Chapter 12:

Space-Time Analysis

Ned Levine
Ned Levine & Associates
Houston, TX

Table of Contents

Table of Contents (continued)

Chapter 12:
Space-Time Analysis

In this chapter, we discuss three techniques that are used to analyze the relationship between space and time. Up to this point, we have analyzed the distribution of incidents irrespective of the order in which they appeared or in which the time frame in which they appeared. The only temporal analysis that was conducted was in Chapter 4 where several spatial description indices, including the standard deviational ellipse, were compared for different time periods.

As police departments usually know, however, the spatial patterning of incidents does not occur uniformly throughout the year, but instead are often clustered together during short time periods. At certain times, a rash of incidents will occur in certain neighborhoods and the police often have to respond quickly to these events. In other words, there is both clustering in time as well clustering in space. This area of research has been developed mostly in the field of epidemiology (Knox, 1963, 1988; Mantel, 1967; Mantel and Bailar, 1970; Besag and Newell, 1991; Kulldorf and Nargawalla, 1995; Bailey and Gattrell, 1995). However, most of these techniques are applicable to crime analysis and criminal justice research as well.

CrimeStat includes four space-time techniques: the Knox index, the Mantel index, the Spatial-temporal moving average, and Correlated Walk Analysis. Figure 12.1 shows the Space-Time Analysis screen.

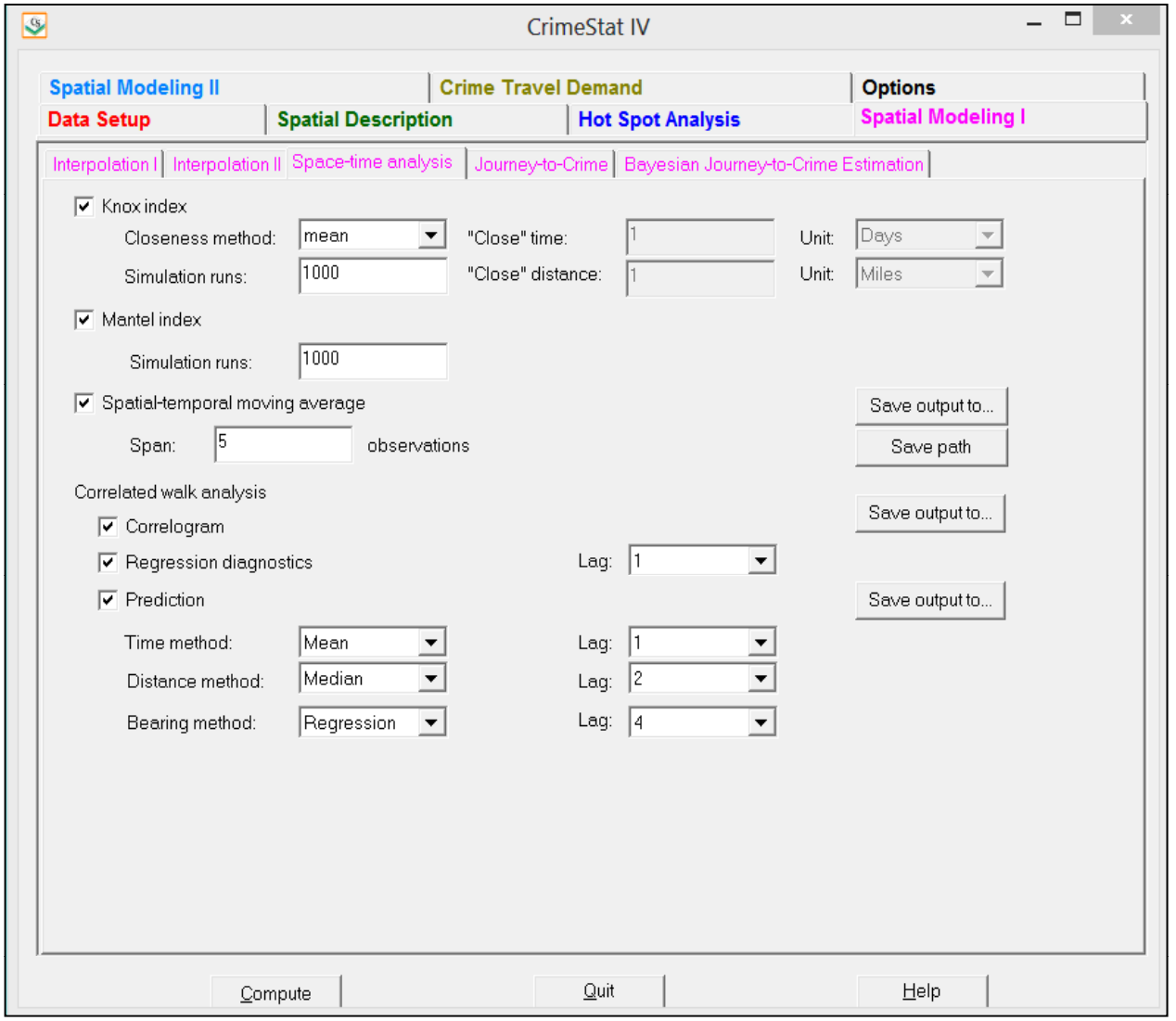
Measurement of Time in *CrimeStat*

Time can be defined as hours, days, weeks, months, or years. The default is days. However, please note that for any of these techniques, in *CrimeStat* time must be measured as an ***integer*** (or *real*) variable, as mentioned in Chapter 3. Time **cannot** be defined by a formatted date code (e.g., 11/06/01, July 30, 2002). Each of the three space-time routines require that time be an integer or real variable (e.g., 1, 2, 34527, 2.8). If given formatted dates, the routines will calculate an answer, but the result will not be correct.

If the time unit is days, a simple transformation is to use the number of days since January 1, 1900. Most spreadsheet and data base programs usually assign an integer number from this reference point. For example, November 12, 2001 has the integer value of 37207 while January 30, 2002 has the integer value of 37286. These are the number of days since January 1, 1900. Any spreadsheet program (e.g., Excel) can convert a date format into a real number with the Value function. Also, any arbitrary numbering system will work (e.g., 1, 2, 3).

Figure 12.1:

Space-Time Analysis Screen



Space-Time Interaction

There are different types of interaction that could occur between space and time. Four distinctions can be made. First, there could be ***spatial clustering all the time***. Certain communities are prone to certain events. For example, robberies often are concentrated in particular locations as are vehicle thefts. The hot spot methods that were discussed in chapters 7, 8 and 9 are useful for identifying these concentrations. In this case, there is no space-time interaction since the clustering occurs all the time.

Second, there could be ***spatial clustering within a specific time period***. Hot spots could occur during certain time periods. For example, motor vehicle crashes tend to occur with much higher frequencies in the late afternoon and early evening, often as a by-product of congestion on the roads. Crash hot spots will tend to appear at certain times because of the congestion. At most other times, the concentration does not occur because the congestion levels are lower.

Third, there could be ***episodic space-time clustering***. A number of events could occur within a short time period within a concentrated area. This type of effect is very common with motor vehicle thefts. A car thief gang may decide to attack a particular neighborhood. After a binge of car thefts, they move on to another neighborhood. In this instance, there are a number of theft incidents that are occurring within a limited period in a limited location. The cluster moves from one location to another. In this case, there is an interaction between space and time in that spatial hot spots appear at particular times, but are temporary. The ability to detect this type of shift is very important to police departments since it affects their ability to respond.

Fourth, there could be ***periodic space-time interaction*** in which the relationship between space and time occurs at certain times but not others and is somewhat predictable. The interaction could be concentrated, as in the spatial clustering mentioned above, or it could follow a more complex pattern. For example, there could be a diffusion of drug sales from a central location to a more dispersed area. Whereas initially, the drug dealing is concentrated in a few locations, it starts to diffuse to other areas. However, the diffusion may occur at different times of the year (e.g., Christmas and New Years). Alternatively, vehicle thefts may shift towards seaside communities during the summer months when the number of vacationers increases and then shift back to the city at other times of the year. We saw an example of this in Chapter 4 where the ellipse of motor vehicle thefts shifted between June and July to the communities along the Chesapeake River near Baltimore. This type of diffusion is not clustering *per se*, in that it may be spread over a very large coastline. But it is a distinct space-time interaction.

The importance of these distinctions is that many space-time tests that exist only measure gross space-time interaction, rather than space-time clustering. For example, the Knox and Mantel tests that are discussed below test for spatial interaction. The interaction could be the result of spatial clustering, but does not necessarily have to be. The interaction could occur in a very complex way that would not easily lend itself to more focused intervention by the police. Still, the ability to identify the interaction is an important first step in planning an intervention strategy.

Knox Index

The Knox Index is a simple comparison of the relationship between incidents in terms of distance (space) and time (Knox, 1963; 1964). That is, each individual pair is compared in terms of distance and in terms of time interval. Since each pair of points is being compared, there are N*(N-1)/2 pairs. The distance between points is divided into two groups - Close in distance and Not close in distance, and the time interval between points is also divided into two groups - Close in time and Not close in time. The definitions of 'close' and 'Not close' are left to the user.

A simple 2 x 2 table is produced that compares closeness in distance with closeness in time. The number of pairs that fall in each of the four cells is compared (Table 12.1).

Table 12.1:
Logical Structure of Knox Index

	Close in time	**Not close in time**	**TOTAL**
Close in distance	O_1	O_2	S_1
Not close in distance	O_3	O_4	S_2
TOTAL	S_3	S_4	N

where $N = O_1 + O_2 + O_3 + O_4$

$S_1 = O_1 + O_2$

$S_2 = O_3 + O_4$

$S_3 = O_1 + O_3$

$S_4 = O_2 + O_4$

The actual number of pairs that falls into each of the four cells are then compared to the expected number if there was no relationship between closeness in distance and closeness in time. The expected number of pairs in each cell under strict independence between distance and the time interval is obtained by the cross-products of the columns and row totals (Table 12.2).

Table 12.2:
Expected Frequencies for Knox Index

	Close in time	**Not close in time**
Close in distance	E_1	E_2
Not close in distance	E_3	E_4

where $\mathbf{E_1 = S_1 * S_3 / N}$
$\mathbf{E_2 = S_1 * S_4 / N}$
$\mathbf{E_3 = S_2 * S_3 / N}$
$\mathbf{E_4 = S_2 * S_4 / N}$

The difference between the actual (observed) number of pairs in each cell and the expected number is measured with a Chi-square statistic (equation 12.1):

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} \qquad (12.1)$$

Monte Carlo Simulation of Critical Chi-square Values

Unfortunately, the usual probability test associated with the Chi-square statistic cannot be applied since the observations are not independent. The interaction between space and time tends to be compounded when calculating the Chi-square statistic. For example, we have noticed that the Chi-square statistic tends to get larger with increasing sample size, a condition that would normally not be true with independent observations.

To handle the issue of interdependency, there is a Monte Carlo simulation of the Chi-square value for the Knox Index under spatial randomness (Dwass, 1957; Barnard, 1963). This is known as *randomization* since it assumes that any location within the study area could be available for an event. If the user selects a simulation, the routine randomly selects M pairs of a

distance and a time interval where M is the number of pairs in the data set $M = N\frac{(N-1)}{2}$ and calculates the Knox Index and the Chi-square test. Each pair of a distance and a time interval are selected from the range between the minimum and maximum values for distance and time interval in the data set using a uniform random generator.

An alternative simulation is to assume that the spatial location of the events are fixed and cannot change. This would occur, for example, if one was measuring a unique set of individuals who do not change or certain neighborhoods only or even applying the statistic to grouped data where the groups do not change In this case, a *permutation* simulation would be appropriate, similar to the simulations used in the spatial autocorrelation indices (see Chapters 5 and 9). For this version of *CrimeStat*, we only use a randomization simulation.

Output of simulation

The randomization simulation is repeated *K* times, where *K* is specified by the user. Usually, it is wise to run the simulation 1000 or more times. The output includes:

1. The sample size
2. The number of pairs
3. The calculated chi-square value of the Knox Index from the data
4. The minimum chi-square value of the Knox Index from the simulation
5. The maximum chi-square value of the Knox Index from the simulation
6. Ten percentiles from the simulation:

    a. 0.5%
    b. 1%
    c. 2.5%
    d. 5%
    e. 10%
    f. 90%
    g. 95%
    h. 97.5%
    i. 99%
    j. 99.5%

Methods for Dividing Distance and Time

In the *CrimeStat* implementation of the Knox Index, the user can divide distance and time interval based on the three criteria:

1. The mean (mean distance and mean time interval). This is the default.

2. The median (median distance and median time interval)

3. User-defined criterion for distance and time separately.

There are advantages to each of these methods. The mean is the center of the distribution; it denotes a balance point. The median will divide both distance and time interval into approximately equal numbers of pairs. The division is approximate since the data may not easily divide into two equal numbered groups. A user-defined criterion can fit a particular need of an analyst. For example, a police department may only be interested in incidents that occur within two miles of each other within a one week period. Those criteria would be the basis for dividing the sample into 'Close' and 'Not close' distance and time intervals.

Example of the Knox Index

For an example, vehicle thefts in Baltimore County for 1996 were taken. There were 1855 vehicle thefts for which a date was recorded in the data base. The data base was further broken down into twelve separate monthly subsets. Using the median as the criterion for dividing the data into 'Close' and 'Not close' for both distance and time interval, the Knox Index was calculated for the entire set of 1855 incidents. Then, using the median distance for the entire year but a month-specific median time interval, the Knox Index was calculated for each of the twelve months. Table 12.3 presents the Chi-square values and their pseudo-significance levels.

To produce a better test of the significance of the results, 1000 random simulations were calculated for the vehicle theft for the entire year. Table 12.3 below shows the results. Because an extreme value could be obtained by chance with a random distribution, reasonable cut-off points are usually selected from the simulation. In this case, we want a cut-off point that approximates a 5% significance level. Since the Knox Index is a one-tailed test (i.e., only a high chi-square value is indicative of spatial interaction), we adopt an upper threshold of the 95 percentile. In other words, only if the observed Chi-square test for the Knox Index is larger than the 95^{th} percentile will the null hypothesis of a random distribution between space and time be rejected.

Table 12.3:

Knox Index for Baltimore County Vehicle Thefts

Median Split

(N = 1,855 with 1,719,585 comparisons)

Month	Actual Chi-square	95 Percentile Simulation Chi-square	Approx. p
January	0.26	6.95	n.s.
February	0.00	6.61	n.s.
March	0.00	6.86	n.s.
April	0.50	6.56	n.s.
May	1.04	7.25	n.s.
June	0.01	6.02	n.s.
July	9.96	9.05	.05
August	5.91	5.55	.05
September	0.27	5.41	n.s.
October	3.33	6.43	n.s.
November	10.79	8.91	.01
December	0.00	6.87	n.s.
All of 1996	**8.69**	**41.89**	**n.s.**

For the entire year, there was not a significant clustering between space and time. Approximately, 26.7% of the incidents were both close in distance (i.e., closer than the median distance between pairs of incidents) and close in time (i.e., closer than the median time interval between pairs of incidents). However, when individual months are examined, three show significant relationships: July, August, and November. During these months, there is an interaction between space and time. Typically, this indicates that, during those months, incidents that cluster together spatially tend also to cluster together temporally. However, it could be the opposite (i.e., events that cluster together temporally tend to be far apart spatially).

The next step would to identify whether there are particular clusters that occur within a short time period. Using one of the 'hot spot' analysis methods discussed in Chapters 7 and 8, an analyst could take the events for the three months and try to identify whether there is spatial clustering during those three months that does not normally occur. We did not do that here, but the point is that the Knox Index is useful to identify *when* there is spatial clustering.

Problems with the Knox Index

The Knox Index is a simple measure of space-time clustering. But there are potential problems with it. First, because it is only a 2 x 2 table, different results can be obtained by varying the cut-off points for distance or time. For example, using the mean as the cut-off, the overall Chi-square statistic for all vehicle thefts was 8.67, reasonably close. However, when a cut-off point for distance of 1000 meters and a cut-off point for time of 80 days was used, the Chi-square statistic dropped to 3.16. In other words, the Knox Index will produce different results for different cut-off points.

A second problem has to do with the interpretation. As with any Chi-square test, differences between the observed and expected frequencies could occur in any cell or any combination of cells. Finding a significant relationship does not automatically mean that events that were close in distance were also close in time; it could have been the opposite relationship. However, a simple inspection of the table can indicate whether the relationship is as expected or not. In the above example, all the significant relationships showed a higher proportion of events that were both close in distance and close in time.

Mantel Index

The Mantel Index resolves some of the problems of the Knox Index. Essentially, it is a correlation between distance and time interval for pairs of incidents (Mantel, 1967). More formally, it is a general test for the correlation between two *dissimilarity* matrices that summarizes comparisons between pairs of points (Mantel and Bailar, 1970). It is based on a simple cross-product of two interval variables (e.g., distance and time interval):

$$T = \sum_{i=1}^{N} \sum_{j \neq i=1}^{N-1} \left(X_{ij} - \bar{X}\right)\left(Y_{ij} - \bar{Y}\right) \qquad (12.2)$$

where X_{ij} is an index of similarity between two observations, i and j, for one variable (e.g., distance) while Y_{ij} is an index of similarity between the same two observations, i and j, for another variable (e.g., time interval). The comparison is between two observations and does not include a comparison of an observation with itself. Hence, *j* is incremented up to *N-1*.

The cross-product is then normalized by dividing each deviation by its standard deviation:

$$r = \sum_{i=1}^{N} \sum_{j \neq i=1}^{N-1} \frac{(X_{ij} - \bar{X})}{S_X} \frac{(Y_{ij} - \bar{Y})}{S_Y} \qquad (12.3)$$

$$= \frac{1}{(N-1)} \sum_{i=1}^{N} \sum_{i \neq j=1}^{N-1} Z_X Z_y \quad (12.4)$$

where X_{ij} and Y_{ij} are the original variables for comparing two observations, i and j, and Z_x and Z_y are the normalized variables.

Monte Carlo Simulation of Confidence Intervals

Even though the Mantel Index is a Pearson product-moment correlation between distance and time interval, the observations are not independent and, in fact, are highly interdependent. That is, the correlation is between observations for the same distance and time variable rather than between the two variables by themselves. Thus, the values of the Mantel *r* tend to be very low.

Further, the usual significance test for a correlation coefficient is not appropriate. Instead, the Mantel routine offers a simulation of the confidence intervals around the index. If the user selects a simulation, the routine randomly selects M pairs of a distance and a time interval where M is the number of pairs in the data set $M = N\frac{(N-1)}{2}$ and calculates the Mantel Index. Each pair of a distance and a time interval are selected from the range between the minimum and maximum values for distance and time interval in the data set using a uniform random generator. As with the Knox Index simulation discussed above, the simulation is a randomization where every location within the study area is possible for an event to occur, compared to a permutation simulation where the spatial locations are fixed. For this version of *CrimeStat*, we only use a randomization simulation.

The random simulation is repeated *K* times, where *K* is specified by the user. Usually, it is wise to run the simulation 1000 or more times. The output includes:

1. The sample size
2. The number of pairs
3. The calculated Mantel Index from the data
4. The minimum Mantel value from the simulation
5. The maximum Mantel value from the simulation
6. Ten percentiles from the simulation:
    a. 0.5%
    b. 1%
    c. 2.5%
    d. 5%
    e. 10%
    f. 90%

g. 95%
h. 97.5%
i. 99%
j. 99.5%

Example of the Mantel Index

In *CrimeStat*, the Mantel Index routine calculates the correlation between distance and time interval. To illustrate, Table 12.4 examines the Mantel correlation for the 1996 vehicle thefts in Baltimore County that was illustrated above. As seen, the correlations are all low. However, as with the Knox Index, July, August and November produce relatively higher correlations. As mentioned above, the correlations tend to be very low because the test is between observations for the same variables, rather than between variables.

Table 12.4:
Mantel Index for Baltimore County Vehicle Thefts
Median Split
(N = 1,855 and 1,719,585 Comparisons)

Simulation Month	Simulation r	Approx. 2.5%	97.5%	p-level
January	-.0047	-0.033	0.033	n.s.
February	-.0023	-0.037	0.042	n.s.
March	-.0245	-0.032	0.039	n.s.
April	0.0077	-0.040	0.041	n.s.
May	0.0018	-0.038	0.043	n.s.
June	0.0043	-0.035	0.041	n.s.
July	0.0348	-0.034	0.033	.025
August	0.0544	-0.034	0.035	.01
September	0.0013	-0.044	0.046	n.s.
October	0.0409	-0.037	0.043	n.s.
November	0.0630	-0.042	0.040	.001
December	0.0086	-0.035	0.038	n.s.
All of 1996	**0.0015**	**-0.009**	**0.010**	**n.s.**

To test whether these correlations are significant or not, 1000 random simulations were calculated for each month using the same sample size as the monthly vehicle theft totals. Table 12.4 above shows the results. Because an extreme value could be obtained by chance with a random distribution, reasonable cut-off points are usually selected from the simulation. In this

case, we want cut-off points that approximate a 5% significance level. Since the Mantel Index is a two-tailed test (i.e., one could just as easily get dispersion between space and time as clustering), we adopt a lower threshold of the 2.5 percentile and an upper threshold of 97.5 percentile. Combined, the two cut-off points ensure that approximately 5% of the cases will be either lower than the lower threshold or higher than the upper threshold under random conditions.[1] In other words, only if the observed Mantel Index is smaller than the lower threshold or larger than the upper threshold will the null hypothesis of a random distribution between space and time be rejected.

In Table 12.4, for the entire year, the observed Mantel Index (correlation between space and time) was 0.0015. The 2.5^{th} percentile was -.009 and the 97.5^{th} percentile was 0.01. Since the observed value is between these two cut-off points, we cannot reject the null hypothesis of no relationship between space and time. However, for the individual months, again, July, August and November have correlations above the upper cut-off threshold. Thus, for those three months *only*, the amount of space-time clustering in the vehicle theft data is most likely greater than what would be expected on the basis of a chance distribution. One would, then, have to explore the data further to find out where those vehicle thefts were occurring, using one of the hot spot routines in Chapters 7 or 8.

Limitations of the Mantel Index

The Mantel Index is a useful measure of the relationship between space and time. But it does have limitations. First, because it is a Pearson-type correlation coefficient, it is prone to the same types of problems that befall correlations. Extreme values of either space or time could distort the relationship, either positively, if there are one or two observations that are extreme in *both* distance in time interval, or negatively, if there are only one or two observations that are extreme in *either* distance or in time interval.

Second, because the test is a comparison of all pairs of observations, the correlations tend to be very small, as noted above. This makes it less intuitive as a measure than a traditional correlation coefficient that varies between -1 and +1 and in which high values are expected. For most analysts, it is not very intuitive to have an index where 0.05 is a high value. This does not fault the statistic as much make it a little non-intuitive for users.

---

[1] It would be possible to make a one-tailed test with the simulation. For example, if one is only interested in the degree of clustering, one could adopt the 95 percentile as the threshold. An observed Mantel value that was lower than this threshold would be consistent with the null hypothesis.

Third, as with any correlation coefficient, the sample size needs to be fairly large to produce a stable estimate. In the above, example, one could further break down monthly vehicle thefts by week or, even, day. However, the number of cases will decrease considerably. In the above example, with 1,855 vehicle thefts over a year, the weekly average would be around 36, which is a small sample. Intuitively, a crime analyst wants to know when space-time clustering is occurring and a short time frame is critical for detection. A week would be the largest time interval that would be useful.

However, as the sample size gets small, the index becomes unstable. The sample size makes the index volatile. While the Monte Carlo simulation will adjust for the sample size, the range of the cut-off thresholds will vary considerably from one week to another with small sample sizes. The analyst will have to run the simulation a large number of times to adjust for the varying sample sizes. Also, the shortened time frame allows fewer distinctions in time. If one takes a very narrow time frame (e.g., a day), there will be virtually no time differences observed because there is not enough data to produce reliable estimates.

One way to get around this is to have a moving average where the time frame is adjusted to fit a constant number of days (e.g., a 14 day moving average). The advantage is that the sample size tends to remain constant; one could therefore reduce the number of recalculations of the cut-off thresholds since they would not vary much from one day to another. To make this work, however, the data base must be set up to produce the appropriate number of incidents for a moving average analysis.

Nevertheless, the Mantel Index remains a useful tool for analysts. It is still widely used for space-time analysis and it has been generalized to many other types of dissimilarity analyses than just space and time. If used carefully, the index can be a powerful tool for detection of clusters that are also concentrated in time.

Spatial-Temporal Moving Average

The Spatial-Temporal Moving Average is a simple statistic. It is the moving mean center of *M* observations where *M* is a sub-set of the total sample, *N*. By 'moving', the observations are sequenced in order of occurrence. Hence, there is a time dimension associated with the sequence. The *M* observations are called the *span* and the default span is 5 observations. The span is centered on each observation so that there are an equal number on both sides. Because there are no data points prior to the first event and after the last event, the first few mean centers will have fewer observations than the rest of the sequence. For example, with a span of 5, the first and last mean centers will have only three observations, the second and next-to-last will have 4 observations, while all others will have 5. In general, it is a good idea to choose an odd

number since the middle of the span will be centered on a real observation rather than having to fall between two in the case of an even span.

Though simple, the Spatial-Temporal Moving Average is very useful for detecting changes in behavior by serial offenders. In the next chapter, we will examine journey-to-crime models that attempts to estimate the likely origin location of a serial offender based on the distribution of incidents committed by the offender. However, if the serial offender has either moved residences or else moved the field of operation, then the technique will error because it is assuming a stable field of operations when, in fact, it is not. The moving average can suggest whether the offender's behavior is stable or not.

As an example, figure 12.2 below shows the Spatial-Temporal Moving Average of an offender who committed 12 offenses before being arrested. The individual committed eight thefts from vehicles, two thefts from stores, one residential burglary and one highway robbery. The actual incidents are shown in red circles with the sequence number displayed. The moving average is shown in blue squares with the sequence number displayed. The path of the moving average is shown as a green line.

As seen, there is a definite shift in the field of operation by this offender. The mean center moved about a mile during this period but the consistency of the trend suggests that something fundamental changed by the offender, either the person moved residences or the nature of the committed crimes changed. In using the Journey-to-crime tools, an analyst would probably want to focus on the latter events since these are more geographically circumscribed. Notice that the last two moving averages are relatively close to the actual residence location of the offender when arrested (less than three-quarters of a mile away).

In short, the Spatial-Temporal Moving Average simply plots the changes in the mean center of the span and is useful for detecting changes in the behavior pattern of serial offenders.

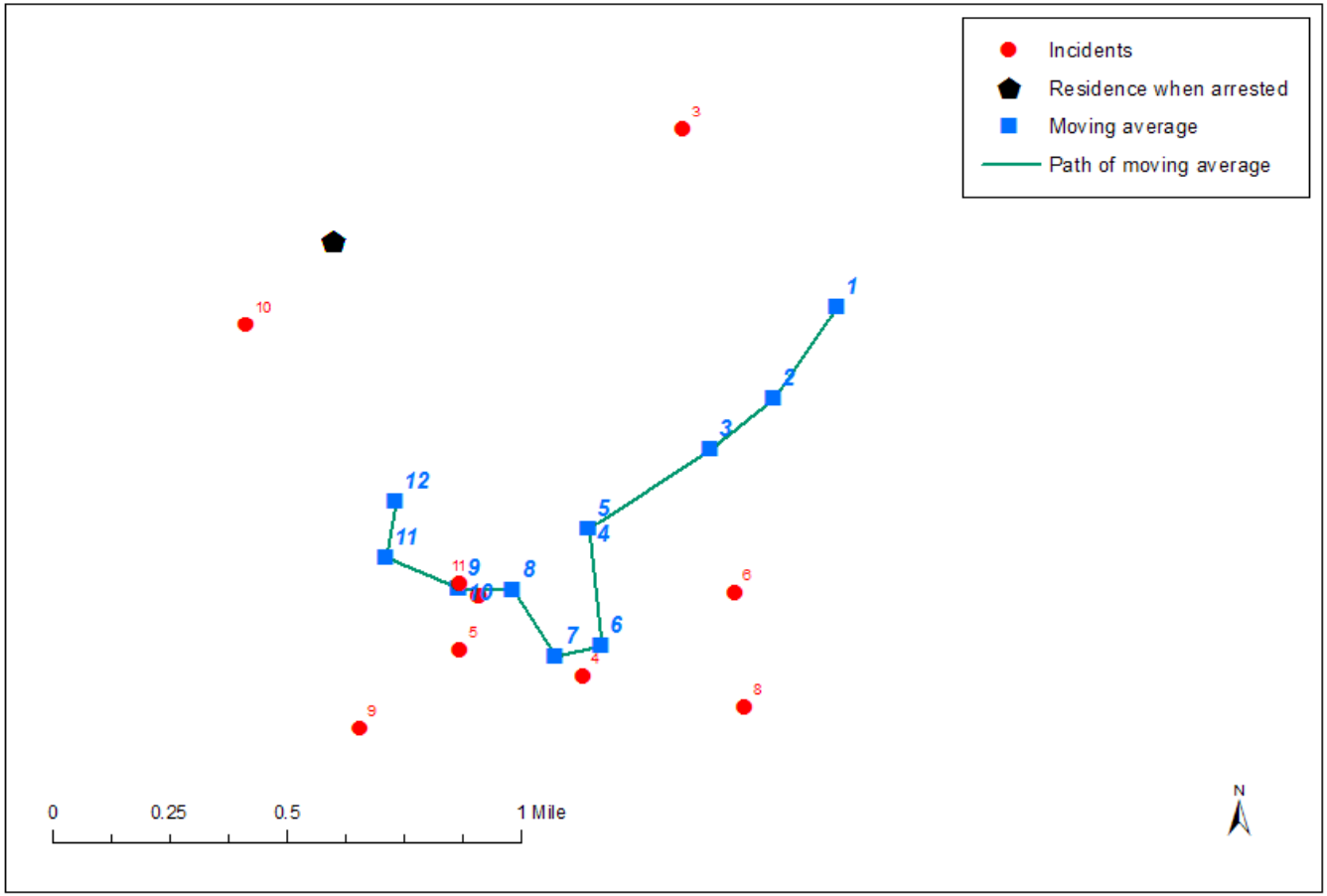
Correlated Walk Analysis

Correlated Walk Analysis (CWA) is a tool that is aimed at analyzing the spatial and temporal *sequencing* of incidents committed by a single serial offender. In this sense, it is the 'flip side' of Journey to crime analysis (see Chapter 13). Whereas journey to crime analysis makes guesses about the likely origin location for a serial offender, based on the spatial distribution of the incidents committed by the offender, the CWA routine makes guesses about the time and location of a next event, based on both the spatial distribution of the incidents and the temporal sequencing of them. In effect, it is a Spatial-Temporal Moving Average with a prediction of a next event.

Figure 12.2:

Moving Path of Serial Offender:

Sequence of 12 Crimes



The statistical origin of CWA is Random Walk Theory. Random Walk Theory has been developed by physicists to explain the distribution of molecules in a rapidly changing environment (e.g., the movements of a particle in a gas which is diffusing - Brownian movement). Sometimes called a 'drunkard's walk', the theory starts with the premise that movement is random in all directions. From an arbitrary starting point, a particle (or person) moves in any direction in a series of steps. The direction of each step is independent of the previous steps. After each step, a random decision is made and the person moves in a random direction. This process is repeated *ad infinitum* until an arbitrary stopping point is selected (i.e., the observer quits looking). It has been shown mathematically that all one and two dimensional random walks must eventually return to their original starting point (Spitzer, 1963; Henderson, Renshaw, & Ford, 1983; see endnote *i*). This is called a *recurrent random walk*. On the other hand, independent random walks in more than two dimensions are not necessarily recurrent, a state called *transient random walk*.

Figure 12.3 illustrates a random walk of 2000 steps. For a large number of steps in a two-dimensional walk, the likely distance of a person (or particle) from the starting point is:

$$E(d) = d_{RMS}\sqrt{N} \tag{12.4}$$

where d_{RMS} is the *root mean square* of distance.

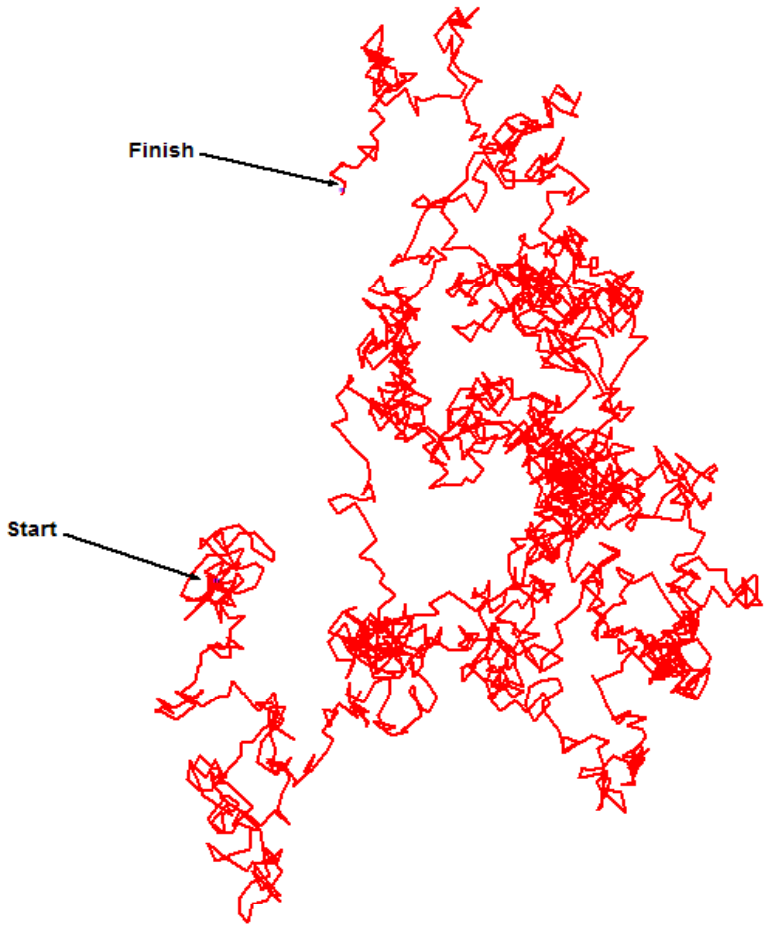
There are a number of different types of random walks. The simplest is a movement of uniform distance only along a grid cell (i.e., a Manhattan geometry). The person can only move North, South, East or West for a unit distance of 1. A more complex random walk allows angular distances and an even more complex random walk allows varying distances (e.g., normally distributed random distances, uniformly random distances). The walk in Figure 12.3 was of this latter type. X and Y values were selected randomly from a range of -1 to +1 using a uniform random number generator. For a conceptual understanding of Random Walk Theory, see Chaitin (1990) and, for a mathematical treatment, see Spitzer (1976). Malkiel (1999) applied the concepts of Random Walk Theory to stock price fluctuations in a book that has now become a classic.

Henderson, Renshaw and Ford (1984) have introduced the concept of a *correlated random walk*. In a correlated random walk, momentum is maintained. If a person is moving in a certain direction, they are more likely to continue in that direction than to reverse direction or travel orthogonally. In other words, at any one decision point, the probabilities of traveling in any direction are not equal; the same direction has a higher probability than an orthogonal change (i.e., turning 90 degrees) and those, in turn, have a higher probability than completely reversing direction.

Figure 12.3:

A Random Walk

2000 Random Steps of -1.0 to +1.0 in X and Y Direction



By implication, the same is true for distance and distance. A longer step than average is likely to be followed by another longer step than average while a shorter step than average is likely to be followed by another short step. Similarly, there is consistency in the time interval between events; a short interval is also likely to be followed by a short interval. In other words, a correlated random walk is a random walk with momentum (Chen & Renshaw, 1992; 1994). These authors have applied the theory to the analysis of the branching of tree roots (Henderson, Ford, Renshaw, & Deans, 1983; Renshaw, 1985).

Correlated Walk Analysis Routine

Correlated Walk Analysis is a set of tools that can help an analyst understand the sequencing of sequential events in terms of time interval, distance and direction. In *CrimeStat*, there are three CWA routines. The first two help the analyst understand whether there are *repeating* patterns in time, distance or direction while the last routine allows the analyst to make a guess about the next likely event, when it will occur and where it will occur. The three routines are:

1. CWA - Correlogram
2. CWA - Diagnostics
3. CWA - Prediction

CWA - Correlogram

The CWA - *Correlogram* routine calculates the correlation in time interval, distance, and bearing (direction) between events. It does this through *lags*. A lag is a separation in the intervals between events. The difference between the first and second event is the first interval. The difference between the second and third events is the second interval. The difference between the third and fourth events is the third interval, and so forth. For each successive interval, there is a time difference; there is a distance and there is a direction. One could extend this to all the intervals, comparing each interval with the next one; that is, we compare the first interval with the second, the second interval with the third, the third interval with the fourth, and so on until the sample is complete. When comparing successive intervals, this is called a *lag of 1*. It is important to keep in mind the distinction between an event (e.g., an incident) and an interval. It takes two events to create an interval. Thus, for a lag of 1, there are M=N-1 intervals where *M* is the number of intervals and *N* is the number of events (e.g., for 3 incidents, there are 2 intervals).

A lag of two compares every other event. Thus, the first interval is compared to the third interval; the second interval is compared to the fourth; the third interval is compared to the fifth; and so on until there are no more intervals left in the sample. Again, the comparison is for time

difference, distance, and direction separately. We can extend this logic to a lag of 3 (every third event), a lag of 4 (every fourth event), and so forth.

The CWA - Correlogram routine calculates the Pearson Product-Moment correlation coefficient between successive events. For a lag of 1, it compares successive events and correlates the time interval, distance, and bearing separately for these successive events. For a lag of 2, it compares every other event and correlates the time interval, distance, and bearing separately for these successive events. The routine does this until it reaches a maximum of 7 lags (i.e., every seventh event). However, if the sample size is very small, it may not be able to calculate all lags. It will require 12 incidents (events) to calculate all seven lags since it requires at least four observations per lag (i.e., N - L - 4 where N is the number of events and L is the maximum number of lags calculated).

Adjusted Correlogram

The Correlogram calculates the raw Pearson correlation coefficient between intervals by lag for time, distance, and bearing. One of the problems that may appear, especially with small samples, is that the correlation with higher-order lags are very high, either positive or negative. There are probably two reasons for this. For one thing, with each lag, the sample size decreases by one; with a very small sample size, correlations can become very volatile, jumping from positive to negative, and from low to high. Another reason is that periodicity in the data set is compounded with higher-order lags in the form of 'echos'. For example, if a lag of 2 is high, then a lag of 4 will also be somewhat high since there is a compounding of the lag 2 effect. When combined with a small sample size, it is not uncommon to have higher-order lags with very high correlations, sometimes approaching +/- 1.0. The user must be careful in selecting a higher-order lag because there is an apparent effect that may be due to the above reasons, rather than any real predictability. One of the key signs for spurious higher-order effect is a sudden jump in the strength of the correlation from one lag to the next (though sometimes a high higher-order lag can be real; see examples below).

To minimize these effects, the output also includes an adjusted correlogram that adjusts for the loss of degrees of freedom. The formula is:

$$A = \frac{M-L-1}{M-1} \tag{12.5}$$

where M is the number of intervals (N-1) and L is the number of lags. For example, for a sample size of 13, there will be 12 intervals (M). For a lag of 1, the adjustment will be:

$$A = \frac{12-1-1}{12-1} = \frac{10}{11} = 0.909 \quad (12.6)$$

The effect of the adjustment is to reduce the correlation for higher-order lags. It will not completely eliminate the effect, but it should help minimize spurious effects. As will be shown below, however, sometimes high correlations for higher-order lags are real.

CWA - Correlogram Output

The CWA - Correlogram routine outputs 10 parameters:

1. The sample size (number of events);
2. Number of intervals;
3. Information on the units of time, distance, and bearing;
4. Final distance to origin in meters (distance between last and first event);
5. Expected random walk distance from origin (if sequence was strictly random);
6. Drift (the ratio of actual distance from origin to expected random walk distance);
7. Final bearing from origin (direction between last event and first event);
8. Expected random walk bearing. Defined as 0 because there is no expected direction.
9. Correlations by lag for time, distance, and bearing (up to 7 lags); and
10. Adjusted correlations by lag for time, distance, and bearing (up to 7 lags).

The aim of the CWA - Correlogram is to examine repetitive sequences, whether for time interval, distance or direction. It is possible to have separate repetitions for time, distance and direction. For example, an offender may commit crimes every 7 days or so, say, on the weekend. In this case, the individual is repeating himself/herself about once every week. Similarly, an individual may alternate directions, first going East then going West, then going back to the East, and so forth. In other words, what we're asking with the routine is whether there are any repetitions in the sequence of incidents committed by a serial offender. Does he/she repeat the crimes in time? If so, what is the *periodicity* (the repititious sequence? Does he/she repeat the crimes in distance? Is so, what is the periodicity? Finally, does he/she repeat the crimes in direction? If so, what is the periodicity? The CWA - Correlogram, therefore, analyzes the sequence of incidents committed by an individual and does this separately for time interval, distance, and direction.

Offender repetition

Why is this important? Most crime analysis is predicted on the assumption that offenders (people in general) repeat themselves, consciously or unconsciously. That is, individuals have

specific behavior patterns that tend to be repeated. If an individual acts in a certain way (e.g., committing a burglary), then, most likely, the person will repeat himself/herself again. There is no guarantee, of course. But, because human beings do not behave spatially or temporally random but tend to operate in somewhat consistent ways, there is a likelihood that the individual will act in a similar manner again.

This assumption is the basis of profiling which aims at understanding the MO of an offender. If offenders were totally random in their behavior, detection and apprehension would be made much more difficult than it already is. So, between the two extremes of a totally random individual (the 'random walk person') and a totally predictable individual (the 'algorithmic person'), we have the bulk of human behavior, at least in terms of time, distance and direction.

CWA - Diagnostics

The Diagnostics routine is similar to the CWA - Correlogram except that it calculates an Ordinary Least Squares autoregression for a particular lag. That is, for a variable the routine regresses each interval against a previous interval. The user enters the lag number (the default is 1) and the routine produces three regression models for the successive event as the dependent variable against the prior event as the independent variable. There are three equations, for time interval, distance, and bearing separately. The output includes:

1. The sample size (number of events);
2. The number of intervals;
3. Information on the units of time, distance, and bearing;
4. The multiple correlation coefficient;
5. The squared multiple correlation coefficient (i.e., R^2);
6. The overall standard error of estimate;
7. The regression coefficient for the constant and for the prior event;
8. The standard error of the regression coefficients;
9. The t-values for the regression coefficients;
10. The p-value (two-tail) for the regression coefficients;
11. An analysis of variance test for the full model. This includes sum of squares for the regression term and for the residual;
12. The ratio of the regression sum of squares to the residual sum of squares (the F-ratio); and
13. The p-value associated with the F-value.

What the regression diagnostics provides is an indicator of the amount of predictability in the lag. It has the same information as the Correlogram (since the square of the correlation, r^2, is

the same as R^2 for a single independent variable regression equation), but it is easier to interpret. Essentially, it is argued below that, unless the R^2 in the regression equation is sufficiently high, that one is better off using the mean or median lag for prediction. Conversely, if the R^2 is very high, then the user should be suspicious about the data.

CWA - Prediction

Finally, after having analyzed the sequential pattern of events, the user can make a prediction about the time and place of the next event. There are three methods for making a prediction, each with a separate lag:

1. Mean difference
2. Median difference
3. Regression equation

The method is applied to the last event in the data set. The *mean difference* applies the mean interval of the data for the specified lag to the last event. For example, for time interval and a lag of 1, the routine calculates the time interval between each event and takes the average. It then applies the mean time interval to the last time in the data set as the prediction. The *median difference* applies the median interval of the data for the specified lag to the last event. For example, for bearing and a lag of 1, the routine calculates the direction (bearing) between each event, calculates the median bearing, and applies that median to the location of the last event in the data set as the predicted value.

The *regression equation* calculates a regression coefficient and constant for the specified lag and uses the data value for the last *interval* as input into the regression equation; the result is the predicted value. For example, for distance and a lag of 1, the routine calculates the regression coefficient and constant for a regression equation in which each event is compared to the previous event. The last distance in the data set (i.e., between the last event and the previous event) is used as an input for the regression equation and the predicted distance is marked off from the coordinates of the last event.

In other words, the routine takes the time and location of the last event and adds a time interval, a direction, and a distance as a predicted next event (next time, next location). The method by which this prediction is made can be the mean interval, the median interval, or the regression equation. If the user species a lag other than 1, that lag is applied to the last event. For example, for time with a mean difference and a lag of 2, the routine calculates the time interval between each event and every other event, calculates the average, and applies that average to the last event in the data set.

CWA - Prediction Graphical Output

The CWA - Prediction routine outputs five graphical objects in 'shp', 'mif, 'kml' (if the coordinates are spherical) or various Ascii formats. The routine adds five prefixes to the file name of the output object:

1. Events - a line indicating the sequence of events. If the user also brings in the points in the data set, it will be possible to number each of these steps;
2. PredDest - the predicted location for the next event;
3. PW - a line from the last location in the data set to the predicted location;
4. POrigL - a point representing the center of minimum distance of the data set. The center of minimum distance is taken as a proxy for the origin location of the offender; and
5. Path - a line from the expected origin to the predicted destination

For example, if the user provides the file name 'NightRobberies' and specifies a 'shp' output, there will be five objects output:

EventsNightRobberies.shp
PredDestNightRobberies.shp
PathNightRobberies.shp
POrigLNightRobberies.shp
PWNightRobberies.shp

Example 1: A Completely Predictable Individual

The simplest way to illustrate the logic of the CWA is to start with a completely predictable individual. This individual commits crimes on a completely systematic basis. Table 12.5 illustrates the behavior of this individual.

Starting at an arbitrary origin with an X coordinate of 1 and a Y coordinate of 1 on day 1, the individual commits 13 incidents in total. In the table, these are numbered events 1 through 13. From the origin, the individual always travels in a Northeast direction of 45 degrees (clockwise from due North - 0 degrees). The individual's second incident is at coordinate X=2, Y=2. Thus, the individual traveled at 45 degrees from the previous incident and for a distance of 1.4142 (the hypotenuse of the right angle created by traveling one unit in the X direction and one unit in the Y direction). For the third incident, the individual commits this at X=4, Y=4. Thus, the direction is also at 45 degrees from the previous location but the distance is now 2.8284 (or the square root of 8 which comes from a step of 2 along the X axis and a step of 2 along the Y axis). For the fourth incident, the individual commits the crime at X=7, Y=7. Again, the

direction is 45 degrees, but the distance is 4.2426 (or the square root of 18 which comes from a step of 3 along the X axis and a step of 3 along the Y axis).

Table 12.5:

Example of a Predictable Serial Offender: 1

(N = 13 incidents)

Event	X	Y	Distance	Days	Time Interval
1	1	1	-	-	
2	2	2	1.4142	3	2
3	4	4	2.8284	7	4
4	7	7	4.2426	9	2
5	8	8	1.4142	13	4
6	10	10	2.8284	15	2
7	13	13	4.2426	19	4
8	14	14	1.4142	21	2
9	16	16	2.8284	25	4
10	19	19	4.2426	27	2
11	20	20	1.4142	31	4
12	22	22	2.8284	33	2
13	25	25	4.2426	37	4
Logical prediction for next event	14	26	26	1.4142 39	2

For the fifth incident, again the individual traveled at 45 degrees to the previous incident, but repeated himself/herself with a step of only 1 unit in both the X and Y directions. The individual then continued the sequence, always traveling in a 45 degree orientation due North. For distance, a step of 1 in both the X and Y directions is followed by a step of 2 in both directions, and is followed by a step of 3 in both directions. In other words, the individual repeats direction every time and repeats distance every third time. There is a periodicity of 1 for direction and 3 for distance.

For time interval, this individual repeats him/herself every other time. The second event occurs 2 days after the first event. The third event occurs 4 days after the second event; the fourth event occurs 2 days after the third event; the fifth events occurs 4 days after the fourth event; and so forth. In other words, for time interval, the individual repeats him/herself every other interval (i.e., the periodicity is 2).

Figure 12.4 illustrates the sequence; the number at each event location is the number of the day that the individual committed the offense (starting at an arbitrary day 1).

Since this fictitious individual is completely predictable, we can easily guess when and where the next event will occur (see Table 12.5 above). The direction will, of course, be at 45 degrees from the previous location. Looking at the last known event (event 13), the distance traveled was 4.2426. Thus, we predict that the individual will revert to a move of 1 in the X direction and 1 in the Y direction, or coordinates X=26, Y=26. Finally, for time interval, since the last known time interval was 4 days, then this individual will commit the next event 2 days later, or day number 39.

Example 1: Analysis

The first step is to analyze the sequencing of the events. There are 13 events and 12 intervals. The CWA - Correlogram produces the output shown in Table 12.6 below.

Table 12.6:
Correlogram of Predictable Serial Offender: 1
(N=13 Incidents and M=12 Intervals)

Correlated Walk Analysis -- Correlogram:

Sample size .....…....: 13
Measurement type ...: Direct
Input units …….....: Feet
Time units ....…….....: Days
Distance units ...…...: Feet
Bearing units ....…....: Degrees

Correlation

Lag	Time	Distance	Bearing
0	1.00000	1.00000	1.00000
1	-1.00000	-0.42105	1.00000
2	1.00000	-0.56522	1.00000
3	-1.00000	1.00000	1.00000
4	1.00000	-0.38462	1.00000
5	-1.00000	-0.58824	1.00000
6	1.00000	1.00000	1.00000
7	-1.00000	-0.28571	1.00000

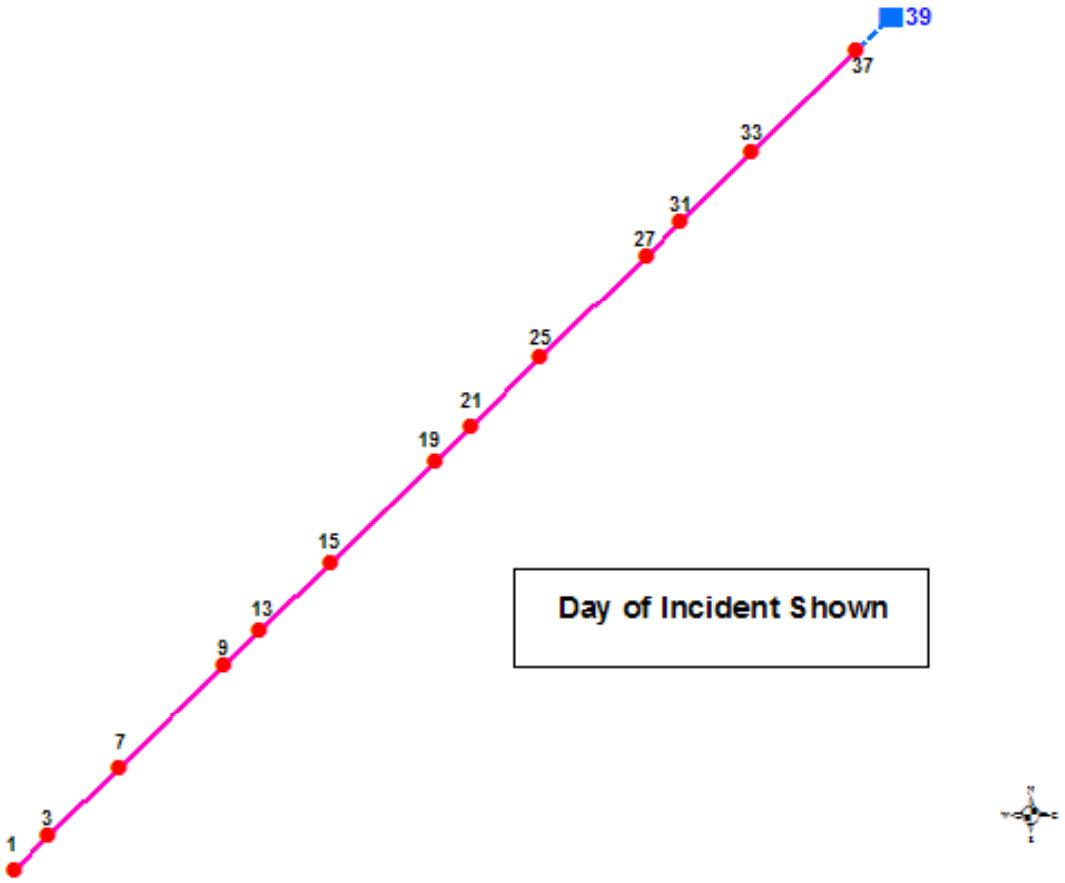
Adjusted Correlation

Lag	Time	Distance	Bearing
0	1.00000	1.00000	1.00000
1	-0.90909	-0.38278	0.90909
2	0.81818	-0.46245	0.81818
3	-0.72727	0.72727	0.72727
4	0.63636	-0.24476	0.63636
5	-0.54545	-0.32086	0.54545
6	0.45455	0.45455	0.45455
7	-0.36364	-0.10390	0.36364

Figure 12.4:

Example of a Predictable Serial Offender: I

(N=13 Incidents)



Looking at the unadjusted correlations, it can be seen that time shows an alternating pattern of perfect correlations. The first repeating positive 1.0 correlation is for lag 2, which is the exact periodicity that was specified in the example. This offender repeats the time sequence every other time. Thus, if the individual alternates between committing offenses 2 and 4 days after the last, then knowing the time interval for the last offense, it can be assumed that the next event will repeat the next-to-the-last time interval.

For distance, the highest correlation is for a lag of 3. This offender repeated himself/herself every third time, which is exactly what was programmed into the example. Knowing the location of the last event, it can be assumed that the individual will choose the same distance for the next interval as three earlier. Finally, all lags show a perfect 1.0 correlation for bearing. The lowest one is taken, which is a lag of 1. That is, this individual repeats the direction every single time (i.e., he/she always travels in the same direction). In summary, the CWA - Correlogram shows that the individual repeats the time interval every other time, the distance every third time, and the direction every time.

The CWA - Diagnostics routine merely confirms these correlations. The regression equations yield an R^2 of 1.0 (unadjusted) for each of three variables, for the appropriate lag. For example, Table 12.7 below shows the regression results for distance for a lag of 3

Table 12.7:
Regression Results for Serial Offender 1: Distance

Variable:	distance	Standard error of estimate:	0.00000
Multiple R:	1.00000	Squared multiple R:	1.00000

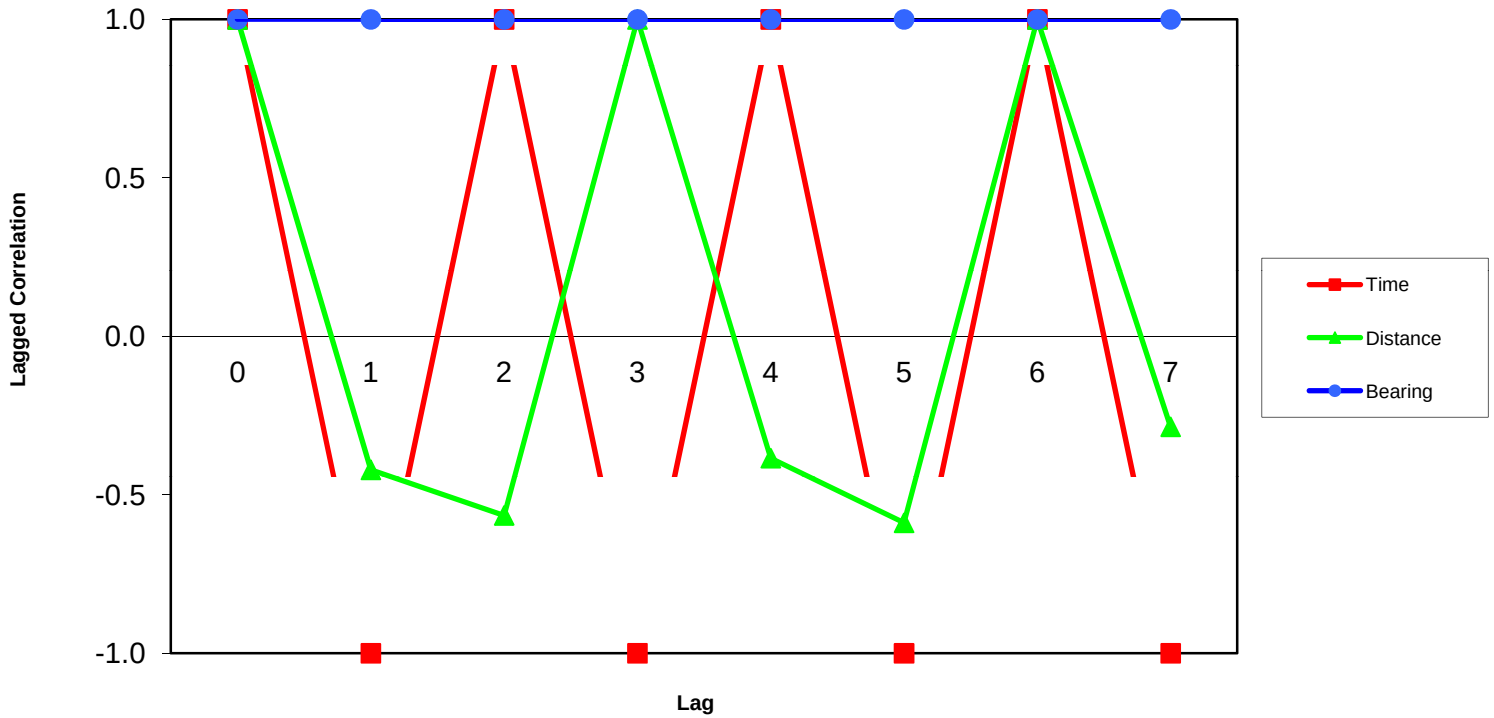
	Coefficient	***Std Error***	***t***	***p(2 Tail)***
Constant	0.000000	0.00000	0.00000	0.00000
Coefficient	1.000000	0.00000	0.00000	0.00000

Analysis of Variance

Source	Sum-of-Squares	df	Mean-Square	F-ratio	P
Regression	12.00000	1	12.00000	0.00000	0.00000
Residual	0.00000	8	0.00000		
Total	12.00000	9			

Figure 12.5:

Correlogram of a Predictable Offender: 1



The adjusted CWA - Correlogram shows a similar pattern, though the absolute correlations have been reduced. The best decision would still be for a lag of 2 for time, a lag of 3 for distance, and a lag of 1 for bearing. Figure 12.5 shows a graph of the correlogram. *CrimeStat* has a built-in graph function for the CWA - Correlogram and CWA-Adjusted correlogram.

Example 1: Prediction

Finally, for prediction, it is apparent that the best method would be to use a regression equation with lags of 2 for time, 3 for distance, and 1 for bearing. Table 12.8 shows the output. As can be seen, the routine predicts exactly the next time and location. The next event for this completely predictable serial offender will be on day 39 at the location with coordinates X=26, Y=26.

Table 12.8:
Predicted Results for Serial Offender 1
(Regression Equation with Lags of 2 for Time, 3 for Distance, 1 for Bearing)

Variable	Predicted Value	From Event	Method	Lag
Time interval	2.00000	13	Regression	2
Distance interval	1.41421	13	Regression	3
Bearing interval	44.99997	13	Regression	1
Predicted time	39.00000			
Predicted X coordinate	26.00000			
Predicted Y coordinate	26.00000			

The regression equation is the best model in this case. The other methods produce reasonably close approximations, however. Table 12.9 shows the results of using other methods for prediction. As seen, a model where all three components (time, distance, bearing) were lagged by 1 as well as a model where all three components were lagged by 3 also produces the expected correct answer. The mean interval and median interval methods also produce reasonably close, though not exact, answers. In this particular case, the regression method with the best lags produced the optimal solution.

Table 12.9:
Comparison of Methods for Predictable Serial Offender 1

	EVENT	X	Y	DISTANCE	DAYS	TIME INTERVAL
Logical Prediction for next event	**14**	**26**	**26**	**1.4142**	**39**	**2**
PREDICTION:						
Mean (lag=1)	14	27.0	27.0	2.8	40.0	3.0
Median (lag=1)	14	27.0	27.0	2.8	41.0	4.0
Regression:						
Lag=1	14	26.6	26.6	2.3	39.0	2.0
Lag=2	14	27.0	27.0	2.9	39.0	2.0
Lag=3	14	26.0	26.0	1.4	39.0	2.0
Optimal (t=2,d=3,b=1)	14	26.0	26.0	1.4	39.0	2.0

Example 2: Another Completely Predictable Individual

A second example is also a perfectly predictable individual. This time, the directional component changes from event to event. The directional trend is northward, but with changes in angle every third event. The time pattern is completely consistent with subsequent events occurring every two days. Table 12.10 presents the pattern and the logical next event while figure 12.6 displays the pattern.

The CWA - Correlogram reveals that both distance and bearing repeat themselves every third event while the time interval is repeated every time. The regression diagnostics show that there is perfect predictability for time and for distance, and high predictability for bearing (not shown). Finally, a regression model is used for prediction with lags of 1 for time, 3 for distance, and 3 for bearing. The model correctly predicts the expected time (days=25) and location (X=3, Y=25). Table 12.11 shows the results.

Table 12.10:

Example of a Predictable Serial Offender: 2

(N = 14 incidents)

Time Event	X	Y	Distance	Days	Interval
1	3	1	-	1	-
2	1	3	2.8284	3	2
3	1	5	2.0000	5	2
4	3	7	2.8284	7	2
5	1	9	2.8284	9	2
6	1	11	2.0000	11	2
7	3	13	2.8284	13	2
8	1	15	2.8284	15	2
9	1	17	2.0000	17	2
10	3	19	2.8284	19	2
11	1	21	2.8284	21	2
12	1	23	2.0000	23	2
Logical prediction for next event	**13**	**3**	**25**	**2.8284**	**25**	**2**

Methodology for CWA

These two examples illustrate what the CWA routine is doing. There are three steps. First, the sequential pattern is analyzed with the CWA - Correlogram. This analysis shows which lags have the strongest correlations between lags for time, distance, and bearing separately. Second, the pattern is tested with a regression model. The purpose is to determine how strong a relationship can be obtained for any particular model. As will be suggested below, if a model is too weak or, conversely, too strong, it most likely will not predict very well. Third, a prediction model is selected. The user can utilize the regression model or use the mean interval or median interval. Fourth, and finally, the prediction is made.

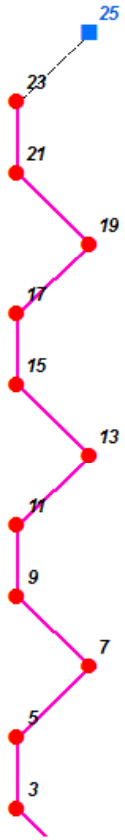
Example 3: A Real Serial Offender

How well does the CWA routine work with real serial offenders? People are not as predictable as these examples. The examples are algorithmic and people don't work like

Figure 12.6:

Example of a Predictable Serial Offender: 2

(N=12 Incidents)



Day of Incident Shown

Table 12.11:
Comparison of Methods for Predictable Serial Offender 2

	EVENT	X	Y	DISTANCE	DAYS	TIME INTERVAL	DIRECTION
Logical Prediction for next event	**13**	**3**	**25**	**2.8284**	**25**	**2**	**45**
PREDICTION:							
Mean (lag=1)	13	2.2	25.2	2.5	25.0	2.0	28.6
Median (lag=1)	13	3.0	25.0	2.8	25.0	2.0	45.0
Regression:							
Lag=1	13	3.0	25.0	2.8	25.0	2.0	45.0
Lag=2	13	1.9	25.2	2.4	25.2	2.0	22.5
Lag=3	13	3.0	25.0	2.8	25.0	2.0	45.0
Optimal (t=1,d=3,b=3)	13	3.0	25.0	2.8	25.0	2.0	45.0

algorithms. But, to the extent to which there is some predictability in human behavior, the CWA routine can be a useful tool for crime analysis, detection, and apprehension.

To illustrate this, a serial offender was identified from a large data set obtained from Baltimore County. The individual committed 16 offenses between 1992 and 1997 when he was eventually apprehended. The profile of crimes committed by this individual were quite diverse. There were 11 larceny incidents (shoplifting and bicycle theft), 1 residential burglary, 1 commercial burglary, 2 assaults, and 1 robbery.

To test the model, the first 15 incidents were used to predict the 16^{th}. This allowed the error between the observed and predicted values for time and location to be used for evaluation. Figure 12.7 shows the sequencing of actions of the first 15 incidents committed by this individual, most of which occurred in the eastern part of Baltimore County.

The CWA - Correlogram revealed a complicated pattern (Figure 12.8). The adjusted matrix was used because of the high correlations at higher-order lags. Nevertheless, the optimal lags appeared to be 1 for time, 3 for distance, and 6 for bearing. A regression model was used to test these parameters. Figure 12.7 also shows the predicted location for the next likely location (the red plus sign) and the location where the individual actually committed the 16^{th} event (green triangle). The error in prediction was good. The distance between the actual and predicted

Figure 12.7:

Likely Location for Next Crime Serial Offender in Baltimore County

N=16 Incidents

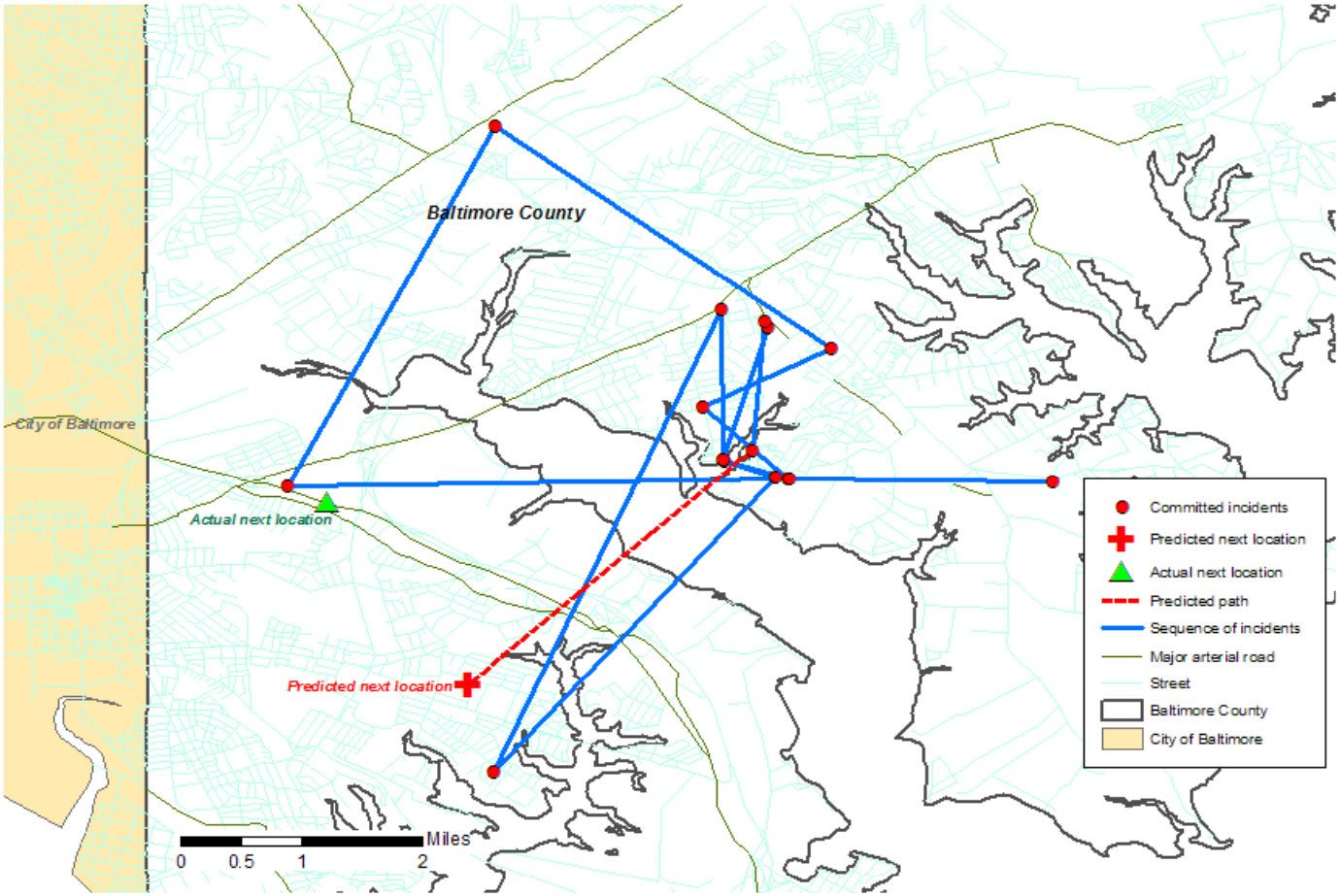
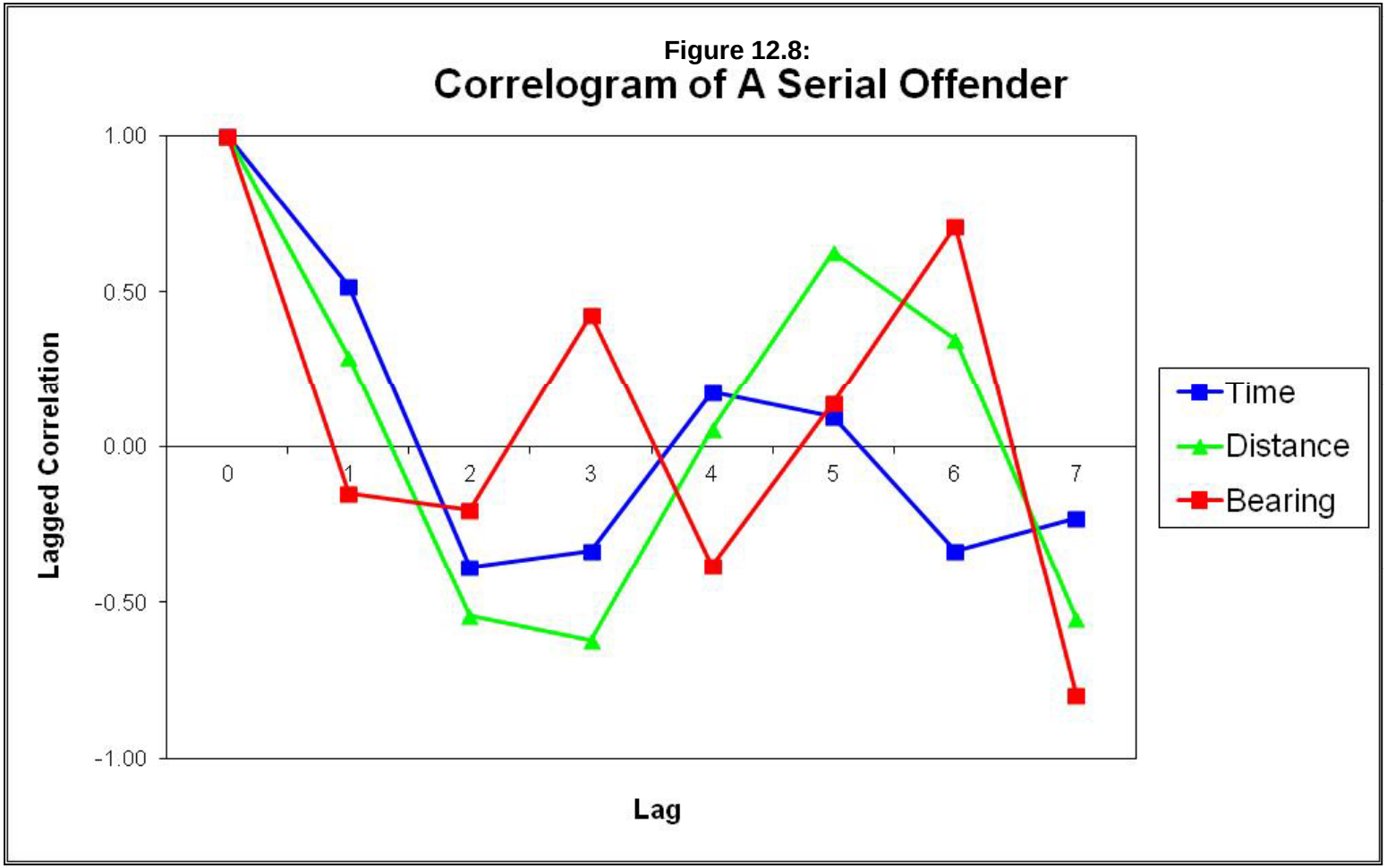


Figure 12.8:


Correlogram of A Serial Offender
1.00
0.50
0.00
-0.50
-1.00
Lagged Correlation
0
1
2
3
4
5
6
7
Lag
Time
Distance
Bearing

locations was 1.8 miles and the error in predicting the time of the next location was 3.9 days. Overall, the model did quite well for this individual.

Event Sequence as an Analogy to a Correlated Walk

Nevertheless, there are problems in the model for this case. First, this is not a true sequence of actions, but a pseudo-sequence. The individual doesn't go from the first event to the second event to the third event, and so forth. A considerable time may elapse between events. Similarly, distance and direction are conceptual only, not real. For example, in figure 12.7, the individual did not actually travel across the inlets of the Chesapeake Bay as the lines indicate. Distance between the events was actually much greater than estimated by the model and direction was more complex. Nevertheless, to the extent to which an individual makes a spatial decision about where to go, implicitly he or she is making a directional and distance decision. In other words, the decision making process may take into account prior locations. In this case, the CWA routines would be useful.

Example 4: A Second Real Serial Offender

A second real example confirms that the method can produce reasonably close predictions. An offender committed 13 crimes, including three incidents of shoplifting, eight incidents of theft from a vehicle, one residential burglary, and one highway robbery. The correlogram showed that a lag of 1 was strongest for time, distance, and bearing (figure 12.9). The R-squares were moderate (0.45 for time; 0.18 for distance; 0.18 for bearing). Using the regression method with a lag of 1 for each component, the likely location of the next event was predicted (Figure 12.10). The error between the predicted event and the actual event was, again, reasonable with a difference in time of 3.3 days and a difference in distance of 2.4 miles.

Accuracy of Predictions

However, it is important not to be overly optimistic about the technique. It is always possible to find cases that fit a method very well. The above mentioned cases appear to do that. Unfortunately, the method is not a magic elixir for predicting serial offenders. Like any method, it has error. It is also a fairly new tool in crime analysis so that we do not have a lot of experience with it. One example of its use was by Helms (2005), who was also is cautious about its utility.[2]

2 Personal communication from Dan Helms, National Law Enforcement Corrections and Technology Center, Denver, CO.

Figure 12.9:

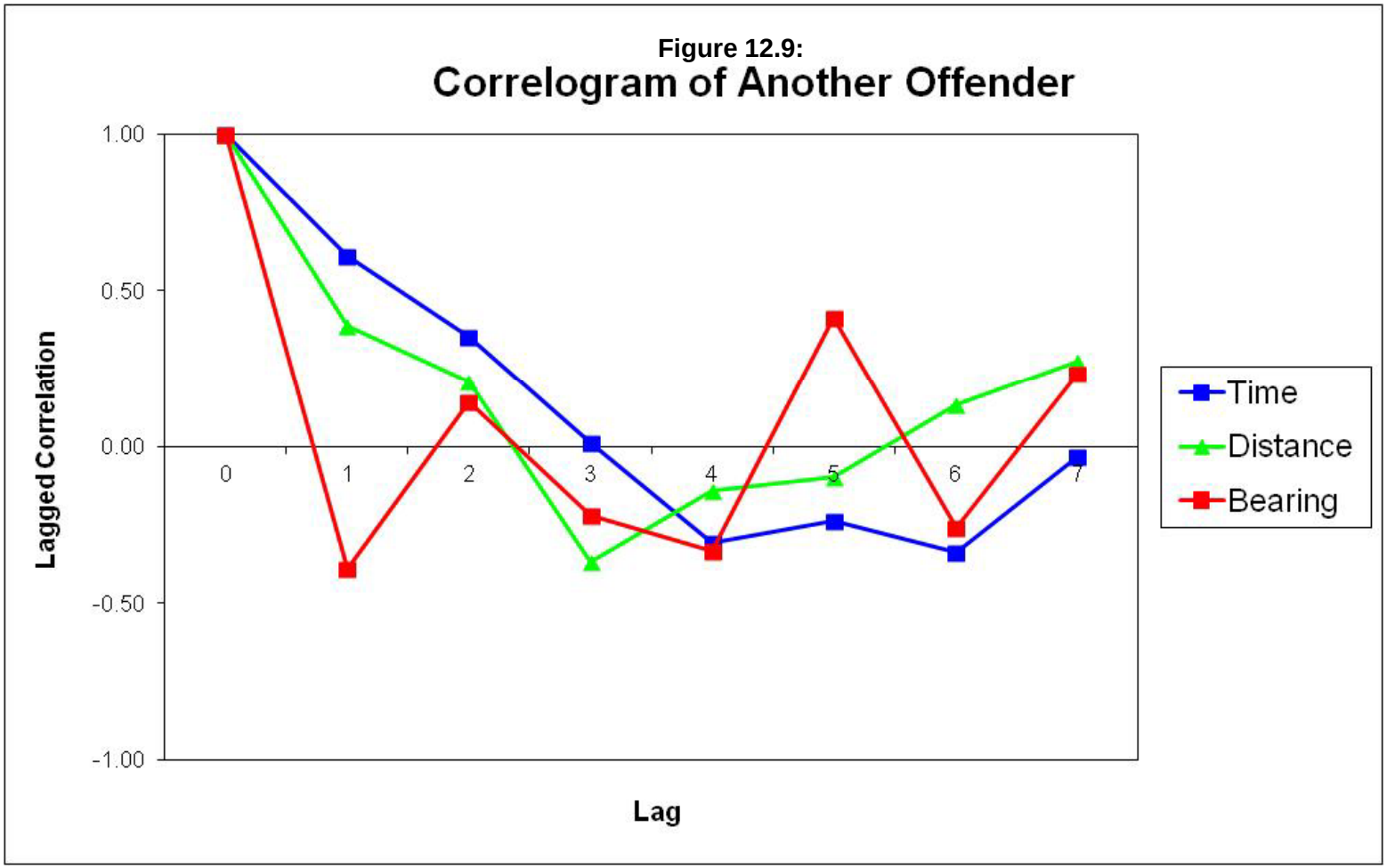
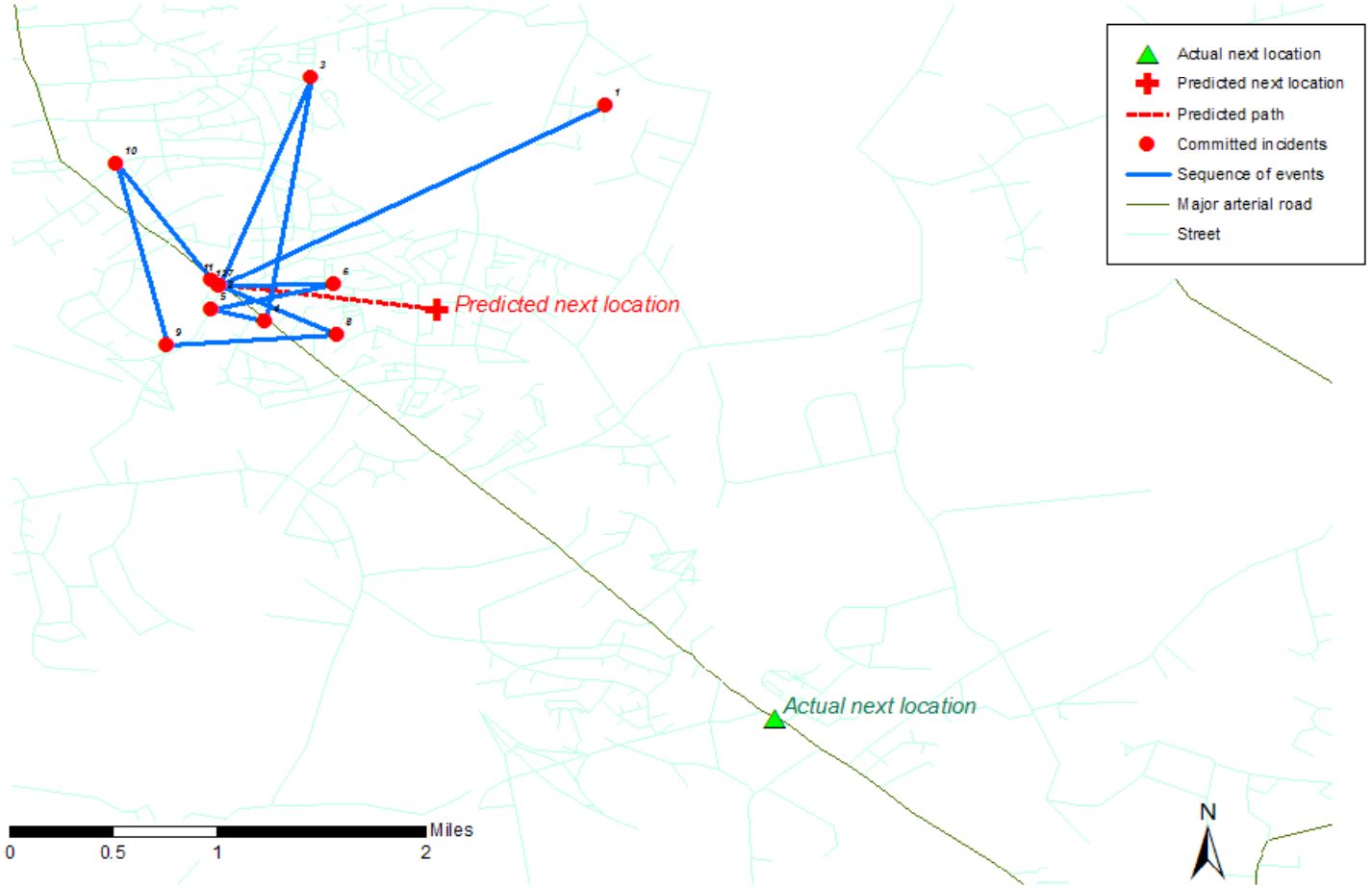

Correlogram of Another Offender
1.00
0.50
0.00
-0.50
-1.00
Lagged Correlation
0
1
2
3
4
5
6
7
Lag
Time
Distance
Bearing

Figure 12.10:

Likely Location for Next Crime
Another Serial Offender in Baltimore County
N=13 Incidents



To explore the accuracy of the method, 50 serial offenders were identified from a large data base of more than 41,000 incidents in Baltimore County between 1993 and 1997 (see Chapter 10 for details). The 50 offenders were identified based on knowing the dates on which they committed crimes, or at least on which they committed crimes for which they were charged and eventually tried. The number of incidents varied from a low of 7 incidents to a high of 38 incidents. An attempt was made to produce balance in the number of incidents, though the actual distribution of cases did reflect the availability of candidates in the data base. For the fifty individuals, the distribution of incidents was 7 (five individuals), 8 (four individuals), 9 (six individuals), 10 (two individuals), 11 (five individuals), 12 (five individuals), 13 (six individuals), 14 (three individuals), 15 (six individuals), 17 (two individuals), and one individual each for 20, 21, 24, 29 and 38 incidents.

To test the CWA model, the last event committed by these individuals was removed so that N-1 events could be used to predict event N. In this way, it is possible to evaluate the accuracy of the method.

Ten methods were compared:

1. The optimal regression method for time with the lag having the strongest relationship being selected;
2. The optimal regression method for location (distance and bearing) where the with the lags for distance and bearing having the strongest relationship being selected;
3. A regression model for time with a lag of 1;
4. A regression model for location with a lag of 1 (for both distance and bearing);
5. The mean interval for time;
6. The mean interval for location (distance and bearing);
7. The median interval for time;
8. The median interval for location (distance and bearing);
9. The mean center of the incidents (for location only); and
10. The center of minimum distance of the incidents (for location only).

The latter two methods were used for reference. For journey to crime estimation, the center of minimum distance is the best at predicting the origin location of serial offenders (see Chapter 13). The reason is because this statistic *minimizes the distance* to all incident locations. The mean center was close behind, though not quite as good. As an estimate, the center of minimum distance is a very good index when there is a single origin that is being predicted. On the other hand, where the purpose is to predict the location of a next event, the center of minimum distance and mean center may be less than useful since they will not generally predict the actual next location. They minimize error, but are rarely accurate. For example, in the above mentioned cases (two theoretical and two real), these statistics did not predict accurately the location of the

next event. Instead, they identified a point in the middle of the distribution where the sum of the distances to all incident locations was small.

Error analysis

Each of the models was compared to the actual time and location of the last, removed incident. For time, the error measure was in days (the absolute difference between the actual day and the predicted day). For location, the error measure was in miles (i.e., absolute distance between the actual and predicted location). The results were mixed. Overall, error was moderate. Table 12.12 summarizes the overall error.

Overall, the center of minimum distance and the mean center do produce, as expected, smaller errors for distance than any of the CWA methods; as noted above, locations in the middle of the distribution of incidents will minimize error, but they will not predict accurately the location of a next event nor indicate in which direction it will occur from the last event. On the other hand, the CWA methods are not particularly accurate, either. They work very well for a completely predictable offender, as was seen in the examples above, but not necessarily for real offenders.

Among the CWA methods, the mean interval, median interval and the lag 1 regression appears to give better results for time than the optimal regression. Overall, the median interval produces the lowest median error, which is about a month and half. In terms of location, the mean interval and median intervals produce slightly better results than the optimal regression, though the lag 1 regression was just as good.

Comparison of CWA Methods

At this point, it is unclear as when it is best to use this technique. Three variables seem to explain part of the error variation.

First, a larger sample size leads to better prediction, as would be expected (Table 12.13). For time, there is definitely an improvement in predictability with larger sample sizes. Among these methods, the mean interval and lag 1 regression show the smallest error for the largest samples (14 cases). For distance, on the other hand, generally, the error increases with increasing sample size. The one exception is for the optimal regression method where medium-sized samples (10-13 cases) produce the lowest error.

Table 12.12:
Average and Median Error for CWA Methods
(50 Serial Offenders)

Method	Average Error	Median Error
Time (days)		
Optimal regression: time	112.2	79.8
Lag 1 regression: time	88.1	70.0
Mean interval: time	89.7	64.9
Median interval: time	91.2	45.5
Distance (miles)		
Optimal regression: location	6.4	5.4
Lag 1 regression: location	5.7	4.2
Mean interval: location	5.8	4.7
Median interval: location	5.3	3.9
Reference Location (miles)		
Mean center	3.3	1.7
Center of minimum distance	3.1	1.2

Factors Affecting Predictability

Long time span

There are a variety of reasons for these results, but one reason may be the time span of the events. Some of these offenders committed crimes over a long period, up to five years. Sample size is intrinsically related to the time span (r=0.55). The longer the time span that an offender commits crimes, the more incidents he/she will perpetrate. With increasing time, the individual's behavior patterns may change (e.g., he/she may move residences).

For those offenders with many incidents, a separate analysis was conducted of the events occurring within the last year. Many of these individuals appeared to have moved their base of operation over time, so the isolation of the most recent events was done in order to produce a clearer behavior pattern. The results, while promising, were not dramatic. Accuracy was improved a little compared to using the full sequence, particularly spatial accuracy. However, even with the last few events, these frequently occurred over a long time period (up to two years).

Consequently, the idea of isolating a 'clean' set of events did not materialize, at least with these data. On the other hand, with a data set of only recent events, it may be possible to improve predictability.

Table 12.13:
Sample Size and Prediction Error
(Average Error)

Time (days)

Sample Size	Optimal Regression	Lag 1 Regression	Mean Interval	Median Interval
6-9	143.4	108.5	116.4	120.8
10-13	108.2	86.8	83.4	79.5
11+	79.8	65.1	65.7	71.2

Distance (miles)

Sample Size	Optimal Regression	Lag 1 Regression	Mean Interval	Median Interval
6-9	7.4	5.2	5.0	4.4
10-13	5.5	6.0	5.7	5.5
11+	6.1	5.9	6.8	6.1

Centrographic: Distance (miles)

Sample Size	Mean Center	Center of Minimum Distance
6-9	2.9	2.4
10-13	2.9	3.1
11+	4.3	4.1

Strength of predictability

A second variable that appears to have an effect is the strength of predictability, based on the first N-1 cases. For the diagnostics routine, as the overall R-square for the regression equation increases, the regression equation does better. However, with very high R-square coefficients, the error is worse. Table 12.14 shows the relationship.

The lowest error is obtained with moderate R-square coefficients, for both time and distance. This is why one has to be careful with very high lagged correlations in the correlogram and high R-squares in the diagnostics. Unless one is dealing with a perfectly predictable individual (as the two theoretical examples illustrated), high correlations may be a result of a very small sample size, rather than any inherent predictability.

Table 12.14:
Regression Diagnostics and Prediction Error
Comparison of CWA Regression Methods

	Time (days)		*Distance (miles)*	
R-Square	**Optimal Regression**	**Lag 1 Regression**	**Optimal Regression**	**Lag 1 Regression**
0-0.29	93.7	90.9	6.7	6.3
0.30-0.59	89.3	33.8	6.0	5.0
0.60+	164.3	122.7	6.3	5.2

Limitations of the Technique

In short, users should be careful about using the CWA technique. It can be useful for identifying repeating patterns by an offender, but it won't necessarily predict accurately the offender's next actions. There are a variety of reasons for the lack of predictability. First, there may be intermediate events that are unknown. With each of these offenders in the Baltimore County data base, there is always the possibility that the individuals committed other crimes for which they were not charged. The sequential analysis assumes that all the events are known. But this may not be the case.

A simulation on several cases was conducted by removing events and then re-running the correlogram and prediction models. Removing one event did not appreciably alter the relationship, but removing more than one event did. In other words, if there are unknown events, the true sequential behavior pattern of the offender may not be properly identified. Considering that most offenders commit fewer than 10 incidents before they get caught, the statistical effect of missing information may be critical.

A second reason has been alluded to already. In applying the model to crime events, it is not a true sequential model, but a *pseudo-sequential* model since much time may intervene between events. Distance and direction are conceptual in the sense that the individual doesn't directly orient from one event to the other, but returns to his/her living patterns. Thus, what may

appear to be a repeating pattern may not be. Here, the issue of sample size is critical. If there are only a few incidents on which to base an analysis, one could see a pattern which actually doesn't exist. One has to be careful about drawing inferences from very small samples.

A third reason is that people are inherently unpredictable. The two algorithmic examples produced excellent results, but few persons are that systematic about their behavior. Therefore, we must be cautious in expecting too much out of the model.

Conclusion

Nevertheless, the model has utility. First, it can help police identify whether there is a pattern in an offender's behavior. Knowing that there is a pattern can help in planning an arrest strategy. Even if the strategy does not pay off every time, it may improve police effectiveness. In short, the CWA can help a police department analyze the sequential behavior of an offender they are trying to catch. They may be able to anticipate a new event and may be able to warn people who are more likely to be attacked by this individual. If used carefully, the model can be useful for crime analysis and detection.

Second, it can encourage the development of additional predictor tools for individuals. As mentioned above, the center of minimum distance produces a 'best guess' estimate in the sense that it minimizes the distance to the next event. It usually doesn't predict the next event, but it does produce a minimal error. If used in conjunction with the CWA, it may be possible to narrow the search area for the next event.

Third, the CWA model can stimulate research into crime prediction. Police are always trying to predict the next event by an offender and will use multiple techniques and a lot of intuition in trying to 'out-guess' an offender. It is hoped that the CWA model will stimulate more research into predicting the sequence of offender behavior as well into how those sequences aggregate into a large spatial pattern. Most of this text has been devoted to analyzing the spatial patterns of a large number of events. The statistics have, perhaps naively assumed that each of those events were independent. In reality, they are not since many crimes are committed by the same individuals. In theory, a distribution of crime incidents could be disaggregated into a distribution of *sequences of events* committed by the same offenders, if we had enough information. Understanding how aggregate distributions is a by-product of the behavior of a limited number of individuals is an important research goal that needs to be addressed.

References

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

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

Besag, J. & Newell, J. (1991). The detection of clusters in rare diseases. Journal of the Royal Statistic Society A, 154, Part I, 143-155.

Chaitin, G. (1990). *Information, Randomness and Incompleteness* (second edition). World Scientific: Singapore.

Chen, A. & Renshaw, E. (1994) The general correlated random walk. *Journal of Applied Probability*, 31, 869-884.

Chen, A. & Renshaw, E. (1992). The Gillis-Domb-Fisher correlated random walk. *Journal of Applied Probability*, 29, 792-813.

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

Henderson, R., E., Ford, D., Renshaw, E. & Deans, J. D. (1983). Morphology of the structural root system of Sitka Spruce 1. Analysis and Quantitative Description. *Forestry*, 56 (2), 121-135.

Henderson, R., Renshaw, E., & Ford, D. (1984). A correlated random walk model for two-dimensional diffusion. Journal of Applied Probability, 21, 233-246.

Henderson, R., Renshaw, E., & Ford, D. (1983). A note on the recurrence of a correlated random walk. Journal of Applied Probability, 20, 696-699.

Knox, E. G. (1988). Detection of clusters. In Elliott, P. (ed), *Methodology of Enquiries into Disease Clustering*, London School of Hygiene and Tropical Medicine: London.

Knox, E. G. (1964). The detection of space-time interactions. *Applied Statistics*, 13, 25-29.

Knox, E. G. (1963). Detection of low intensity epidemicity: application in cleft lip and palate. *British Journal of Preventive and Social Medicine*, 18, 17-24.

References (continued)

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

Malkiel, B. G. (1999). *A Random Walk Down Wall Street* (revised edition). W. W. Norton & Company: New York.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27, 209-220.

Mantel, N. & Bailar, J. C. (1970). A class of permutational and multinomial test arising in epidemiological research, *Biometrics*, 26, 687-700.

Renshaw, E. (1985). Computer simulation of sitka spruce: spatial branching models for canopy growth and root structure. *Journal of Mathematics Applied in Medicine and Biology*, 2, 183-200.

Spitzer, F. (1976). *Principles of Random Walk* (second edition). Springer: New York.

Endnotes

i. Henderson, Renshaw and Ford (1983) defined the correlated walk as a two-dimensional walk where the sum of the probabilities in four directions along a lattice are:

$$P = p + q + 2r = 1$$

where *P* is the total probability (1), *p* is the probability of continuing in the same direction, *q* is the probability of moving in an opposite direction, and *r* is the probability of moving one unit to the right or to the left. The advantage of this formulation is that the probabilities do not have to be equal (i.e., p could exceed q or r). Nevertheless, the individual steps can be considered a special case of a correlated random walk in the plane (Henderson, 1981).

The non-lattice two dimensional case can also be considered a recurrent random walk since a step in any direction (not just along a lattice) can be considered the result of two steps, one in the X direction and one in the Y (or, alternatively, a pairing of all steps in the X direction with all steps in the Y direction). Unfortunately, this logic does not apply to more than two dimensions. Such multi-dimensional walks do not have to return to their origin. However, Spitzer (1963) has shown that an independent walk is recurrent if the second moment around the origin is finite.

Attachments

Tracking a Burglary Gang with the Correlated Walk Analysis

Bryan Hill
Glendale Police Department
Glendale, AZ

The space-time analysis tools provided with *CrimeStat* add an important element to an analyst's review of a tactical prediction effort. Although the method for calculating the Correlated Walk Analysis (CWA) is still more experimental than proven, it allows the analyst to see potential patterns in relation to a suspect's crime travel in terms of time, distance, and direction. In a recent burglary series involving several jurisdictions in our county, the CWA technique was used as part of an aggregate process referred to as the Probability Grid Method. That method combines results from several models to predict the next likely area for a new hit in a crime series. One of the most confusing aspects of these burglaries was the fact that several jurisdictions were involved and the offenders seemed to bounce back and forth from one jurisdiction to the next.

There were also 219 offenses in the series, providing considerable complexity. Because there were so many events, the distances could be anywhere from 0.5 miles to 20 miles, I could never really put my finger on what direction or distance the offender would hit next, but was confident a pattern existed and was likely changing over time. The following map shows the probability grid areas predicted and the CWA points predicted. The triangles shown represent the last four hits. The first hit was near the probability grid prediction in the northern portion of the map; however the subsequent hits were all very close to where the CWA routine predicted they would be. This was also a brand new area for these offenders and was a surprise to the department investigating these incidents. This area was not what was expected based on the SD ellipses and other methods used to predict the next event. The CWA tool requires more testing to determine the accuracy of its predictions, however it may turn out to be a valuable tool in a crime analyst's arsenal.