Chapter 5:

Spatial Autocorrelation Statistics

Ned Levine[1]
Ned Levine & Associates
Houston, TX

[1] The author would like to thank Dr. David Wong for help with the Getis-Ord 'G' and local Getis-Ord statistics.

Table of Contents

Table of Contents (continued)

Chapter 5:

Spatial Autocorrelation Statistics

This chapter discusses statistics for describing spatial autocorrelation that are applicable to zonal data. A good grasp of basic statistics is a requirement for reading this chapter. Figure 5.1 shows the Spatial Autocorrelation page within the Spatial Description section. This includes global tests of spatial autocorrelation for zone data or point data in which an attribute can be associated with the coordinates. The section includes six tests for global spatial autocorrelation:

1. Moran's "I" statistic
2. Geary's "C" statistic
3. Getis-Ord "G" statistic
4. Moran Correlogram
5. Geary Correlogram
6. Getis-Ord Correlogram

These indices would typically be applied to zonal data where an attribute value can be assigned to each zone. Six spatial autocorrelation indices are calculated. All require an intensity variable in the Primary File.

The discussion in the chapter will concentrate on defining the indices and demonstrating how they can be used. Specific instructions for running the routines are given at the end of the chapter while detailed information is provided in Chapter 2.

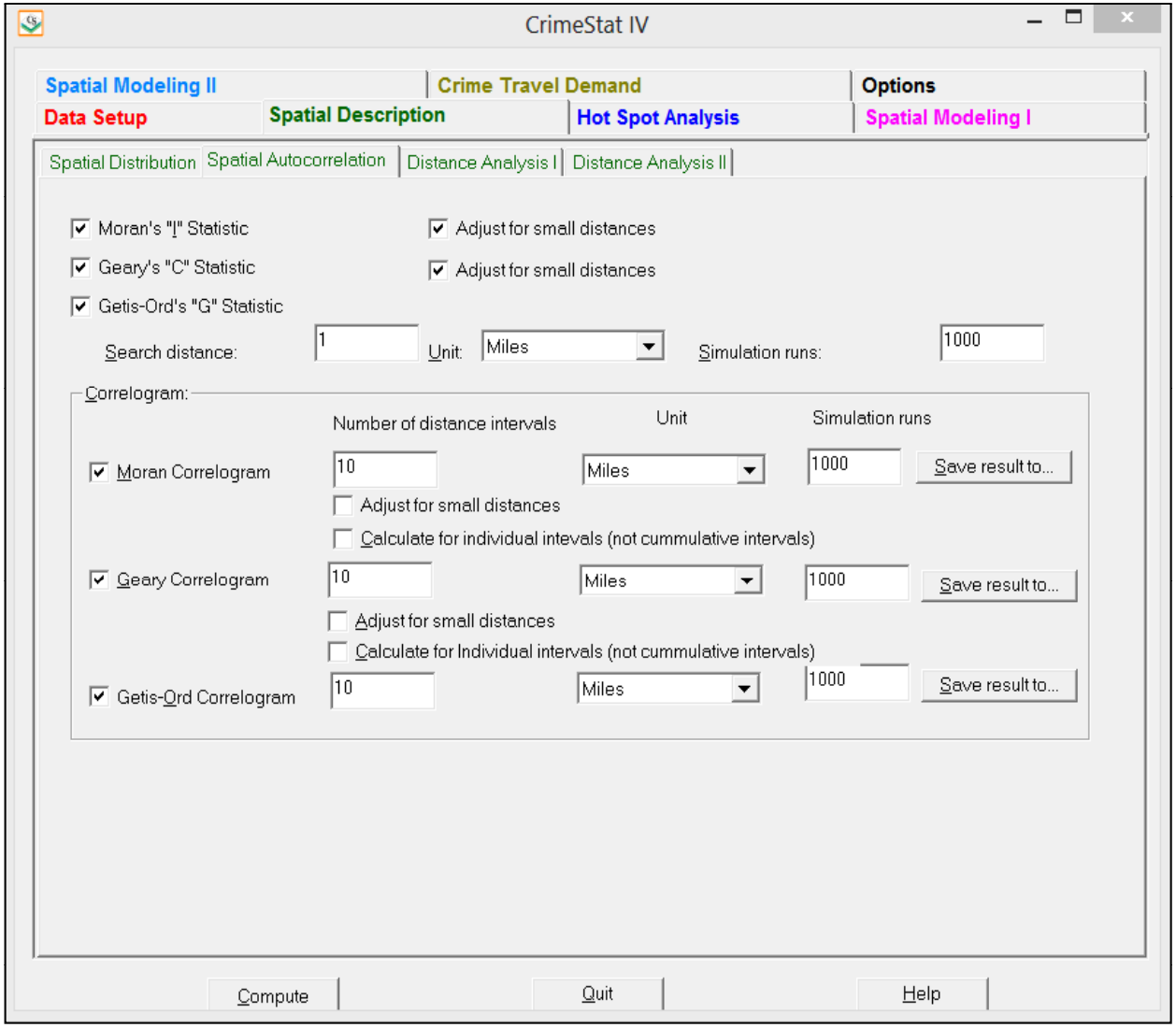
Spatial Autocorrelation

The concept of *spatial autocorrelation* is one of the most important in spatial statistics in that it implies a lack of spatial *independence*. Classical statistics assumes that observations are independently chosen and are spatially unrelated to each other. The intuitive concept is that the location of an incident (e.g., a street robbery, a burglary) is unrelated to the location of any other incident. The opposite condition - spatial autocorrelation, is a spatial arrangement of incidents such that the locations where incidents occur are related to each other; that is, they are not statistically independent of one another. In other words, spatial autocorrelation is a spatial arrangement where spatial independence has been violated.

When events or people or facilities are clustered together, we refer to this arrangement as *positive* spatial autocorrelation. Conversely, an arrangement where people, events or facilities

Figure 5.1:

Spatial Autocorrelation Statistics



are extremely dispersed is referred to as *negative* spatial autocorrelation; it is a rarer arrangement, but does exist (Levine, 1999).

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

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

Statistically, however, non-spatial independence indicates that many statistical tools and inferences are inappropriate. For example, the use of a correlation coefficient or Ordinary Least Squares regression (OLS) model to predict a consequence (e.g., correlates or predictors of burglaries) assumes observations are randomly selected. If, however, the observations are spatially clustered, the estimates obtained from the correlation coefficient or OLS estimator will be biased and overly precise. The coefficients will be biased because areas with a higher concentration of events will have a greater impact on the model estimate and precision will be overestimated because concentrated events tend to have fewer independent observations than are being assumed. The spatial autocorrelation concept underlies almost all of *CrimeStat*'s spatial statistics tools.

Indices of Spatial Autocorrelation

Assigning Point Data to Zones

If a user has information on the location of individual events (e.g., robberies), then it is better to utilize that information with the point statistics discussed in Chapter 4 and the hot spot tools that will be discussed in Chapters 7 and 8. The individual-level information will contain all the uniqueness of the events.

However, sometimes it is not possible to analyze data at the individual level. The user may need to aggregate the individual data points to spatial areas (zones) in order to compare the events to data that are only obtained for zones, such as census data, or to model environmental correlates of the data points or may find that individual data are not available (e.g., when a police

department releases information by police beats but not individual streets). In this case, the individual data points are allocated to zones by, first, spatially assigning them to the zones in which they fall and, second, counting the number of points assigned to each zone. A user can do this with a GIS program or with the "Assign Primary points to Secondary Points" routine that will be discussed in Chapter 6.

In this case, the zone becomes the unit of analysis instead of the individual data points. All the incidents are assigned to a single geographical coordinate, typically the *centroid* of the zone, and the number of incidents in the zone (the count) becomes an *attribute* of the zone (e.g., number of robberies per zone; number of motor vehicle crashes per zone).

It should be obvious that when individual data points are assigned to zones, information is lost. Instead of capturing the unique locations of the individual events, all events that occur within a zone are assigned a single location. Thus, the distance between zones is a singular value for all the points in those zones whereas there is much greater variability with the distances between individual events.

Further, zones have attributes which are properties of the zone, not of the individual events. The attribute can be a *count* or a continuous variable for a distributional property of the zone (e.g., median household income; percentage of households below poverty level).[2]

Analysis then proceeds on the basis of the zonal information. The results will be different than for an analysis of the individual event information since the spatial characteristics are measured by single points for each zone (e.g., the centroid) and the attribute information is measured by a property of the zone, not the individual events (e.g., the count of events in the zone; a characteristic of the zone such as income level).

In other words, the user must realize that an analysis of zonal data is quite different from an analysis of individual data and that the conclusions might be different. Aggregating data to zones creates properties that may be different than those of individual events and that the relationships between variables at the zonal level also might be different than at the individual level. This is called an *ecological* relationship and there is a large literature on ecological inference and fallacies (see Freedman, 1999; Langbein & Lichtman, 1979).

Individual level data can also have attributes. For example, Levine and Lee (2013) analyzed journey-to-crime distances for offenders in Manchester, England. In this case, the attribute variable was the distance traveled and the statistics discussed in this chapter are

[2] There is no fundamental difference between a count variable and a continuous interval or ratio variable since a real number can be converted into a count by multiplying by a power of 10 (e.g., $1.23 = 123 \times 10^{-2}$). The statistics discussed in this chapter are applicable to either count or continuous data.

appropriate for analyzing that attribute data. Other examples of individual level data with attributes would be the age of the offender, the number of prior convictions, or the number of years of formal education. The key criterion is that the records must have an attribute which is either a count or an interval variable.

Spatial Autocorrelation Statistics for Attribute Data

There are a number of formal statistics that attempt to measure spatial autocorrelation at the zonal level or for individual level data with count or interval attributes. These statistics include simple indices, such as the Moran's I", Geary's C or the Getis-Ord "G" statistic, the application of these statistics to individual zones or records (discussed in Chapter 9), and multivariate indices such as the Markov Chain Monte Carlo spatial regression models (discussed in Chapter 19). The simple indices attempt to identify whether spatial autocorrelation exists for a single variable while the more complicated indices attempt to estimate variability in spatial autocorrelation in a study area of the effect of spatial autocorrelation on a particular attribute variable.

CrimeStat includes three global indices - Moran's I statistic, Geary's C statistic, and the Getis-Ord "G" statistic. It also includes *Correlograms* that apply each of these indices to different distance intervals. Moran, Geary, and Getis-Ord are *global* in that they represent a summary value for all the data points. In Chapter 9, we will present some local indicators of spatial autocorrelation that apply the Moran, Geary and Getis-Ord statistics to individual zones. But, for now, we are focused on describing the entire study area.

Moran's "I" Statistic

Moran's "I" statistic (Moran, 1950) is one of the oldest indicators of spatial autocorrelation. It is applied to zones or points that have attribute variables associated with them (intensities). For any continuous variable, X_i, a mean, $\bar{X}$, can be calculated and the deviation of any one observation from that mean, s_X, can also be calculated. The statistic then compares the value of the variable at any one location with the value at all other locations (Ebdon, 1988; Griffith, 1987; Anselin, 1992). Formally, it is defined as:

$$I = \frac{N \sum_i \sum_j W_{ij} \left(X_i - \bar{X}\right)(X_j - \bar{X})}{(\sum_i \sum_j W_{ij}) \sum_i (X_i - \bar{X})^2} \qquad (5.1)$$

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

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

Note that in adopting a distance-based weight, there are advantages and disadvantages. Contiguity (or adjacency) is a property of a zone, not a point. Thus, adjacency defines whether one zone is next to another zone whereas distance is the distance between single points that represent the zones (e.g., centroids). If two zones are, say, 0.25 miles apart, it is not known whether they are adjacent or not. In other words, in adopting a distance-based weight, information about adjacencies is lost. On the other hand, a distance-based weight is standardized. If two zones are adjacent, it is not known how far apart they are separated. Adjacencies can be misleading since they don't indicate the size of the adjacent zones whereas a specified distance is always constant.

The weighted Moran's I is similar to a correlation coefficient in that it compares the sum of the cross-products of values at different locations, two at a time, weighted by the inverse of the distance between the locations and with the variance of the variable. Like a correlation coefficient, it typically varies between -1.0 and + 1.0. However, this is not absolute as an example later in the chapter will show. When nearby points have similar values, their cross-product is high. Conversely, when nearby points have dissimilar values, their cross-product is low. Consequently, an "I" value that is high indicates more spatial autocorrelation than an "I" that is low.

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

$$E(I) = -\frac{1}{N-1} \tag{5.2}$$

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

Adjust for Small Distances

CrimeStat calculates the weighted Moran's I formula using equation 5.1. However, there is one problem with this formula that can lead to unreliable results. The distance weight between

two locations, W_{ij}, is defined as the reciprocal of the distance between the two points, consistent with Moran's original formulation:

$$W_{ij} = \frac{1}{d_{ij}} \tag{5.3}$$

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

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

$$W_{ij} = \frac{one\ mile}{one\ mile + d_{ij}} \tag{5.4}$$

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

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

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

Testing the Significance of Moran's "I"

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

$$Z(I) = \frac{I - E(I)}{S_{E(I)}} \tag{5.5}$$

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

There are several interpretations of the theoretical standard deviation that affect the particular statistic used for the denominator as well as the interpretation of the significance of the statistic (Anselin, 1992). The most common assumption is that the standardized variable, Z(I), has a sampling distribution which follows a standard normal distribution, that is with a mean of 0 and a variance of 1. This is called the *normality* assumption.[3] A second interpretation assumes that each observed value could have occurred at any location, that is the location of the values and their spatial arrangement is assumed to be unrelated. This is called the *randomization* assumption and has a slightly different formula for the theoretical standard deviation of 5.13.[4] *CrimeStat* outputs the Z-values and p-values for both the normality and randomization assumptions.

Example: Testing Houston Burglaries with Moran's "I"

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

Figure 5.2 shows 2006 burglaries in the City of Houston by individual Traffic Analysis Zones (TAZ). TAZ's are groupings of census blocks but designed to equalize the number of trips to and from the zone in the base year. They are typically very small in downtown Houston (typically a block in size) and much larger in the suburban parts of the City. With a GIS program, 26,480 burglary locations were overlaid on top of a map of 1,179 TAZ's and the number of burglaries within each TAZ were counted and then assigned to the TAZ as a variable (see the 'Assign primary points to secondary points' routine in Chapter 6).[5] The numbers varied from 0 burglaries (for 250 TAZ's) up to 284 burglaries incidents (for 1 TAZ). The map shows the plot of the number of burglaries per TAZ.

---

3 The theoretical standard deviation of "I" under the assumption of normality is (Ebdon, 1985):

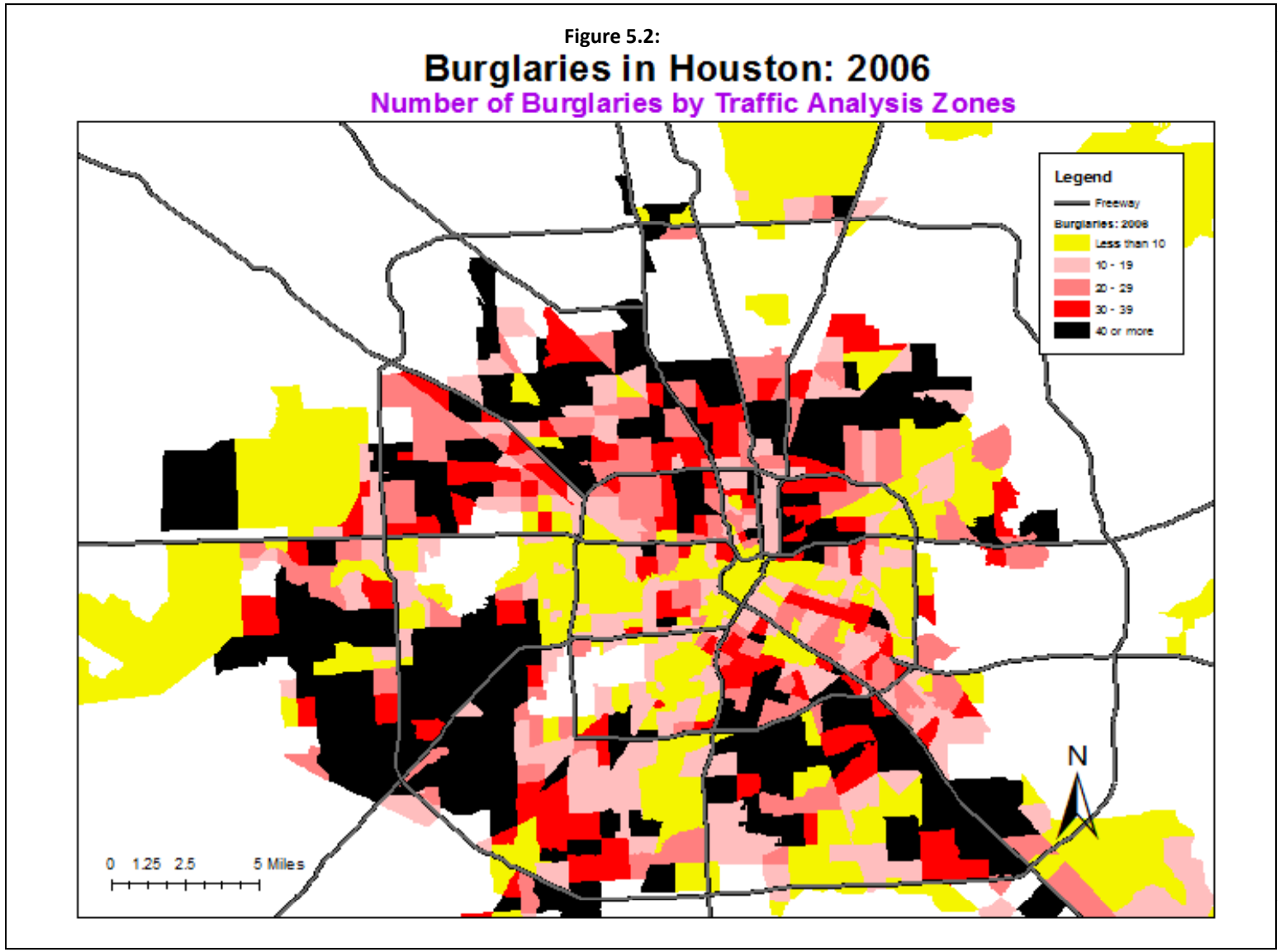
$$S_{E(I)} = \sqrt{\frac{N^2(\sum_i \sum_j W_{ij}^2)+3(\sum_i \sum_j W_{ij})^2 - N\sum_i(\sum_j W_{ij})^2}{(N^2-1)(\sum_i \sum_j W_{ij})^2}}$$

4 The formula for the theoretical standard deviation of "I" under the randomization assumption is (Ebdon, 1985):

$$S_{E(I)} = \sqrt{\frac{N[(N^2+3-3N)(\sum_I \sum_J W_{IJ}^2)+3(\sum_i \sum_j W_{ij})^2 - N\sum_i(\sum_j W_{ij})^2] - k[(N^2-N)\sum_i \sum_j W_{ij}^2+6(\sum_i \sum_j W_{ij})^2-2N(\sum_i(\sum_j W_{ij})^2]}{(N-1)(N-2)(N-3)(\sum_i \sum_j W_{ij})^2}}$$

5 The TAZ data were obtained from the Houston-Galveston Area Council, the Metropolitan Planning Organization for the Houston metro area.

Figure 5.2:


Burglaries in Houston: 2006
Number of Burglaries by Traffic Analysis Zones
Legend
Freeway
Burglaries: 2008
Less than 10
10 - 19
20 - 29
30 - 39
40 or more
N
0 1.25 2.5 5 Miles

Clearly, aggregating incident locations to zones, such as TAZ's, eliminates some information since all incidents within a block are assigned to a single location (the centroid of the block). The use of Moran's I, however, requires the data to be in this format. Using data in this form, Moran's I was calculated using the small distance adjustment because many TAZ's are very close together, especially in downtown Houston.

Figure 5.3 shows the output of the "I" in *CrimeStat.* "I" was 0.251790, the theoretical value of "I" as -0.000849, and the standard error of "I" as 0.002796. The test of significance using the normality assumption gave a Z-value of 213.20, a highly significant value. Below are the calculations for burglaries by TAZ:

$$Z\ (I_{veh}) = \frac{I_{veh} - E(I)}{S_{E(I)}} = \frac{0.251795 - (-0.000849)}{0.002796} = 213.20\ (p \leq .0001)$$

Comparing Moran's "I" for Two Distributions

Figure 5.4 shows the distribution of households in the city by TAZ. The calculations for the "*I*" of households are similar (not shown). It turns out that the "I" of households is 0.298117 while the theoretical "I" and the standard error of "I" are the same as for burglaries (because of the same zonal geography). One can compare an "I" value for one distribution with the "I" value for another distribution. For example, a Z-test can then be made of whether the "*I*" value of burglaries is statistically different than that of households. The calculations are shown below:

$$Z\left(I_{difference}\right) = \frac{I_{burg} - I_{hh}}{S_{E(I)}} = \frac{0.251795 - (0.298117)}{0.002796} = -16.57\ (p \leq .001)$$

where I_{burg} is the "I" value for burglaries, I_{hh} is the "*I*" value for households, and $S_{E(I)}$ is the standard deviation of "I" for households under the assumption of normality. The Z-test of the difference is -16.57, a highly significant difference. The high Z-value suggests that burglaries are even more clustered than the clustering of households. To put it another way, they are more clustered than would be expected based on the household distribution. As mentioned, this is an approximate test since the joint distribution of "I" for two empirical distributions of "I" is not known.

Geary's C Statistic

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

Figure 5.3:

Moran's I Statistic Output

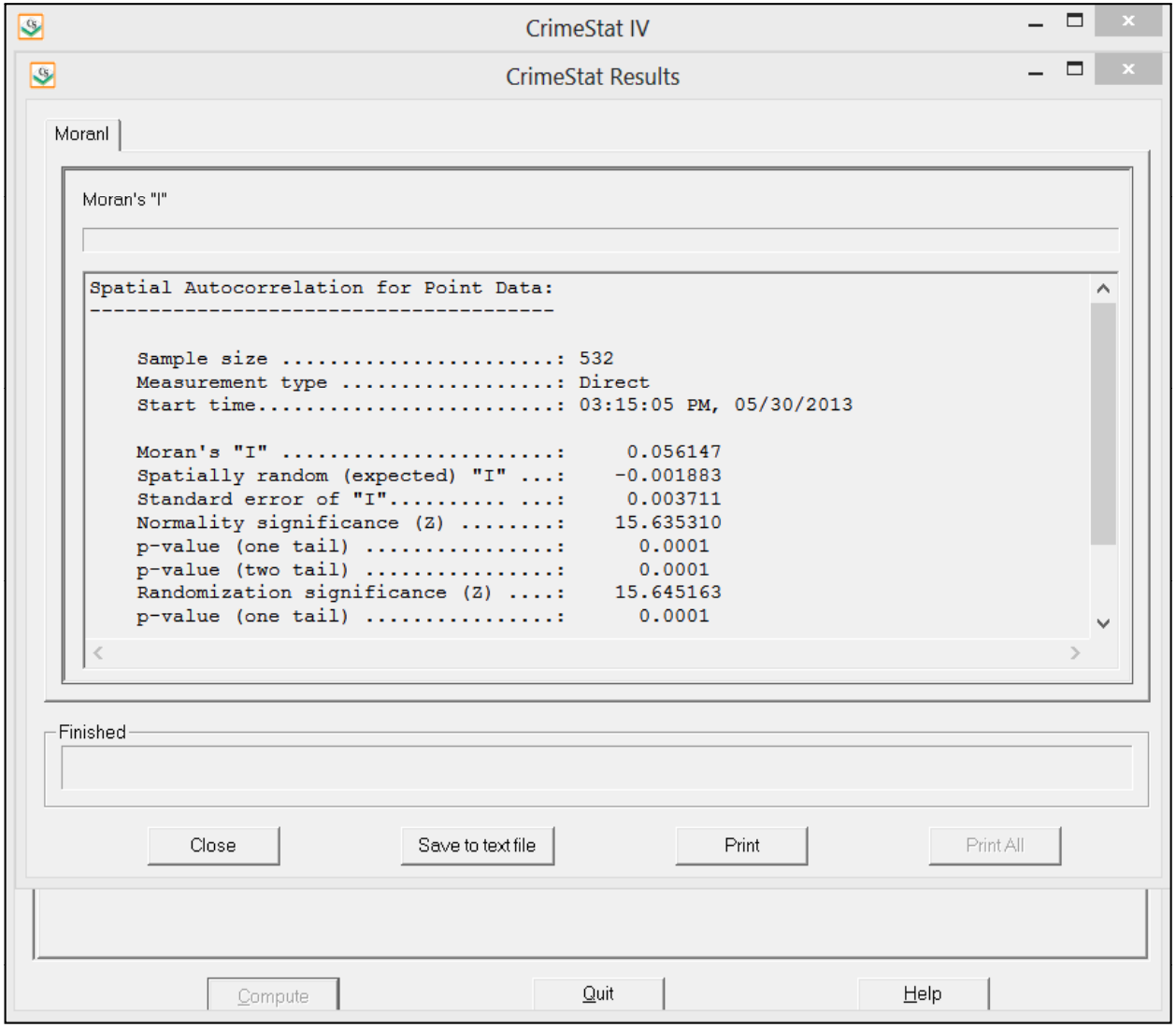
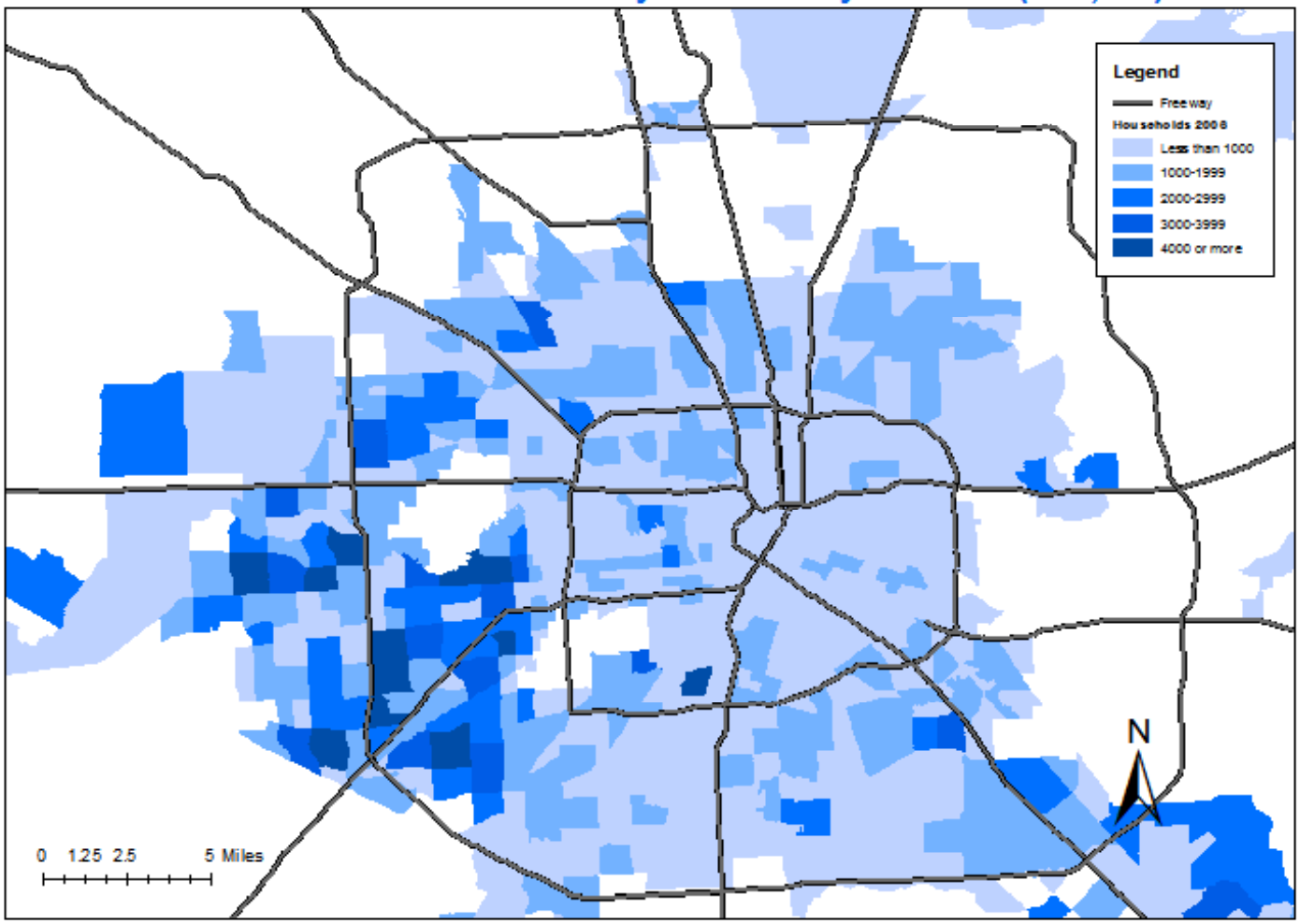


Figure 5.4:

Households in Houston: 2006

Number of Households by Traffic Analysis Zones (N=1,179)



$$C = \frac{(N-1)\sum_i \sum_j W_{ij}(X_i - X_j)^2}{2(\sum_i \sum_j W_{ij})\sum_i (X_i - \bar{X})^2} \quad (5.6)$$

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

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

Adjusted "C"

A more intuitive interpretation of *"C"* can be obtained by calculating an adjusted*"C"*:

$$Adjusted\ C = 1 - C \quad (5.7)$$

In this case, the adjusted *"C"* will be on the same scale as Moran's "I". An adjusted *"C"* value that is positive indicates positive spatial autocorrelation while an adjusted *"C"* value that is negative indicates negative spatial autocorrelation. An adjusted *"C"* of 0 indicates no spatial autocorrelation and is also the expected adjusted "C". *CrimeStat* calculates both the regular and adjusted "C" values.

Adjust for Small Distances

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

$$W_{ij} = \frac{1}{d_{ij}} \quad (5.3)\ \text{repeat}$$

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

$$W_{ij} = \frac{one\ mile}{one\ mile + d_{ij}} \qquad (5.4)\ \text{repeat}$$

The adjustment scales the distances to one mile in the distance units specified on the Primary file page (miles, feet, kilometers, meters, or nautical miles). This is the default condition although the user can calculate all weights as the reciprocal distance by turning off the small distance adjustment.

Testing the Significance of Geary's "C"

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

$$Z(C) = \frac{C - E(C)}{S_{E(C)}} \qquad (5.8)$$

where *C* is the empirical "C", *E(C)* is the theoretical mean of a random distribution and $S_{E(C)}$ is the theoretical standard deviation of E(C). The usual test is to assume that the sample Z follows a standard normal distribution with mean of 0 and variance of 1 (normality assumption), though it is possible to calculate the standard error under a randomization assumption (Ripley, 1981).[6]

Note that for testing, the regular "C" value should be used since an adjusted standard error of "C" is not easily calculated. The adjusted "C" is useful for a quick intuitive appraisal as well as for the Geary Correlogram (see below).

Example: Testing Houston Burglaries with Geary's "C"

Using the same data on burglaries in the City of Houston, figure 5.5 illustrates the output. The regular "C" value for burglaries was 0.625702 with a Z-value of -20.00 ($p \leq .0001$). The "C" value of burglaries is *smaller* than the theoretical "C" of 1. Converting this measure an adjusted "C" gives 0.374298 and indicates positive spatial autocorrelation. That is, the index suggests that TAZ's with a high number of burglaries are adjacent to TAZ's also with a high number of burglaries. Thus, Geary's C confirms the evidence for positive spatial autocorrelation identified by Moran's "I".

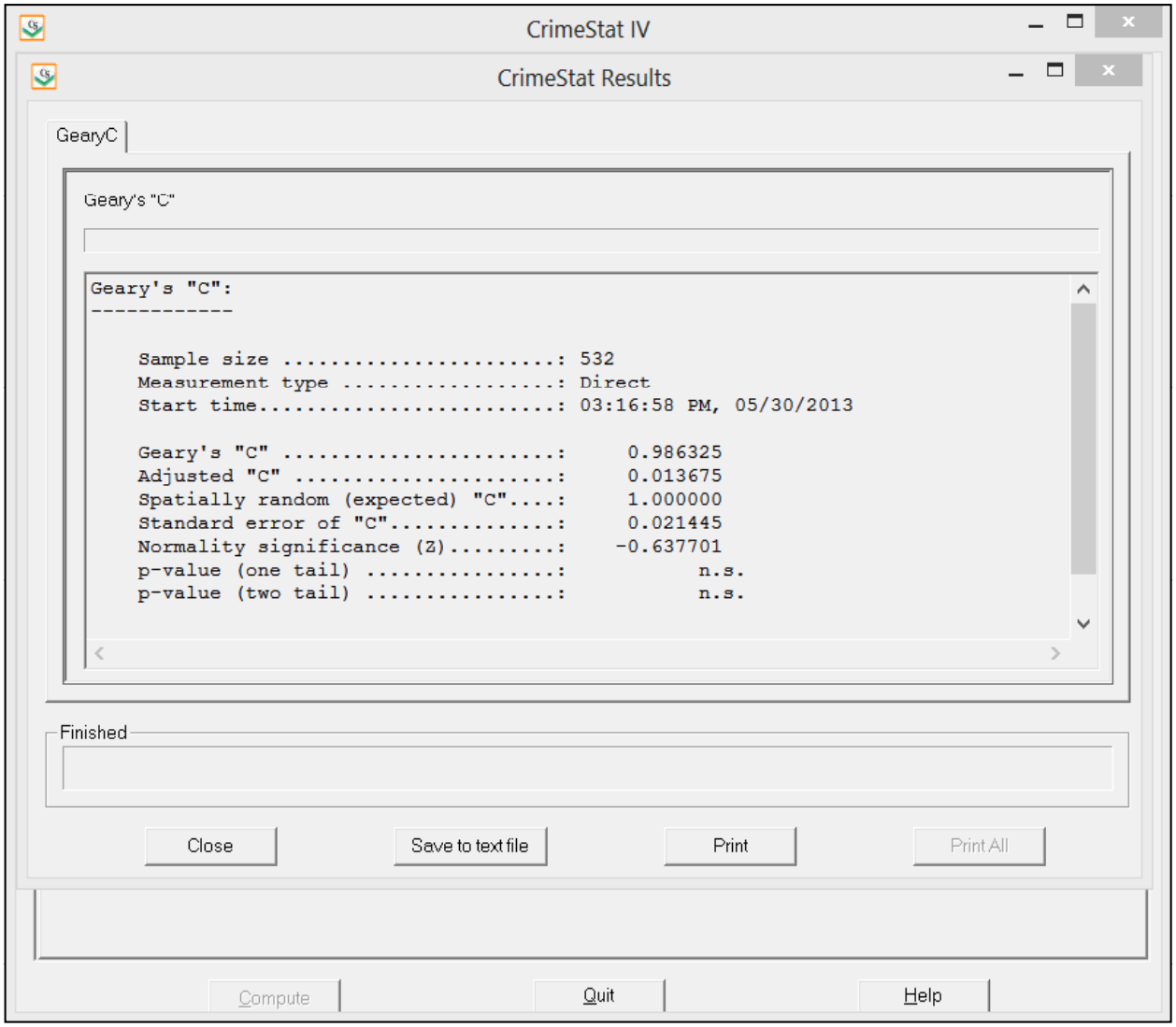
---

[6] The theoretical standard deviation for C under the normality assumption is (Ripley, 1981):

$$S_{E(I)} = \sqrt{\frac{2\sum_i \sum_j (W_{ij}^2) + \sum_i (\sum_j W_{ij})^2 (N-1) - 4(\sum_i \sum_j W_{ij})^2}{2(N+1)(\sum_i \sum_j W_{ij})^2}}$$

Figure 5.5:

Geary's C Statistic Output



Comparing this to the distribution of households in Houston, the C value of households is also below the theoretical "C" of 1 and points to positive spatial autocorrelation ("C" = 0.643120 with a Z-value of -19.07; p≤.0001). Since both the regular "C" for burglaries and for households are below 1 (hence, indicating positive spatial autocorrelation), let us test the difference between the two as indicated by a Z-test of the difference.

$$\left(C_{difference}\right) = \frac{C_{burg} - C_{hh}}{S_{E(C)}} = \frac{0.625702 - (0.643120)}{0.018717} = -0.930598\ (n.s.)$$

In this case, there is no statistical difference between the distribution of burglaries and the distribution of households. Though both distributions show evidence of positive spatial autocorrelation, the Geary test cannot show a difference between the two whereas the Moran's "I" did show a difference.

Typically, Geary's "C" will be consistent with Moran's "I" though there are slight differences between the indices, as we see in this example. Because of the nature of the weighting, the Geary index is more sensitive to local clustering (second-order effects) than the Moran index, which is better seen as measuring first-order spatial autocorrelation. This illustrates how these indices have to be used with care and cannot be generalized by themselves. Each of them emphasizes slightly different information regarding spatial autocorrelation, yet neither is sufficient by itself. They should be used as part of a larger analysis of spatial patterning.[7]

Getis-Ord "G" Statistic

The Getis-Ord "G" statistic is also an index of global spatial autocorrelation but for values that fall within a specified distance of each other (Ord & Getis, 1995; Getis & Ord, 1992). When compared to an expected value of "G" under the assumption of no spatial association, it has the advantage over other two global spatial autocorrelation measures in that it can distinguish between 'hot spots' and 'cold spots', which neither Moran's "I" nor Geary's "C" can do.

The "G" statistic calculates the spatial interaction of the value of a particular variable in a zone with the values of that same variable in nearby zones, similar to Moran's "I" and Geary's

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

"C". Thus, it is also a measure of spatial association or interaction. Unlike the other two measures, it ***only*** identifies ***positive*** spatial autocorrelation, that is, where zones have similar values to their neighbors. It cannot detect negative spatial autocorrelation where zones have different values to their neighbors. But, unlike the other two global measures, it can distinguish between positive spatial autocorrelation where zones with high values are near to other zones with high values (***high positive spatial autocorrelation***) from positive spatial autocorrelation which where zones with low values are near to other zones also with low values (***low positive spatial autocorrelation***). Further, the "G" value is calculated with respect to a specified search distance (defined by the user) rather than to an inverse distance, as with the Moran's "I" or Geary's "C".

The formulation of the general "G" statistic presented here is taken from Lee and Wong (2005). It is defined as:

$$G(d) = \frac{\sum_i \sum_j W_j(d) X_i X_j}{\sum_i \sum_j X_i X_j} \tag{5.9}$$

for a variable, X. This formula indicates that the cross-product of the value of X at location "i" and at another zone "j" is weighted by a distance weight, $w_j(d)$ which is defined by either a '1' if the two zones are equal to or closer than a threshold distance, d, or "0" otherwise. The cross-product is summed for all other zones, j, over all zones, i. Thus, the numerator is a sub-set of the denominator and can vary between 0 and 1. If the distance selected is too small so that no other zones are closer than this distance, then the weight will be 0 for all cross-products of variable X. Hence, the value of G(d) will be 0. Similarly, if the distance selected is too large so that all other zones are closer than this distance, then the weight will be 1 for all cross-products of variable X. Hence, the value of G(d) will be 1.

There are actually two "G" statistics. The first one, G*, includes the interaction of a zone with itself; that is, zone "i" and zone "j" can be the same zone. The second one, G, does not include the interaction of a zone with itself. In *CrimeStat*, we only include the "G" statistic (i.e., there is no interaction of a zone with itself) because, first, the two measures produce almost identical results and, second, the interpretation of "G" is more straightforward than with G*. Essentially, with G, the statistic measures the interaction of a zone with nearby zones (a 'neighborhood'). See articles by Getis and Ord (1996) and by Khan, Qin and Noyce (2006) for a discussion of the use of G*.

Testing the Significance of "G"

By itself, the "G" statistic is not very meaningful. Since it can vary between 0 and 1, as the threshold distance increases, the statistic will always approach 1.0. Consequently, "G" is compared to an expected value of "G" under no significant spatial association. The expected "G" for a threshold distance, d, is defined as:

$$E[G(d)] = \frac{W}{N(N-1)} \tag{5.10}$$

where W is the sum of weights for all pairs and N is the number of cases. The sum of the weights is based on *symmetrical* counts of those zones within the threshold distance. That is, if zone 2 is within the threshold distance of zone 1, then zone 2 contributes a weight of 1 to zone 1. However, zone 1 contributes a weight of 1 to zone 2 as well. In other words, if two zones are within the threshold (search) distance, then they both contribute 2 to the total weight.

Note that, since the expected value of "G" is a function of the sample size and the sum of weights which, in turn, is a function of the search distance, it will be the same for all variables of a single data set in which the same search distance is specified. However, as the search distance changes, so will the expected "G" change.

Theoretically, the "G" statistic is assumed to have a normally distributed standard error. If this is the case (and we often do not know if it is), then the standard error of "G" can be calculated and a simple significance test based on the normal distributed be constructed. The variance of G(d) is defined as:

$$Var[G(d)] = E(G^2) - E(G)^2 \tag{5.11}$$

where

$$E(G)^2 = \frac{1}{(m_1^2 - m_2)^2 n^4}[B_0 m_2^2 + B_1 m_4 + B_2 m_1^2 m_2 + B_3 m_1 m_3 + B_4 m_1^4 \tag{5.12}$$

and where:

$$m_1 = \sum_i X_i \tag{5.13}$$
$$m_2 = \sum_i X_i^2 \tag{5.14}$$
$$m_3 = \sum_i X_i^3 \tag{5.15}$$
$$m_4 = \sum_i X_i^4 \tag{5.16}$$
$$N^4 = N(N-1)(N-2)(N-3) \tag{5.17}$$

$$S_1 = 0.5 \sum_i \sum_j (W_{ij} + W_{ji})^2 \quad (5.18)$$

$$S_2 = \sum_i (\sum_J W_{ij} + \sum_J W_{ji})^2 \quad (5.19)$$

$$B_0 = (N^2 - 3N + 3)S_1 - NS_2 + 3W^2 \quad (5.20)$$

$$B_1 = -[(N^2 - N)S_1 - 2NS_2 + 6W^2] \quad (5.21)$$

$$B_2 = -[2NS_1 - (N + 3)S_2 + 6W^2] \quad (5.22)$$

$$B_3 = 4(N - 1)S_1 - 2(N + 1)S_2 + 8W^2 \quad (5.23)$$

$$B_4 = S_1 - S_2 + W^2 \quad (5.24)$$

where ***i*** is the zone being calculated, ***j*** is all other zones, and *N* is the sample size (Lee and Wong, 2005). Note that this formula is different than that written in other sources (e.g., see Lees, 2006) but is consistent with the formulation by Getis and Ord (1992; 1993).

The standard error of G(d) is the square root of the variance of G. Consequently, a Z-test can be constructed by:

$$S.E.[G(d)] = \sqrt{Var[G(d)]} \quad (5.25)$$

$$Z[G(d)] = \frac{G(d) - E[G(d)]}{S.E.[G(d)]} \quad (5.26)$$

Relative to the expected value of G, a positive Z-value indicates spatial clustering of high values (high positive spatial autocorrelation or 'hot spots') while a negative Z-value indicates spatial clustering of low values (low positive spatial autocorrelation or 'cold spots'). A "G" value around 0 typically indicates either no positive spatial autocorrelation, negative spatial autocorrelation (which the Getis-Ord cannot detect), or that the number of 'hot spots' more or less balances the number of 'cold spots'.

Note that the value of this test will vary with the search distance selected. One search distance may yield a significant spatial association for "G" whereas another may not. In other words, the statistic is useful for identifying distances at which spatial autocorrelation exists.

> In practice, one should use a small search distance to identify local spatial autocorrelation.

Also, and this is an important point, the expected value of "G" as calculated in equation 5.10 is only meaningful if the variable is positive. For variables with negative values, such as residual errors from a regression model, one cannot use equation 5.10 but, instead, must use a simulation to estimate confidence intervals.

Simulating Confidence Intervals for "G"

One of the problems with this test is that "G" may not actually follow a normal standard error. That is, if "G" was calculated for a specific distance, d, with random data, the distribution of the statistic may not be normally distributed. This would be especially true if the variable of interest is a skewed variable with some zones having very high values while the majority of zones having low values.

Consequently, the user has an alternative for estimating the confidence intervals using a Monte Carlo simulation. In this case, a *permutation* type simulation is run whereby the original values of the intensity variable, Z, are maintained but are randomly re-assigned for each simulation run (Anselin, 2008). This will maintain the distribution of the variable Z but will estimate the value of "G" under random assignment of this variable. The user can take the usual 95% or 99% confidence intervals based on the simulation.

> Keep in mind that a simulation may take time to run especially if the data set is large or if a large number of simulation runs are requested.

Example: Testing Simulated Data with the Getis-Ord "G"

To understand how the Getis-Ord "G" works and how it compares to the other two global spatial autocorrelation measures - Moran's "I" and the adjusted Geary's "C", three simulated data sets were created. In the first, a random pattern was created (Figure 5.6). In the second, a data set showing positive spatial autocorrelation was created (Figure 5.7) and, in the third, a data set showing negative spatial autocorrelation was created (Figure 5.8).

Table 5.1 compares the three global spatial autocorrelation statistics on the three distributions. For the Getis-Ord "G", both the actual "G" and the expected "G" are shown. A one mile search distance was used for the Getis-Ord "G".

Figure 5.6:

Random Distribution

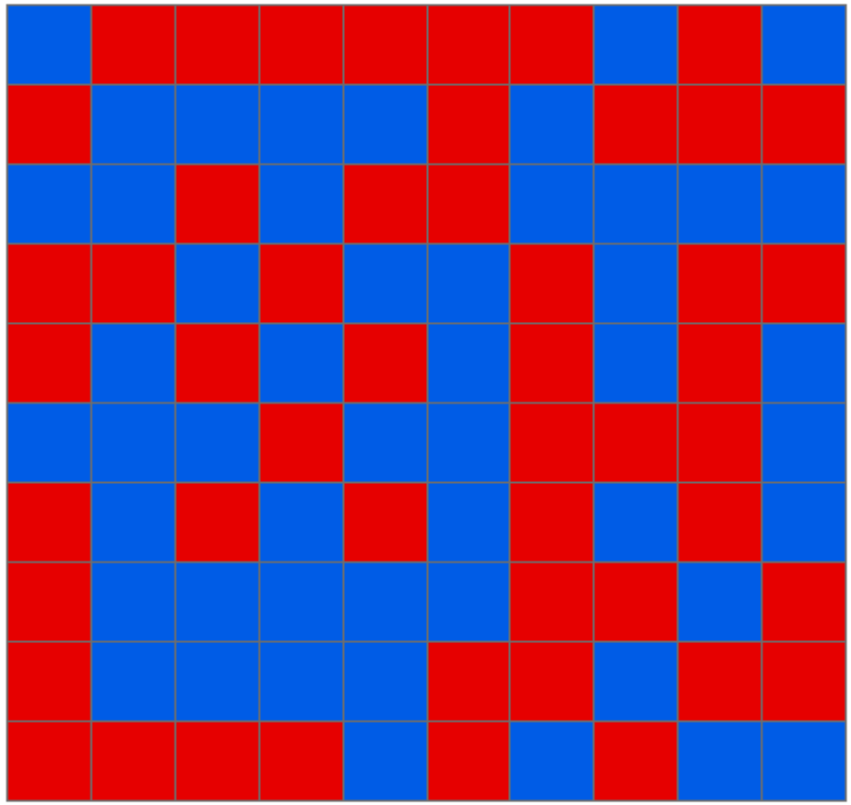


Figure 5.7:

Positive Spatial Autocorrelation

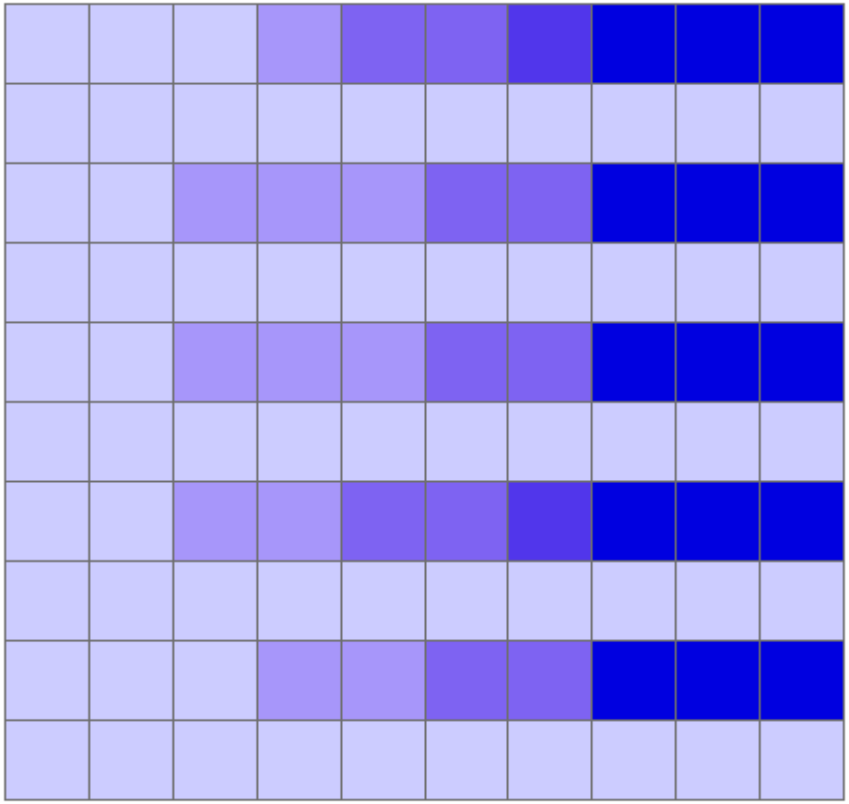


Figure 5.8:

Negative Spatial Autocorrelation

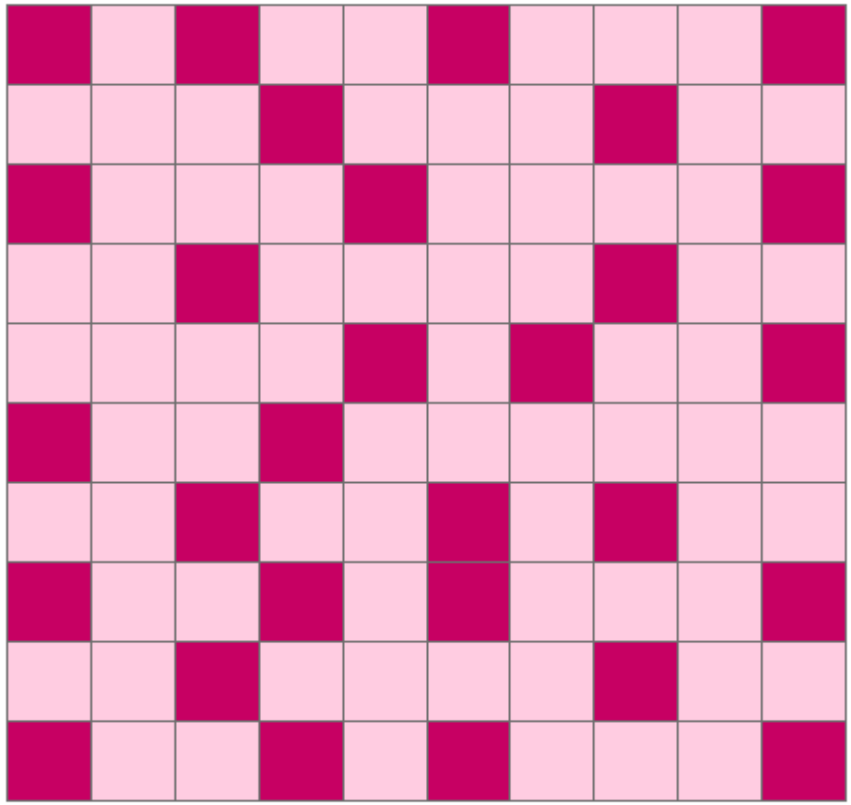


Table 5.1:
Global Spatial Autocorrelation Statistics for Simulated Data Sets
N = 100 Grid Cells

			----------*Getis-Ord "G"*----------	
Pattern	**Moran's "I"**	**Adjusted Geary's "C"**	**Observed "G"** (1 mi search)	**Expected "G"** (1 mi search)
Random	-0.007162 $^{n.s.}$	0.034722$^{n.s}$	0.151059$^{n.s}$	0.159596
Positive spatial autocorrelation	0.061449****	0.166734****	0.230586***	0.159596
Negative spatial autocorrelation	-0.041719*	-0.016479$^{n.s.}$	0.140833$^{n.s.}$	0.159596

n.s	not significant
*	p≤.05
**	p≤.01
***	p≤.001
****	p≤.0001

The random pattern is not significant for all three measures. That is, neither the Moran "I", the adjusted Geary"C", nor the Getis-Ord "G" are significantly different than the expected value under a random distribution. This is what would be expected since the data were assigned randomly.

For the positive spatial autocorrelation pattern, on the other hand, all three measures show highly significant differences with a random distribution. Moran's "I" is highly positive. The adjusted Geary's "C is above 0, indicating positive spatial autocorrelation and the Getis-Ord "G" has a "G" value that is significantly higher than the expected "G" based on the theoretical standard error. The Getis-Ord "G", therefore, indicates that the type of spatial autocorrelation is high positive.

Finally, the negative spatial autocorrelation pattern (Figure 5.8 above) shows different results for the three measures. Moran's "I" shows negative spatial autocorrelation and is significant (p≤.05). Geary's "C also shows negative spatial autocorrelation but is not significant. Finally, the Getis-Ord "G" is slightly smaller than the expected "G", which indicates low positive spatial autocorrelation, but it is not significant.

In other words, all three statistics can identify positive spatial correlation. Of these, Moran's "I" is a more powerful test than either Geary's "C" or the Getis-Ord "G". By 'power' is meant the ability to correctly reject a false null hypothesis (or, in statistical language, to avoid a Type II error). A data set for which Moran's "I" is barely statistically significant might very well fail with Geary's "C" or the Getis-Ord "G" since the Geary and Getis-Ord indices are not a powerful as the Moran index.

However, only Moran's "I" and Geary's "C" are able to detect negative spatial autocorrelation, the latter barely. On the other hand, only the Getis-Ord "G" can distinguish between high positive and low positive spatial autocorrelation. The Moran and Geary tests would show these conditions to be identical, as the example below shows.

Example: Testing Houston Burglaries with the Getis-Ord "G"

Now, let us take the 26,480 burglaries in the City of Houston for 2006 aggregated to 1,179 traffic analysis zones (figure 5.2 above). To compare the Getis-Ord "G" statistic with the Moran's "I" and the regular Geary's "C", the three spatial autocorrelation tests were run on this data set. The Getis-Ord "G" was tested with a search distance of 1 mile and 1000 simulation runs were made on the "G". Table 5.2 shows the three global spatial autocorrelation statistics for these data.

The Moran and Geary tests show that the Houston burglaries have significant positive spatial autocorrelation (zones have values that are similar to their neighbors). Moran's "I" is significantly higher than the expected "I" and the adjusted Geary's "C" is also significantly higher than the adjusted expected "C". However, the Getis-Ord "G" is lower than the expected "G value and is significant whether using the theoretical Z-test or the simulated confidence intervals (notice how the "G" is lower than the 2.5 percentile). This indicates that, in general, zones with low values are nearby other zones with low values. In other words, there is low positive spatial autocorrelation, suggesting a number of 'cold spots'.

Uses and Limitations of the Getis-Ord "G"

The advantage of the "G" statistic over the other two spatial autocorrelation measures is that it can distinguish between 'hot spots' and 'cold spots'. With Moran's "I" or Geary's "C", an indicator of positive spatial autocorrelation means that zones have values similar to their neighbors. However, the positive spatial autocorrelation could be caused by many zones with low values being concentrated, too. In other words, one cannot tell from those two indices whether the concentration is a hot spot or a cold spot. The Getis-Ord "G" can do this.

Table 5.2:

Global Spatial Autocorrelation Statistics for City of Houston Burglaries: 2001

N = 1,179 Traffic Analysis Zones

	Moran's "I"	Adjusted Geary's "C"	Getis-Ord "G"
			(1 mile search)
Observed	0.251790	0.374298	0.007063
Expected	-0.000849	0.000000	0.061753
Observed –Expected	0.252639	0.374298	-0.054690
Standard Error	0.002796	0.018717	0.007491
Z-test	90.36	20.00	-7.30
p-value	****	****	****
Based on simulation:			
2.5 percentile:	---	---	0.048664
97.5 percentile:	---	---	0.076445

n.s	not significant
*	p≤.05
**	p≤.01
***	p≤.001
****	p≤.0001

The main limitation of the Getis-Ord "G" is that it cannot detect negative spatial autocorrelation, a condition that, while rare, does occur. With the haphazard pattern above (Figure 5.8), this test could not detect that there was negative spatial autocorrelation. For this condition, Moran's "I" or possibly Geary's "C" would be more appropriate tests. In the example, Geary's "C" did not detect it but Moran's "I" did

Moran Correlogram

Moran's "I", Geary's "C", and the Getis-Ord "G" indices are summary tests of global autocorrelation. That is, they summarize all the data with respect to spatial autocorrelation but do not distinguish different subsets. For examining particular sub-sets of data that are spatially autocorrelated, such as 'hot spots', 'cold spots' or space-time clusters, a different approach is required. Chapter 9 discusses the local Moran and local Getis-Ord statistics.

An alternative approach is to calculate the spatial autocorrelation statistics by different distance intervals. The *Moran* Correlogram calculates the "I" value by different distance intervals (or bins). When graphed, the plot indicates how concentrated or distributed is the spatial autocorrelation (Cliff and Haggett, 1988; Bailey and Gatrell, 1995). Essentially, a series of concentric circles is overlaid on the points and the Moran's I statistic is calculated for only

those points falling within each circle. The radius of the circle changes from a small circle to a very large one. As the circle increases, the "I" value approaches the global value.

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

Adjust for Small Distances

If the 'Adjust for small distances' box is checked, small distances are adjusted so that the maximum weighting is 1 (equation 5.4 above). This ensures that the "I" values for individual distances will not become excessively large or excessively small for points that are close together. The default value is no adjustment.

Simulation of Confidence Intervals

A permutation Monte Carlo simulation can be run to estimate approximate confidence intervals around the "I" value. Each simulation inputs random data and calculates the "I" value. The distribution of the random "I" values produce an approximate confidence interval for the actual (empirical) "I". To run the simulation, specify the number of simulations to be run (e.g., 100, 1000, 10000). The default is no simulations. The output percentiles are the 0.5^{th}, 2.5^{th}, 97.5^{th} and 99^{th}. Pairing the 2.5^{th} with the 97.5^{th} or the 0.5^{th} with the 99.will create approximate 95% or 99% confidence intervals.

Example: Moran Correlogram of Baltimore County Vehicle Theft and Population

For the three correlograms, we will use a different example than Houston burglaries. These are 1996 data on vehicle thefts from Baltimore County, MD. Figure 5.9 shows the distribution of 1996 vehicle thefts by Traffic Analysis Zones (TAZ) while figure 5.10 shows the Moran Correlogram for these thefts. Also shown in the graph are the maximum and minimum values from a Monte Carlo simulation of 1000 runs and the 2.5th and 97.5th percentiles to simulate approximate 95% confidence intervals (called 'credible intervals').

As seen, the "I" value at zero distance is about 0.60. As the distance between zones increase (i.e., the search circle radius gets larger), the "I" value drops off slowly until about 19 miles whereupon it approaches the global "I" value. Further, the curve for the "I" values is always higher than the 97^{th} percentile curve from the random simulation and indicating that vehicle thefts are more clustered than what would be expected on the basis of chance for all

Figure 5.9:

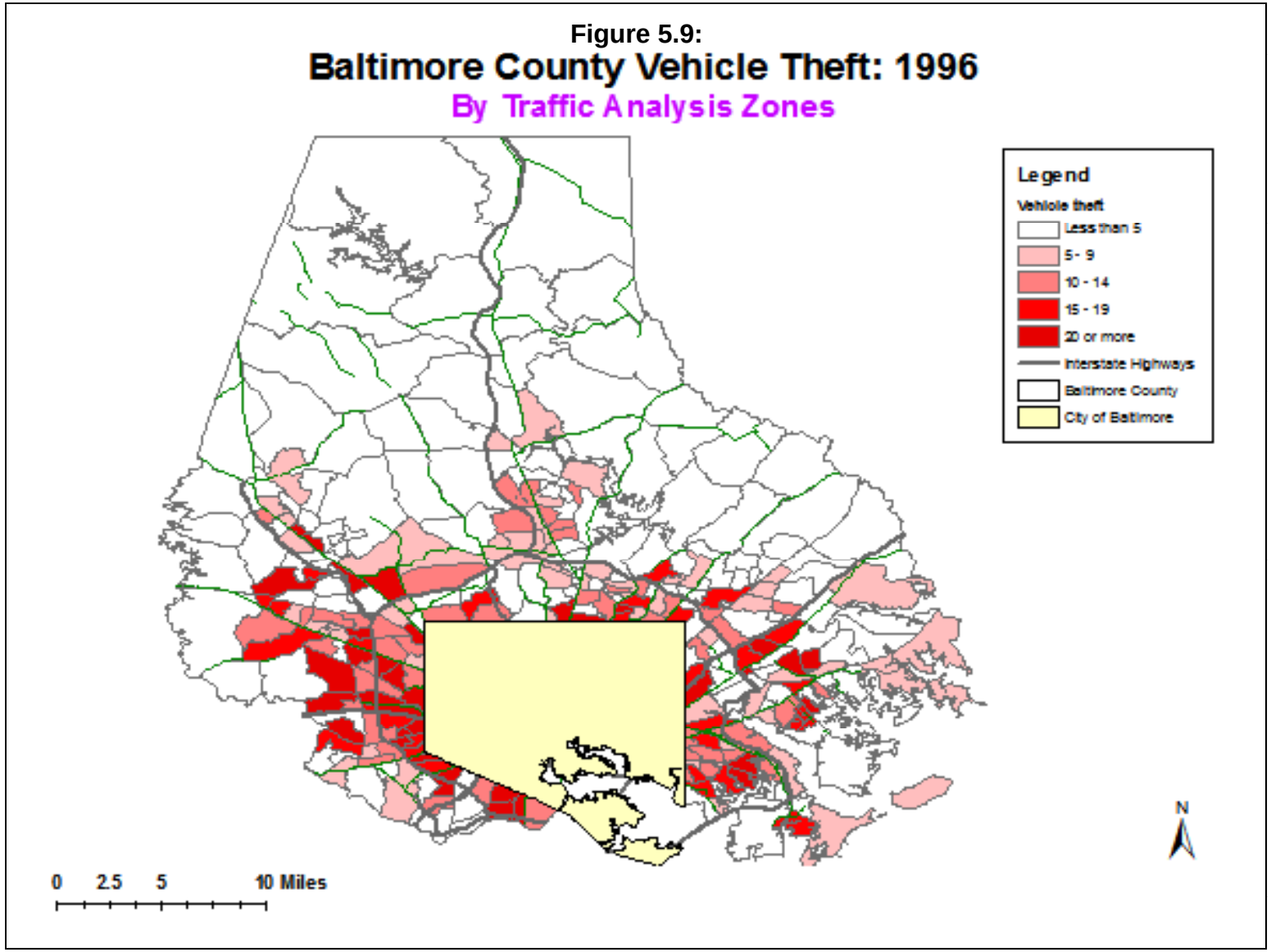
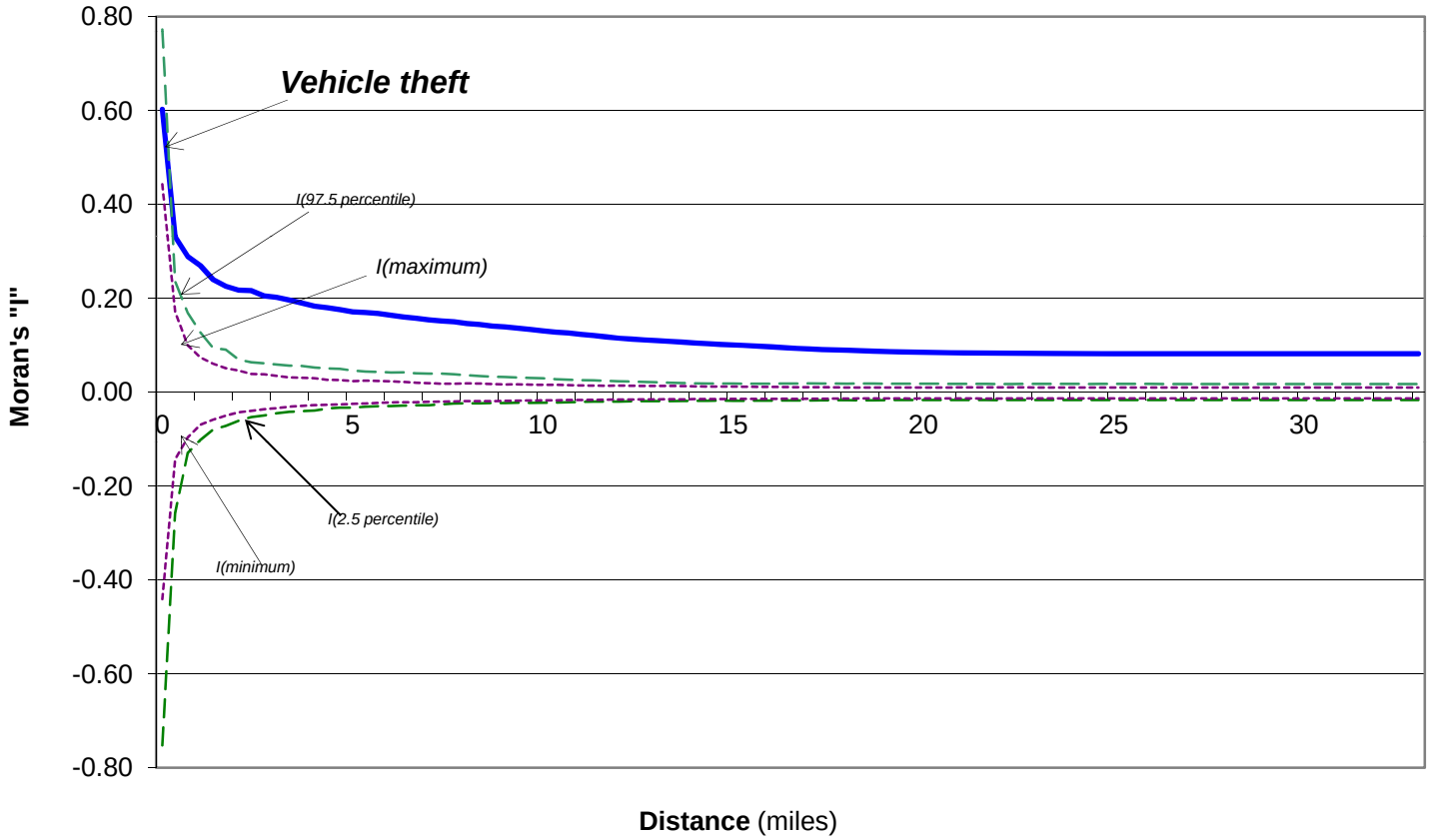

Baltimore County Vehicle Theft: 1996
By Traffic Analysis Zones
Legend
Vehicle theft
Less than 5
5 - 9
10 - 14
15 - 19
20 or more
Interstate Highways
Baltimore County
City of Baltimore
N
0 2.5 5 10 Miles

Figure 5.10:

Moran Correlogram: Baltimore County Vehicle Theft: 1996

"I" with 95% Confidence Intervals (N=1000 Monte Carlo Simulations)



distance separations. In other words, vehicle thefts appear to be highly clustered, much more so than would be expected by chance.

Now, compare this distribution with that the 1996 population (Figure 5.11). The 1996 data were estimated by the Baltimore Metropolitan Council, the regional planning agency. Comparing this map with Figure 5.9, intuitively it can be seen that population is more dispersed than vehicle thefts. Consequently, the Moran Correlogram shows much less spatial autocorrelation. The "I" value for zero distance is 0.39, lower than the 0.60 for vehicle thefts. The graph then drops off very quickly and approaches the global "I" value at about 3 miles. Further, from about 2 miles on, the "I" value is not different than what might be expected by chance since the curve falls between the 2.5^{th} percentile and the 97.5^{th} percentile. In other words, nearby TAZ's tend to have similar population levels, but there is no relationship between the population of TAZ's and those farther away.

Figure 5.13 compares the Moran Correlogram of vehicle theft with that of population by looking at only the positive "I" values. As seen, vehicle theft has a much higher "I" value for short distances than for population. The reason is most likely that a disproportionate number of vehicle thefts occur in commercial areas which, in turn, are more concentrated than the distribution of population.

Uses and Limitations of the Moran Correlogram

In other words, the Moran Correlogram provides information about the scale of spatial autocorrelation, whether it is more concentrated (as with the vehicle theft example) or more diffuse (as with the population example). This can be useful for gauging the extent to which 'hot spots' are truly isolated concentrations of incidents or whether they are by-products of spatial clustering over a larger area. In Chapter 7, we will examine a clustering algorithm that examines a hierarchy of clusters (e.g., first-order clusters that are within larger second-order clusters which, in turn, are within even larger third-order clusters). The Moran Correlogram provides a quick snapshot of the extent of spatial autocorrelation as a function of scale.

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

Figure 5.11:

Baltimore County Population: 1996 (estimated)

By Traffic Analysis Zones

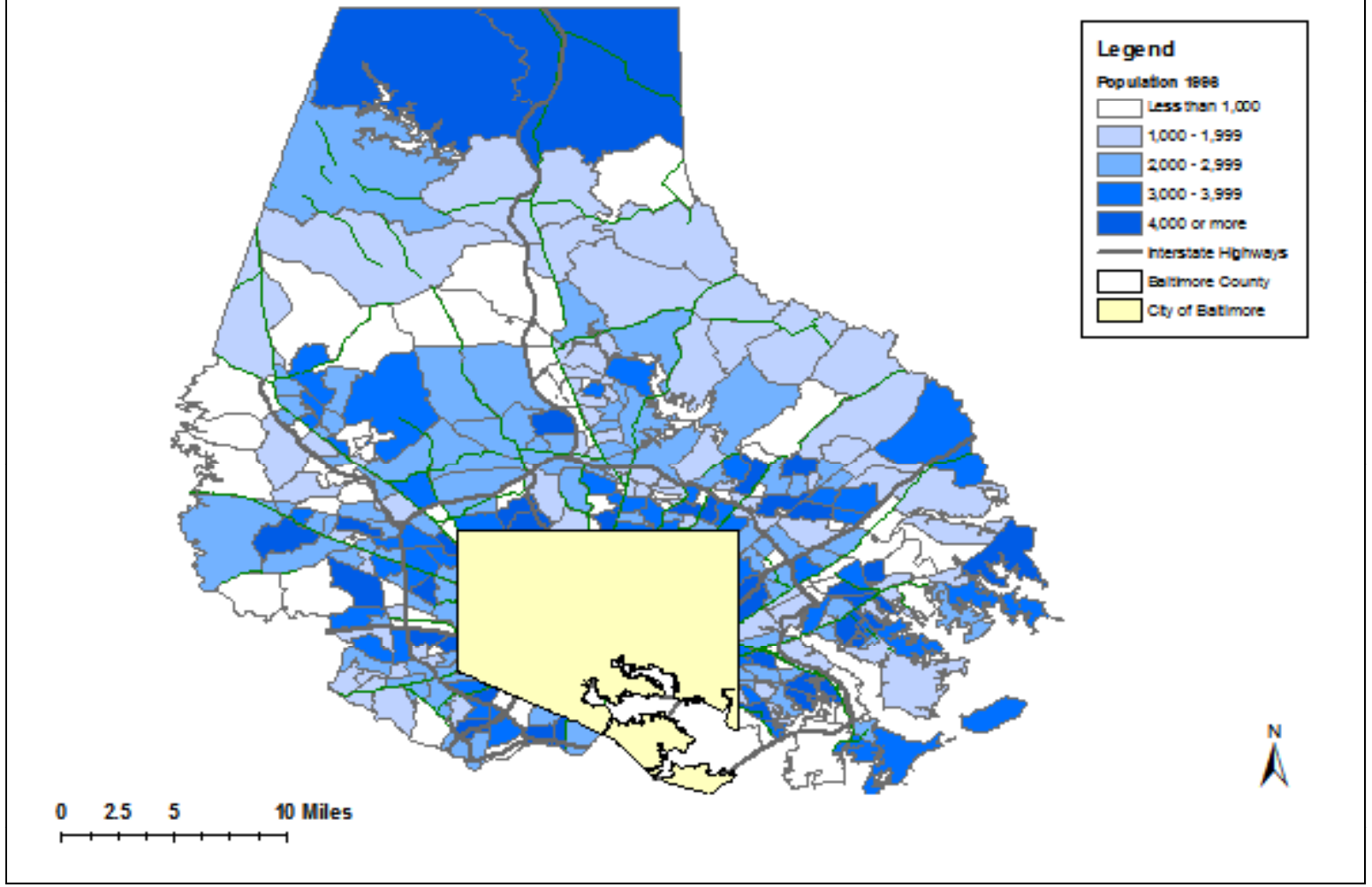


Figure 5.12:

Moran Correlogram: Baltimore County Population: 1996

"I" with 95% Confidence Intervals (N=1000 Monte Carlo Simulations)

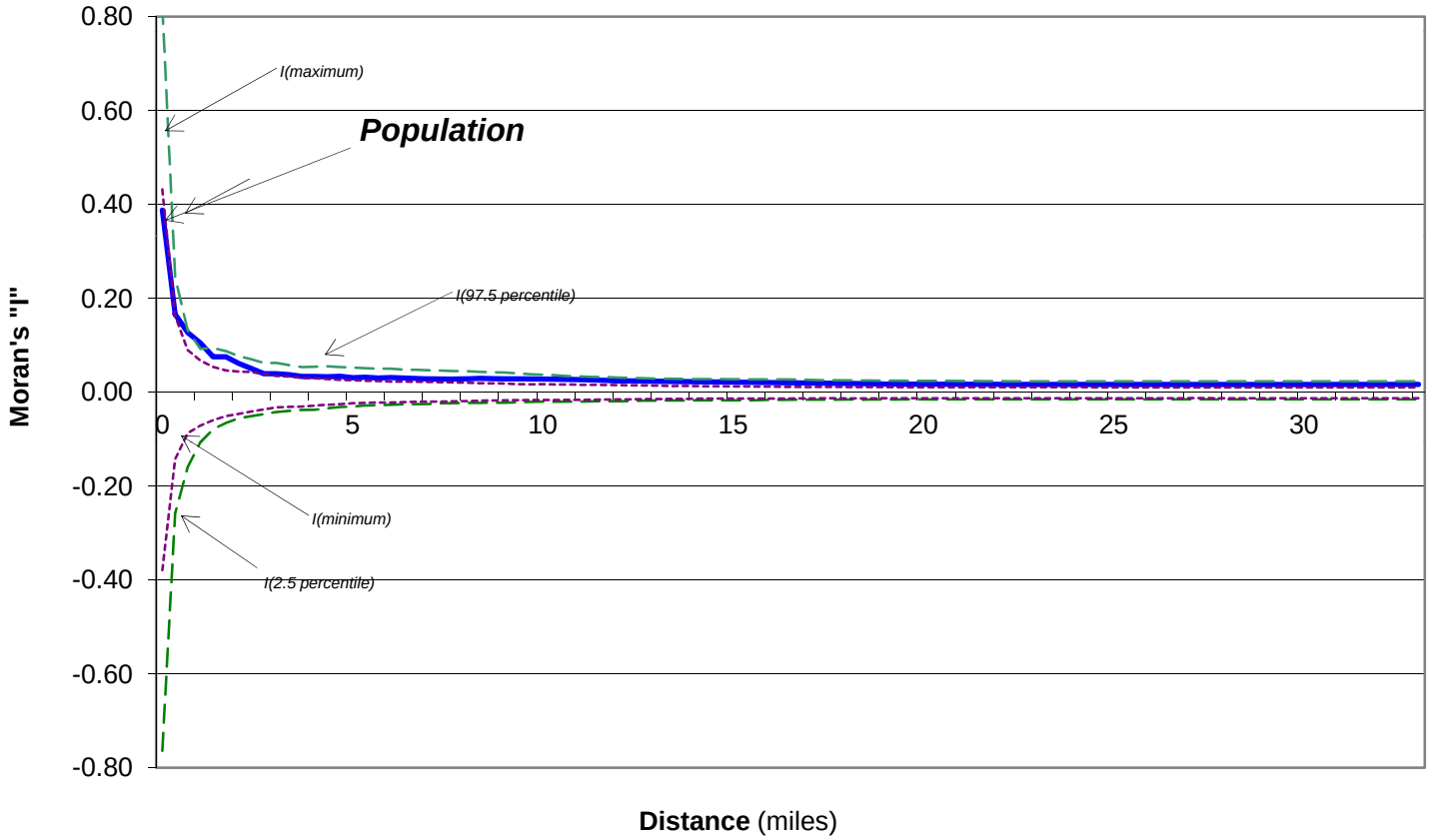
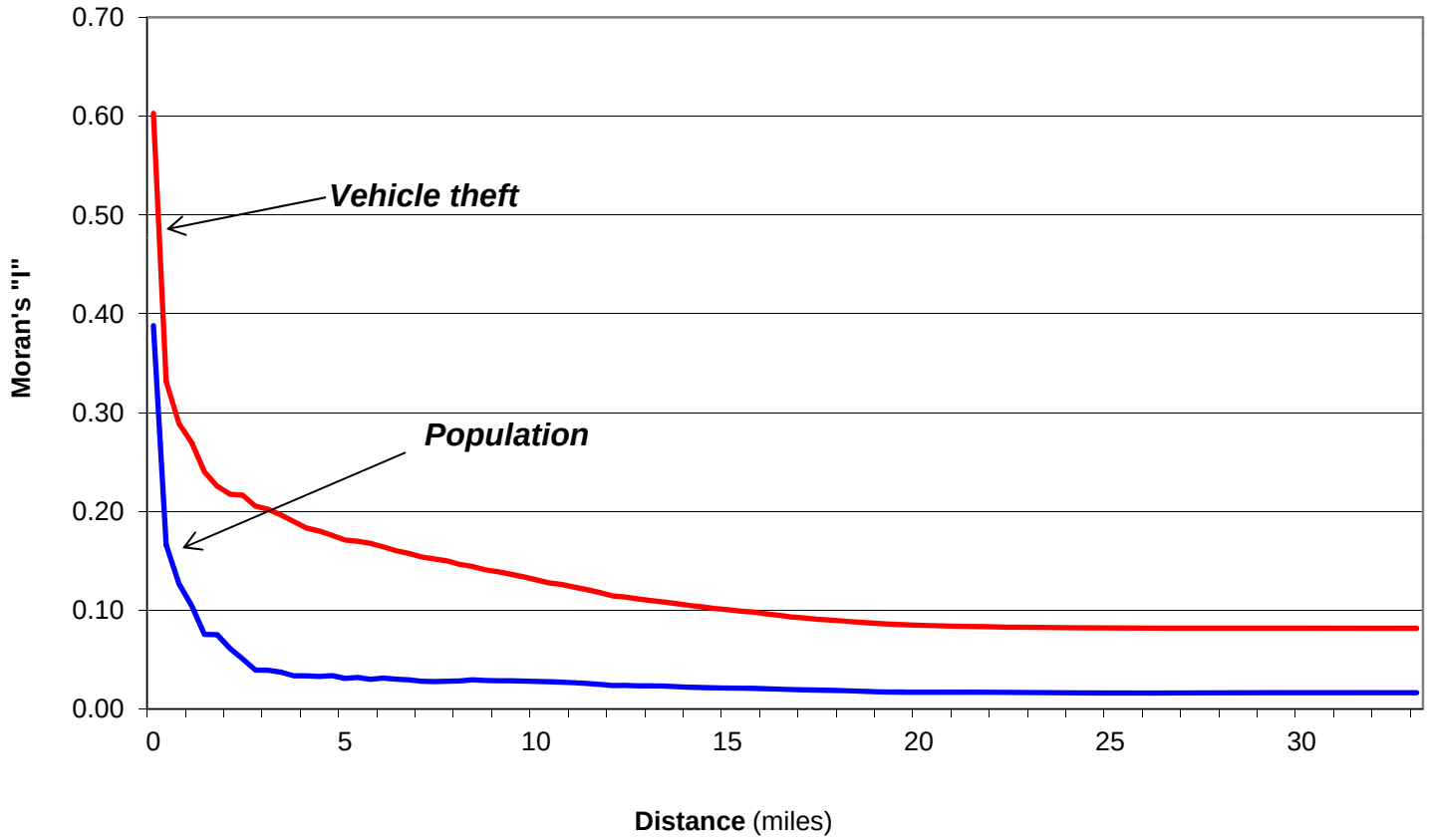


Figure 5.13:

Two Moran Correlograms of Baltimore County Vehicle Theft & Population: 1996



A third use for the Moran Correlogram is in identifying the degree of decline in spatial autocorrelation with distance (sometimes called *distance decay*) in choosing an appropriate parameter for spatial regression models. Chapter 19 will discuss this methodology.

On the other hand, like all global spatial autocorrelation statistics, the Correlogram will not indicate where there is clustering or dispersion, only that it exists. For that, we will have to examine tools that focus on concentrated events (or the opposite, the lack of concentration).

Geary Correlogram

The Geary Correlogram is similar to the Moran Correlogram in that it calculates the Geary "C" index for different distance intervals/bins. The user can select any number of distance intervals. The default is 10 distance intervals. The size of each interval is determined by the maximum distance between zones and the number of intervals selected. The output includes both the regular "C" and the adjusted "C". The graph presented on the results tab show the adjusted "C" since this is more intuitive and can be compared to the Moran Correlogram.

Adjust for Small Distances

If the 'Adjust for small distances' box is checked, small distances are adjusted so that the maximum weighting is 1 (see equation 5.4 above.) This ensures that the "C" values for individual distances won't become excessively large or excessively small for points that are close together. The default value is no adjustment.

Geary Correlogram Simulation of Confidence Intervals

Since the Geary's "C" statistic may not be normally distributed, the significance test is frequently inaccurate. Instead, a permutation Monte Carlo simulation is run whereby the original values of the variable, Z, are maintained but are randomly re-assigned for each simulation run. This will maintain the distribution of the variable Z but will estimate the value of "C" under random assignment of this variable. Specify the number of simulations to be run (e.g., 1000, 5000, 10000). Note, a simulation may take time to run especially if the data set is large or if a large number of simulation runs are requested.

Example: Geary Correlogram of Baltimore County Vehicle Thefts

Using the same data set on the Baltimore County vehicle thefts as shown in Figure 5.9 above, the Geary Correlogram was run with 100 intervals (bins). The routine was also run with 1000 simulations to estimate confidence intervals around the "C" value. Because it is more intuitive visually, the adjusted "C" was used instead of the regular "C".

Figure 5.14 illustrates the distance decay of the adjusted "C" as a function of distance along with the simulated 95% confidence interval. The theoretical adjusted "C" under random conditions is also shown. As seen, the "C" values are above 0 for all distances tested. However, when compared with the 2.5th and 97.5th percentiles from the simulated rescaled "C" for all intervals, the adjusted "C" values are not outside these percentiles for the very short distances but are from about 1.5 miles separation or greater. In other words, the graph suggests that the distribution of "C" for nearby zones is not different than what would be expected by chance. Only with increasing distance is the distribution clearly more clustered than chance.

This illustrates a subtle difference between the Geary and Moran indices. The Geary is more sensitive to local variations while the Moran reacts more to global variations. The Geary shows that there is positive spatial autocorrelation in vehicle theft for the immediate neighborhood around zones, but it is not much different than might be expected on chance. However, with increasing distance, positive spatial autocorrelation is shown. This suggests a type of sub-regional clustering of vehicle thefts; local clustering is limited but the events tend to be concentrated in only part of Baltimore County. As seen in Figure 5.9 above, the TAZ's nearer the border with the City of Baltimore had much higher vehicle theft numbers than the rural parts of the County.

The Geary Correlogram can also be used for comparison to other distributions, such as the comparison of vehicle theft with population as shown in Figure 5.13. This example will not be repeated here for the Geary Correlogram, but it does show that vehicle theft has higher "C" values than population over most distances, similar to the Moran Correlogram.

Uses and Limitations of the Geary Correlogram

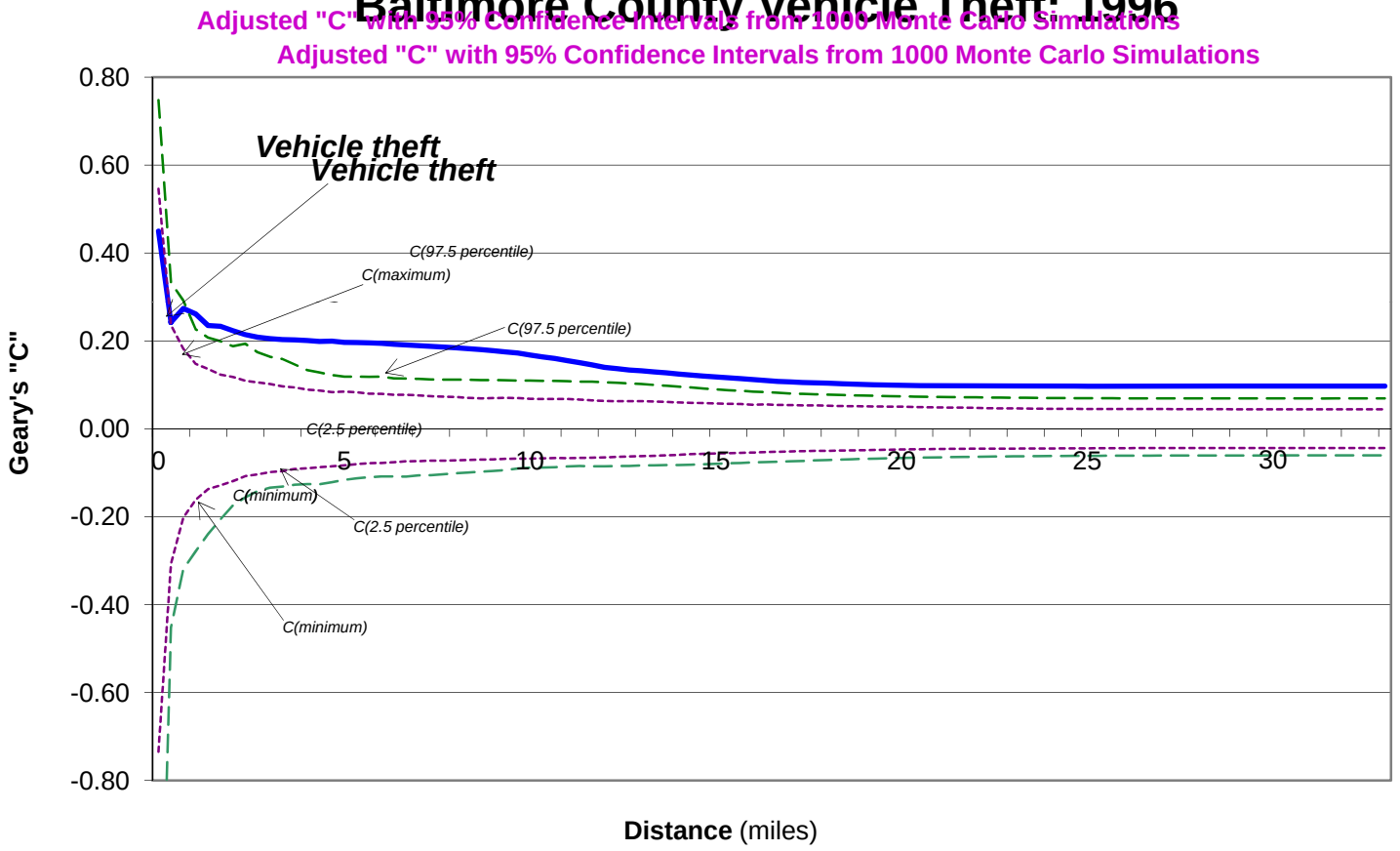
Similar to the Moran and the Getis-Ord correlograms (see below), the Geary Correlogram is useful in order to determine the degree of spatial autocorrelation and how far away from each zone it typically extends. Since it is an average over all zones, it is a general indicator of the spread of the spatial autocorrelation. This can be useful for defining limits to search distances in other routines, such as the single kernel density interpolation routine where a fixed bandwidth would be defined to capture the majority of spatial autocorrelation. Its biggest limitation is that it is not as powerful a test as the Moran Correlogram.

Getis-Ord Correlogram

The Getis-Ord Correlogram calculates the Getis-Ord "G" index for different distance intervals/bins. The statistic requires an intensity variable in the primary file and calculates the Getis-Ord "G" index for different distance intervals/bins. The user can select any number of

Figure 5.14:

Geary Correlogram:
Baltimore County Vehicle Theft: 1996



distance intervals. The default is 10 distance intervals. The size of each interval is determined by the maximum distance between zones and the number of intervals selected.

Getis-Ord Correlogram Simulation of Confidence Intervals

Since the Getis-Ord "G" statistic may not be normally distributed, the significance test is frequently inaccurate. Instead, a permutation Monte Carlo simulation is run whereby the original values of the intensity variable, Z, are maintained but are randomly re-assigned for each simulation run. This will maintain the distribution of the variable Z but will estimate the value of "G" under random assignment of this variable. The user should specify the number of simulations to be run (e.g., 100, 1000, 10000). Note, a simulation may take time to run especially if the data set is large or if a large number of simulation runs are requested.

If a simulation is run, percentiles for the 0.5^{th}, 2.5^{th}, 97.5^{th} and 99^{th} percentiles are provided. Pairing the 2.5^{th} with the 97.5^{th} or the 0.5^{th} with the 99.will create approximate 95% or 99% confidence intervals. For the three correlograms, these statistics are provided for each of the distance bins.

Example: Getis-Ord Correlogram of Baltimore County Vehicle Thefts

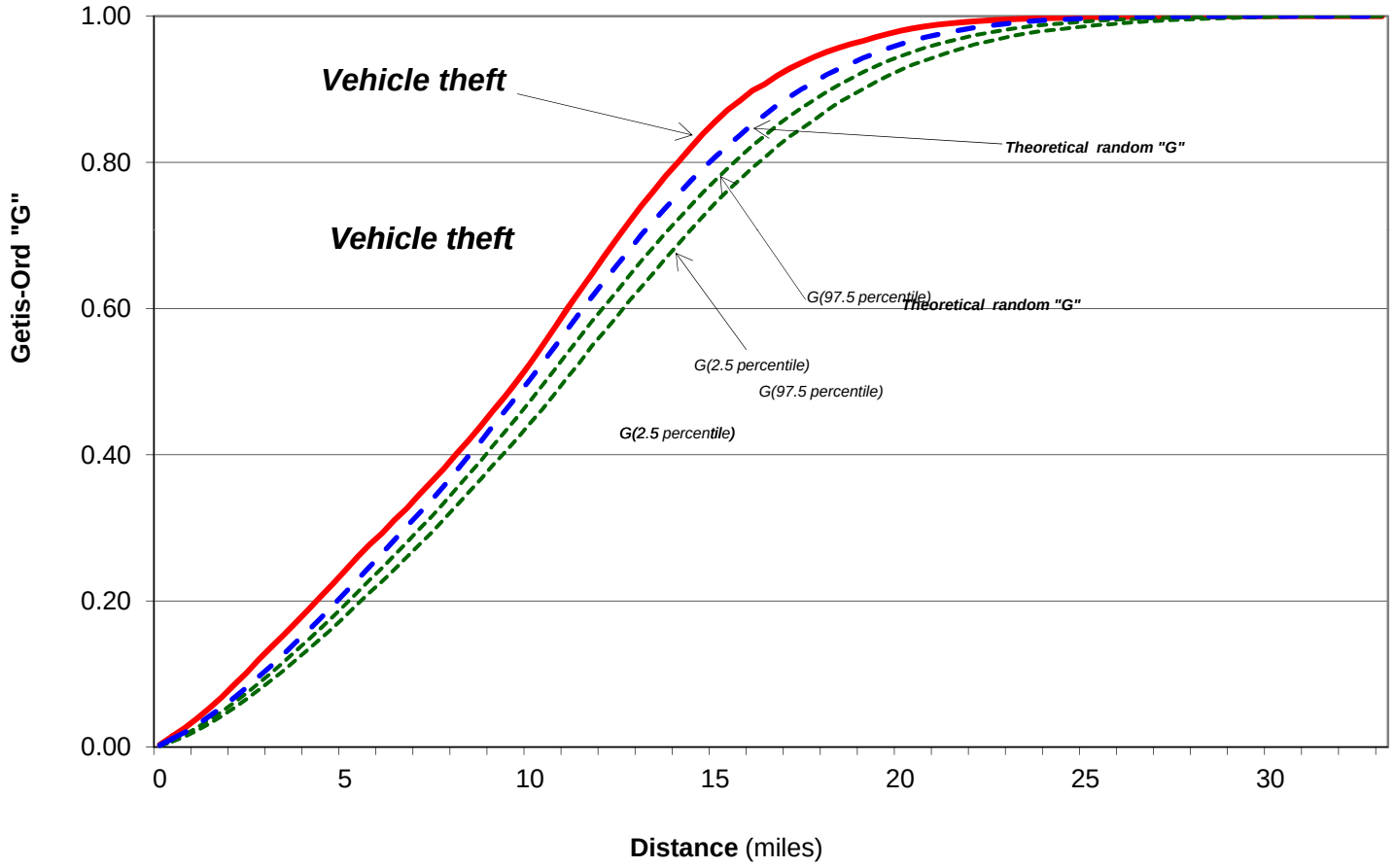
Using the same data set on the Baltimore County vehicle thefts as in figure 5.9, the Getis-Ord Correlogram was run. The routine was run with 100 intervals and 1000 Monte Carlo simulations in order to simulate 95% confidence intervals around the "G" value. The output was then brought into Excel to produce a graph. Figure 5.15 illustrates the distance decay of the "G", the expected "G", and the 2.5 and 97.5 percentile "G" values from the simulation.

Note that the "G" value *increases* with distance from close to 0 to close to 1 at the largest distance, around 33 miles. The actual "G" is higher than the expected "G" for all distances until the maximum, indicating that there is consistent high positive spatial autocorrelation in the data set. Since the Getis-Ord can distinguish a hot spot from a cold spot, the excess of "G" over the expected "G" indicates that there are somes zones with substantial numbers of vehicle thefts. Notice how the expected "G" also falls above the 97.5 percentile suggesting that there are more 'hot spots' than 'cold spots'. That is, if the zones were spatially re-arranged, then would not expect as much concentration as actually occurred.

Figure 5.15:

Getis-Ord Correlogram: Baltimore County Vehicle Theft: 1996

"G" with 95% Confidence Intervals from 1000 Monte Carlo Simulations



Uses and Limitations of the Getis-Ord Correlogram

Similar to the Moran Correlogram and the Geary Correlogram, the Getis-Ord Correlogram is useful in order to determine the degree of spatial autocorrelation and how far away from each zone it typically extends. Since it is an average over all zones, it is a general indicator of the spread of the spatial autocorrelation. This can be useful for defining limits to search distances in other routines, such as the single kernel density interpolation routine or the MCMC spatial regression module (see Chapters 10 and 19).

Unlike the other two correlograms, however, it can distinguish hot spots from cold spots. In the example above, there are more hot spots than cold spots since the "G" is greater than the expected "G" for all distances. The biggest limitation for the Getis-Ord Correlogram is that it cannot detect negative spatial autocorrelation whereby zones have different values from their neighbors. For that condition, which is rare, the other two correlograms should be used.

Running the Spatial Autocorrelation Routines

The six routines are defined on the Spatial Autocorrelation tab under spatial description. With the Moran and Geary routines, the user simply checks the box for each routine. If distance is to be adjusted for small distances, the user must check the appropriate box. For the Getis-Ord "G" routine, the user must specify a search distance and a unit of distance measurement (the default is 1 mile). For the three correlograms, the user must specify the number of intervals and the number of simulations that are to be run, if any.

The output for the six routines is somewhat similar. For the three global indices, statistics are provided on the index ("I", "C" or"G") and the expected value. For the three correlograms, these statistics are provided for each of the distance bins. If a simulation is run, percentiles for the 0.5^{th}, 2.5^{th}, 97.5^{th} and 99^{th} percentiles are provided. Pairing the 2.5^{th} with the 97.5^{th} or the 0.5^{th} with the 99.will create approximate 95% or 99% confidence intervals.

Guidelines for Examining Spatial Autocorrelation

To summarize, a number of indices for examining spatial autocorrelation have been presented. These indices are used with data in which there is an attribute variable, a count or interval variable associated with specific locations. Typically, the indices are used with data on zones since zonal information is published by many different agencies. However, the indices could also be used with individual data if there are attributes associated with the individual records.

While there is no single way to utilize these indices, the following are suggestions for using them. First, identify whether there is positive spatial autocorrelation using Moran's "I" and Geary's "C". Positive spatial autocorrelation indicates that zones are located near to other zones with similar values, either zones with high values on the variable being located near to zones also with high values or the opposite condition (low values nearby other low values).

If both the Moran "I" and Geary "C" (either regular or adjusted values) are both significant, this is strong evidence that there is sizeable spatial autocorrelation in the data. Whether the spatial autocorrelation is due to global (regional) factors or local clustering cannot be easily determined from the indices. On the other hand, if the Moran is significant, but the Geary is not, this could indicate that the clustering is a function of global concentration rather than local concentration since the Moran index is more sensitive to region-wide variation in the variable.

If there is negative spatial autocorrelation, which does occasionally happen, this indicates that zones with high values are located near to zones with low values, or the opposite. The user is advised to use one of the hot spot techniques described in Chapters 7, 8 and 9 to see if the hot spots can be isolated.

Second, if there is positive spatial autocorrelation, identify the type using the Getis-Ord "G" statistic. The Getis-Ord "G" is only applicable for positive spatial autocorrelation but can distinguish a predominance of high positive or low positive. High positive means that there are more zones with high values located near to other zones also with high values whereas low positive means the opposite (low near to low). The index is a type of average that weights the predominance of these types. In practice, there will be both types but the index indicates which is stronger. Since the Getis-Ord "G" requires a search distance, the user may have to run the Getis-Ord Correlogram first in order to identify a distance for which the positive spatial autocorrelation is most distinguishable from the theoretical random "G".

Third, examine the decline of the spatial autocorrelation with distance by using the three correlograms. While the Moran and Geary correlograms can be used for both positive and negative spatial autocorrelation, the Getis-Ord correlogram can only be used with positive spatial autocorrelation. The three correlograms will indicate how spatial autocorrelation varies by distance from each zone, on average. They can provide useful information about whether the concentration is very large, such as concentrated in the center of a metropolitan area, in which case the spatial autocorrelation is primarily a function of global factors. Alternatively, if the indices fall off very quickly, this suggests neighborhood (or local) effects rather than a dominant global pattern. In practice, there will be both types of factors, but the correlograms can indicate which is most important.

As with the global indices, the correlograms can provide useful information about the rate of decline in spatial autocorrelation (distance decay) for the kernel density routines (Chapter 10), the journey-to-crime routine (Chapter 13), the spatial regression routines (Chapter 19), or the trip distribution module of the Crime Travel Demand Model (Chapter 28).

In other words, identifying whether there is spatial autocorrelation and, if so, the type is important with zonal data (or with individual records having attributes) in that it is a first step in understanding where and why that spatial autocorrelation occurs. It is a necessary step in conducting hot spot analysis and in modeling the predictive factors that cause the spatial autocorrelation to occur. Chapter 9 examines hot spot identification routines appropriate for zonal data or individual data with attributes while Chapter 19 examines various regression tools for modeling the predictors of the spatial autocorrelation.

References

Anselin, L. (2008). Personal note on the testing of significance of the local Moran values.

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

Anselin, L.. (1992). *SpaceStat: A Program for the Statistical Analysis of Spatial Data*. Santa Barbara, CA: National Center for Geographic Information and Analysis, University of California.

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

Cliff, A. D. & Haggett, P. (1988). *Atlas of Disease Distributions*. Blackwell Reference: Oxford.

Cliff, A. & Ord, J. (1973). *Spatial Autocorrelation*. Pion: London.

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

Freedman, D. A. (1999). Ecological inference and ecological fallacy. *International Encyclopedia of the Social and Behavioral Sciences*, Technical Report No. 549, October. http://www.stanford.edu/class/ed260/freedman549.pdf. Accessed March 26, 2012.

Geary, R. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5, 115-145.

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

Getis. A, & Ord, J. K. (1993) Erratum, *Geographical Analysis*, 25, 276.

Getis, A. & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics, *Geographical Analysis*, 24, 189-206.

Griffith, D. A. (1987). *Spatial Autocorrelation: A Primer*. Resource Publications in Geography, The Association of American Geographers: Washington, DC.

References (continued)

Khan, G., Qin, X. & Noyce, D. A. (2006). Spatial analysis of weather crash patterns in Wisconsin. 85th Annual meeting of the Transportation Research Board: Washington, DC.

Langbein, L. I. & Lichtman, A. J. (1978). *Ecological Inference*. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-010. Beverly Hills and London: Sage Publications.

Lee, J. & Wong, D. W. S. (2005). *Statistical Analysis with ArcView GIS and ArcGIS*. J. Wiley & Sons, Inc.: New York.

Lees, B. (2006). The spatial analysis of spectral data: Extracting the neglected data, *Applied GIS*, 2 (2), 14.1-14.13.

Levine, N. (1999). The effects of local growth management on regional housing production and population redistribution in California, *Urban Studies*. 1999. 36 12, 2047-2068.

Levine, N. & Lee, P. (2013). Crime travel of offenders by gender and age in Manchester, England. Leitner, M. (ed), *Crime Modeling and Mapping Using Geospatial Technologies,* Springer. 145-178.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37, 17-23.

Ord, J. K. & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional Issues and an Application. *Geographical Analysis*, Vol. 27, 1995, 286-306.

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.

Attachments

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

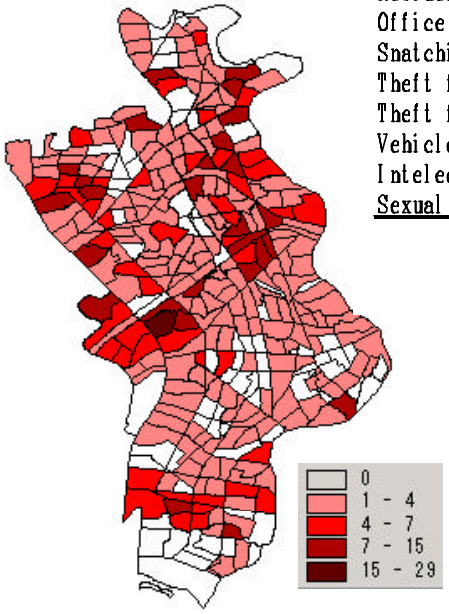
Takahito Shimada
National Research Institute of Police Science
National Police Agency, Chiba, Japan

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

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

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

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



Calculated Moran's I by Crime Types

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

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

Preliminary Statistical Tests for Hotspots: Examples from London, England

Spencer Chainey
Jill Dando Institute of Crime Science
University College
London, England

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

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

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

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

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