CrimeStat III

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® III. 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
Figure 4.1: Spatial Distribution Screen

<table>
<thead>
<tr>
<th>Spatial Distribution</th>
<th>Distance Analysis I</th>
<th>Distance Analysis II</th>
<th>'Hot Spot' Analysis I</th>
<th>'Hot Spot' Analysis II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean center and standard distance (Mcsd)</td>
<td>Save result to...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard deviational ellipse (Sde)</td>
<td>Save result to...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median Center (MdnCntr)</td>
<td>Save result to...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Center of minimum distance (Mcmd)</td>
<td>Save result to...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Directional mean and variance (DMean)</td>
<td>Save result to...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convex Hull (CHull)</td>
<td>Save result to...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Spatial autocorrelation

<table>
<thead>
<tr>
<th>Moran's &quot;I&quot; statistic (Moran's &quot;I&quot;)</th>
<th>Geary's &quot;C&quot; statistic (Geary's &quot;C&quot;)</th>
<th>Moran Correlogram</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Simulation runs:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of distance intervals:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unit</td>
</tr>
</tbody>
</table>

Adjust for small distances: 
Simulation runs: 0
Number of distance intervals: 100
Unit: Miles
Save result to...
Figure 4.2: Selecting Centrographic Statistics

- Mean center and standard distance (Mcsd)
- Standard deviational ellipse (Sde)
- Median Center (MdnCnt)
- Center of minimum distance (Mcmd)
- Directional mean and variance (DMean)
- Convex Hull (CHull)

Spatial autocorrelation:
- Moran's "I" statistic (Moran's "I")
- Geary's "C" statistic (Geary's "C")
- Moran Correlogram

Simulation runs: 100
Unit: Miles
Number of distance intervals: 10

Save result to...
center of minimum distance (see below). However, the mean center can be thought of as a 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:

\[
X = \frac{\sum_{i=1}^{N} X_i}{N}, \quad Y = \frac{\sum_{i=1}^{N} Y_i}{N}
\]  

(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.

\[
X = \frac{\sum_{i=1}^{N} W_i X_i}{\sum_{i=1}^{N} W_i}, \quad Y = \frac{\sum_{i=1}^{N} W_i Y_i}{\sum_{i=1}^{N} W_i}
\]  

(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). 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>).

4.6
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
Figure 4.7: Mean Center and Standard Distance Deviation Output

Sample size: 14853
Measurement type: Direct
Start time: 10:45:43 AM, 09/06/2004
Unit: Degrees

Variable: X Y
Minimum: -76.870500 39.202700
Maximum: -76.355400 39.696700
Mean: -76.627194 39.318392
Standard Deviation: 0.071488 0.043419
Geometric Mean: -76.627161 39.318960
Harmonic Mean: -76.627127 39.318344

Average Density: 0.000006 points per sq. m
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:

\[
\text{Center of Minimum Distance} = C = \sum_{i=1}^{N} d_{ic} \text{ is a minimum}
\]  

(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 that 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. 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 #10...

Center of minimum distance:
-----------------------------
Sample size........: 14853
Start time..........: 10:47:10 AM, 09/06/2004
Mean X.............: -76.627194
Mean Y.............: 39.318892
Iterations..........: 10
Tolerance...........: 0.000001
DeltaX...............: 0.000000
DeltaY...............: 0.000001
Median Center X.....: -76.626354
Median Center Y.....: 39.312280
End time............: 10:47:10 AM, 09/06/2004

Finished
Figure 4.10: 1996 Metropolitan Baltimore Auto 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:

\[
S_X = \sqrt{\frac{\sum_{i=1}^{N-1}(X_i - \bar{X})^2}{N}} \quad (4.4)
\]

\[
S_Y = \sqrt{\frac{\sum_{i=1}^{N-1}(Y_i - \bar{Y})^2}{N}} \quad (4.5)
\]

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

\[
S_{XY} = \sqrt{\frac{\sum_{i=1}^{N-2}(d_{MC})^2}{N-2}} \quad (4.6)
\]
Figure 4.11: 1996 Metropolitan Baltimore Auto Thefts
Mean Center and Standard Deviations of X and Y Coordinates
where $d_{mc}$ 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).

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

$$Bivariate\;Distribution = \text{SQRT} \left[ \frac{\sigma_x^2 + \sigma_y^2}{2} \right] \quad (4.7)$$

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, $\theta$, where

$$\theta = \text{ARCTAN} \left( \frac{\sum(X_i - \bar{X})^2 - \sum(Y_i - \bar{Y})^2}{\sum(X_i - \bar{X})^2 + 4(\sum(X_i - \bar{X})(Y_i - \bar{Y}))^{1/2}} \right)$$

$$\sum^{(4.8)}$$

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

4.17
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.

\[
S_x = \sqrt{\frac{\sum (X_i - \bar{X}) \cos \theta - \sum (Y_i - \bar{Y}) \sin \theta}{N-2}}
\]

\[
S_y = \sqrt{\frac{\sum (X_i - \bar{X}) \sin \theta - \sum (Y_i - \bar{Y}) \cos \theta}{N-2}}
\]

(4.9)

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

\[
\text{Length}_x = 2S_x
\]

\[
\text{Length}_y = 2S_y
\]

(4.11)

(4.12)

4. The area of the ellipse is

\[
A = \pi S_x S_y
\]

(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:

\[
\text{Geometric Mean of } X = \text{GM}(X) = \prod_{i=1}^{N} (X_i)^{1/N}
\]

(4.14)

\[
\text{Geometric Mean of } Y = \text{GM}(Y) = \prod_{i=1}^{N} (Y_i)^{1/N}
\]

(4.15)
Figure 4.13: **Standard Deviational Ellipse Output**

Sample size: 14853
Measurement type: Direct
Start time: 10:48:31 AM, 09/06/2004

Clockwise angle of Y-axis rotation: 9.16656 degrees
Ratio of long to short axis: 1.2909

SD along new Y axis: 6767.27 m, 22202.32 ft, 4.20498 mi
SD along new X axis: 8735.91 m, 28661.13 ft, 5.42825 mi

Y axis length: 13534.53 m, 44404.63 ft, 8.40997 mi
X axis length: 17471.83 m, 57322.27 ft, 10.85649 mi

Area of ellipse: 185725473.40 sq m
Figure 4.14: 1996 Metropolitan Baltimore Auto Thefts

Mean Center and Standard Deviational Ellipse
where $\Pi$ 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.

\[
\ln[\text{GM}(X)] = \frac{1}{N} \left[ \ln(X_1) + \ln(X_2) + \ldots + \ln(X_N) \right] = \frac{1}{N} \sum \ln(X_i) \tag{4.16}
\]

\[
\ln[\text{GM}(Y)] = \frac{1}{N} \left[ \ln(Y_1) + \ln(Y_2) + \ldots + \ln(Y_N) \right] = \frac{1}{N} \sum \ln(Y_i) \tag{4.17}
\]

\[
\text{GM}(X) = e^{\ln(\text{GM}(X))} \tag{4.18}
\]

\[
\text{GM}(Y) = e^{\ln(\text{GM}(Y))} \tag{4.19}
\]

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

\[
\text{Harmonic mean of } X = \text{HM}(X) = \frac{N}{\sum (1/X_i)} \tag{4.20}
\]

\[
\text{Harmonic mean of } Y = \text{HM}(Y) = \frac{N}{\sum (1/Y_i)} \tag{4.21}
\]

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:

4.22
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
- **MdnCnt<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.

### 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 centriographic 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
Figure 4.17: *MapInfo* Output Options

- Mean center and standard distance (Mcsd)
- Standard deviational ellipse (Sde)

Save output options:

- 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 radiiuses 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
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

Features
- Standard Distance Deviation of Commercial Burglaries
- Standard Distance Deviation of Residential Burglaries
- County
- Commercial Burglaries
- Residential Burglaries

Miles
0 2 4

SDD Commercial Burglaries
SDD Residential Burglaries
Baltimore County
City of Baltimore
Figure 4.24:

Identifying Serial Burglars

Standard Deviational Ellipse of Incidents for Two Serial Offenders

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° due North) and all locations are defined by angular deviations from this reference vector. By convention, angles are defined as deviations from 0°, clockwise through 360°. Note the measurement scale is a circle which returns back on itself (i.e. 0° is also 360°). 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°, 10°, 15°, 345°, and 350° from the reference vector (moving clockwise from due North), the statistical mean will produce an erroneous estimate of 144°. 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, \( \theta_i \), is calculated by

\[
\theta_i = \arctan\left(\frac{\text{Abs}(X_i - X_o)}{\text{Abs}(Y_i - Y_o)}\right)
\]

(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, \( \theta_i \), is in radians and can be converted to polar coordinate degrees using:

\[
\theta_i \text{ (degrees)} = \theta_i \text{ (radians)} \times \frac{180}{\pi}
\]

(4.23)

**Third Quadrant**

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

\[
\theta_i = \pi + \arctan\left(\frac{\text{Abs}(X_i - X_o)}{\text{Abs}(Y_i - Y_o)}\right)
\]

(4.24)

where the angle, \( \theta_i \), is again in radians. Since there are \( 2\pi \) radians in a circle, \( \pi \) radians is 180°. 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, \( \theta_i \), differently.
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.
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.
\[
\theta_i = 0.5\pi + \arctan \left( \frac{\text{Abs}(Y_i - Y_0)}{\text{Abs}(X_i - X_0)} \right)
\] (4.25)

\[
\theta_i = 1.5\pi + \arctan \left( \frac{\text{Abs}(Y_i - Y_0)}{\text{Abs}(X_i - X_0)} \right)
\] (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

\[
\text{Mean angle} = \bar{\theta} = \frac{\sum d_i \sin \theta_i}{\sum d_i \cos \theta_i}
\] (4.27)

where the summation of sines and cosines is over the total number of points, \(i\), defined by their angles, \(\theta_i\). Each angle, \(\theta_i\), 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 \(\sum \sin \theta_i > 0\) and \(\sum \cos \theta_i > 0\), then \(\bar{\theta}\) can be used directly as the mean angle
2. If \(\sum \sin \theta_i > 0\) and \(\sum \cos \theta_i < 0\), then the mean angle is \(\pi/2 + \bar{\theta}\).
3. If \(\sum \sin \theta_i < 0\) and \(\sum \cos \theta_i < 0\), then the mean angle is \(\pi + \bar{\theta}\).
4. If \(\sum \sin \theta_i < 0\) and \(\sum \cos \theta_i > 0\), 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).

\[
R = \text{SQRT} \left[ \left( \sum d_i \sin \theta_i \right)^2 + \left( \sum d_i \cos \theta_i \right)^2 \right]
\]  

(4.28)

where \( d_i \) is the length of vector, \( i \), with an angle (bearing) for the vector of \( \theta_i \). 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.

\[
\frac{\bar{R}}{N}
\]

(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

\[
\text{Circular variance} = \frac{1}{D} \frac{\bar{R}}{N} \left\{ D - \frac{\bar{R}}{N} \right\} = \frac{(D - \bar{R})/D}{1 - \frac{\bar{R}}{D}} = 1 - \frac{\bar{R}}{D}
\]

(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.

**Convex Hull**

The convex hull is a boundary drawn around the distribution of points. It is a relatively simple concept, at least on the surface. Intuitively, it represents a polygon that circumscribes all the points in the distribution such that no point lies outside of the polygon.

The complexity comes because there are different ways to define a convex hull. The most basic algorithm is the *Graham scan* (Graham, 1972). Starting with one point known to be on the convex hull, typically the point with the lowest X coordinate, the algorithm sorts the remaining points in angular order around this in a counterclockwise manner. If the angle formed by the next point and the last edge is less than 180 degrees, then that point is added to the hull. If the angle is greater than 180 degrees, then the chain of nodes starting from the last edge must be deleted. The routine proceeds until the hull closes back on itself (de Berg, van Kreveld, Overmans, and Schwarzkopf, 2000).

Many alternative algorithms have been proposed. Among these are the ‘gift wrap’ (Chand and Kapur, 1970; Skiena, 1997), the Quick Hull, the “Divide and conquer” (Preparata and Hong, 1977), and the incremental (Kallay, 1984) algorithms. Even more complexity has been introduced by the mathematics of fractals where an almost infinite number of borders could be defined (Lam and De Cola, 1993). In most implementations, though, a simplified algorithm is used to produce the convex hull.

*CrimeStat* implements a ‘gift wrap’ algorithm. Starting with the point with the lowest Y coordinate, A, it searches for another point, B, such that all other points lie to the left of the line AB. It then finds another point, C, such that all remaining points lie to the left of the line BC. It continues in this way until it reaches the original point A again. It is like ‘wrapping a gift’ around the outside of the points.

The routine outputs three statistics:

1. The sample size;
2. The number of points in the convex hull
3. The X and Y coordinates for each of the points in the convex hull

The convex hull can be saved as an ArcView 'shp', MapInfo 'mif', or Atlas*GIS 'bna' file with a 'Chull' prefix.

Figure 4.26 shows the convex hull of Baltimore County robberies for 1996. As seen, the hull occupies a relatively smaller part of Baltimore County. Figure 4.27, on the other
Figure 4.26:
Convex Hull of Baltimore County Robberies: 1996
Figure 4.27:
Convex Hull of Baltimore County Burglaries: 1996
hand, shows the convex hull of 1996 Baltimore County burglaries. As seen, the convex hull of the burglaries cover a much larger area than for the robberies.

**Uses and Limitations of a Convex Hull**

A convex hull can be useful for displaying the geographical extent of a distribution. Simple comparisons, such as in figures 4.26 and 4.27, can show whether one distribution has a greater extent than another. Further, as we shall see, a convex hull can be useful for describing the geographical spread of a crime hot spot, essentially indicating where the crimes are distributed.

On the other hand, a convex hull is vulnerable to extreme values. If one incident is isolated, the hull will of necessity be large. The mean center, too, is influenced by extreme values but not to the same extent since it averages the location of all points. The convex hull, on the other hand, is defined by the most extreme points. A comparison of different crime types or the same crime type for different years using the convex hull may only show the variability of the extreme values, rather than any central property of the distribution. Therefore, caution must be used in interpreting the meaning of a hull.

**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 that 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 that are included in CrimeStat.

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 global indices - Moran’s I statistic and Geary’s C statistic,
and an application of Moran’s I to different distance intervals. Moran and Geary are global
in that they represent a summary value for all the data points. They are also 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, 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
\[ I = \frac{N \sum_i \sum_j W_{ij} (X_i - \bar{X})(X_j - \bar{X})}{(\sum_i \sum_j W_{ij}) \sum_i (X_i - \bar{X})^2} \]  \hspace{1cm} (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 \( i \neq j \)), \( \bar{X} \) 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, \( W_{ij} \), was 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 typically varies between -1.0 and +1.0. However, this is not absolute as an example later in the chapter will show. 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 that is high indicates more spatial autocorrelation than an “I” that 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.

\[ E(I) = \frac{1}{N-1} \]  \hspace{1cm} (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:
\[ W_{ij} = \frac{1}{d_{ij}} \]  

(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 that 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

\[
W_{ij} = \frac{\text{one mile}}{\text{one mile} + d_{ij}}
\]  

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 4.28), 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

\[
Z(I) = \frac{I - E(I)}{S_{E(I)}}
\]  

(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 \( E(I) \).
Figure 4.28: Selecting Spatial Autocorrelation Statistics
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. 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. CrimeStat outputs the Z-values and p-values for both the normality and randomization assumptions (figure 4.29).

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 zones and count the number of incidents per zone.

Figure 4.30 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 (see the 'Assign primary points to secondary points' routine in chapter 5). 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.

\[
\begin{align*}
Z(I) &= \frac{I - E(I)}{S_{E(I)}} = \frac{0.012464 - (-0.000076)}{0.000100} = 125.13 \ (p \leq .001)
\end{align*}
\]

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.
Figure 4.29: Moran's I Statistic Output

Moran's Spatial Autocorrelation Index for Point Data

Spatial Autocorrelation for Point Data:

Sample size .........................: 532
Measurement type ....................: Direct
Start time .........................: 10:53:50 AM, 09/06/2004

Moran's "I" .........................: 0.024742
Spatially random (expected) "I" ...: -0.001883
Standard deviation of "I" ..........: 0.002519
Normality significance (Z) .......: 10.567837
p-value (one tail) .................: 0.0001
p-value (two tail) ................: 0.0001
Randomization significance (Z) ...: 10.570922
p-value (one tail) .................: 0.0001
p-value (two tail) ...............: 0.0001

Finished

Close  Save to text file  Print  Print All
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 4.31). 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:

\[
Z(I) = \frac{I_{MV} - I_p}{S_E(I)} = \frac{0.012464 - 0.001659}{0.000100} = 108.05 \ (p < .001)
\]

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 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

\[
C = \frac{(N - 1) \sum_i \sum_j w_{ij} (X_i - X_j)^2}{2 \sum_i \sum_j w_{ij} \sum_i (X_i - \bar{X})^2} \quad (4.36)
\]

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
1996 Baltimore Region Motor Vehicle Thefts

Number of Vehicle Thefts Per Block Group

Figure 4.30: 1996 Baltimore Region Motor Vehicle Thefts

- Auto Thefts
  - 10 or fewer thefts
  - 11-20 thefts
  - 21-30 thefts
  - 31-40 thefts
  - 41-50 thefts
  - 51 or more thefts

Map showing the distribution of vehicle thefts across Baltimore County, with a legend indicating the number of thefts per block group.
Figure 4.31: 1990 Baltimore Population Density
Number of Persons Per Square Mile by Block

Howard County
Baltimore County
City of Baltimore
Harford County

Persons Per Sq Mi

- Less than 1,000
- 1,000-1,999
- 2,000-2,999
- 3,000-3,999
- 4,000 or more

Miles

0 2 4
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:

\[ W_{ij} = \frac{1}{d_{ij}} \]  \hspace{1cm} (4.33)

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

\[ W_{ij} = \frac{\text{one mile}}{\text{one mile} + d_{ij}} \]  \hspace{1cm} (4.34)

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.

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

\[ Z(C) = \frac{C - E(C)}{S_{E(C)}} \]  \hspace{1cm} (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). Figure 4.32 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 < .001) while that for population was 0.924811 with a Z-value of 122.61 (p < .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
Figure 4.32: Geary's C Statistic Output

Geary's "C":

Sample size ......................: 532
Measurement type ................: Direct
Start time ......................: 10:55:20 AM, 09/06/2004

Geary's "C" ......................: 1.023268
Spatially random (expected) "C": 1.000000
Standard deviation of "C".........: 0.015313
Normality significance (Z).......: 1.519516
p-value (one tail) ..............: 0.1
p-value (two tail) ..............: n.s.

End time ........................: 10:55:21 AM, 09/06/2004
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.

<table>
<thead>
<tr>
<th>Spatial Patterns of Residential Burglary: Moran’s I = 0.023, z=7.58</th>
</tr>
</thead>
</table>

### Calculated Moran’s I by Crime Types

<table>
<thead>
<tr>
<th>Crime Type</th>
<th>Original</th>
<th>Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Moran’s I</td>
<td>z</td>
</tr>
<tr>
<td>Felonious Offense</td>
<td>0.018</td>
<td>4.09 **</td>
</tr>
<tr>
<td>Violent Offense</td>
<td>0.030</td>
<td>6.27 **</td>
</tr>
<tr>
<td>Residential Burglary</td>
<td>0.055</td>
<td>11.21 **</td>
</tr>
<tr>
<td>Office Burglary</td>
<td>0.028</td>
<td>5.93 **</td>
</tr>
<tr>
<td>Smashing</td>
<td>0.031</td>
<td>6.48 **</td>
</tr>
<tr>
<td>Theft from Vendor</td>
<td>0.030</td>
<td>6.38 **</td>
</tr>
<tr>
<td>Theft from Cars</td>
<td>0.061</td>
<td>16.08 **</td>
</tr>
<tr>
<td>Vehicle Theft</td>
<td>0.047</td>
<td>9.65 **</td>
</tr>
<tr>
<td>Intellectual Offense</td>
<td>0.023</td>
<td>4.99 **</td>
</tr>
<tr>
<td>Sexual Assault</td>
<td>0.060</td>
<td>16.00 **</td>
</tr>
</tbody>
</table>

**p<.01:**; **p<.05**
Preliminary Statistical Tests for Hotspots:  
Examples from London, England

Spencer Chainey
Jill Dando Institute of Crime Science
University College
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.

<table>
<thead>
<tr>
<th>Crime type</th>
<th>Number of crime records</th>
<th>Standard distance</th>
<th>NN Index</th>
<th>z-score (test statistic)</th>
<th>Evidence of Clustering?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robbery</td>
<td>1132</td>
<td>3119.5 m</td>
<td>0.47</td>
<td>-34.2</td>
<td>Yes</td>
</tr>
<tr>
<td>Residential burglary</td>
<td>3104</td>
<td>3664.6 m</td>
<td>0.46</td>
<td>-57.5</td>
<td>Yes</td>
</tr>
<tr>
<td>Vehicle crime</td>
<td>9314</td>
<td>3706.2 m</td>
<td>0.26</td>
<td>-137.0</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Crime type</th>
<th>Moran’s I</th>
<th>Geary’s C</th>
</tr>
</thead>
<tbody>
<tr>
<td>All crime</td>
<td>0.0067</td>
<td>1.14</td>
</tr>
<tr>
<td>Robbery</td>
<td>0.0078</td>
<td>1.15</td>
</tr>
<tr>
<td>Residential burglary</td>
<td>0.014</td>
<td>0.99</td>
</tr>
<tr>
<td>Vehicle crime</td>
<td>0.0082</td>
<td>1.08</td>
</tr>
</tbody>
</table>

With the point statistics, all three crime types show evidence of clustering. Vehicle crime shows the more dispersed 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 (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.
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 4.30 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

Moran Correlogram

Moran’s I and Geary’s C indices are summary tests of global autocorrelation. That is, they summarize all the data and don’t distinguish between spatial autocorrelation for different subsets. In subsequent chapters, we will examine particular sub-sets of the data that are spatially autocorrelated, such as ‘hot spots’, ‘cold spots’ or space-time clusters.

One simple application of Moran’s I is a plot of the “I” by different distance intervals (or bins). Called a Moran Correlogram, the plot indicates how concentrated or distributed is the spatial autocorrelation (Cliff and Haggett, 1988; Bailey and Gatrell, 1995). Essentially, a series of concentric circles is overlaid over the points and the Moran’s I statistic is calculated for only those points falling within the circle. The radius of the circle changes from a small circle to a very large one. As the circle increases, the “I” calculation approaches the global value.

In CrimeStat, the user can specify how many distance intervals (i.e., circles) are to be calculated. The default is 10, but the user can choose any other integer value. The routine takes the maximum distance between points and divides it into the number of specified distance intervals, and then calculates the “I” value for those points falling within that radius.

Adjustment for Small Distances

If checked, small distances are adjusted so that the maximum weighting is 1 (see p. 49 above). This ensures that the “I” values for individual distances won’t become excessively large or excessively small for points that are close together. The default value is no adjustment.
Simulation of Confidence Intervals

A Monte Carlo simulation can be run to estimate approximate confidence intervals around the "I" value. Each simulation inputs random data and calculates the "I" value. The distribution of the random "I" values produce an approximate confidence interval for the actual (empirical) "I". To run the simulation, specify the number of simulations to be run (e.g., 100, 1000, 10000). The default is no simulations.

Output

The output includes:

1. The sample size
2. The maximum distance
3. The bin (interval) number
4. The midpoint of the distance bin
5. The "I" value for the distance bin (I[B])

and if a simulation is run:

6. The minimum "I" value for the distance bin
7. The maximum "I" value for the distance bin
8. The 0.5 percentile for the distance bin
9. The 2.5 percentile for the distance bin
10. The 97.5 percentile for the distance bin
11. The 99.5 percentile for the distance bin.

The two pairs of percentiles (2.5 and 97.5; 0.5 and 99.5) create an approximate 5% and 1% confidence interval. The minimum and maximum "I" values create an envelope. The tabular results can be printed, saved to a text file or saved as a '.dbf' file. For the latter, specify a file name in the "Save result to" in the dialogue box. The dbf file can be imported into a spreadsheet or graphics program to make a graph.

Graphing the 'I' values by Distance

A quick graph is produced that shows the "I" value on the Y-axis by the distance bin on the X-axis. Click on the "Graph" button. The graph displays the reduction in spatial autocorrelation with distance. The graph is useful for selecting the type of kernel in the Single- and Duel-kernel interpolation routines when the primary variable is weighted (see Interpolation).

Example: Moran Correlogram of 2000 Baltimore Population

I'll illustrate the Moran correlogram with the 2000 Baltimore regional population. Unlike figure 4.31 above, data by Traffic Analysis Zones (TAZ) were used. These are zones used typically for travel demand modeling (see chapter 12). The reason for using TAZ's,
however, is that data on both employment and population are available and it’s possible to compare them. The TAZ data were obtained from the Baltimore Metropolitan Council, the Metropolitan Planning Organization for the Baltimore Metropolitan area.

Figure 4.33 shows a map of the 2000 Baltimore population by TAZ’s. There is a higher concentration of population in the City of Baltimore, though some of the outlying TAZ’s also have a large population (primarily because they are large in area). Nevertheless, the distribution of population by TAZ’s falls off at a relatively slow rate from the center.

Figure 4.34 shows the Moran correlogram for the 2000 TAZ population and compares it to the maximum and minimum values from a Monte Carlo simulation of 100 runs. As seen, the “I” value at short distances of less than a mile is quite high, 0.78. As the distance between zones increase (i.e., the search circle radius gets larger), the “I” value drops off until about 8 miles whereupon it approaches the global “I” value. However, for all distance intervals, the empirical “I” value is higher than the maximum simulated “I” value with random data. In other words, it is highly unlikely that the “I” values obtained for each of the distance intervals was due to chance based on the distribution of random “I” values.

Now, let’s look at the distribution of employment (figure 4.35). In this case, employment is much more concentrated in a handful of TAZ’s. In most metropolitan areas, employment is usually more concentrated than population. A number of TAZ’s in downtown Baltimore have a high concentration of employment as does a corridor leading northward along Charles Street. In Baltimore County, there are stretches of higher employment but, again, they tend to be limited to a handful of TAZ’s. In other words, compared to the distribution of population, the distribution of employment is more clustered.

Figure 4.36 compares the Moran correlogram of employment with that of population. As seen, employment has a very high “I” value for short distances, much higher than for population. As mentioned above, the Moran I typically falls between -1.00 and +1.00, but this is not guaranteed. If the differences in values between zones is much greater than the average distance within zones, then the “I” value can exceed 1.0. In the case of figure 4.36, it approaches 3.0. Nevertheless, as the distance increases, the “I” value drops quickly and becomes lower than population for larger distance separations.

**Uses and Limitations of the Moran Correlogram**

In other words, the Moran correlogram provides information about the scale of spatial autocorrelation, whether it is diffuse over a larger area (e.g., as with the population example) or is more concentrated (e.g., as with the employment example). This can be useful for gauging the extent to which ‘hot spots’ are truly isolated concentrations of incidents or whether they are by-products of spatial clustering over a larger area. In chapter 6, we will examine a hierarchical clustering algorithm that examines a hierarchy
Figure 4.33:
Baltimore Region Population: 2000
By Traffic Analysis Zones
Figure 4.34:

"I" value

Distance interval

Population
Minimum random "I"
Maximum random "I"
Figure 4.35: Baltimore Region Employment: 2000
By Traffic Analysis Zones
Figure 4.36: Moran Correlogram of Baltimore Employment & Population: 2000
of clusters (e.g., first-order clusters which are within larger second-order clusters which, in turn, are within even larger third-order clusters). The Moran correlogram provides a quick snapshot of the extent of spatial autocorrelation as a function of scale.

Another use for the Moran correlogram is to estimate the type of kernel function that will be used for interpolation. In chapter 8, this methodology will be explained in detail. But, the key decision is to select a mathematical function that will interpolate data from point locations to grid cells. The shape of the Moran correlogram and the spread is a good indicator of the type of mathematical function to use.

On the other hand, like all global spatial autocorrelation statistics, the correlogram will not indicate where there is clustering or dispersion, only that it exists. For that, we’ll have to examine tools that are more focused on the location of concentrations of events (or the opposite, the location of a lack of events).

To explore this further, we will next examine properties of distances between points. Chapter 5 will examine tools for measuring second-order effects using the properties of the distances between incident locations.
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

      $$K_{it} = W_i e^{-d_{ict}}$$

      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

      $$X_{t+1} = \frac{\sum K_{it} X_i}{\sum K_{it}}$$ for $i=1...n$

      $$Y_{t+1} = \frac{\sum K_{it} Y_i}{\sum K_{it}}$$ 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

\[
\begin{align*}
\text{ABS}|X^{i+1} - X^i| & \leq 0.000001 \\
\text{ABS}|Y^{i+1} - Y^i| & \leq 0.000001
\end{align*}
\]

a. If either the \( X \) or \( Y \) coordinates have changed by greater than 0.000001 between iterations, substitute \( X^{i+1} \) for \( X^i \) and \( Y^{i+1} \) for \( Y^i \) and repeat steps B through D.

b. If both the change in \( X \) and the change in \( Y \) is less than or equal to 0.000001, then the estimated \( X_i \) and \( Y_i \) 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

\[
S_{XY} = \sqrt{\frac{\sum w_i (d_{MC})^2}{(\sum w_i)^{-2}}}
\]

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 are 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):

\[
\text{SDD} = \sqrt{\frac{\sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2}{N-2}}
\]

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

\[
\text{SDD} = \sqrt{\frac{2 \sum (X_i - \bar{X})^2}{N-2}}
\]

Similarly, the formula for the transformed axes are (4.9, 4.10):
\[
S_x = \sqrt{\frac{\sum (X_i - \bar{X}) \cos \theta - \sum (Y_i - \bar{Y}) \sin \theta)^2}{N-2}}
\]

\[
S_y = \sqrt{\frac{\sum (X_i - \bar{X}) \sin \theta - \sum (Y_i - \bar{Y}) \cos \theta)^2}{N-2}}
\]

However, if \( S_x = S_y \), then \( \theta = 0 \), \( \cos \theta = 1 \), \( \sin \theta = 0 \) and, therefore,

\[
S_x = S_y = \sqrt{\frac{\sum (X_i - \bar{X})}{N-2}}
\]

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 Table Import <Mapinfo interchange file>. With Atlas*GIS, the command is File Open <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):

\[
S_{E(I)} = \sqrt{\frac{N^2 \sum w_i^2 + 3(\sum w_i)^2 - N \sum (\sum w_i)^2}{(N^2 -1)(\sum w_i)^2}}
\]

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

\[
S_{E(I)} = \sqrt{\frac{N(N^2+3N)\sum w_i^2+3(\sum w_i)^2-N\sum (\sum w_i)^2-k((N^2-N)\sum w_i^2+6(\sum w_i)^2)-2N(\sum w_i)^2}{(N-1)(N-2)(N-3)(\sum w_i)^2}}
\]
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):

\[
S_{E(I)} = \sqrt{\frac{2 \sum_{ij} w_{ij}^2 + \sum_{ij} (\sum_{j} w_{ij})^2 (N-1) - 4(\sum_{ij} w_{ij})^2}{2(N+1) (\sum_{ij} w_{ij})^2}}
\]

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).