.

6. STAC combines features of a hierarchical and partitioning search methods and adapts itself to the size of the clusters.

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

Limitations of STAC

There are also some limitations to using STAC:

1. The distribution of incidents within an ellipse is not necessarily uniform. The user should be careful not to assume that it is. A mapped theme of *CrimeStat's* Mode routine (see above) according to number of incidents or the single kernel density interpolation (see chapter 8) overlaid with STAC ellipses are good ways to overcome this problem (figure 7.5 above and figure 7.6 below).

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

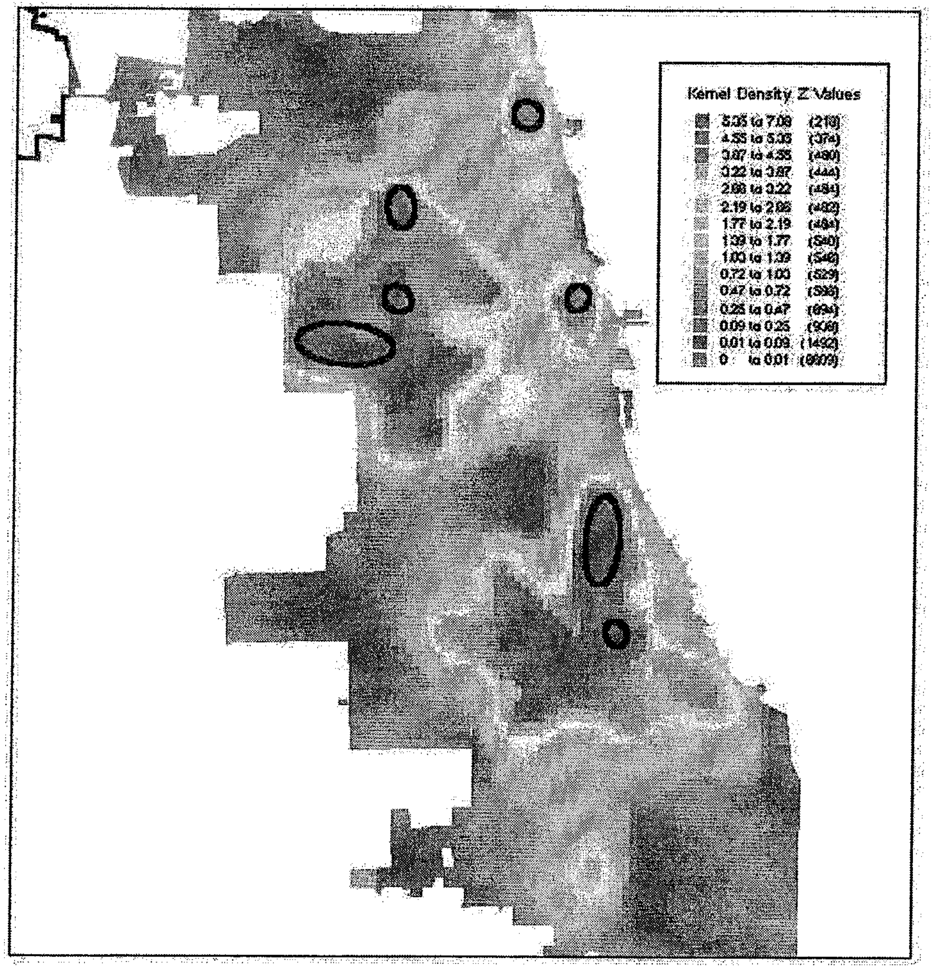
3. Small changes in the STAC study area boundary (reference file) can result in quite different depictions of the ellipses. This is true of any clustering routine. Retaining the same reference file over repeated analyses alleviates this problem. The analysis should also be documented for the analysis parameters.

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

K-Means Partitioning Clustering

The *K-means* clustering routine (Kmeans) is a partitioning procedure where the data are grouped into *K* groups defined by the user. A specified number of seed locations, *K*, are defined by the user (Fisher, 1958; MacQueen, 1967; Aldenderfer and Blashfield, 1984; Systat, 2000). The routine tries to find the best positioning of the *K* centers and then assigns each point to the center that is nearest. Like the Nnh routine, the Kmeans assigns points to one, and only one, cluster. However, unlike the nearest neighbor hierarchical (Nnh) procedure, all points are assigned to clusters. Thus, there is no hierarchy in the routine, that is there are no second- and higher-order clusters.

Figure 7.6: STAC Robbery Hot Spots and Kernel Density Estimation



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

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

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

$$\frac{6000!}{20!\,5980!} = 1.456 \text{ x } 10^{57}$$

combinations. No computer can solve that number and few spreadsheets can calculate the factorial of *N* greater than about 127.[5] In other words, it is almost impossible to solve computationally.

Practically, therefore, the different implementations of the K-means routine all make initial guesses about the *K* locations and then optimize the seating of this location in relation to the nearby points. This is called *local optimization*. Unfortunately, each K-means routine has a different way to define the initial locations so that two K-means procedures will usually not produce the same results, even if *K* is identical (Everitt, 1974; Systat, Inc., 1994).

CrimeStat K-means Routine

The K-means routine in *CrimeStat* also makes an initial guess about the *K* locations and then optimizes the distribution locally. The procedure that is adopted makes initial estimates about location of the K clusters (seeds), assigns all points to its nearest seed location, re-calculates a center for each cluster which becomes a new seed, and then repeats the procedure all over again. The procedure stops when there are very few changes to the cluster composition.[6]

The default K-means clustering routine follows an algorithm for grouping all point locations into one, and only one, of these K groups. There are two general steps: 1) the identification of an initial guess (seed) for the location of the K clusters, and 2) local optimization which assigns each point to the nearest of the K clusters. A grid is overlaid on the data set and the number of points falling within each grid cell is counted. The grid cell with the most points is the initial first cluster. Then, the second initial cluster is the grid cell with the next most points that is separated by at least:

$$\text{Separation} = \quad t * 0.5 * \text{SQRT}\left[\frac{A}{N}\right] \qquad (7.1)$$

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

The algorithm then conducts *local optimization*. It assigns each point to the nearest of the K seed locations to form an initial cluster. For each of the initial clusters, it calculates the center of minimum distance and then re-assigns all points to the nearest cluster, based on the distance to the center of minimum distance. It repeats this process until no points change clusters. Finally, it calculates the standard deviational ellipse of each cluster and outputs the results as a graphical object. To increase the flexibility of the routine, the grid that is overlaid on the data points is re-sized to accommodate different cluster structures, increasing or decreasing in size to try to find the K clusters. After iterating through different grid sizes, the code makes sure that the final seeds are from the "best" grid or the grid that produces the most clusters.

Control over Initial Selection of Clusters

Changing the separation between clusters

The problem with this approach is that in highly concentrated distributions, such as with most crime incidents in a metropolitan area, the separation between clusters may not be sufficiently large to detect clusters farther away from the concentration; the algorithm will tend to sub-divide concentrated groupings of incidents into multiple clusters rather than seek clusters that are less concentrated and, usually, farther away. To increase the flexibility of the routine, *CrimeStat* allows the user to modify the initial selection of clusters since this has a large effect on the final grouping (Everett, 1974). There are two ways the initial selection of cluster centers can be modified. The user can increase or decrease the separation factor. Formula 7.1 is still used to separate each of the initial clusters, but the user can either select a t-value from 1 to 10 from the drop down menu or write in any number for the separation, including fractions, to increase or decrease the separation between the initial clusters. The default is set at 4.

Figure 7.7 shows a simulation of eight clusters, four of which have higher concentrations than the other two. Two partitions of the data set into eight groups are shown, one using a separation of 4 (dashed green ellipses) and one with a separation of 15 (solid blue ellipses). As seen, the partition with the larger separation captures the eight clusters better. With the smaller separation, the routine will tend to sub-divide more concentrated clusters because that reduces the distance of each point from the cluster center. Depending on the purpose of the partitioning, a greater or lesser separation may be desired.

Selecting the initial seed locations

Alternatively, the initial clusters can be modified to allow the user to define the actual locations for the initial cluster centers. This approach was used by Friedman and Rubin (1967) and Ball and Hall (1970). In *CrimeStat,* the user-defined locations are entered with the secondary file which lists the location of the initial clusters. The routine reads the secondary file and uses the number of points in the file for K and the X/Y coordinates of each point as the initial seed locations. It then proceeds in the same way with local optimization. When eight points that were approximately in the middle of the eight clusters in figure 7.7 were input as the secondary file, the K-means routine immediately identified the eight clusters (results not shown). Again, depending on the purpose the user can test a particular clustering by requiring the routine to consider that model, at least for the initial seed location. The routine will conduct local optimization for the rest of the clustering, as in the above method.

The K-means output is similar for both routines. It includes the parameters for the standard deviational ellipse of each cluster. 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. In general, use a 1X standard deviational ellipse since 1.5X and 2X standard deviations can create an exaggerated view of the underlying cluster. The ellipse, after all, is an abstraction from the points in the cluster which may be arranged in an irregular manner. On the other hand, for a regional view, a one 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.

Mean squared error

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

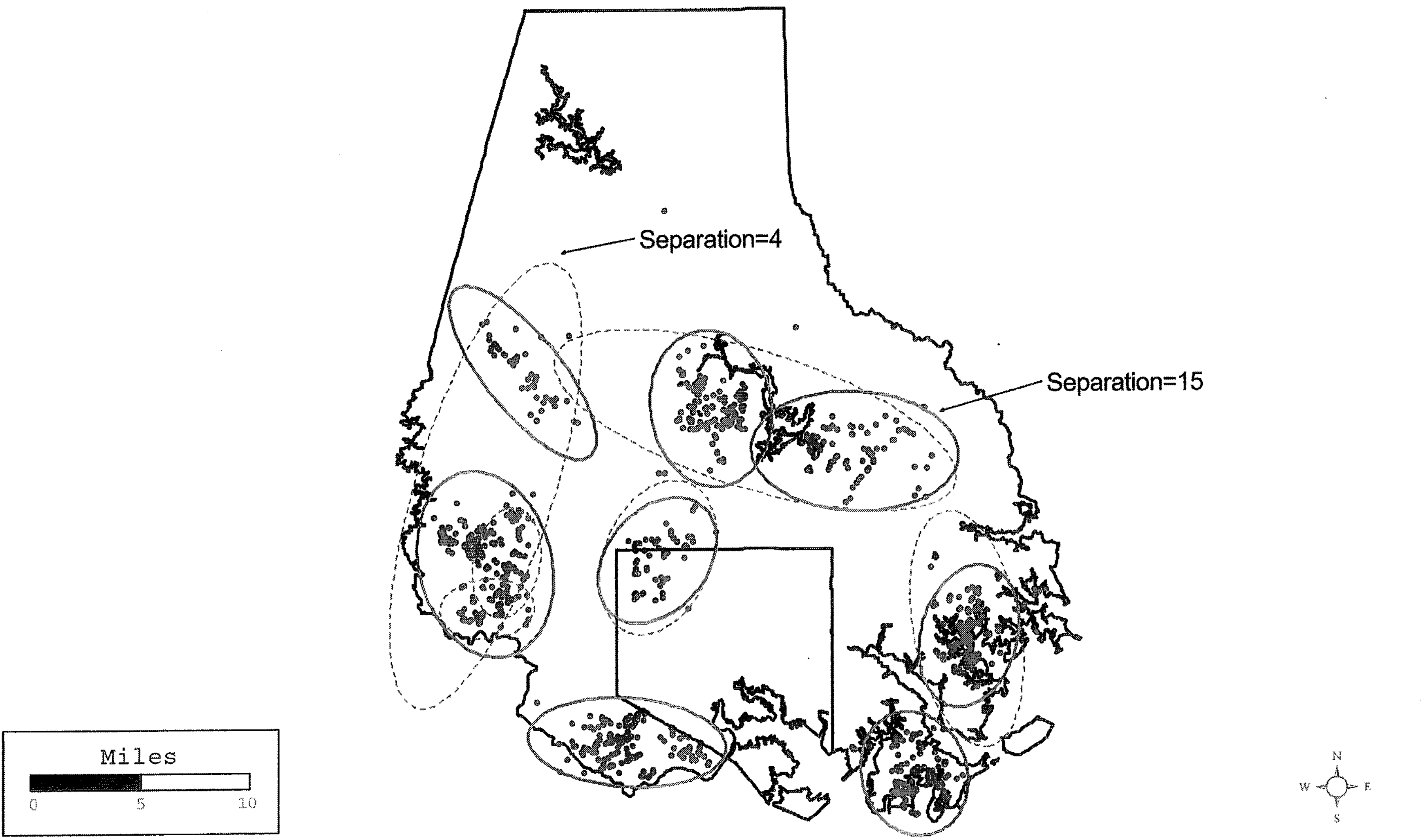
Sum of squares
of cluster C $= SSE_C = \sum_{i=1}^{N_C} \{ [(X_{iC} - MeanX_C]^2 + [Y_{iC}-MeanY_C]^2 \}$ (7.2)

Mean squared
error of cluster C $= MSE_C = SSE_C \;/\, (N_C - 1)$ (7.3)

Figure 7.7:

Separated Data and K-Means Solution

K=8 Paritions with Two Separations of Initial Seed Locations



where X_{iC} is the X value of a point that belongs to cluster C, Y_{iC} is the Y value of a point that belongs to cluster C, MeanX_C is the mean X value of cluster C (i.e., of only those points belonging to C), MeanY_C is the mean Y value of cluster C, and N_C is the number of points in cluster C. There is also a total sum of squares and a total mean square error which is summed over all clusters

$$\text{Total Sum of Squares} = \sum_{C} SSE_C \tag{7.4}$$

$$\text{Total Mean Squared Error} = \sum_{C} SSE_C/(N-K-1) \tag{7.5}$$

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

The sum of squares (or sum of squared errors) is frequently used as a criteria for identifying 'goodness of fit' (Everett, 1974; Aldenderfer and Blashfield, 1984; Gersho and Gray, 1992). In general, for a given number of clusters, K, those with a smaller sum of squares and, correspondingly, smaller mean square error are better defined than clusters with a larger sum of squares and larger mean squared error. Similarly, a K-means solution that produces a smaller overall sum of squares is a tighter grouping than a grouping that produces a larger overall sum of squares.

But, there can be exceptions. If there are points which are 'outliers', that is which don't obviously fall into one cluster or another, re-assigning them to one or another cluster can distort the sum of squares statistics. Also, in highly concentrated distributions, such as with crime incidents, a smaller sum of squares criteria can be obtained by splitting the concentrations rather than clustering less central and less dense groups of incidents (such as in figure 7.7); the results, while minimizing the sum of squared errors from the cluster centers, will be less desirable because the peripheral clusters are ignored. Thus, these statistics are presented for the user's information only. In assigning points to clusters, *CrimeStat* still uses the distance to the nearest seed location, rather than a solution that minimizes the sum of squared distances.

Visualizing the Clusters with Ellipses

Finally, the K-means clustering routine (Kmeans) outputs clusters as ellipses, similar to the other clustering routines. The user can choose between 1X, 1.5X, and 2X standard deviations to display the ellipses. The prefix 'Km' is used to designate the ellipses of clusters irrespective of the number of standard deviations. It should be noted, however, that the ellipses are an abstraction of the cluster. The clusters are *not* necessarily arranged in ellipses. They are for visualization purposes only.

K-means Output Files

The naming system for the K-means outputs is simpler than the Nnh routine since there are no higher-order clusters. The final seed locations are displayed in the output table and can be saved as a '.dbf' file. A slide-bar allows ellipses to be defined for 1X, 1.5X, and 2X standard deviations and can be output in *ArcView* '.shp', *MapInfo* '.mif' or *Atlas*GIS* '.bna' formats. Each file is named

Km<*username*>

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

KmEll<N><*username*>

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

KmEll3robbery

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

KmEll12burglary

is the 12th ellipse for the file called 'burglary'.

Example 2: K-means Clustering of Street Robberies

In *CrimeStat*, the user specifies the number of groups to sub-divide the data. Using the 1996 robbery incidents for Baltimore County, the data were partitioned into 10 groups with the K-means routine (figure 7.8). As can be seen, the clusters tend to fall along the border with Baltimore City. But there are three more dispersed clusters, one concentrated in the central eastern part of the county and two north of the border with the City. Because these clusters are very large, a finer mesh clustering was conducting by partitioning the data into 35 clusters (figure 7.9). Though the ellipses are still larger than those produced by the nearest neighbor hierarchical procedure (see figure 6.7 in chapter 6), there is some congruency; clusters identified by the nearest neighbor procedure have corresponding ellipses using the K-means procedure.

Figure 7.10 shows a section of southwest Baltimore County with five full ellipses and one partial ellipse visible. Looking at the distribution, several ellipses make intuitive sense while a couple of others do not. For example, two ellipses highlight a concentration along a major arterial (U.S. Highway 40). Similarly, the ellipse in the lower right appears to capture incidents along two arterials. However, the other three full ellipses do not appear to capture meaningful clusters and appear somewhat arbitrary.

Figure 7.8:

Baltimore County Robbery 'Hot Spots'

Using K-Means Routine with K=10 Clusters

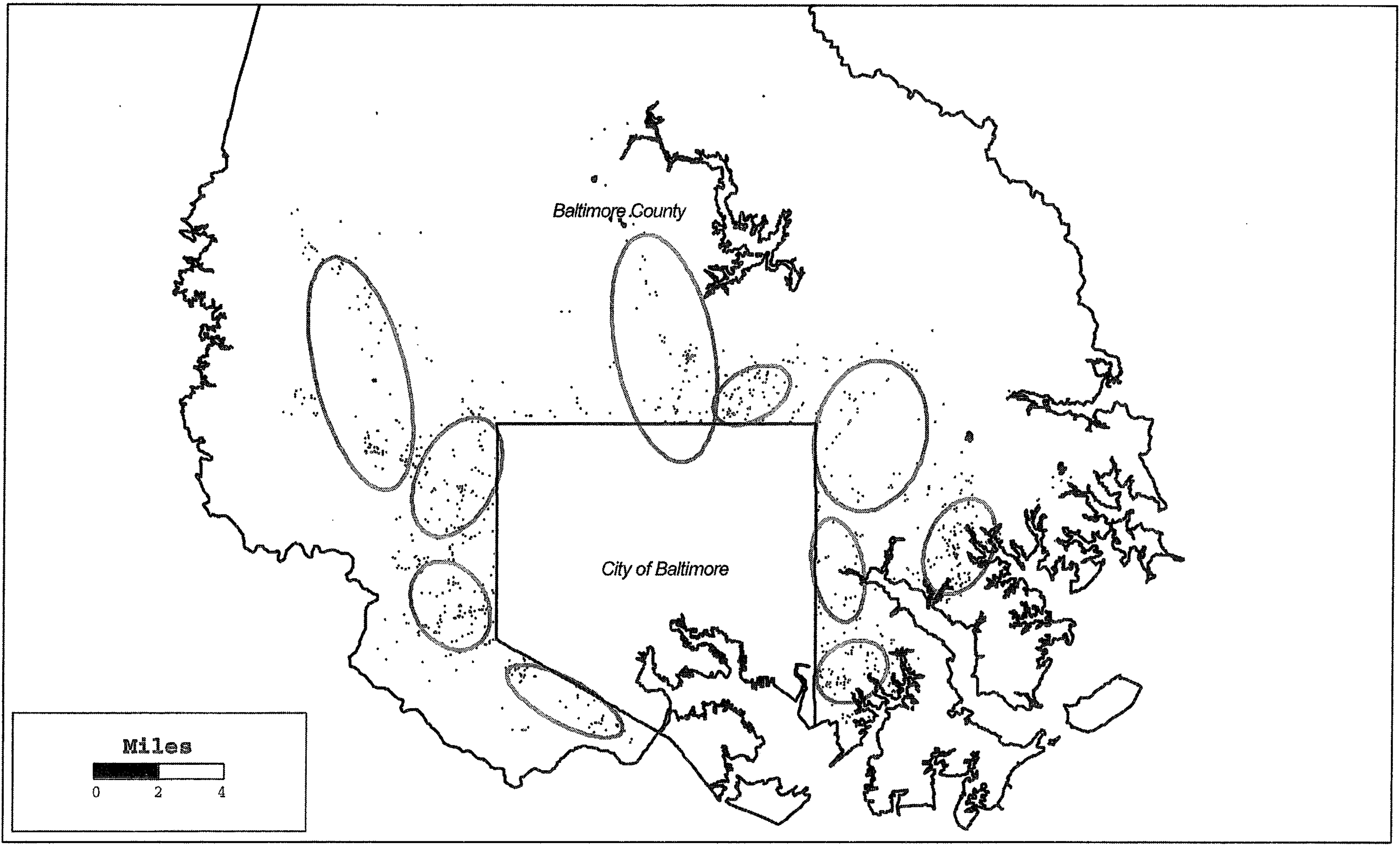


Figure 7.9:

Baltimore County Robbery 'Hot Spots'

Using K-Means Routine with K=35 Clusters

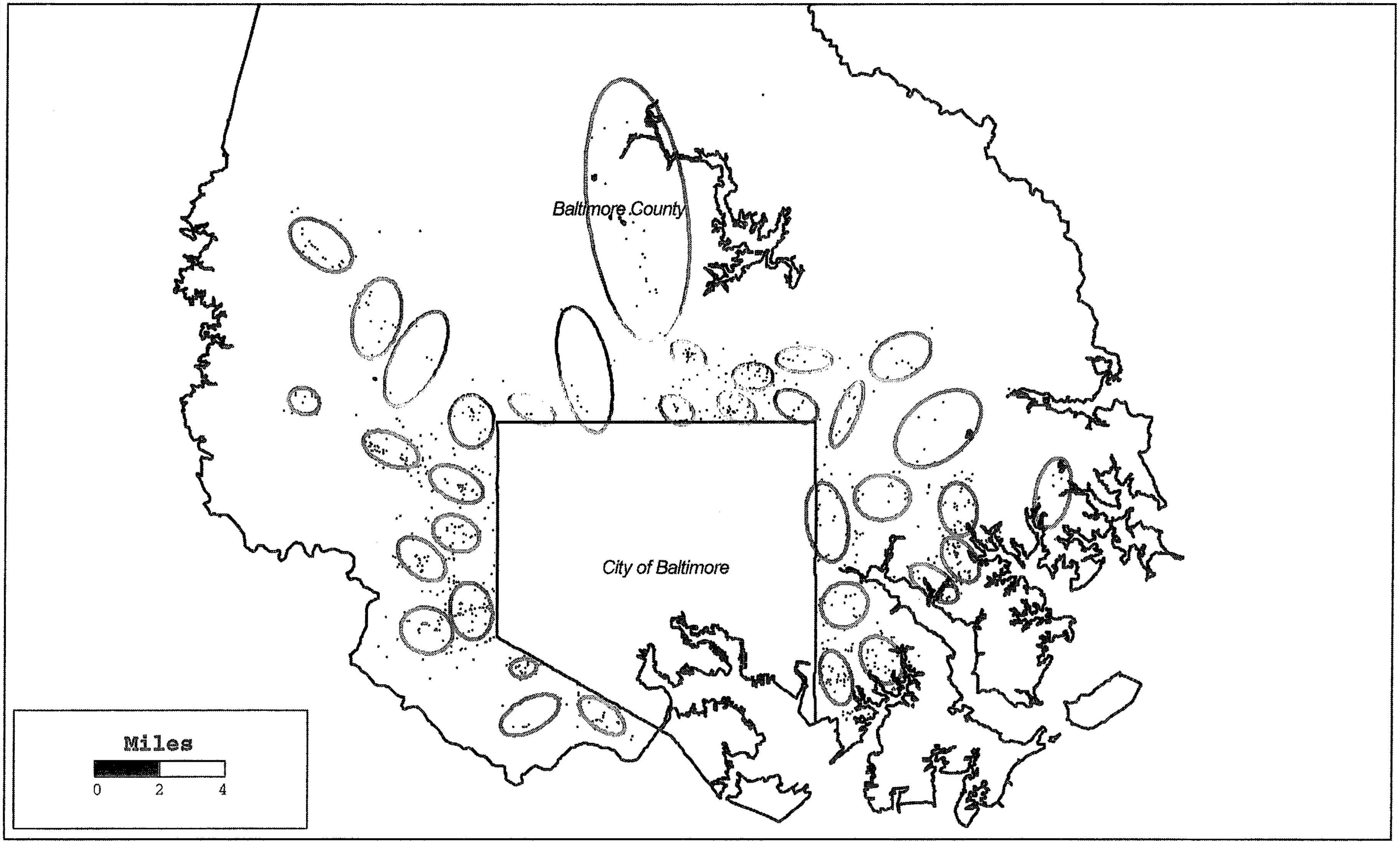
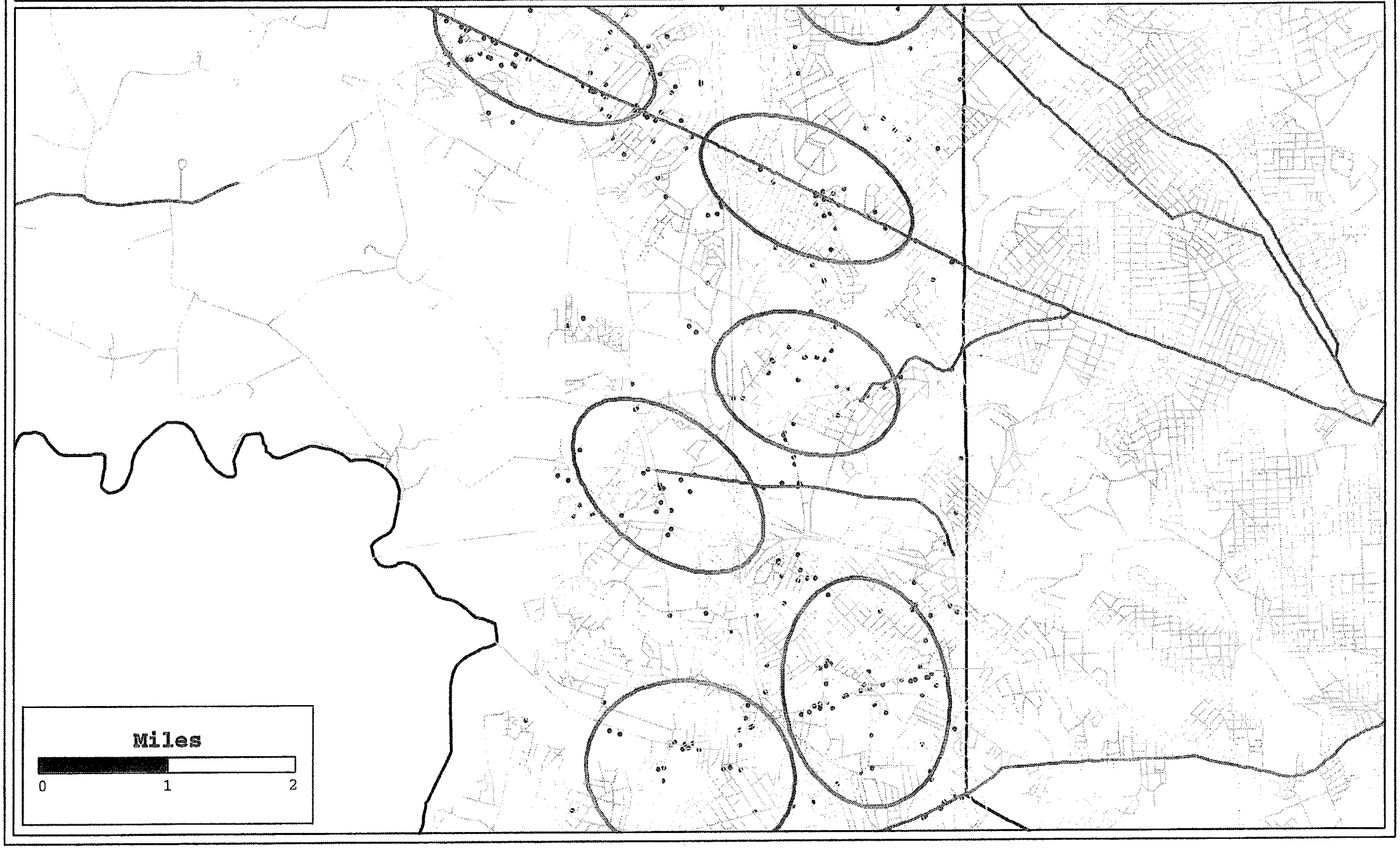


Figure 7.10:

Southwest Baltimore County Robbery 'Hot Spots'

Using K-Means Routine with K=35 Clusters



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

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

Advantages and Disadvantages of the K-means Procedure

In short, the K-means procedure will divide the data into the number of groups specified by the user. Whether these groups make any sense or not will depend on how carefully the user has selected clusters. Choosing too many will lead to defining patterns that don't really exist whereas choosing too few will lead to poor differentiation among neighborhoods that are distinctly different.

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

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

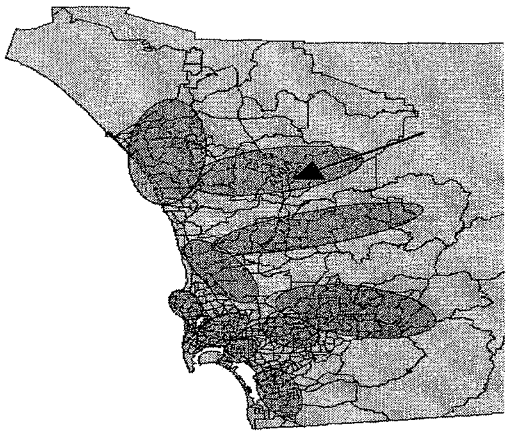
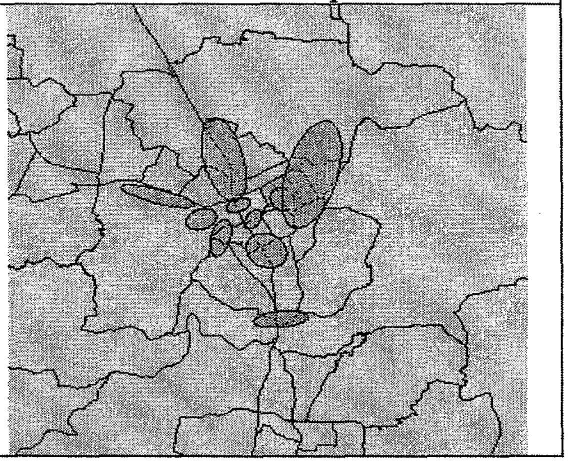
The technique is, therefore, better seen as both an exploratory tool as well as a tool for refining a 'hot spot' search. If the user has a good idea of where there should be 'hot spots', based on community experience and the reports of beat officers, then the technique can be used to see if the incidents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

K-Means Clustering as an Alternative Measure of Urban Accessibility

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

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

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

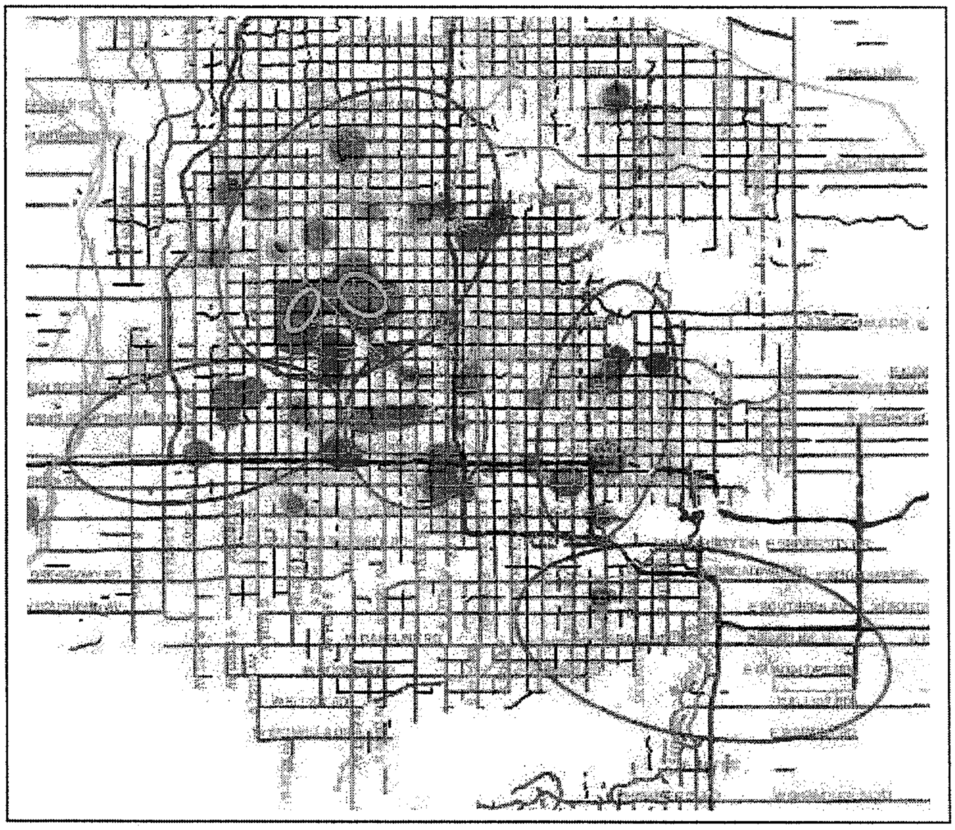
Regional Hot Spots	Local Hot Spots
	

Hot Spot Verification in Auto Theft Recoveries

Bryan Hill
Glendale Police Department
Glendale, AZ

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

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



In the map, the K-means (red ellipses), Nnh (green ellipses) and *Spatial Analyst* grid (red-yellow grid cells) all showed that the area was a high density or clustering of stolen vehicle recoveries. Although this information was not new, it did help verify our conclusion and aided in organizing a response statistical procedure. Experience and sensitivity are needed to know whether an identified 'hot spot' is real or not.

Anselin's Local Moran Statistic (LMoran)

The last 'hot spot' technique in *CrimeStat* is a zonal technique called the *Anselin's Local Moran* statistic and was developed by Luc Anselin (1995). Unlike the nearest neighbor hierarchical and K-means procedures, the local Moran statistics require data to be aggregated by zones, such as census block groups, zip codes, police reporting areas or other aggregations. The procedure applies Moran's I statistic to individual zones, allowing them to be identified as similar or different to their nearby pattern.

The basic concept is that of a *local indicator of spatial association* (*LISA*) and has been discussed by a number of researchers (Mantel, 1967; Getis, 1991; Anselin, 1995). For example, Anselin (1995) defines this as any statistic that satisfies two requirements:

1. The *LISA* for each observation indicates the extent to which there is significant spatial clustering of similar values around that observation; and

2. The sum of the *LISA*s for all observations is proportional to the global indicator of spatial association.

$$L_i = f(Y_i, Y_{J_i}) \tag{7.6}$$

where L_i is the local indicator, Y_i is the value of an intensity variable at location i, and Y_{J_i} are the values observed in the neighborhood J_i of i.

In other words, a *LISA* is an indicator of the extent to which the value of an observation is similar or different from its neighboring observations. This requires two conditions. First, that each observation has a variable value that can be assigned to it (i.e., an intensity or a weight) in addition to its X and Y coordinates. For crime incidents, this means data that are aggregated into zones (e.g., number of incidents by census tracts, zip codes, or police reporting districts). Second, the *neighborhood* has to be defined. This could be either adjacent zones or all other zones negatively weighted by the distance from the observation zone.

Once these are defined, the *LISA* indicates the value of the observation zone in relation to its neighborhood. Thus, in neighborhoods where there are 'high' intensity values, the *LISA* indicates whether a particular observation is similar (i.e., also 'high') or different (i.e., low) and, conversely, in neighborhoods where there are 'low' intensity values, the *LISA* indicates whether a particular observation is similar (i.e., also 'low') or different (i.e., 'high'). That is, the *LISA* is an indicator of similarity, not absolute value of the intensity variable.

Formal Definition of Local Moran Statistic

The I_i statistic

Anselin (1995) has applied the concept to a number of spatial autocorrelation statistics. The most commonly used, which is included in *CrimeStat*, is Anselin's Local Moran statistic, I_i, the use of Moran's I statistic as a *LISA*. The definition of I_i is (from Getis and Ord, 1996):

$$I_i = \frac{(Z_i - \bar{Z})}{S_Z^2} * \sum_{j=1}^{N} [W_{ij} * (Z_j - \bar{Z})] \qquad (7.7)$$

where $\bar{Z}$ is the mean intensity over all observations, Z_i is the intensity of observation i, Z_j is intensity for all other observations, j (where $j \neq i$), S_Z^2 is the variance over all observations, and W_{ij} is a distance weight for the interaction between observations i and j. Note, the first term refers only to observation i, while the second term is the sum of the weighted values for all other observations (but not including i itself).

Distance weights

The weights, W_{ij}, can be either an indicator of the adjacency of a zone to the observation zone (i.e., '1' if adjacent; 0 if not adjacent) or a distance-based weight which decreases with distance between zones i and j. Adjacency indices are useful for defining near neighborhoods; the adjacent zones have full weight while all other zones have no weight. Distance weights, on the other hand, are useful for defining spatial interaction; zones which are farther away can have an influence on an observation zone, although one that is much less. *CrimeStat* uses distance weights, in two forms.

First, there is a traditional distance decay function:

$$W_{ij} = \frac{1}{d_{ij}} \qquad (7.8)$$

where d_{ij} is the distance between the observation zone, i, and another zone, j. Thus, a zone which is two miles away has half the weight of a zone that is one mile away.

Small distance adjustment

Second, there is an adjustment for small distances. Depending on the distance scale used (miles, kilometers, meters), the weight index becomes problematic when the distance falls below 1 (i.e., below 1 mile, 1 kilometer); the weight then increases as the distance decreases, going to infinity for $d_{ij} = 0$. To correct for this, *CrimeStat* includes an adjustment for small distances so that the maximum weight can be never be greater than

1.0 (see chapter 4). The adjustment scales distances to one mile. When the small distance adjustment is turned on, the minimal distance is scaled automatically to be one mile. The formula used is

$$W_{ij} = \frac{\text{one mile}}{\text{one mile} + d_{ij}} \qquad (7.9)$$

in whichever units are specified.

Similarity or dissimilarity

An exact test of significance has not been worked out because the distribution of the statistic is not known. The expected value of I_i and the variance of I_i are somewhat complicated (see endnote 7 for the formulas).[7] Instead, high positive or high negative standardized scores of I_i, $Z(I_i)$, are taken as indicators of similarity or dissimilarity. A high *positive* standardized score indicates the spatial clustering of similar values (either high or low) while a high *negative* standardized score indicates a clustering of dissimilar values (high relative to a neighborhood that is low or, conversely, low relative to a neighborhood that is high). The higher the standardized score, the more the observation is similar (positive) or dissimilar (negative) to its neighbors.

In other words, the Local Moran statistic is a good indicator of either 'hot spots' or 'cold spots', zones which are different from their neighborhood. 'Hot spots' would be seen where the number of incidents in a zone is much higher than in the nearby zones. 'Cold spots' would be seen where the number of incidents in a zone is much lower than in the nearby zones. The Local Moran statistic indicates whether the zone is similar or dissimilar to its neighbors. A user must then look at the absolute value of the zone (i.e., the number of incidents in the zone) to see whether it is a 'hot spot' or a 'cold spot'.

For each observation, *CrimeStat* calculates the Local Moran statistic and the expected value of the Local Moran. *If* the *variance* box is checked, the program will also calculate the variance and the standardized Z-value of the Local Moran. The default is for the variance not to be calculated because the calculations are very intense and may take a long time. Therefore, a user should test how long it takes to calculate variances for a small sample on a particular computer before running the variance routine on a large sample.

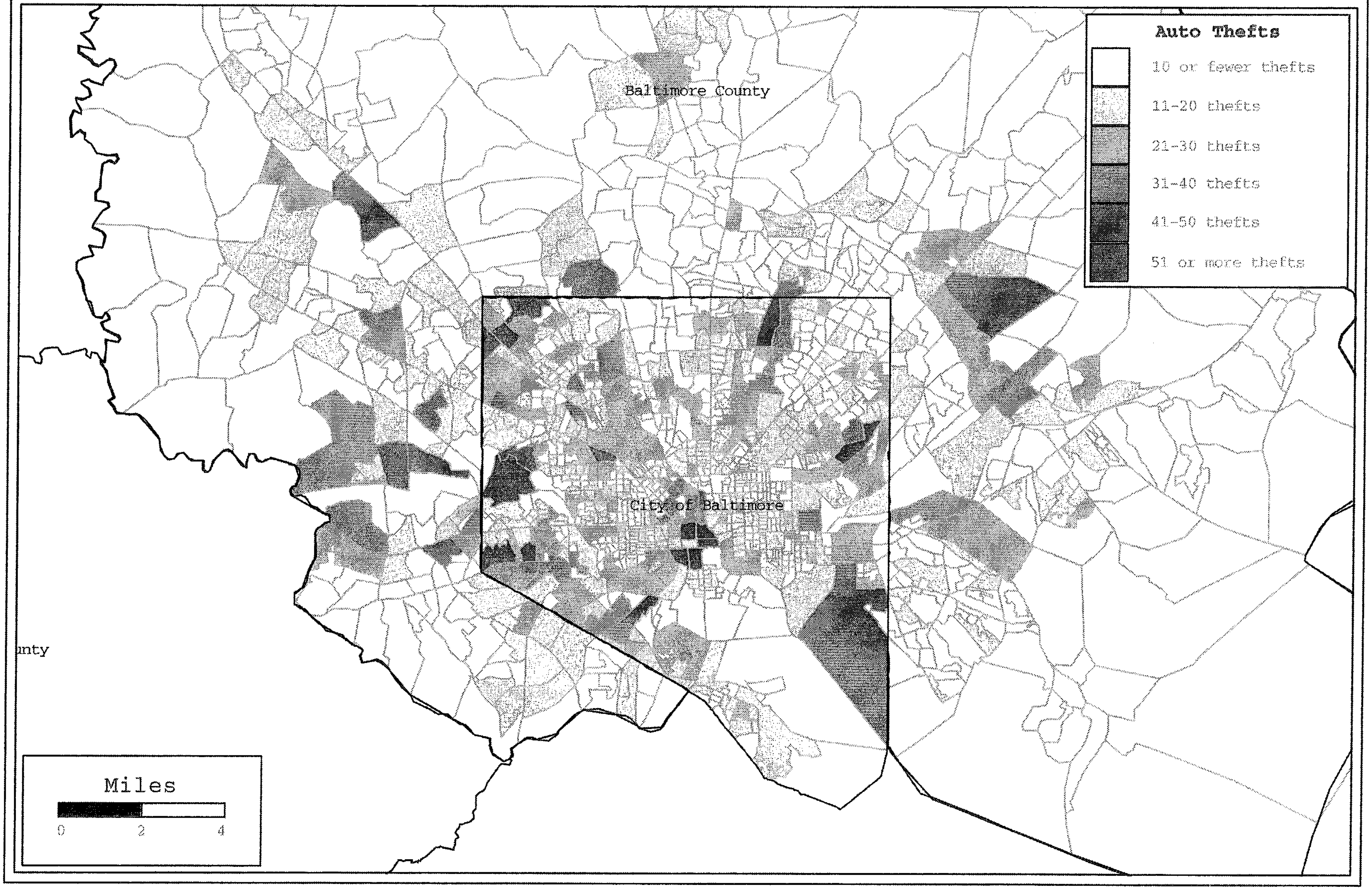
Example 3: Local Moran Statistics for Auto Thefts

Using data on 14,853 motor vehicle thefts for 1996 in both Baltimore County and Baltimore City, the number of incidents occurring in each of 1,349 census block groups was calculated with a GIS (Figure 7.11). As seen, the pattern shows a higher concentration towards the center of the metropolitan area, as would be expected, but that the pattern is not completely uniform. There are many block groups within the City of Baltimore with very low number of auto thefts and there are a number of block groups within the County with a very high number.

Figure 7.11:

1996 Motor Vehicle Thefts

Number of Auto Thefts Per Block Group



This document is a research report submitted to the U.S. Department of Justice. This report has not been published by the Department. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Using these data, *CrimeStat* calculated the Local Moran statistic with the variance box being checked and the small distance adjustment being used. The range of I_i values varied from -37.26 to +180.14 with a mean of 5.20. The pseudo-standardized Local Moran 'Z' varied from -12.71 to 50.12 with a mean of 1.61. Figure 7.12 maps the distribution. Because a negative I_i value indicates dissimilarity, these values have been drawn in red, compared to blue for a positive I_i value. As seen, in both the City of Baltimore and the County of Baltimore, there are block groups with large negative I_i values, indicating that they differ from their surrounding block groups. For example, in the central part of Baltimore City, there is a small area of about eight block groups with low numbers of auto thefts, compared to the surrounding block groups. These form a 'cold spot'. Consequently, they appear in dark tones in figure 7.12 indicating that they have high I_i values (i.e., negative autocorrelation). Similarly, there are several block groups on the western side of the County which have relatively high numbers of auto thefts compared to the surrounding block groups. They form a 'hot spot'. Consequently, they also appear in dark tones in figure 7.12 because this indicates negative spatial autocorrelation, having values that are dissimilar to the surrounding blocks.

Another use of Anselin's Local Moran statistic is to identify 'outliers', zones that are very different from their neighbors. In this case, zones with a high negative I value (e.g., with an I smaller than two standard deviations below the mean, -2) are indicative of outliers. They either have a high number of incidents whereas their neighbors have a low number or, the opposite, a low number of incidents amidst zones with a high number of incidents. Identifying the outliers can focus on zones which are unique (and which should be studied) or, in multivariate analysis, on zones which need to be statistically treated different in order to minimize a large modeling error (e.g., creating a dummy variable for the extreme outliers in a regression model).

In short, the Local Moran statistic can be a useful tool for identifying zones which are dissimilar from their neighborhood. It is the only statistic that is in *CrimeStat* that demonstrates dissimilarity. The other 'hot spot' tools will only identify areas with high concentrations. To use the Local Moran statistic, however, requires that the data be summarized into zones in order to produce the necessary intensity value. Given that most crime incident databases will list individual events without intensities, this will entail additional work by a law enforcement agency.

Some Thoughts on the Concept of 'Hot Spots'

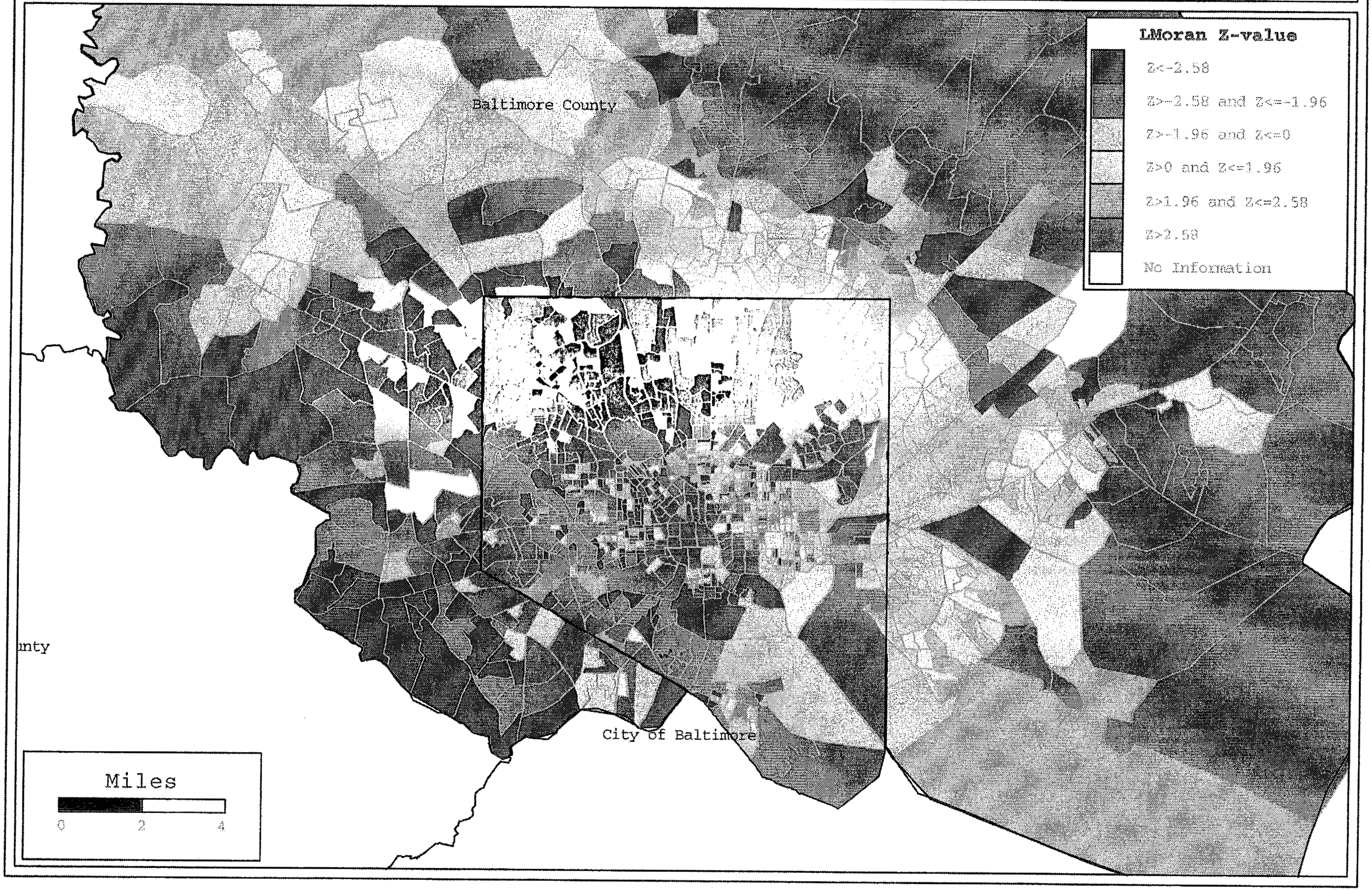
Advantages

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

Figure 7.12:

Local Spatial Autocorrelation of 1996 Vehicle Thefts

Local Moran Z-Value of Block Groups

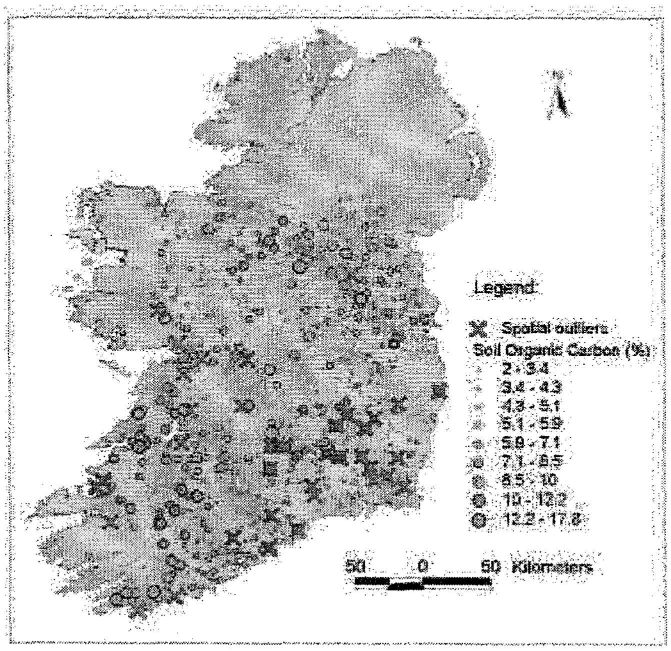


Using Local Moran's I to Detect Spatial Outliers in Soil Organic Carbon Concentrations in Ireland

Chaosheng Zhang[1] David McGrath[2]
Lecturer in GIS Research Officer
[1] Department of Geography, National University of Ireland, Galway, Ireland
[2] Teagasc, Johnstown Castle Research Centre, Wexford, Ireland

One objective in the study of soil organic carbon concentrations is to produce a reliable spatial distribution map. A geostatistical variogram analysis was applied to study the spatial structure of soils in Ireland for the purpose of carrying out a spatial interpolation with the Kriging method. The variogram looks at similarities in organic carbon concentrations as a function of distance. In the analysis, a relatively poor variogram was observed, and one of the main reasons was the existence of spatial outliers. Spatial outliers make the variogram curve erratic and hard to interpret, and impair the quality of the spatial distribution map.

CrimeStat was used to identify the spatial outliers. The parameter of the standardized Anselin's Local Moran's I (*z*) was used. When $z < -1.96$, the sample was defined as a spatial outlier. Out of 678 soil samples, a total of 39 samples were detected as spatial outliers, and excluded in the spatial structure calculation. As a consequence, the variogram curve was significantly improved. This improvement made the final spatial distribution map more reliable and trustable.



Spatial outliers are clearly different from the majority of samples nearby. Compared with the samples nearby, high value spatial outliers are found in the southeastern part, and low value spatial outliers are located in the western and northern parts of the country.

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

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

A third advantage is that these techniques are visual, particularly when used with a GIS. The two cluster analysis routines output ellipses that can be displayed in a GIS while the Local Moran technique can be adapted for thematic mapping (as Figure 7.12 demonstrates). Visual information can help crime analysts and officers to understand the distribution of crime in an areas, a necessary step in planning a successful intervention. We should never underestimate the importance of visualization in any analysis.

Limitations

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

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

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

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

Risk, which is defined as the number of incidents relative to the number of potential victims/targets, is only loosely correlated with the volume of incidents. Yet, 'hot spots' are usually defined by volume, rather than risk. The risk-adjusted hierarchical nearest neighbor clustering routine, discussed in chapter 6, is the only tool among these that identifies risk, rather than volume. It is clear that more tools will be needed to examine hot spot locations that are more at risk.

The final problem with the 'hot spot' concept is more theoretical. Namely, given a concentration of incidents, how do we explain it? To identify a concentration is one thing. To know how to intervene is another. It is imperative that the analyst discover some of the underlying causes that link the events together in a systematic way. Otherwise, all that is left is an empirical description without any concept of the underlying causes. For one thing, the concentration could be random or haphazard; it could have happened one time, but never again. For another, it could be due to the concentration of the population *at risk*, as discussed above. Finally, the concentration could be circumstantial and not be related to anything inherent about the location.

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

Endnotes for Chapter 7

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

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

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

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

5. The total number of ways for selecting K distinct combinations of *N* incidents, irrespective of order, is (Burt and Barber, 1996, 155):

$$\frac{N!}{K!\,(N-K)!}$$

6. The steps are as follows:

Global Selection of Initial Seed Locations

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

B. A separation distance is defined, which is

$$\text{Separation} = \quad t * 0.5 \text{ SQRT } \left[\frac{A}{N} \right]$$

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

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

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

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

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

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

Local Optimization of Seed Locations

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

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

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

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

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

7. The formulas are as follows as follows. The expected value of the Local Moran is:

$$E(I_i) = \frac{- \sum_{j=1}^{N} W_{ij}}{N - 1}$$

where W_{ij} is a distance weight for the interaction between observations i and j (either an adjacency index or a weight decreasing with distance). The variance of the Local Moran is defined in three steps:

A. First, define b_2.

$$b_2 = \Sigma \left\{ \frac{(X_i - \bar{X})^4}{N} \right\} \; / \; \left[\Sigma \left\{ \frac{(X_i - \bar{X})^2}{N} \right\}\right]^2$$

This is the fourth moment around the mean divided by the squared second moment around the mean.

B. Second, define $2w_{i(kh)}$:

$$2w_{i(kh)} = \Sigma \Sigma W_{ik} W_{ih} \qquad \text{where } k \neq i \text{ and } h \neq i$$

This term is twice the sum of the cross-products of all weights for i with themselves, using k and h to avoid the use of identical subscripts. Since each pair of observations, i and j, has its own specific weight, a cross-product of weights are two weights multiplied by each other (where $i \neq j$) and the sum of these cross-products is twice the sum of all possible interactions irrespective of order (i.e., $W_{ij} = W_{ji}$). Because the weight of an observation with itself is zero (i.e., $W_{ii} = 0$), all terms can be included in the summation.

C. Third, define the variance, standard deviation, and an approximate (pseudo) standardized score of I_i:

$$\text{Var}(I_i) = \frac{(\Sigma w_{ij}^2)*(n - b_2)}{(n-1)} + \frac{2w_{i(kh)}(2b_2 - n)}{(n-1)(n-2)} + \frac{(\Sigma w_{ij})^2}{(n-1)^2}$$

$$S(I_i) = \sqrt{[\text{ Var } (I_i)]}$$

$$Z(I_i) = [I_i - E(I_i)] / S(I_i)$$

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