Chapter 6:

Distance Analysis I and II

Ned Levine
Ned Levine & Associates
Houston, TX

Table of Contents

Table of Contents (continued)

Chapter 6:
Distance Analysis I and II

In this chapter, the 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 the points are centered, how far they spread out, and whether there is any orientation to the dispersion. *Second-order* (or *local*) properties, on the other hand, refer to sub-regional or 'neighborhood' patterns within the overall distribution. If there are distinct 'hot spots' where many crime incidents cluster together, their distribution is spatially related to something unique in the sub-region or neighborhood, and less to the global distribution Second-order characteristics indicate how particular environments concentrate crime incidents.

There are two distance analysis pages. In Distance analysis I, various second-order statistics are provided, including:

1. NN
2. Linear NN
3. Ripley
4. Assign primary points to secondary points

In Distance analysis II, there are four routines for calculating and outputting distance matrices. This chapter will discuss both sets of routines.

Distance Analysis I

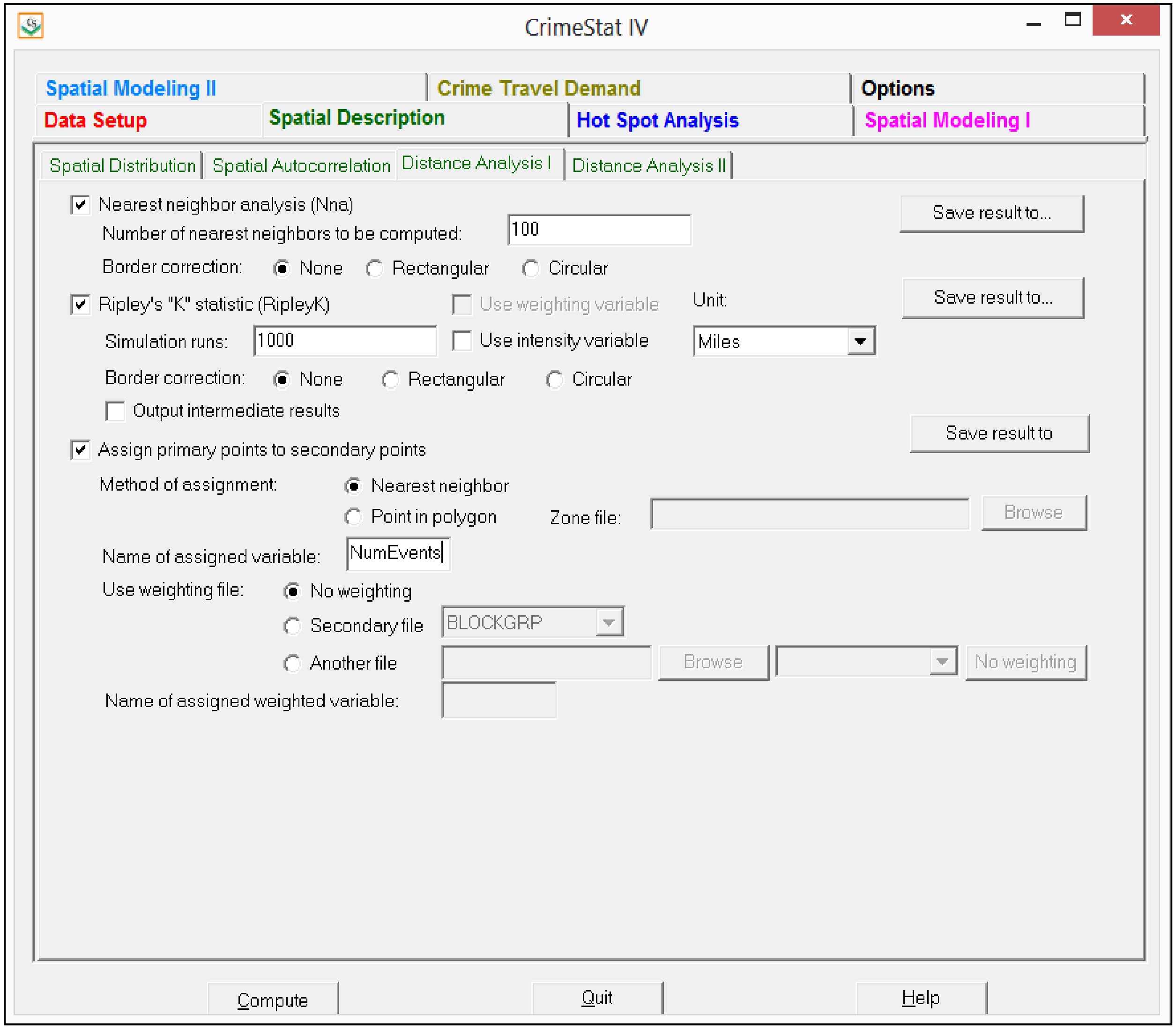
Figure 6.1 shows the Distance analysis I screen and the distance statistics on that page that are calculated by *CrimeStat*.

Nearest Neighbor Index

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

Figure 6.1:

Distance Analysis I 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 Unit:

Simulation runs: 1000 □ Use intensity variable Miles

Border correction: ● None ○ Rectangular ○ Circular

□ Output intermediate results

☑ Assign primary points to secondary points

Method of assignment: ● Nearest neighbor

○ Point in polygon Zone file:

Name of assigned variable: NumEvents

Use weighting file: ● No weighting

○ Secondary file BLOCKGRP

○ Another file

Name of assigned weighted variable:

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,i, the distance to every other point, j, is calculated and minimum selected (the nearest neighbor). The nearest neighbors are then averaged over all points:

$$d_{NN} = \sum_{i=1}^{N} \sum_{i \neq j=1}^{N-1} \frac{Min(d_{ij})}{N} \tag{6.1}$$

where 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.

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:

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

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

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

$$d_{dispersed} = \frac{\sqrt{2}}{3^{1/4}\sqrt{\frac{N}{A}}}$$

where N is the number of points and A is the area. 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.

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

$$NNI = \frac{d_{NN}}{d_{NN(ran)}} \tag{6.3}$$

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:

$$Z = \frac{d_{NN} - d_{NN(ran)}}{SE_{d(ran)}} \tag{6.4}$$

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

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

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 6.4 and 6.5 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 11 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)
11. The p-values associated with a one tail and two tail significance 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 Baltimore County Street Robberies

In 1996, there were 1,181 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 6.1 with the NNA routine. The mean nearest neighbor distance was 0.116 miles while the mean nearest neighbor distance under randomness was 0.358. The nearest neighbor index (the ratio of the actual to the random nearest neighbor distance) is 0.3236. The Z-value of -44.4672 is highly significant. In other words, the distribution of the nearest neighbors of street robberies in Baltimore County is significantly smaller than what would be expected randomness.

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.

Table 6.1:
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
p-value (one tail)	≤.0001
p-value (two tail)	≤.0001

Example 2: The Nearest Neighbor Index for Baltimore County 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 6.2).

The distribution of residential burglaries is also highly significant. Now, suppose we want to compare the distribution of street robberies (table 6.1) with that of residential burglaries

[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

(table 6.2). 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.

Table 6.2:
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
p-value (one tail)	≤.0001
p-value (two tail)	≤.0001

However, comparing the relative nearest neighbor indices can be meaningful,

$$Relative\ NN\ Comparison = \frac{NNI_A}{NNI_B} \qquad (6.6)$$

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:

$$\frac{NNI_A}{NNI_B} = \frac{NNI_{robberies}}{NNI_{burglaries}} = \frac{0.3236}{0.4256} = 0.7603 \qquad (6.7)$$

In other words, the distribution of street robberies relative to an expected random distribution appears to be more concentrated than that of burglaries. There is not a simple significance test of this comparison since the standard error of the joint distributions is not known.[3] But the relatively greater concentration of robberies suggests that they are more likely to have ‘hot spots’.

[3] It could be tested with a Monte Carlo simulation. Two separate random samples of 1181 ‘robberies’ and 6051 ‘burglaries’ would be drawn. The nearest neighbor distance for each sample would be calculated and the ratio of the two would be taken. The simulation would be repeated many times (e.g., 1000) to yield an approximate 95% credible interval. However, we have not implemented this simulation at this point.

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.

Use of Network Distance

In calculating the nearest neighbor index, network distance can be used to calculate the distance between points (see chapter 3). However, unless the data set is very small or you have a lot of patience, I highly recommend that you **do not** do this. Network calculations are very slow and will take a long time to complete for a large file.

K-Order Nearest Neighbor

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 calculating the second nearest neighbor, or the third nearest neighbor, or the 10th nearest neighbor? *CrimeStat* can construct K-order nearest neighbor indices. On the distance analysis page, the user specifies the number of nearest neighbor indices to be calculated.

The K-order nearest neighbor routine returns four columns:

5. The order, starting from 1
6. The mean nearest neighbor distance for each order (in meters)
7. The expected nearest neighbor distance for each order (in meters)
8. 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:

$$d_{K(ran)} = \frac{K(2K)!}{(2^K K!)^2 \sqrt{\frac{N}{A}}} \qquad (6.8)$$

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.[4]

Nevertheless, the K-order nearest neighbor distance and index can be useful for understanding the overall spatial distributions. Figure 6.2 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).

Graphing the K-order Nearest Neighbor

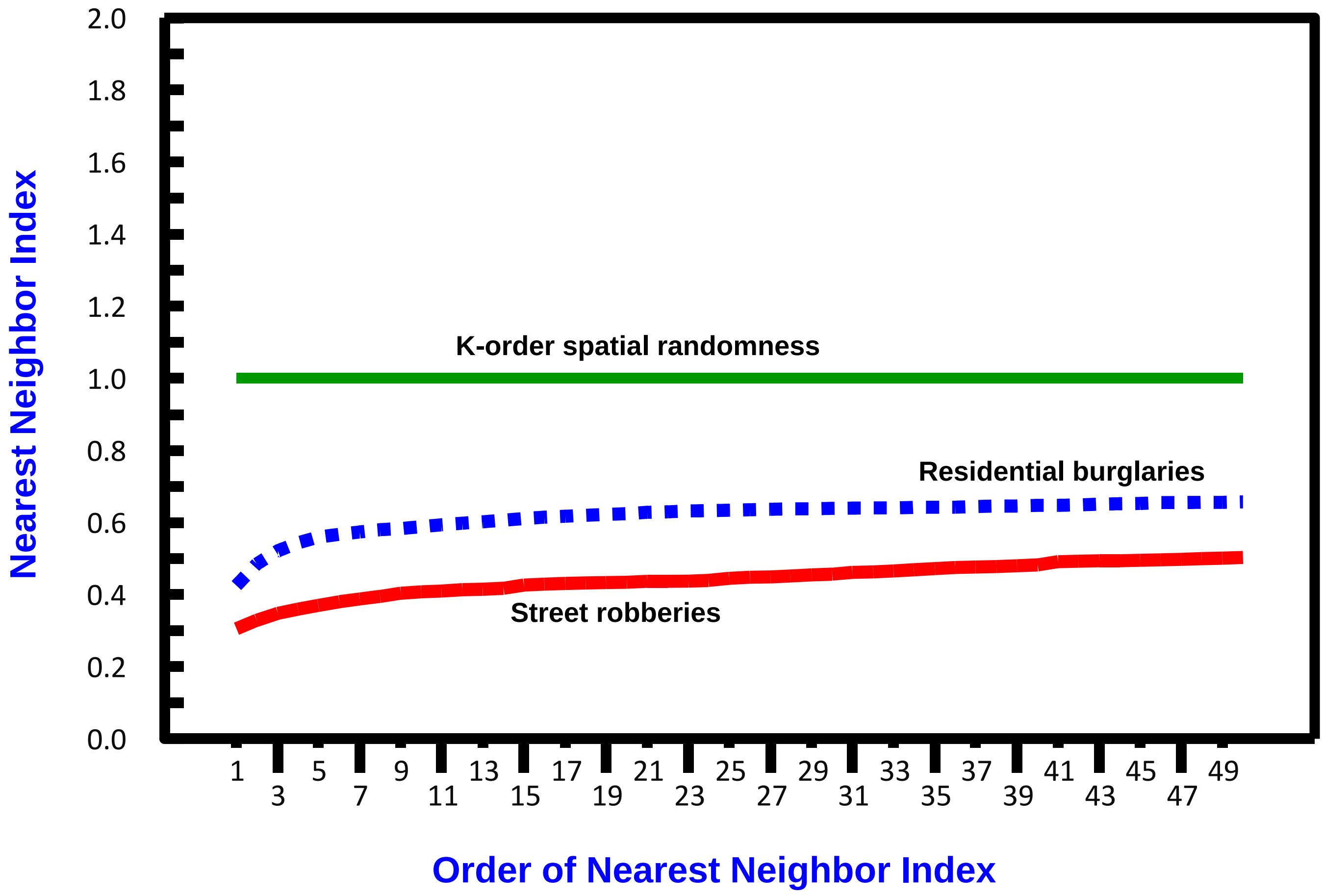
On the output page, there is a quick graph function that displays a curve similar to figure 6.2. This is useful for quickly examining the trends. However, a better graph is made by importing the 'dbf' file output into a spreadsheet or graphics program.

[4] There is not a hard-and-fast rule about how many K-order nearest neighbor distances should be calculated. Cressie (1991, p. 613) showed 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 showed that even after only 25 nearest neighbors, the uncorrected measure yields opposite conclusions about clustering from the corrected measures. So, as a rough rule, orders no greater than 2.5% of the cases should be calculated.

Figure 6.2:

K-Order Nearest Neighbor Indices

1996 Street Robberies and Residential Burglaries



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). In *CrimeStat*, the Kth-order nearest neighbor can be adjusted for boundary (edge) effects.

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. 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

$$R = \text{SQRT}\,[A / \pi\,] \tag{6.9}$$

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 6.9.

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

$$R_{iC} = R - R_i \tag{6.10}$$

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 6.3 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.

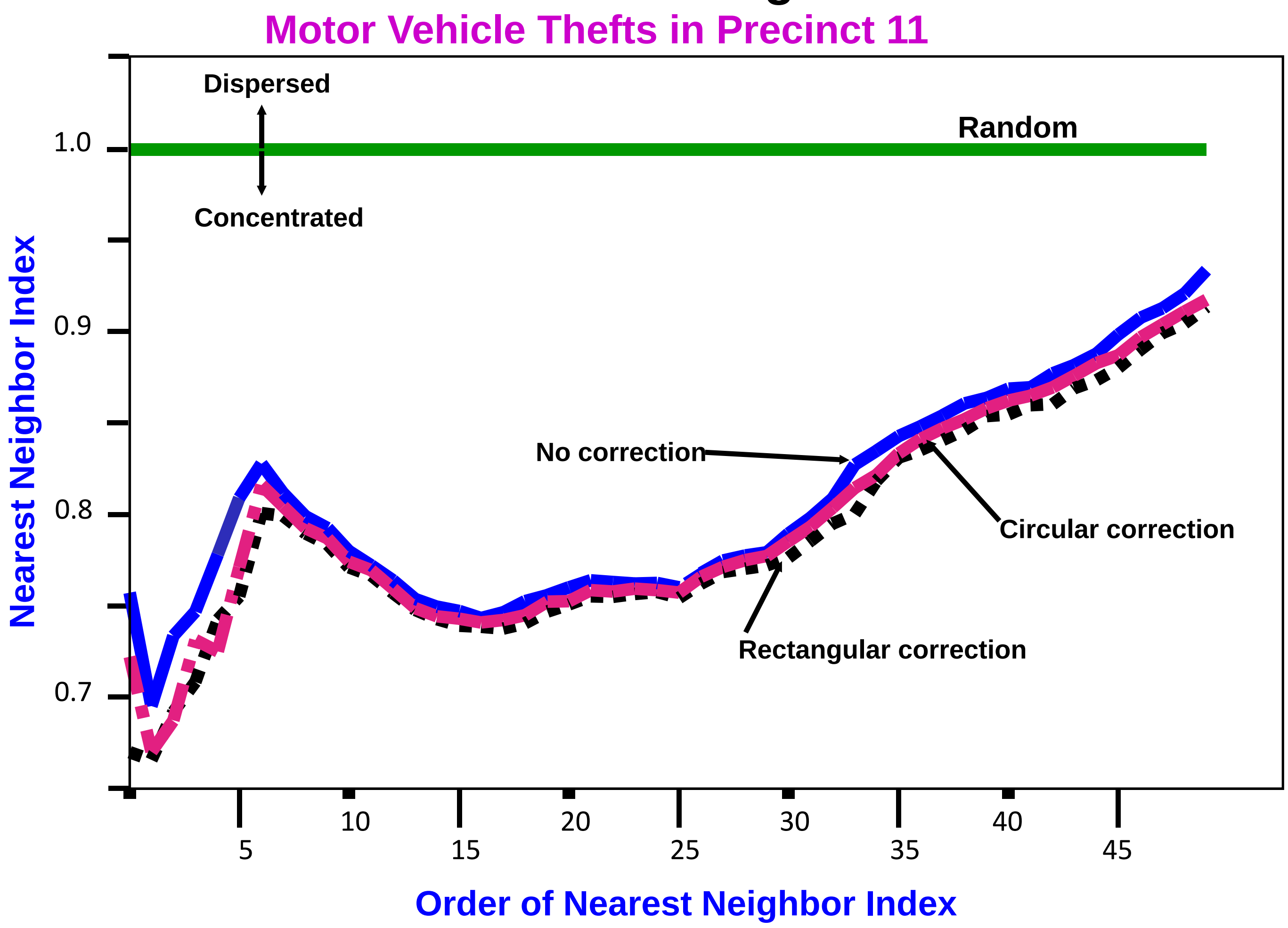
The user has to decide whether either of these corrections is meaningful or not. Depending on the shape of the study area, either correction may or may not be appropriate. If the study area is relatively rectangular, then the rectangular model may provide a good approximation. Similarly, if the study 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 or both of these corrections may produce more distortion than the raw nearest neighbor index. One has to use these corrections with judgment. 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.

Linear Nearest Neighbor Index

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.

Figure 6.3:

Correction of Nearest Neighbor Indices



The theory of linear nearest neighbors comes from Hammond and McCullagh (1978). The observed linear nearest neighbor distance, $L_{d(NN)}$, 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:

$$L_{d(ran)} = 0.5\frac{L}{N-1} \qquad (6.11)$$

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:

$$LNNI = \frac{L_{d(NN)}}{L_{d(ran)}} \qquad (6.12)$$

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:

$$s_{Ld(NN)} \cong \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N-1}[Min(d_{ij})-Ld(NN)]^2}{N-1}} \qquad (6.13)$$

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

$$SE_{L_{d(NN)}} \cong \frac{s_{Ld(NN)}}{\sqrt{N}} \qquad (6.14)$$

An approximate significance test can be obtained by:

$$t = \frac{L_{d(NN)}-L_{d(ran)}}{SE_{L_{d(NN)}}} \qquad (6.15)$$

where $L_{d(NN)}$ is the average linear nearest neighbor distance, $_{Ld(ran)}$ is the expected linear nearest neighbor distance (equation 6.11), and $SE_{L_{d(NN)}}$is the approximate standard error of the linear nearest neighbor distance (equation 6.14). Since the empirical standard deviation of the linear nearest neighbor is being used instead of a theoretical value, the test is a "t" 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, indirect or network. If the measurement type is direct or network, 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 10 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)
10. The p-values associated with a one tail and two tail significance test.

Example 3: Auto Thefts Along Two Baltimore County 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 6.4). 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.

Table 6.3 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.

Figure 6.4:

Vehicle Thefts in Baltimore County: 1996

Incident Distribution on State Highways 26 and 150

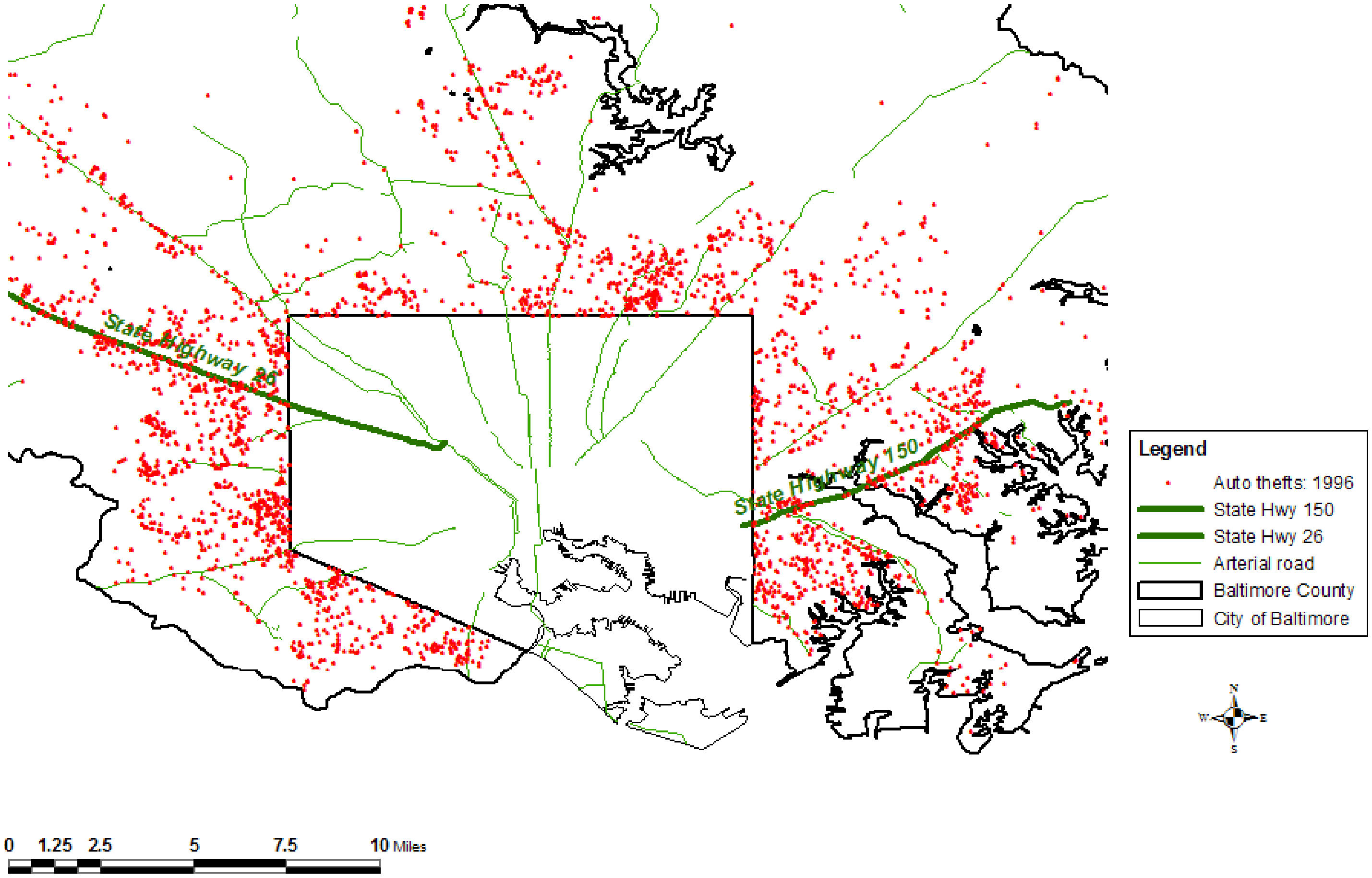


Table 6.3:

Comparison of 1996 Baltimore County Vehicle 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***		
			“Relative to random”			*“Relative to itself”*
Where Incidents Occurred	**Number of Incidents**	**Expected Number *if* Random**	**Ratio of Frequency**	**Average Linear Nearest Neighbor Distance**	**Average Random Linear Nearest Neighbor Distance**	**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 highways	607	272.8	2.2	0.13 mi	0.20	0.64 (p≤.001)
All roads	3,774	3,774.0	1.0	0.09 mi	0.44	0.21 (p≤.001)

The analysis is done proportional to the road network (i.e., all roads) and proportional to the same road. 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 6.11). Using this estimate, Table 6.3 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 highways, 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. Thus, relative to the entire network, 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 was a spatial pattern to the incidents, such as 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* calculated that for Highway 26, the average linear nearest neighbor distance was 0.05 miles which was 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 was 0.079 miles which, again, was almost identical to the average random linear nearest neighbor distance (0.084 miles); the nearest neighbor index was 0.94 and the t-value was -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.[5]

[5] 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 6.4. 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

On the other hand, the distribution of vehicle thefts along all major highways was not random in 1996 nor was the distribution of vehicle thefts along all roads. For those two high volume highways, however, unfortunately, the distribution of auto thefts was random and the clustering that is evident on all highways and all roads is apparently occurring at other locations. Not every test shows clustering and an analyst should be able to recognize a distribution that is no different than random.

Linear K-Order Nearest Neighbor

In *CrimeStat*, 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)
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 6.11 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.

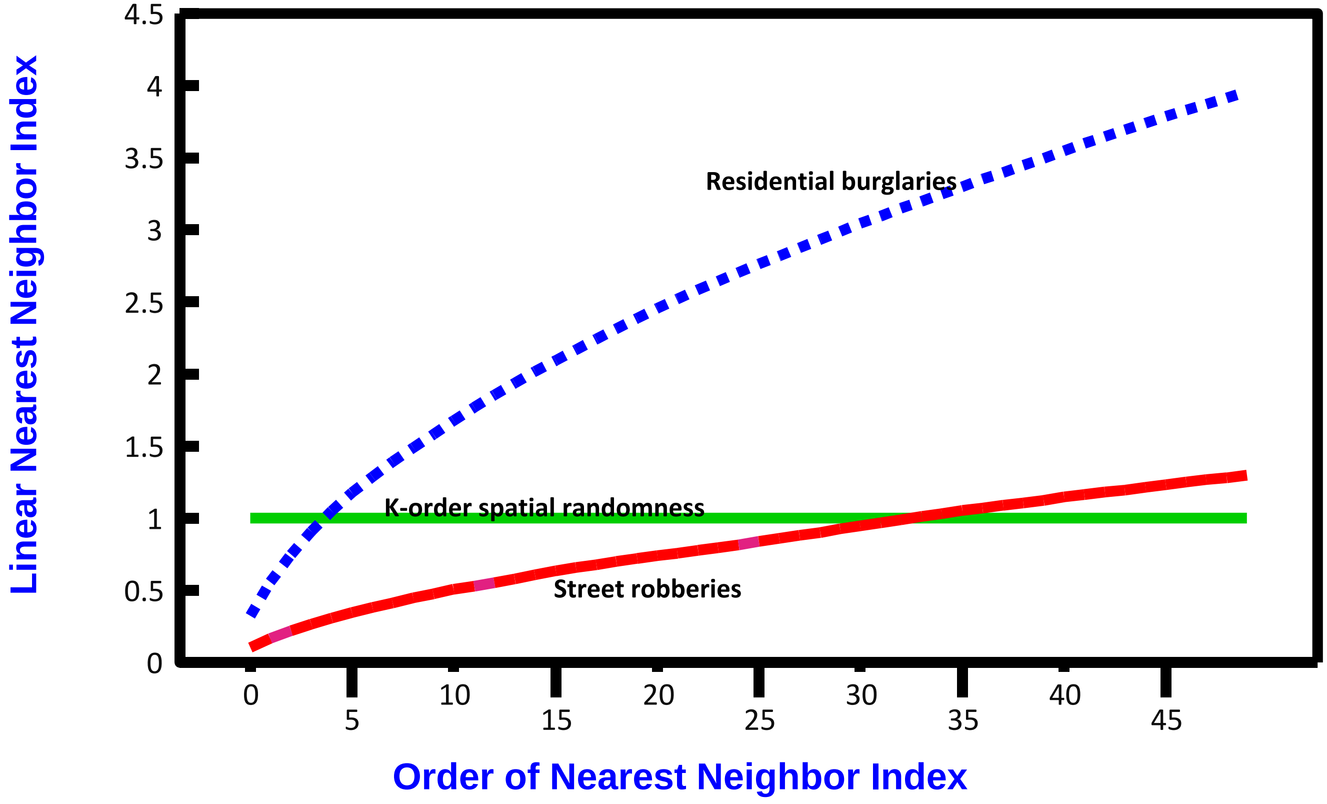
Nevertheless, like the areal k-order nearest neighbor index, the k-order linear nearest neighbor index can provide insights into the distribution of the points, even if the first-order is random. Figure 6.5 shows a graph of 50 linear nearest neighbors for 1996 residential burglaries and street robberies for Baltimore County. As with the areal k-order nearest neighbors (see figure 6.3) 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

decreased slightly from 0.05843 miles to 0.05061 miles and the linear nearest neighbor index became 0.8354 (t=-.91; n.s.). 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 very small sample, this could be a major source of error, but will be negligible for a lare sample. However, if a more precise measure is required, then the user should rotate the distribution so that the incidents have a horizontal or vertical orientation as closely as possible. An alternative is to calculate the regular nearest neighbor distance but use a network for distance calculations (see chapter 3).

Figure 6.5:

K-Order Linear Nearest Neighbor Indices

1996 Street Robberies and Residential Burglaries



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 area. 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 falls in this manner, it can provide useful information.

Graphing the Linear K-order Nearest Neighbor

On the output page, there is a quick graph function that displays a curve similar to figure 6.5 below. This is useful for quickly examining the trends.

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 some specified limit. 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, t_s, are drawn around each point, where s is the order of radii from the smallest to the largest, and the number of other points that are found within the circles are counted and then summed over all points (allowing for duplication), then the expected number of points under *complete spatial randomness* (csr) within that radius are:

$$E_{Id_i} = \frac{N}{A} K(t_s) = \frac{\pi t_s^2}{A} N \tag{6.16}$$

where N is the sample size, A is the total study area, and K(t_s) is the area of a circle defined by radius, t_s. For example, if the cumulative 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 one or more circles. Notice that individual points can be counted in multiple circles but the total number of points counted (excluding duplicates) is proportional to the cumulative area of the circle relative to the total area.

On the other hand, if the total number of points found within the circles for a particular radius placed over each point, in turn, is greater than that found in equation 6.16, 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 total number of points found within the circles for a particular radius placed over each point is, in turn, less than that found in equation 6.16, then 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 total number 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:

$$K(t_s) = \frac{A}{N^2} \sum_{i=1}^{N} \sum_{i \neq j}^{N-1} I(t_{ij}) \tag{6.17}$$

where I(t_{ij}) is the number of other points, j, found within distance, t_s, summed over all points, i. That is, a circle of radius, t_s, is placed over each point, i. Then, the number of other points, j, within the circle is 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. Note, the count does *not* include itself, only other points.

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 100 intervals by which the statistic can be counted.

One can graph K(t_s) against the distance, t_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, Vega, Cardoso, & Shelly, 1998). Consequently, K(t_s) is transformed into a square root function, L(t_s), to make it more linear. L(t_s) is defined as:

$$L(t_s) = \sqrt{\frac{K(t_s)}{\pi}} - t_s \qquad (6.19)$$

That is, K(t_s) is divided by π and then the square root is taken. Then the distance interval (the particular radius), t_s, is subtracted from this.[6] In practice, only the L statistic is used even though the name of the statistic, *K*, is based on the K derivation.

Because the L(t_s) is a measure of second-order clustering, it is usually analyzed for only a short distance. In *CrimeStat*, the distance is set at one-third the side of a square defined by the area,$\frac{\sqrt{A}}{3}$, and 100 intervals (radii) are used. Figure 6.6 shows a graph of L(t) against distance for 1996 robberies in Baltimore County. As can be seen, L(t) increases up to a distance of about 3 miles whereupon it decreases again. A "pure" random distribution, known as *complete spatial randomness* (CSR), is shown as a horizontal line at L=0.

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(t_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.[7]

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

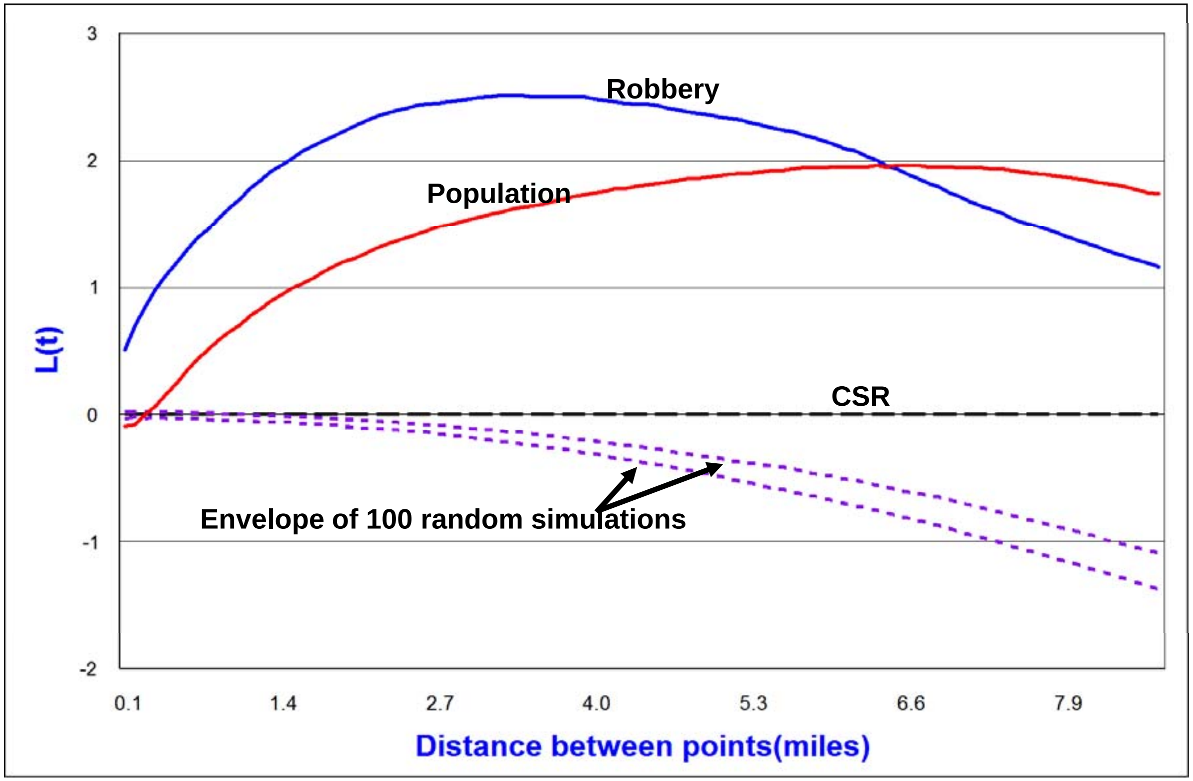
[7] Note that, since there is not a formal test of significance, the comparison with an envelope produced from a number of simulations provides only approximate confidence about whether the distribution differs from chance or not.

Figure 6.6:

"K" Statistic For 1996 Robberies

Compared to Random and 2000 Population Distributions

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



In figure 6.6, the L envelope of random data is much less concentrated than that for robberies, indicating that it is highly unlikely the concentration of robberies was due to chance.

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 in the box (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 0.5th, 2.5th, 5th, 95th, 97.5th and 99th percentile intervals; these latter statistics only make sense if many simulation runs are conducted (e.g. 1000). Approximate 95% credible intervals can be estimated by taking the 2.5th and 97.5th percentiles while approximate 99% credible intervals can be estimated by taking the 0.5th and 99.5th percentiles.[8]

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 *M* simulations of 1181 points randomly distributed over a rectangle, where *M* 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.

Comparison to Baseline Populations

For most social distributions, such as crime incidents, randomness is not a very

[8] With simulations, statisticians usually refer to their percentiles as *credible* intervals rather than *confidence intervals*, preferring to leave the latter term to formal statistical tests where the mathematical distribution of the standard error is known.

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.

Use of Intensity or Weight Variable

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 variable (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 the current version, if there is an intensity, however, no simulation can be run. The reason is that the sampling distribution of the intensity variable is unknown and it would be difficult to find a candidate distribution from which to draw samples. In a future version, we may allow permutation-type simulations whereby the original intensity values are maintained but they are randomly re-assigned to the existing X/Y coordinates. For now, though, there is no simulation when there is an intensity variable.

In Figure 6.6 above, there is an envelope produced from 100 random simulations as well as the L distribution from the 2000 population; the latter variable was obtained by taking the centroid of traffic analysis zones from the 2000 census and using population as the intensity variable. As can be seen, the amount of clustering for robberies is greater than both the random envelope as well as the distribution of population. The robbery function is higher than the population function up to about 6 miles. This indicates that robberies are more concentrated than what would be expected from the population distribution for a fairly large area.

In other words, robberies are more clustered than even what would be expected on the basis of the population distribution and this holds for distances up to about 6 miles, whereupon the distribution of robberies is indistinguishable from a random distribution. For larger distance separations, the L function has little utility since it is usually used to understand localized spatial autocorrelation (Bailey and Gattrell, 1995).

For comparison, figure 6.7 below shows the distribution of 1996 burglaries, again compared to a random envelope and the distribution of population. Burglaries are more clustered than population, but less so than for robberies; the L value is higher for robberies than for burglaries for near distances but becomes more dispersed at about 3 miles; it is still more concentrated than a random distribution, however, as seen by the random envelope.. 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 6.6 and 6.7 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 6.16). 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.[9]

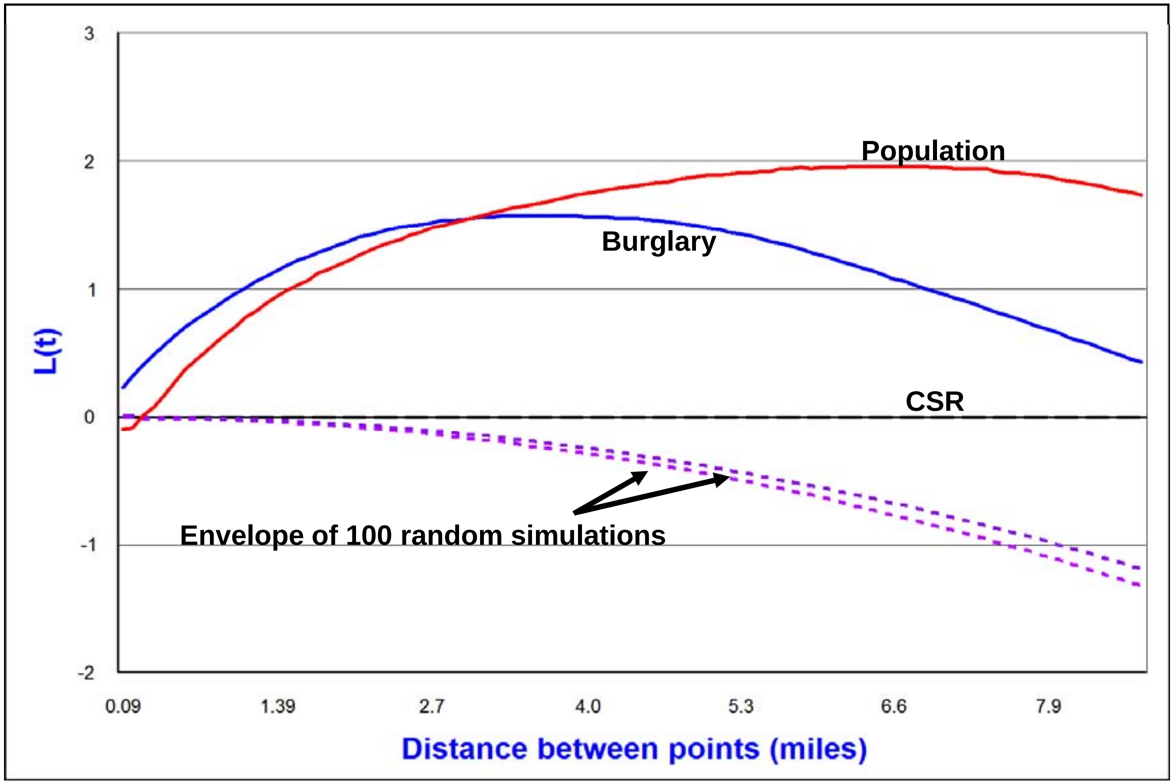
[9] The 'guard rail' concept, while frequently used, is a poor methodology because it involves ignoring data near the boundary of a study area. That is, points within the guard rail are only allowed to be selected by other points and not, in turn, be allowed to select others. This has the effect of throwing 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.

Figure 6.7:

"K" Statistic For 1996 Burglaries

Compared to Random and 2000 Population Distributions

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



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 6.17 is re-written as:

$$K(t_s) = \frac{A}{N^2}\sum_{i=1}^{N}\sum_{i\neq j}^{N-1} W_{ij}^{-1} I(t_{ij}) \tag{6.20}$$

where W_{ij}^{-1} is the inverse of the proportion of the circumference of a circle of radius, t_s, placed over each point that 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. An alternative weighting scheme can be found in Marcon and Puech (2003).

In *CrimeStat*, two possible corrections are conducted. One assumes that the study area is a rectangle while the other assumes that it is a circle.

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 bounding rectangle defined by the minimum and maximum X/Y values 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_{minRX} and d_{minRY}. Third, each of the minimum distances is compared to the search circle radius, R_j:

1. If neither the minimum distance in the X-direction, d_{minRX}, nor the minimum distance in the Y-direction, d_{minRY}, are less than the search circle radius, R_j, then the circle falls entirely within the rectangle and E = 1;

2. If either the minimum distance in the X-direction, d_{minRX}, or the minimum distance in the Y-direction, d_{minRY}, 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 that 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_{minRX}, and the minimum distance in the Y-direction, d_{minRY}, 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.

The formulas used to calculate the rectangular weights are:

Radius does not extend beyond the rectangle

$$W_{ij}^{-1} = k = 1 \qquad (6.21)$$

Radius extends beyond one edge of the rectangle (but not two)

$$W_{ij}^{-1} = k = \frac{2\pi}{2\pi - 2\cos\left\{-1\left[\frac{d(minR)}{R_i}\right]\right\}} \qquad (6.22)$$

Radius extends beyond two edges of the rectangle

$$W_{ij}^{-1} = k = \frac{2\pi}{1.5\pi - \cos\left\{-1\left[\frac{d(minR_x)}{R_i}\right]\right\} - \cos\left\{-1\frac{[d(minR_y)]}{R_i}\right\}} \qquad (6.23)$$

While intuitive, this weight, W_{ij}^{-1}, is prone to cause upward ‘drift’ in the K function, so a log transformation is used:

$$W'^{-1}_{ij} = \ln(W_{ij}^{-1}) + 1 \qquad (6.24)$$

This has the effect of tempering the drift somewhat.

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 has 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 6.9 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 6.9 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

$$R_{jC} = R - R_j \tag{6.25}$$

Third, the search circle radius, R_j, is compared to the nearest edge of the circle, R_{iC}, and the weight will vary from 1 (point and radius totally within the study area) to 2.3834 (point is located exactly on boundary of area circle). The formulas for the circular correction are:

$$\theta = arccos\frac{r^2+t_c-R^2}{2rt_c} \tag{6.26}$$

$$W_{ij}^{-1} = k = \pi / \theta \tag{6.27}$$

where r is the radius of the search circle, R is the radius of the circular study area, and t_c is the distance from the point to the center of the circular study area.

For either correction

During the calculation of Ripley's K, each point is multiplied by the weight and the K and L statistics are calculated as before. The simulation of random point distributions is treated in an analogous way. While intuitive, this weight, W_{ij}^{-1}, is prone to cause upward 'drift' in the K function, so a log transformation is used:

$$W'^{-1}_{ij} = \ln(W_{ij}^{-1}) + 1 \tag{6.28}$$

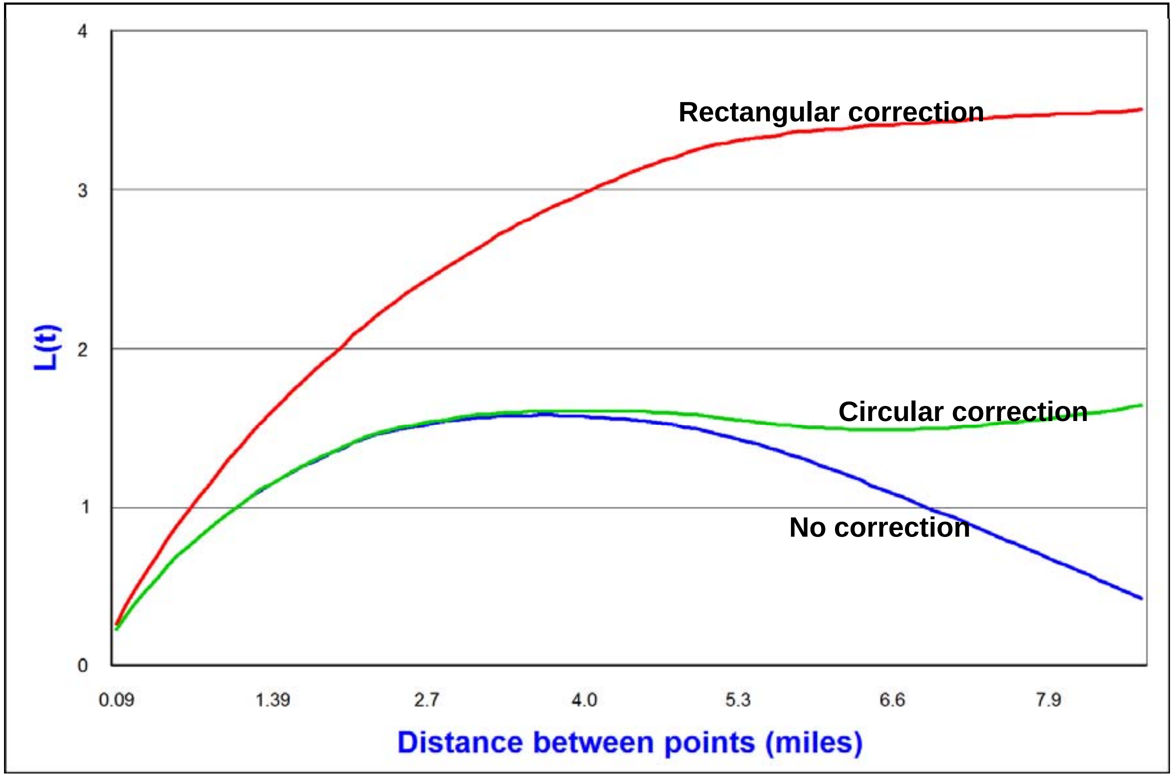
This has the effect of tempering the drift somewhat. Figure 6.8 below shows a Ripley's K distribution for 1996 Baltimore County burglaries, with and without edge corrections. As can be seen, the uncorrected L distribution decreases and falls below the theoretical random count (complete spatial randomness, L=0) after about 7 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.

Figure 6.8:

"K" Statistic For 1996 Burglaries

With Different Types of Corrections

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



Output Intermediate Results

There is a box labeled “Output intermediate results”. If checked, a separate dbf file will be output that lists the intermediate calculations. The file will be called “RipleyTempOutput.dbf”. There are five output fields:

1. The point number (POINT), starting at 0 (for the first point) and proceeding to N–1 (for the Nth point)
2. The search radius in meters (SEARCHRADI)
3. The count of the number of *other* points that are within the search radius (COUNT)
4. The weight assigned, calculated from equations 6.25 or 6.29 above (WEIGHT)
5. The count times the weight (CTIMESW)

This output can be useful for examining the counts for specific points or for trying out alternative weighting schemes.

Some Cautions in Using Ripley’s K

While Ripley’s K is a powerful tool for analyzing spatial autocorrelation (usually clustering, rather than dispersion), like any statistic it is prone to biases. Edge biases have been discussed above, but there are others. First, there is a sample size issue. The routine calculates 100 separate L(t) values, one for each distance bin. However, the precision of any one L(t) value is dependent on the sample size. With a small sample, there is insufficient data to estimate 100 independent values of L(t). While the Monte Carlo simulation partly can account for that bias, it has to be realized that the precision of the interpretation is suspect. For example, in comparing two similar distributions, say robberies and burglaries, unless the sample size is large differences for any one bin could easily be due to chance. One would need a very different type of procedure to estimate the ‘standard error’ of two functions with a small sample. But, I would suspect that there would be many bins for which they would be indistinguishable (shown as the two functions crisscrossing each other).

Users should be very cautious in drawing conclusions about differences in the L function with small samples. Even with sample sizes greater than 100, the imprecision of any one L(t) value is considerable. Until the sample sizes get into the hundreds, precision is an issue for specific L(t) values.

A second caution has to do with the scale of the interpretation. Data sets with strong *first-order* properties (i.e., a high degree of central concentration of incidents) will exert bias on

Ripley's K statistic. Thus, any data set that is correlated with human populations will most likely have a very strong 'central tendency'. Thus, there will be a high degree of concentration in the L values for even near distances. This was seen in the robbery and burglary data shown above. The K statistic was created to estimate *second-order* spatial autocorrelation, namely localized clustering. However, if the first-order effect is so dominant, then it is hard to disentangle it from a second-order effect. In other words, it is often not clear whether the clustering that is observed in Ripley K is due to primary, first-order clustering or actual localized, second-order clustering. That is why it is generally wise to use the K statistic for short distance ranges and not for larger distance separations. For larger distance separations, it is almost impossible to tell whether the effect is due to the large central concentration of the population or whether there are interactions between neighborhoods at a large scale.

There are different ways to handle to problem, none of which are perfect. For example, one can estimate a first-order concentration effect and then apply Ripley's K to the residuals. Alternatively, one can use a baseline population to calculate a rate and test for concentration only in the rates, not the volumes of incidents. In chapters 7 and 9, there will be a discussion of using a baseline population to control for first-order effects. But, whether this is done or not, the user should be aware of the interaction between first-order and second-order (or localized) effects.

The third caution has to do with the shape of the boundaries in interpreting the K statistic. This is particularly true when an edge correction is applied. Unless the study area was an actual rectangle, the correction may alter the interpretation compared to the uncorrected L. 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 or a circle. Thus, the differences in the shape comparisons may account for some variations.

The realism of the corrected function depends on the validity of the underlying assumptions. If it is likely that there are points outside the study area, then a weighting may produce a more realistic interpretation of the L function. On the other hand, if the density of the points outside the study area is lower (e.g., if the study area is a metropolitan area, then the area outside is more likely to be suburban or rural and of low population density), then the weighting will exaggerate the function relative to what it should be. In the extreme case, if the study area is an island (e.g., Honolulu), then there are no points outside the study area and no weighting is justified. Even when weighting would be justified, the actual boundary is probably not a rectangle or a square so that the geometric correction above may distort the L function, too. In short, some understanding of the basis for weighting is necessary to produce a reasonable L function.

Assign Primary Points to Secondary Points

This routine will assign each primary point to a secondary point and then will sum by the number of primary points assigned to each secondary point. The routine is useful for summarizing data. For example, if the primary file represents the number of robberies and the secondary file represents the centroids of census tracts, then the routine will assign all robberies to a census tract and will then sum the number of robberies in each census tract. The result is a count of the number of primary points for each secondary point (zone). Other examples might be to assign students to the nearest school or to assign patients to the nearest hospital. There are many uses for summarizing data by another data reference. In the Trip Generation module (under Crime Travel Demand - see Chapter 27), a model is developed for the number of crimes originating in each zone and a separate model for the number of crimes ending in each zone. The "Assign primary points to secondary points" routine is a good way to summarize the number of crimes by zones.

There are two methods for assigning the primary points to the secondary.

Nearest Neighbor Assignment

This routine assigns each primary point to the secondary point to which it is closest. It goes through all the primary points and sums the number assigned to each secondary point. Thus, the logical operation is 'nearest to'. If there are two or more secondary points that are exactly equal, the assignment goes to the first one on the list.

Point-in-polygon Assignment

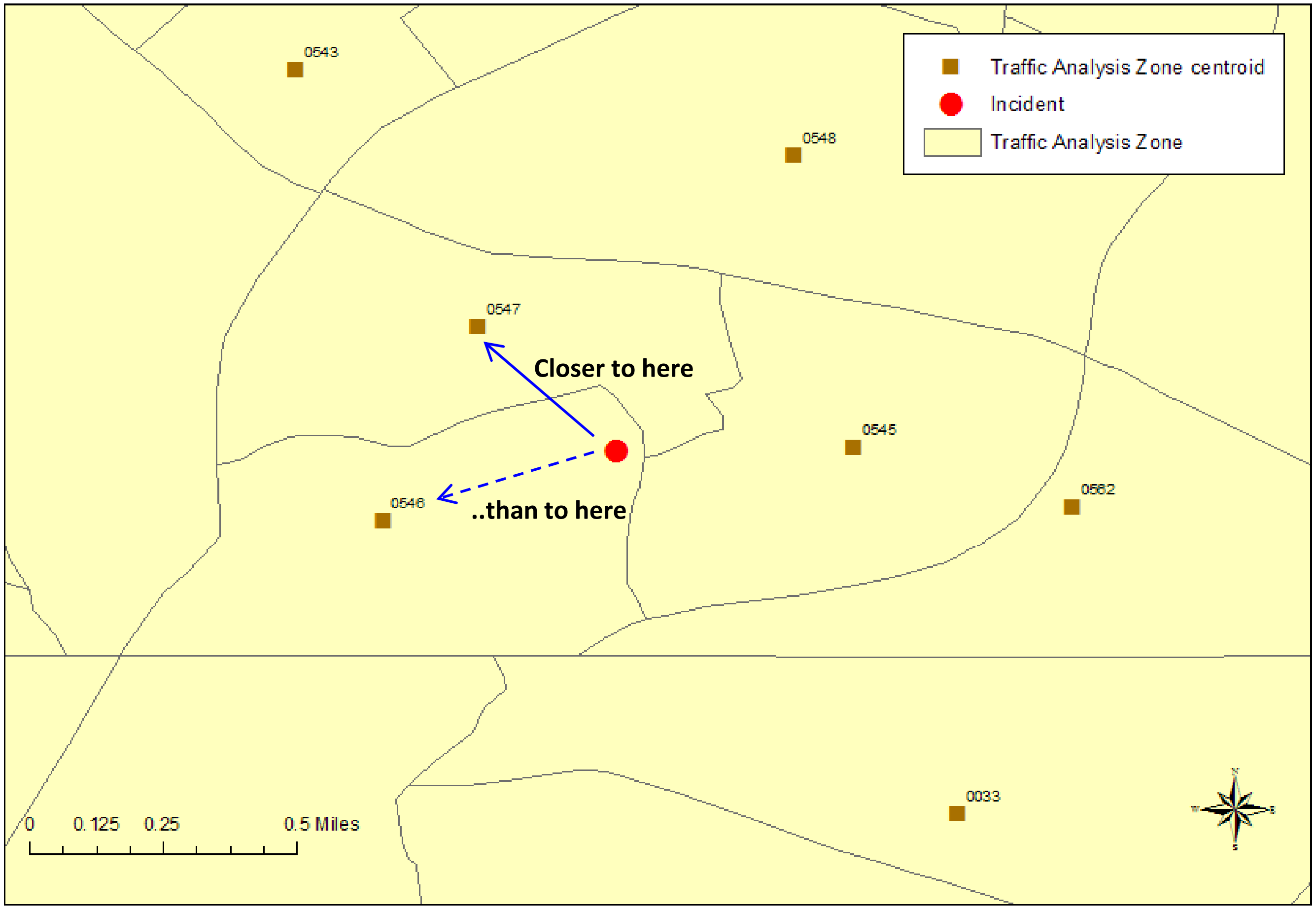
This routine assigns each primary point to the secondary point for which it falls within its polygon (zone). The point-in-polygon assignment reads a zonal boundary file (in ArcGIS 'shp' file format) and determines which zone each primary point falls within. In this case, the logical operation is 'belongs to'. A zone (polygon) shape file must be provided and the routine checks which secondary zone each primary point falls within.

Most GIS packages can do a point-in-polygon operation but few allow a nearest neighbor assignment. In general, the two are similar though there will be differences due to the irregular shape of zone boundaries. For example, figure 6.9 below shows an incident that is within Traffic Analysis Zone (TAZ) 0546, but is actually closer to the centroid of TAZ 0547. The characteristics associated with this incident are more likely to be associated with the characteristics of the second zone than the zone to which it belongs. The decision on which criteria to use in assigning the incident to a zone depends on how integral is the zone to which it

Figure 6.9:

Incident Assignment

Point in Relation to Traffic Analysis Zone Boundaries and Centroids



belongs. If the zones are bounded by major arterials, then travel behavior within the zone will be defined by those arterials; in this case, it would probably be prudent to use the point-in-polygon assignment. On the other hand, if the zone boundaries are not a fundamental separation, then the nearest neighbor assignment would probably produce a better fit to the incident since the characteristics of the closer zone are liable to hold for the incident. In short, the user must decide on which theoretical basis to assign points.

Zone file

If the point-in- polygon method is used, an *ArcGIS* zonal shape file must be defined under the routine. This is a polygon file that defines the zones to which the primary points are assigned. The zonal shape file correspond to secondary file (see Secondary file), but will be the full shape file as opposed to the 'dbf' portion of the file. For each point in the primary file, the routine identifies which polygon (zone) it belongs to and then sums the number of points per polygon.

On the other hand, if the nearest neighbor method is used, then only the secondary file need be defined.

Name of assigned variable

Specify the name of the summed variable. The default name is FREQ.

Use Weighting File

The primary file records can be weighted by another file. This would be useful for correcting the totals from the primary file. For example, if the primary file were robbery incidents from an arrest record, the sum of this variable (i.e. the total number of robberies) may produce a biased distribution over the secondary file zones because the primary file was not a random sample of all incidents (e.g., if it came from an arrest record where the distribution of robbery arrests is not the same as the distribution of all robbery incidents).

The secondary file or another file can be used to adjust the summed total. The weighting variable should have a field that identifies the ratio of the true to the measured count for each zone. A value of 1 indicates that the summed value for a zone is equal to the true value; hence no adjustment is needed. A value greater than 1 indicates that the summed value needs to be adjusted upward to equal the true value. A value less than 1 indicates that the summed value needs to be adjusted downward to equal the true value.

If another file is to be used for weighting, indicate whether it is the secondary file or, if another file, the name of the other file.

Name of assigned weighted variable

For a weighted sum, specify the name of the variable. The default will be ADJFREQ.

Save Result

For both routines, the output is a 'dbf' file. Define the file name. Note: be careful about using the same name as the secondary file as the saved file will have the new variable. It is best to give it a new name.

A new variable will be added to this file that gives the number of primary points in each secondary file zone and, if weighting is used, a secondary variable will be added which has the adjusted frequency.

Example: Assigning Robberies to Zones

To illustrate the routine, table 6.4 shows the results of summarizing 1,181 robberies that occurred in 1997 in 325 Baltimore County Traffic Analysis Zones. The two methods are compared. Only the first 30 assignments are shown. In general, they give similar results. However, there are differences due to the method. One is that the nearest neighbor method will assign points on the basis of proximity while the point-in-polygon method will not. In the case of the Baltimore County robberies, some of these were assigned to a City of Baltimore TAZ because those TAZ's were closer, rather than to a Baltimore County TAZ. Another is that if a zone is very irregular, points may be assigned to it under the point-in-polygon method which may be quite far away.

Thus, the user has to decide which method makes the most sense. If the purpose is to assign incidents to the zone which it is most likely to be related, for example, when developing a data set for zonal modeling (see Chapter 26), then the nearest neighbor method may produce a better representation. The incidents are then assigned to a zone which has characteristics that probably will be related to the factors causing the incidents in the first place. On the other hand, if the object is to assign incidents on the basis of membership (e.g., assigning crimes to police precincts), then the point-in-polygon method will be the most accurate.

Table 6.4:

Assigning Incidents to Zones

1997 Robberies (N=1181) and Traffic Analysis Zones (M=325)

TAZ	Point-in-Polygon	Nearest Neighbor
0401	0	0
0402	0	0
0403	1	1
0404	0	0
0405	0	0
0406	0	0
0407	0	0
0408	0	0
0409	0	0
0410	0	0
0411	0	0
0412	0	0
0413	0	0
0414	1	1
0415	0	0
0416	0	0
0417	0	0
0418	0	0
0419	0	0
0420	0	0
0421	0	0
0422	0	1
0423	0	0
0424	1	0
0425	3	0
0426	2	2
0427	3	2
0428	0	0
0429	5	5
0430	0	0

Distance Analysis II

The remaining distance analysis routines are on the Distance Analysis II page. Figure 6.10 shows the page.

Distance Matrices

CrimeStat has the capability for outputting distance matrices. There are four types of matrices that can be output.

1. 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).

2. 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 Primary File Points to Secondary File Points* matrix (Imatrix).

3. Third, if there is a reference file defined, the distance from each primary point to each grid cell can be computed. This is called the *From Primary File Points to Grid* matrix (PGMatrix).

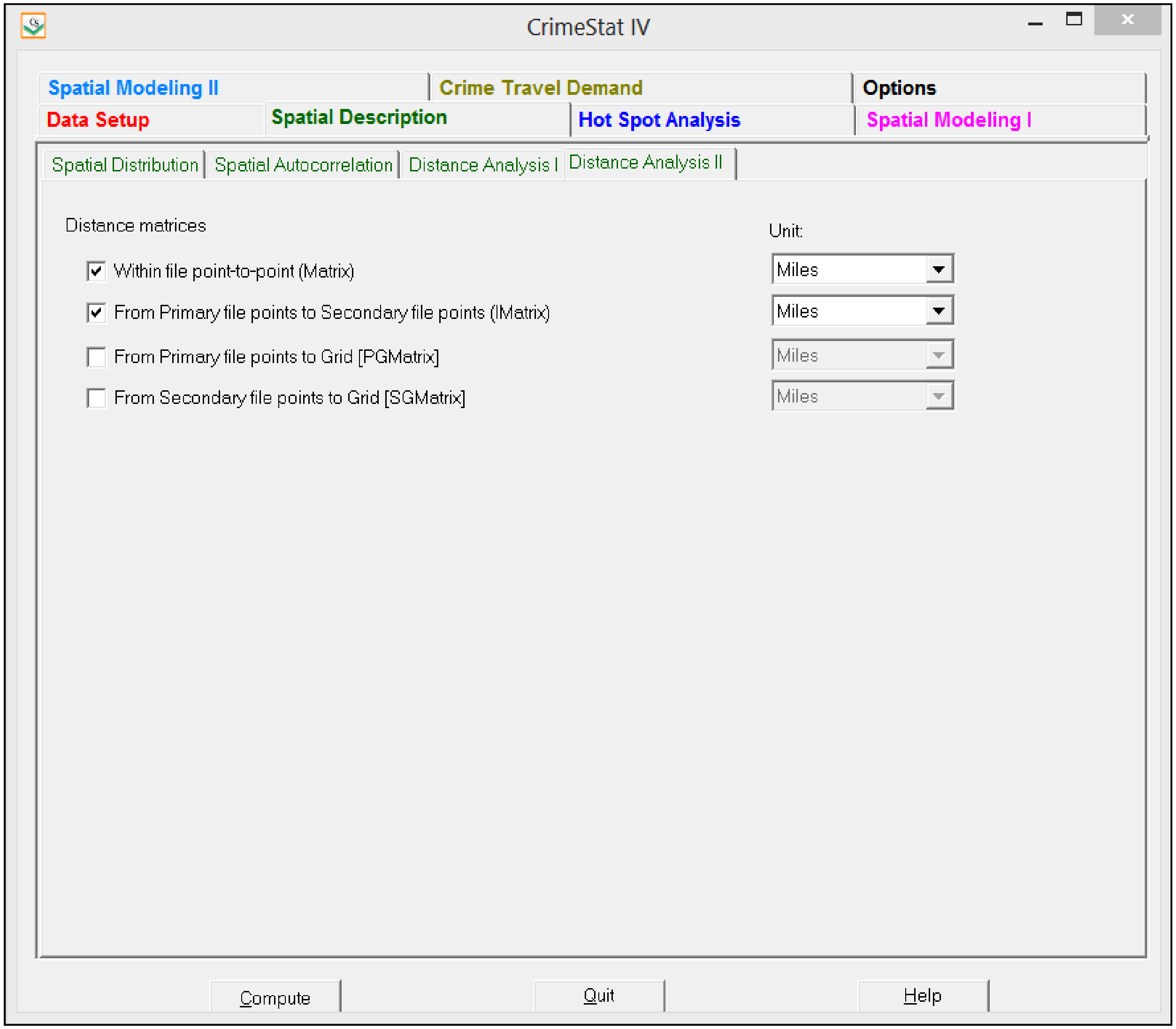
4. Fourth, if there is also a secondary file and a reference file, the distance from each secondary point to each grid cell can be computed. This is called the *From Secondary File Points to Grid* matrix (SGMatrix).

Each of these types of matrices can be displayed or saved to an Ascii 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 Ascii ‘.txt’ matrix* can also

Figure 6.10:

Distance Analysis II Screen



Distance matrices

Unit:

☑ Within file point-to-point (Matrix) — Miles

☑ From Primary file points to Secondary file points (Matrix) — Miles

□ From Primary file points to Grid [PGMatrix] — Miles

□ From Secondary file points to Grid [SGMatrix] — Miles 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 distances to the grid cells can be used to compare the distances for different distributions to a central location (e.g., a police station). There are many applications where distances are the primary unit of analysis. However, the user will need other software to read the files.

Be careful in outputting distances, though, because the files will generally be very large. For example, a primary file of 1000 incidents when interpolated to 9000 grid cells (100 columns x 90 rows) will produce 9 million paired comparisons. Such a file will take a lot of disk space. For that reason, we only allow output to an Ascii text file.

References

Aplin, G. (1983). *Order-Neighbour Analysis*. Concepts and Techniques in Modern Geography No. 36. Institute of British Geographers, Norwich, England: Geo Books.

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

Clark, P. J. & Evans, F. C. (1954). Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*, 35, 445-453.

Cressie, N. (1991). *Statistics for Spatial Data.* New York: J. Wiley & Sons, Inc.

Ebdon, D. (1988). *Statistics in Geography* (second edition with corrections). Blackwell: Oxford.

Getis, A. & Boots, B. (1978). *Models of Spatial Processes: An Approach to the Study of Point, Line and Area Patterns*. London: Cambridge University Press.

Hammond, R. & McCullagh, P. (1978). *Quantitative Techniques in Geography: An Introduction*. Second Edition. Clarendon Press: Oxford, England.

Kaluzny, S. P., Vega, S. C., Cardoso, T. P., & Shelly, A. A. (1998). *S+ Spatial Stats: User Manual for Windows and Unix*. Springer: New York.

Ripley, B. D (1981). *Spatial Statistics*. John Wiley & Sons: New York.

Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of Applied Probability* 13: 255-66.

Thompson, H. R. (1956). Distribution of distance to nth neighbour in a population of randomly distributed individuals. *Ecology*, 37, 391-394.

Venables, W.N. & Ripley, B. D. (1997). *Modern Applied Statistics with S-Plus (second edition)*. Springer-Verlag: New York.

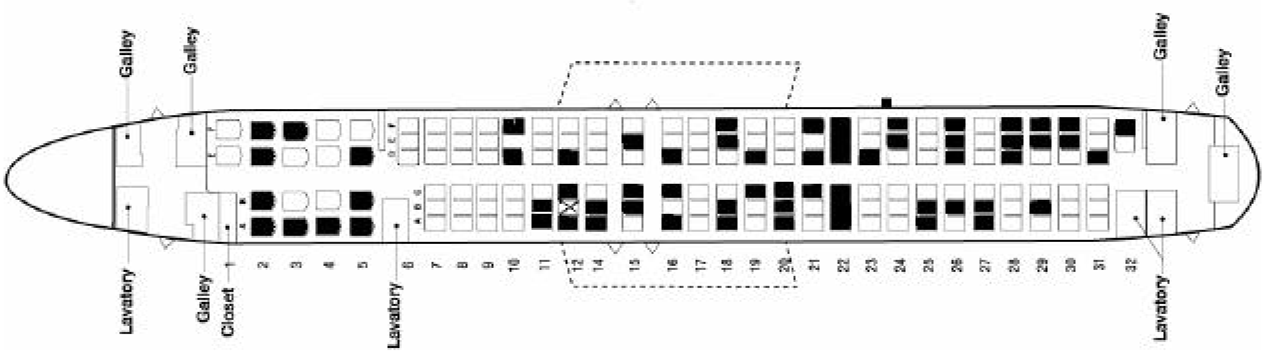
Attachments

SARS and the Distribution of Passengers on an Airplane

Marta A. Guerra
Senior Staff Epidemiologist,
Centers for Disease Control and Prevention
Atlanta, GA

Illness in passengers on board airplanes occurs rather frequently, and investigations are performed to assess whether transmission to other passengers has occurred. During 2002, several passengers with Severe Acute Respiratory Syndrome (SARS) traveled to the United States by airplane while they were infectious. Since transmission of SARS can be airborne, there is concern that it could spread during an airline flight. A survey was undertaken on a flight where a confirmed SARS case was on board. Serum samples of passengers were taken to evaluate if transmission of SARS had occurred during the flight, and whether transmission is related to sitting near the SARS case.

The nearest neighbor index was used to compare the distances between the seats of passengers on this flight to distances expected on the basis of chance. A grid (7 m x 32 m) was superimposed on the airline seat configuration, and each seat was assigned an X, Y coordinate based on the width (x) and the length (y) of the airplane. In the diagram below, the seat location of the SARS index case is indicated by an X, and the passengers' seat locations are shaded in black.



Nearest Neighbor Statistics for Airline Flight with SARS Case

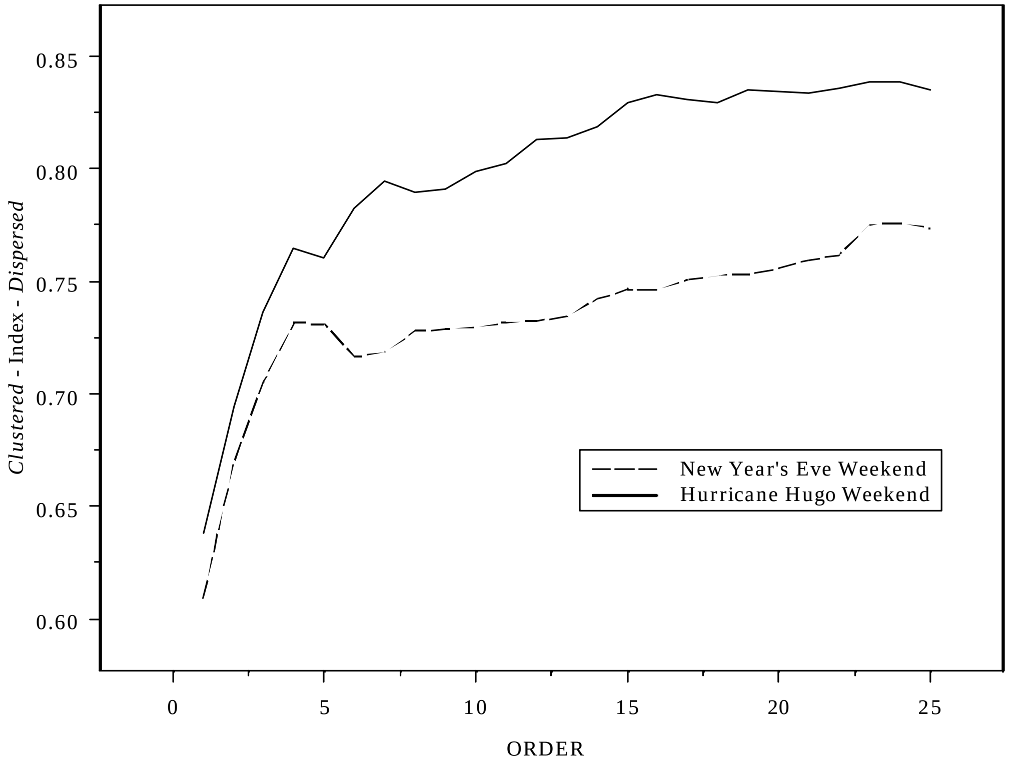
The nearest neighbor index of passengers' seats was 0.931 indicating that the distribution was random, not clustered. This preliminary analysis was important in order to establish that the seating arrangement of the passengers was random and independent, and that the passengers' seats were not clustered around the SARS case. Therefore, if any passengers have positive serum samples for SARS, we would be able to evaluate their locations in relation to the SARS case and assess patterns of transmission. In this survey, however, there was no evidence of transmission since none of the passengers had positive serum samples for SARS.

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



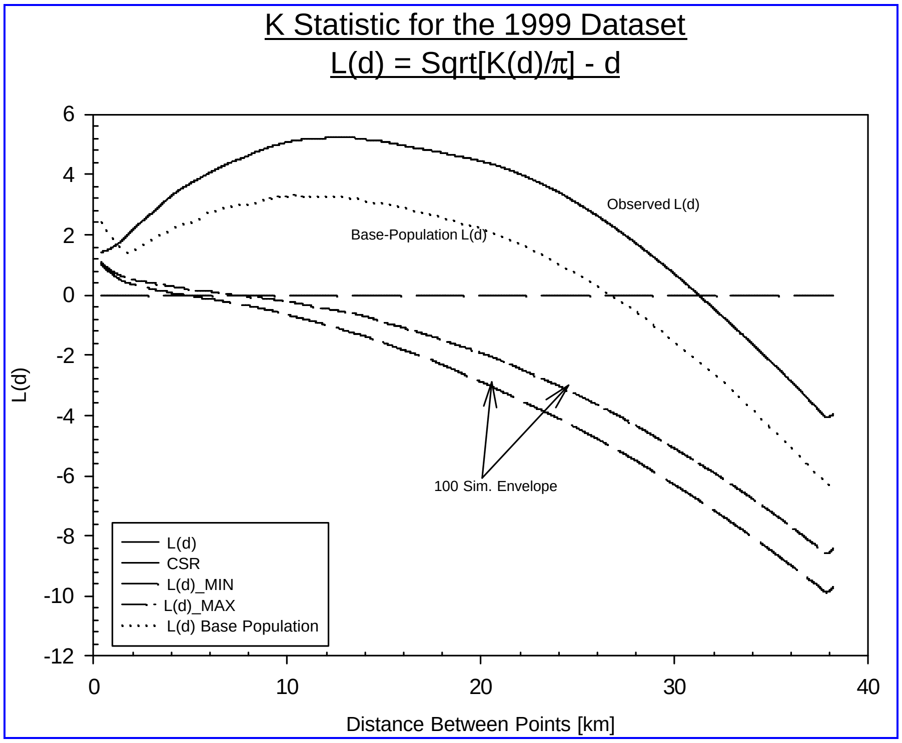
The Nearest Neighbor Index in *CrimeStat* was used to compare the distributions. From the onset, the Hurricane Hugo *Man With a Gun* locations are more dispersed than New Year's Eve. After the 5th nearest neighbor (Order 5) the differences become more pronounced

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,