CHAPTER 5

PATTERN DESCRIPTORS

Human settlements appeared in places where there were resources to support the population and where climate allowed such development. Changes in animal habitats often happened when events that altered their environment occurred. Locations of geographic objects form various spatial patterns according to their characteristics. Changes in spatial patterns over time illustrate spatial processes, as dictated by the underlying environmental or cultural factors.

The spatial patterns of geographic objects are often the result of physical or cultural processes taking place on the surface of the earth. *Spatial patterns* is a static concept since these patterns only show how geographic objects distribute at a given time. However, *spatial processes* is a dynamic concept because these processes show how the distribution of geographic objects changed over time. For any given geographic phenomenon, we often need to study both its spatial patterns and the spatial processes associated with these patterns.

Understanding the spatial patterns allows us to understand how the geographic phenomenon distributes and how it can be compared with others. The ability to describe spatial processes enables us to determine the underlying environmental or cultural factors that are changing the patterns. If the changes were desirable, we would find ways to promote them. If the changes were not desirable, we would need to find ways to correct the problems.

*Spatial statistics* are the most useful tools for describing and analyzing how various geographic objects (or events) occur or change across the study area. These statistics are formulated specifically to take into account the locational attributes of the geographic objects studied. We can use spatial statistics to describe the *spatial patterns* formed by a set of geographic objects so that we can compare them with patterns found in other study areas. For the spatial processes associated with these patterns, we can use spatial statistics to describe their forms, to detect changes, and to analyze how some spatial patterns change over time.

In earlier chapters, we demonstrated the use of descriptive statistics to measure central tendency and dispersion among point-based geographic objects. In this chapter, we will discuss the use of spatial statistics to describe and measure spatial patterns formed by geographic objects that are associated with *areas* or *polygons*. We will use spatial statistics to describe spatial patterns exhibited in a set of polygons according to their characteristics. In addition, we will examine how the measured spatial patterns can be compared by using tests based on these spatial statistics.

5.1 SPATIAL RELATIONSHIPS

When studying a spatial pattern, we may want to compare that pattern with a theoretical pattern to determine if the theory holds in the case of the pattern being studied. Alternatively, we may want to classify spatial patterns into an existing categorical structure of known patterns. If the spatial patterns we study correspond to a particular theoretical pattern, we will be able to apply properties of the theory to interpret the spatial pattern we study. Or, if the spatial pattern can be closely related to a known pattern, we will be able to borrow from experience with and knowledge of the pattern for further study. In either case, it is necessary for us to establish a structure of categories of the spatial pattern.

A spatial pattern can be *clustered*, *dispersed*, or *random*. In Figure 5.1, these three possibilities are shown in hypothetical patterns in the seven counties of northeastern Ohio (Cuyahoga, Summit, Portage, Lake, Geauga, Trumbull, and Ashtabula). In Case 1, the darker shade representing certain characteristics associated with the counties appear to have *clustered* on the western side of the seven-county area. On the eastern side, the lighter shade (the other type of characteristics) prevails among the remaining counties. Perhaps the darker shade indicates the growth of the urban population in recent years, whereas the lighter shade indicates the loss of the urban population during the same period.

In Case 2, counties that are shaded darker appear to be spaced evenly and away from each other. This is often referred to as a *dispersed* pattern. It suggests that the geographic phenomenon displayed may be uniformly distributed across this seven-county area. For example, it may be that the political preferences between the two parties in these counties happen to change between neighboring counties. For an extremely uniform or dispersed pattern, the checkerboard in which black and white cells distribute systematically and change between every pair of neighboring cells is a typical example of a dispersed spatial pattern. A repulsive spatial relationship among neighbors is implied.

The spatial pattern in Case 3 of Figure 5.1 does not appear to be either clustered or dispersed. It may be close to what we typically call the *random* pattern. If a spatial pattern is random, it suggests that there may not be any particular sys-
to determine its closeness to a random pattern so that the pattern can be further studied.

5.2 SPATIAL AUTOCORRELATION

In classifying spatial patterns as either clustered, dispersed, or random, we can focus on how various polygons are arranged. We can measure the similarity or dissimilarity of any pair of neighboring polygons. When these similarities and dissimilarities are summarized for the spatial pattern, we have spatial autocorrelation (Odlund, 1988).

Spatial autocorrelation means that the attribute values being studied are self-correlated and the correlation is attributable to the geographic ordering of the objects.

Many situations exhibit some degree of spatial autocorrelation. When comparing agricultural production levels among all the farms in a region, we would not find that all farms are producing at the same rate. Even though the local climate may be the same for all of these farms, the soil conditions or water supplies within the region may vary. Still, neighboring farms within the region share similar soil and moisture conditions, so these farms probably have similar levels of production.

Among population densities measured in each county, we would see that those counties located in or close to big central cities tend to have higher population densities than those located far away. This is because the outward influence of the central city often makes the surrounding counties its hinterlands. Once the underlying reason is identified, it would be logical to investigate why urban population densities cluster around big central cities or why the production levels of farms cluster or disperse along known soil structures or water supplies.

The basic property of spatially autocorrelated data is that the values are not random in space. Instead, the data values are spatially related to each other, even though they may be related in different ways. Referring back to the three cases in Figure 5.1, Case 1 has a positive spatial autocorrelation, with adjacent or nearby polygons having similar shades (values). In Case 2, the dispersed pattern has a negative spatial autocorrelation, with changes in shade often occurring between adjacent polygons. As for Case 3, it appears to be close to a random pattern in which little or no spatial autocorrelation exists.

In addition to its type or nature, spatial autocorrelation can be measured for its strength. Strong spatial autocorrelation means that the attribute values of adjacent geographic objects are strongly related (either positively or negatively). If the attribute values of adjacent geographic objects appear to have no clear order or relationship, the distribution is said to have a weak spatial autocorrelation or a random pattern. Figure 5.2 shows five different structures with darker and lighter shades within the seven counties in northeastern Ohio. Opposite to the numeric line, the patterns toward the left end are examples of positive spatial autocorrelation, while the patterns toward the right end are examples of negative spatial
5.3 SPATIAL WEIGHTS MATRICES

For measuring spatial autocorrelation in a set of geographic objects, we have to discuss methods for capturing spatial relationships among areal units, or polygons. Recall that spatial autocorrelation measures the degree of sameness of attribute values among areal units within their neighborhood. The concept of neighborhood has to be quantified so that it can be applied in the calculation of spatial autocorrelation statistics. In other words, the neighborhood relationship among areal units has to be captured before we can calculate the statistics.

Assume that we have \( n \) areal units in our study area. Given any predefined method to determine the neighborhood relation for these \( n \) areal units, we have \( n \times n \) pairs of relationship to be captured. Conventionally, we use a matrix (in this case, \( n \times n \) in dimension) to store and organize the spatial relationship among these \( n \) areal units. Each areal unit is represented by a row and a column. Each value in the matrix indicates the spatial relationship between the geographic features represented by the corresponding row and column. However, given different criteria to define the neighborhood relationship, we may derive different matrices (Griffith, 1996). For example, a binary value of 1 or 0 can be used to represent whether or not two polygons are spatially adjacent. Alternatively, actual measures of distances between centroids of polygons can be stored in such a matrix to represent their spatial adjacency in a different way.

In the following subsections, we discuss different ways to specify spatial relationships and their associated matrices. In general, these matrices are called spatial weights matrices because elements in the matrices are often used as weights in the calculation of spatial autocorrelation statistics.

5.3.1 Neighborhood Definitions

There are many ways to define a spatial relationship. Even if we are concerned with the immediate neighbors of an areal unit, there are at least two common methods. Figure 5.3 illustrates the rook’s case and the queen’s case of neighborhood relationships in a grid. In a highly simplified polygon structure similar to a set of grid cells, there are nine areal units, with the one we are concerned about at the center (X). Using the rook’s case as the criterion to define a neighbor, only

![Figure 5.3](image_url)
B, D, E, and G are neighbors because each of them shares a boundary with the polygon X. If we adopt the queen’s case, then all the surrounding areal units can be identified as neighbors of X as long as they touch each other even at a point.

The neighborhood definitions described in Figure 5.3 correspond to the adjacency measure. If any two areal units are next to each other (juxtaposed), then they are neighbors of each other. In the rook’s case, the neighboring units have to share a boundary with length greater than 0. In the queen’s case, the shared boundary can be just a point along the diagonal polygons. The neighboring polygons of X identified by these two criteria are its immediate or first order neighbors. The adjacency measure can be extended to identify neighbors of the immediate neighbors or, more technically, second order neighbors. The concept can be further extended to identify higher order neighbors, even though computationally it may sometimes not be tractable.

Besides adjacency as the criterion used to define a neighborhood, another common measure is distance. Distance is a common measure of a spatial relationship. Corresponding to the geographic concept of distance decay, intensity of spatial activities between two distant places are said to be less than that of activities between places that are closer together, assuming that all places have the same characteristics. Consequently, the distance between any two geographic features can be used as an indicator of their spatial relationship. The distance measure can also be used to define a neighborhood.

If, for a household, neighbors means those houses are less than 1 mile, then its neighborhood can be quantitatively defined as all houses within a radius of 1 mile from the household. With this definition, we can check all houses to see if they fall within this area. The results can be displayed in a binary form (inside or outside of the defined neighborhood).

### 5.3.2 Binary Connectivity Matrix

With different ways of defining neighbors, different matrices can be constructed to capture the spatial relationship among geographic features. Using the simplest adjacency definition of neighborhood, areal units sharing a common boundary (at this point, we put aside the difference between the rook’s and queen’s cases) are neighbors. In the matrix capturing the spatial relationship, we can put a 1 in the cell corresponding to the two geographic features or polygons if they are next to each other. If any two polygons are not adjacent to each other, the corresponding cell will have a value of 0. In the entire matrix, the cell value is either 0 or 1. This type of matrix is sometime called a binary matrix. Because this matrix also shows how pairs of polygons are connected, and because a similar concept can be applied to linear features in a network (Chapter 4), this binary matrix is also called a connectivity matrix. In such a matrix, $c_{ij}$ denotes the value (either 0 or 1) in the matrix in the $i$th row and the $j$th column.

A binary connectivity matrix has several interesting characteristics. First, all elements along the major diagonal, $c_{ij}$ (the diagonal from upper left to lower right), are 0 because it is assumed that an areal unit is not a neighbor of itself. Second, this matrix is symmetric, i.e., the upper triangle of the matrix divided along the major diagonal is a mirror of the lower triangle. Using our notation, $c_{ij} = c_{ji}$. This symmetric property of the matrix basically reflects the reciprocal nature of spatial relationships: areal unit A is a neighbor of B, and B is also a neighbor of A. The symmetric structure of the matrix requires information on the spatial relationship to be stored twice—a redundancy in this type of matrix. Finally, a row in the matrix represents how one areal unit is related spatially to all other units. Therefore, summing the cell values for that row across all columns, also known as the row sum, indicates the number of neighbors that the areal unit has. Using our notation, the row sum is

$$c_i = \sum_j c_{ij}.$$ 

Table 5.1 is a binary connectivity matrix of the seven Ohio counties using the queen’s case. Therefore, Cuyahoga County and Portage County are neighbors. Similarly, Geauga and Summit counties are depicted as neighbors in the matrix. The sum of all these cell values (26) is twice the number of shared boundaries or joints (27).

Apparently, the binary matrix is not a very efficient format to store spatial relationship based upon adjacency. Not only the upper triangle of the matrix duplicates the lower triangle; in most spatial configurations, most of the cell values in the matrix are 0s, indicating that the two areal units are not neighbors. In the example of the seven Ohio counties, this characteristic of the matrix may not be obvious. But still, among these seven areal units with a matrix of 49 cells, there are 23 Os. Imagine creating a binary matrix for the approximately 3,000 counties of the entire United States; the matrix would not only be gigantic ($3,000 \times 3,000 = 9$ million cells), but most of the cell values would be Os. Most of the time, we are only interested in neighbors. Using a lot of space to store nonneighbor information is not efficient.

Instead of using a square matrix ($n \times n$) to record areal units that are neighbors and nonneighbors, a more compact format of the matrix includes only areal units that are neighbors of a given unit. In general, each areal unit has a unique identification number, such as the Federal Information Processing Standards (FIPS) code.

### Table 5.1 Binary Connectivity Matrix (Seven Ohio Counties)

<table>
<thead>
<tr>
<th>ID</th>
<th>Geauga</th>
<th>Cuyahoga</th>
<th>Trumbull</th>
<th>Summit</th>
<th>Portage</th>
<th>Ashtabula</th>
<th>Lake</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geauga</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cuyahoga</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Trumbull</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Summit</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Portage</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ashtabula</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Lake</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
TABLE 5.2 A Sparse Matrix Capturing the Same Information as the Binary Matrix

<table>
<thead>
<tr>
<th>ID</th>
<th>Neighbor 1</th>
<th>Neighbor 2</th>
<th>Neighbor 3</th>
<th>Neighbor 4</th>
<th>Neighbor 5</th>
<th>Neighbor 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geauga</td>
<td>Cuyahoga</td>
<td>Trumbull</td>
<td>Summit</td>
<td>Portage</td>
<td>Ashtabula</td>
<td>Lake</td>
</tr>
<tr>
<td>Cuyahoga</td>
<td>Geauga</td>
<td>Summit</td>
<td>Portage</td>
<td>Lake</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trumbull</td>
<td>Geauga</td>
<td>Portage</td>
<td>Ashtabula</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summit</td>
<td>Geauga</td>
<td>Cuyahoga</td>
<td>Portage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portage</td>
<td>Geauga</td>
<td>Cuyahoga</td>
<td>Trumbull</td>
<td>Summit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ashtabula</td>
<td>Geauga</td>
<td>Trumbull</td>
<td>Lake</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lake</td>
<td>Geauga</td>
<td>Cuyahoga</td>
<td>Ashtabula</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

of a county (which combines the state and county FIPS codes), or a unique label such as the county name (unique only within the state). Using the unique identifier, a rather compact (sparse) matrix listing only the identifiers of neighbors can be constructed. Table 5.2 captures the same information as Table 5.1, but its size is greatly reduced. Instead of 49 cells (7 x 7) with many 0s, the dimension of the matrix is now 7 x 6, with only 26 cells containing information. In a much larger area with more polygons, the reduction in matrix size will be more dramatic.

5.3.3 Stochastic or Row Standardized Weights Matrix

The binary matrix basically has a weight of 1 if the areal unit is a neighbor. Mathematically, this unit weight may not be very effective in modeling the spatial relationship. For instance, we want to analyze how a house value is affected by the values of its surrounding units. Following the normal practice of realtors, we can think of the house value as receiving a fractional influence from each of its neighbors. If there are four neighbors for that house, its value may receive 0.25 influence from each neighboring house.

Recall that the binary matrix consists of 1s and 0s. A 1 indicates that the corresponding areal units represented by the rows and the columns are neighbors. Therefore, for each given row, the row sum, c_i, indicates the total number of neighbors that areal unit has. To find out how much each neighbor contributes to the value of the areal unit of concern when each neighbor exerts the same amount of influence on that areal unit, we would calculate the ratios of each house with respect to the total influence. This gives the weight (w_ij) of each neighboring unit:

\[ w_{ij} = c_{ij}/c_i. \]

Figure 5.4 uses the seven Ohio counties to illustrate how the row-standardized matrix or the stochastic matrix, which is usually denoted as W, can be derived from the binary matrix C. Please note that even the matrix still has a major diagonal of 0s; this matrix is no longer symmetric.
5.3.4 Centroid Distances

Besides using adjacency as a measure to describe the spatial relationship among a set of geographic features and to define a neighborhood among them, another common measure is distance. Using distance as the weight in describing a spatial relationship is very powerful. Recall the first law of geography summarized by Waldo Tobler (1970): all things are related, but closer things are more related. In other words, the relationship between any two geographic features is a function of distance between them. Generally, we expect features close to each other to be related. Therefore, using distance as the weight to depict a spatial relationship is theoretically sound.

The distance between two point features is easy to define. There are several ways to measure the distance between any two polygons. A very popular method, especially in transportation studies, is to use the centroid of the polygon to represent that polygon. The centroid is the geometric center of a polygon. There are different ways of determining the centroid of a polygon. Each of them identifies centroids differently. But in general, the shape of a polygon affects the location of its centroid. Polygons with unusual (geometrically irregular) shapes may generate centroids located in undesirable locations. For instance, certain methods for centroid calculation may generate a centroid of the state of Florida located in the Gulf of Mexico, or the centroid of the state of Louisiana to be in the neighboring state of Mississippi. Still, in most cases and with the advances in algorithms in determining centroids, it is quite common, using the distance between the two centroids of the two polygons, to represent the distance between the polygons themselves.

Because distance is used as the weight, the spatial weights matrix is sometimes labeled as the centroids of areal units i and j. Table 5.3 shows the D matrix using centroid distances for the seven Ohio counties. In modeling spatial processes, the distance weight is often used in an inverse manner, as the strengths of most spatial relationships diminish when distance increases. Therefore, when the distance matrix is used, the weight

\[ w_{ij} = \frac{1}{d_{ij}} \]

**TABLE 5.3 Spatial Weights Matrix Using Centroid Distance (Seven Ohio Counties)**

<table>
<thead>
<tr>
<th>ID</th>
<th>Geauga</th>
<th>Cuyahoga</th>
<th>Trumbull</th>
<th>Summit</th>
<th>Portage</th>
<th>Ashtabula</th>
<th>Lake</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geauga</td>
<td>0</td>
<td>25.1508</td>
<td>26.7057</td>
<td>32.7590</td>
<td>25.0289</td>
<td>26.5899</td>
<td>12.6265</td>
</tr>
<tr>
<td>Cuyahoga</td>
<td>25.1508</td>
<td>0</td>
<td>47.8151</td>
<td>23.4834</td>
<td>31.6155</td>
<td>50.8064</td>
<td>28.2214</td>
</tr>
<tr>
<td>Trumbull</td>
<td>26.7057</td>
<td>47.8151</td>
<td>0</td>
<td>41.8561</td>
<td>24.4759</td>
<td>29.5633</td>
<td>36.7375</td>
</tr>
<tr>
<td>Summit</td>
<td>32.7590</td>
<td>23.4834</td>
<td>41.8561</td>
<td>0</td>
<td>17.8031</td>
<td>58.0869</td>
<td>42.7375</td>
</tr>
<tr>
<td>Portage</td>
<td>25.0389</td>
<td>31.6155</td>
<td>24.4759</td>
<td>17.8031</td>
<td>0</td>
<td>47.5341</td>
<td>37.4962</td>
</tr>
<tr>
<td>Ashtabula</td>
<td>26.5899</td>
<td>50.8064</td>
<td>28.5633</td>
<td>58.0869</td>
<td>45.5341</td>
<td>0</td>
<td>24.7490</td>
</tr>
<tr>
<td>Lake</td>
<td>12.6265</td>
<td>28.2214</td>
<td>36.7375</td>
<td>37.4962</td>
<td>24.7490</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

is an inverse of the distance between features i and j. In other words, the weight is inversely proportional to the distance between the two features. Based upon empirical studies of spatial processes, however, the strength of many spatial relationships has been found to diminish more than proportionally to the distance separating the features. Therefore, the squared distance is sometimes used in the following format:

\[ w_{ij} = \frac{1}{d_{ij}^2} \]

These weighting schemes of inverse distance will be used later for different measures of spatial autocorrelation.

5.3.5 Nearest Distances

With the advances in GIS software algorithms, features other than the distance between centroids can be easily determined. It is also relatively easy to determine the distance between any two geographic features based on the distance of their nearest parts. Selecting the two farthest counties among the seven Ohio counties, the nearest parts between Summit and Ashtabula counties are the northeastern corner of Summit and the southwestern corner of Ashtabula (Figure 5.1). This conceptualization of distance between geographic features is potentially useful if the investigator is interested in spatial contact or diffusion.

An interesting situation involving the distance of nearest parts is when the two features are adjacent to each other. When this is the case, the distance between two neighboring features is 0. In other words, under this distance measurement scheme, 0 distance means that the corresponding features are immediate neighbors. The 1s in a binary connectivity matrix also capture this information. By extracting all the nondiagonal 0 cells from the distance matrix, we can also derive the binary matrix. Since the binary connectivity matrix can be seen as a simplified distance weight matrix, it is sometimes said that the binary matrix is a derivative of the distance matrix based on nearest parts. Please note that the above discus-

**TABLE 5.4 Spatial Weights Matrix Based upon Nearest Parts (Seven Ohio Counties)**

<table>
<thead>
<tr>
<th>ID</th>
<th>Geauga</th>
<th>Cuyahoga</th>
<th>Trumbull</th>
<th>Summit</th>
<th>Portage</th>
<th>Ashtabula</th>
<th>Lake</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geauga</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Cuyahoga</td>
<td>0.0000</td>
<td>0</td>
<td>0.3561</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.3614</td>
<td>0.0000</td>
</tr>
<tr>
<td>Trumbull</td>
<td>0.0000</td>
<td>0.3561</td>
<td>0.0000</td>
<td>0.3705</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1670</td>
</tr>
<tr>
<td>Summit</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.3705</td>
<td>0</td>
<td>0.0000</td>
<td>0.4015</td>
<td>0.2179</td>
</tr>
<tr>
<td>Portage</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0</td>
<td>0.1518</td>
<td>0.2180</td>
</tr>
<tr>
<td>Ashtabula</td>
<td>0.0000</td>
<td>0.3614</td>
<td>0.0000</td>
<td>0.4015</td>
<td>0.1518</td>
<td>0</td>
<td>0.0000</td>
</tr>
<tr>
<td>Lake</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.1670</td>
<td>0.2179</td>
<td>0.2180</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
sions on inverse distance weighting schemes are also applicable to the current distance measure.

Table 5.4 is the distance matrix based on nearest parts of the seven Ohio counties. Compare this matrix with the binary matrix in Table 5.1. The nondiagonal cells in the distance matrix with 0 values correspond to the cells with 1s in the binary matrix.

ArcView Notes

As in previous chapters, a project file was designed for this chapter (ch5_spr). It is located on the companion website to this book.

In addition to the standard ArcView menu categories, a new menu category, Spatial Autocorrelation, has been added. Under this menu category, the first menu item is Creating Weights Matrices. Please be sure that a View window with the interested theme is active (i.e., clicked on). If multiple themes were added to the View window, the procedure assumes that the first theme is the one for which the analyst is interested in performing spatial statistical analysis.

When this menu item is selected, the procedure will first check to see if any feature on the View window has been selected. If no feature has been selected, the procedure will select all features appearing in the View window to construct the spatial weights matrix. In other words, if users prefer to create a weights matrix for a subset of features on the View window, the user should either

1. use the selection tool or QueryBuilder in ArcView to select the features desired before choosing the menu item to create the matrix or
2. choose No when the procedure asks users to select all features when users forgot to select features before proceeding to weights matrix creation.

After selecting the desired features, run the weights matrix procedure again. If only a subset of all features is selected to create the weights matrix, users should remember, in subsequent analyses, to use only the subset of features instead of all features in the View window.

With the features for calculation selected, users will be asked to select a field in the table as the Identification (ID) field. Please note that this ID field has to provide a unique identifier for each feature included in the analysis. In the seven Ohio counties example, the combined state or county FIPS code would be appropriate because no two counties in the United States have the same combined state and county FIPS code. We can even use county names as the ID field in this example because we know

that each of the seven counties has a unique name. In a land use analysis, however, using types of land use as the ID field may be inappropriate because they are probably multiple polygons or areal units with the same land use type.

The procedure will provide users with four choices for weights matrices: binary connectivity, stochastic weight, distance between nearest parts, and centroid distance. Choose one of these options; the procedure will ask for the name of the output file.

Please note that the procedures for weights matrix generation developed here are meant for small data sets. They are not intended to support major research or analysis with large numbers of polygons or large spatial systems.

The computation speed also varies, depending on the type of matrix selected (in addition to other factors, such as hardware configurations) and the speed of the hardware. In general, the procedure is quite fast for the centroid distance matrix. For the matrices on binary connectivity and the distance for nearest parts, the procedure can be completed in a reasonable time frame (less than 1 minute on an average desktop computer) for spatial systems as large as 50 areal units. For larger spatial systems, the time required is not proportional but is exponentially related to the increase in the number of areal units.

The procedure is slowest in constructing the stochastic weights matrix because it requires additional steps to derive the total number of neighbors and then to perform the standardization. Still, these steps add just a few more seconds for a spatial system with 50 areal units. For larger systems, the extra time may be longer. But overall, this spatial weights matrix procedure handles small spatial systems very well.

5.4 TYPES OF SPATIAL AUTOCORRELATION MEASURES AND SOME NOTATIONS

In this section, we will discuss different methods for calculating spatial autocorrelation statistics utilizing one or more of the spatial weights matrices discussed above. Different statistics are used to handle attributes that are measured at different scales. In addition, different statistics capture different aspects of spatial autocorrelation. In this section, several statistics are introduced.

If the spatial attributes or variables to be studied are measured in nominal scale and are binary (i.e., the attribute has only two possible values), then joint count statistics can be used. If the spatial variables are measured in interval or ratio scale, the appropriate spatial autocorrelation statistics are Moran's I index and Geary's C Ratio. Another possible choice is the general G-statistic.
All of these measures can be regarded as global measures of spatial autocorrelation or spatial association because one statistic or value is derived for the entire study area, describing the overall spatial relationship of all the areal units. However, there is no reason to believe that any spatial process is homogeneous within the distribution itself. The magnitude of spatial autocorrelation can vary by locations, and thus a distribution or a spatial pattern can be spatially heterogeneous. To describe the spatial heterogeneity of spatial autocorrelation, we have to rely on measures that can detect spatial autocorrelation at a local scale. The Local Indicator of Spatial Association (LISA) and the local G-statistics are designed for this purpose.

In deriving various spatial autocorrelation statistics and related statistics to test for their significance, several terms derived from the spatial weights matrices will be used repeatedly. Therefore, it is logical to introduce and define them here before we discuss those spatial autocorrelation statistics.

Even though a weight, $w_{ij}$, is often used to represent the cell value of the stochastic weights matrix, $W$, for row $i$ and column $j$, it is also quite common to use $w_{ij}$ to represent the cell value of any weights matrix. For any given weights matrix, summing up all cell values of a given row $i$ across all columns (row sum) is denoted as

$$w_i = \sum_j w_{ij}.$$  

Similarly, a column sum is the sum for a given column $j$ across all rows:

$$w_j = \sum_i w_{ij}.$$  

Sometimes $W$ also represents the sum of all cell values of the weights matrix:

$$W = \sum_i \sum_j w_{ij}.$$  

In testing the significance of several spatial autocorrelation statistics, the weight structure has to be summarized by several parameters, including $SUM_1$ and $SUM_2$. These two terms are defined as

$$SUM_1 = \frac{1}{2} \sum_i \sum_j (w_{ij} + w_{ji})^2$$

and

$$SUM_2 = \sum_i \left( \sum_j w_{ij} + \sum_j w_{ji} \right)^2.$$  

The first term, $SUM_1$, is the sum over the weights. If the weights are binary and the matrix is symmetric (i.e., the C matrix), then $(w_{ij} + w_{ji})^2 = 4$. The $SUM_1$ is simply four times the total number of joints or shared boundaries in the entire study area. The second term, $SUM_2$, is based upon the sums of the weights associated with each areal unit first, but in both directions (i.e., for both $w_{ij}$ and $w_{ji}$). The sums are then added, squared, and summed over all areal units.

Let’s use $n$ to denote the number of areal units in the entire study area. If there are two groups of areal units, for instance, defined by an attribute, which carries two values, $x$ and $y$, conventionally we use $n_x$ and $n_y$ to indicate the number of areal units in the two groups. Similar to the notation but very different in meaning, we can use

$$n^{(x)} = n \ast (n - 1) \ast (n - 2) \ast (n - 3) \ast \ldots \ast (n - x + 1)$$

where $n > x$. For example, if $n = 5$, $n^{(3)} = n(n-1)(n-2) = 5 \times 4 \times 3$ and $n^{(1)} = n$.

If $x_i$ is the attribute value for areal unit $i$, a new parameter, $m_j$, can be derived based upon $x_i$, as

$$m_j = \sum_{i=1}^n x_i^j,$$

where $j = 1, 2, 3, 4.$ Therefore, if $j = 1$, $m_j$ is the sum of $x_i$ of all $i$. If $j = 2$, $m_j$ is the sum of all the squares of $x_i$.

All these terms will be used later in this chapter for the discussions of various spatial autocorrelation measures. Readers should refer back to this section later as needed. For the rest of this chapter, we will discuss different spatial autocorrelation measures, starting with joint count statistics for binary data and proceeding to global measures (including Moran’s I, the Geary Ratio, and the general-G statistic) and local measures (including LISA and local G-statistics) for interval and ratio data. Finally, we will discuss the Moran scatterplot, a graphical technique used to visualize the spatial heterogeneity of spatial autocorrelation.

### 5.5 Joint Count Statistics

The use of joint count statistics provides a simple and quick way of quantitatively measuring the degree of clustering or dispersion among a set of spatially adjacent polygons. This method is applicable to nominal data only. Because the statistics are based on comparing the actual and expected counts of various types of joints between adjacent polygons having the same or different attribute values, the nominal data appropriate for this method are limited to binary data. Binary data are those with only two possibilities, such as arable/nonarable lands, high/low income groups, urban/rural counties, and so on. To simplify our description, let’s use black and white to indicate the two possible attribute values associated with
polygons. Consequently, the various types of joints would be black-black joints, black-white joints, and white-white joints.

If a polygon pattern has a clustered pattern, such as that of the Case 1 in Figure 5.1, we would expect to have more white-white or black-black joints than white-black joints. This situation is also known as having positive spatial autocorrelation, or a pattern in which similar values are close to each other. Alternatively, a dispersed pattern, or negative spatial autocorrelation, has more black-white joints than black-black or white-white joints. Case 2 in Figure 5.1 is an example of a dispersed pattern. Of course, for a random pattern, we would expect the actual counts of various types of joints to be fairly close to those of a typical random pattern.

While the joint count statistics method is limited to binary data, data in other measurement scales can be easily downgraded or converted to binary form. For a set of ordinal data values, one can set a rank as the cutoff level so that those values above this level are counted as one type and those below it are counted as another type. In ranking cities by their population sizes, those cities ranked larger than Cleveland are assigned one value and those cities ranked smaller than Cleveland are assigned another value. Similarly, downgrading interval/ratio data is only a matter of setting a cutoff point to separate the values above and below this level into two binary values. Take average family income as an example. One can consider cities with an average family income above the national average as one type and cities with an average family income below the national average as another type. Therefore, the two essential conditions for using the joint count statistics are as follows:

1. Data are related to area (polygons).
2. Data are measured in binary form (only two possible values exist).

Please note that although it is feasible to convert data from interval/ratio and ordinal scales into binary form, this process also reduces the amount of information captured by the data, so we normally try to avoid this process if possible. Especially if interval or ratio data are converted into nominal scale, the precision of the original data is lost.

The joint count statistics method calculates the difference in the number of black-black, white-white, and black-white joints between the patterns to be tested. With small spatial systems, it is probably feasible to manually count the number of the three different types of joints in order to derive the statistics. But if the data are already in digital formats and the spatial systems are too large for manual operations, the counting processes have to be automated. Below are the general steps used to derive the three joint count statistics in a computational environment.

1. Let $x_i = 1$ if polygon $i$ is a black polygon and $x_i = 0$ if it is white.
2. Then, for black-black joints, $O_{BB} = \frac{1}{2} \sum_i \sum_j (w_{ij} x_i x_j)$.
3. For white-white joints, $O_{WW} = \frac{1}{2} \sum_i \sum_j [w_{ij} (1 - x_i) (1 - x_j)]$.
4. For black-white or white-black joints, $O_{BW} = \frac{1}{2} \sum_i \sum_j [w_{ij} (x_i - x_j)^2]$.

Please note that the weight, $w_{ij}$, can be the binary weight or the row-standardized weight. The three statistics above are the observed joint counts describing the actual pattern. If we observe a large number of $O_{BB}$ or $O_{WW}$ joins, or both, we may postulate that the observed pattern may exhibit positive spatial autocorrelation or clustering. However, we cannot conclude that positive spatial autocorrelation exists until we demonstrate that the observed pattern is different from a random pattern and that the difference is probably not due to chance or coincidence. That is the concept of likelihood.

The user of the joint count statistics method, however, needs to know how to estimate the likelihood that each polygon has a white or a black value (the attribute value). Different ways of estimating attribute values for the polygons will affect the outcome of the joint count statistics process.

If the probability of a polygon's being black or white is based on known theories or a trend derived from a larger region, the method by which the attribute values are estimated is known as free sampling. This means that the probability of a polygon's being white or black is not limited or affected by the total number of black or white polygons in the group. Consequently, this approach is sometimes referred to as normality sampling. Alternatively, if the probability of a polygon's being black or white is limited by or dependent upon the total number of black or white polygons, the method by which the attribute values are estimated is known as nonfree sampling or randomization sampling.

In our seven-county example, the nonfree sampling case can only have three black polygons and four white polygons, no matter how they are rearranged. Since the total number of black and white polygons is fixed, the method is nonfree sampling, or sampling without replacement. Compared to nonfree sampling, free sampling does not limit how many polygons can be black or white; therefore, the method is also known as sampling with replacement.

When using joint count statistics, the choice between normality and randomization sampling is important. As a rule of thumb, the normality sampling approach should not be used whenever references to trends from larger regions or those outside of the study area cannot be used with certainty. This is because randomization sampling requires less rigorous assumptions than free sampling. Normality sampling should be used if the relationship between the study area and the national trend or the trend from a larger region can be established with known theories or by experience.

5.5.1 Normality Sampling

In both normality sampling and randomization sampling, calculating joint count statistics involves estimation of the expected number of black-black, white-white, and black-white joints and their standard deviations. The expected numbers of these joints reflect a random pattern, or a pattern with no significant spatial autocorrelation of any type. The number of black-black and white-white joints indicates the magnitude of positive spatial autocorrelation, while the number of
black-white or white-black joints indicates the magnitude of negative spatial autocorrelation.

These observed values are compared with their expected counterparts to derive their differences. These differences are then standardized by their corresponding standard deviations in order to obtain standardized scores. Using these scores, we can decide if there is a significant positive or negative spatial autocorrelation in the pattern. In other words, three pairs of comparisons have to be conducted. For illustrative purpose, we will only show in detail how negative spatial autocorrelation can be tested. With this example, the other two situations can be repeated easily or derived using the accompanying ArcView project file.

For normality sampling, the equations for the expected number of black-black and white-white joints are

\[ E_{BB} = \frac{1}{2} W p^2 \]

and

\[ E_{WW} = \frac{1}{2} W q^2. \]

Then the equation for expected black-white joints is

\[ E_{BW} = W pq. \]

where

\[ E_{BB}, E_{WW}, \text{and } E_{BW}, \]

are the expected number of black-black, white-white, and black-white joints, respectively,

\[ p \] is the probability that an area will be black, and

\[ q \] is the probability that an area will be white.

The two probabilities must sum to 100%, or \( p + q = 1.0 \). If no other information is available, a common method is to set \( p = nB/n \). But there are other considerations, to be discussed later, in determining \( p \). If the spatial weights matrix is a binary matrix, the expected values can be simplified to:

\[ E_{BB} = J p^2 \]
\[ E_{WW} = J q^2 \]
\[ E_{BW} = 2 J pq, \]

where \( J \) is the total number of joints in the study area.

To test if the measured pattern is significant enough, a statistical test, known as the Z-test, can be applied. To perform this test, the standard deviations of the expected joints are also needed. When a stochastic weights matrix is used, the three standard deviations are

\[ \sigma_{BB} = \sqrt{\frac{1}{4} p^2 q [S_1 q + S_2 p]} \]
\[ \sigma_{WW} = \sqrt{\frac{1}{4} q^2 p [S_1 p + S_2 q]} \]
\[ \sigma_{BW} = \sqrt{\frac{1}{4} [4 S_1 p q + S_2 p q + (1 - 4 pq)]}. \]

If a binary matrix is used, these formulas are reduced to

\[ \sigma_{BB} = \sqrt{p^2 J + p^3 K - p^4 (J + K)} \]
\[ \sigma_{BB} = \sqrt{q^2 J + q^3 K - q^4 (J + K)} \]
\[ \sigma_{BW} = \sqrt{2 p q J + p q K - 4 p^2 q^2 (J + K)}, \]

where \( \sigma \) are the standard deviations of the corresponding types of joints and \( J, p, \) and \( q \) are as defined previously, \( K \) is \( \sum_{i=1}^{n} L_i(L_i - 1) \). The \( n \) in \( \sum_{i=1}^{n} L_i(L_i - 1) \) is the total number of polygons, and \( L_i \) is the number of joints between polygon \( i \) and all the polygons adjacent to it.

Let us focus on testing negative spatial autocorrelation (i.e., black-white joints) using the binary matrix (and its corresponding formulas). When calculating the expected black-white, the probability of any polygon's being black or white must be known, namely, \( p, \) and \( q \). If this probability is known, the only step needed to complete the task is to count the number of joints, \( J \).

Calculating \( J \) is quite straightforward since it only involves counting the number of joints between the polygons or summing up all values of the binary connectivity matrix and then dividing this number in half. For the example in Figure 5.5, the number of joints is 11. Consequently, we will have \( J = 11 \). Please note that all joints are to be counted for \( J \), including all black-black, black-white, and white-white joints.

Two or more polygons may be contiguous at only one point rather than sharing borders. In this case, it will be up to the analyst to decide if this point should be counted as a joint. That is, the analyst may decide if the rock's case or the queen's case should be used in constructing the spatial weights matrix. If one joint of this type is included, it is imperative that all such joins are counted, both the total number of joints and the number of black-white joins. In this illustrative example, we adopt the rock's case in defining neighbors. Therefore, Cuyahoga County and Portage County are not neighbors. Similarly, Geauga and Summit counties are not treated as neighbors.

With regard to the \( p \) and \( q \) ratios, the analyst often needs to research or look for proper values from theories, literature, historical documents, or past experience.
with the subject. For example, if the subject shown by maps in Figure 5.5 is the
ratio of infant mortality exceeding a preset threshold at the county level, it would
make sense to use the statewide trend to derive proper values for and .
Similarly, if the maps in Figure 5.5 show the preferences for either the Democratic
or the Republican party at the county level, it may also make sense to research how
the two parties fare for the entire state and use that trend to determine appropriate
values for and . To illustrate how to calculate and , let’s assume
that and . Please note that should hold for all cases.

The next step is to calculate the value of . While there may be
several ways in which this number can be calculated, we suggest that a table be
set up to calculate values of , then , and then for each polygon. In
this manner, the task of calculating will be less confusing and easy
to manage.

In Figure 5.5d, the polygons are labeled A, B, C, D, E, F, and G. The number
of joints for each polygon is its . As shown in Table 5.5, summing the number of
all joints yields . When each is reduced by 1, the value of can be easily calculated. The value for is derived by adding up all
values after multiplying each and . With the seven polygons in
Figure 5.5, summing of values equals 52, as shown in Table 5.5.

All the necessary values have now been calculated and can be substituted into the
following equations:

\[
E_{BW} = 2 \times 11 \times 0.3 \times 0.7 = 4.62
\]

\[
\sigma_{BW} = \sqrt{[2 \times 11 + 52] \times 0.3 \times 0.7 - 4 \times [11 + 52] \times 0.3^2 \times 0.7^2}
\]

\[
= \sqrt{15.54 - 11.11} = \sqrt{4.43} = 2.1
\]

A single number of black-white joints, say 5.04, will not show if there is significant
negative spatial autocorrelation in the patterns we have observed. We would
need to compare that number with the expected number of joints in each polygon pattern. This can be done by calculating the difference between the observed
number of black-white joints in each polygon pattern. If we define to be the
observed number of black-white joints in a polygon pattern, we have
and for Figure 5.5a, Figure 5.5b, and Figure 5.5c, respectively. With these
numbers, we know that Figure 5.5a is more clustered than a random pattern since
. Figure 5.5b and Figure 5.5c are more dispersed than a random pattern.

The difference between an expected number and an observed number of black-white joints can help us assess a polygon pattern in terms of its clustering or
dispersion. However, we do not know how far away each pattern is from a random
pattern. Measuring how much each pattern differs from a random pattern requires
that we standardize the calculated difference, \(O_{BW} - E_{BW}\).

To standardize the comparison, we will use the Z-score for this procedure. The
Z-score uses the standard deviation of the difference as the denominator to
standardize the difference:

\[
Z = \frac{O_{BW} - E_{BW}}{\sigma_{BW}}
\]

From Figure 5.5, we have

\[
Z = \frac{4 - 4.62}{2.1} = -0.29 \text{ for Figure 5.5a,}
\]

\[
Z = \frac{6 - 4.62}{2.1} = 0.65 \text{ for Figure 5.5b,}
\]
and

\[ Z = \frac{8 - 4.62}{2.1} = 1.61 \text{ for Figure 5.5c.} \]

According to the probability distribution of the Z-score, any Z value that is less than -1.96 or greater than 1.96 is less likely to happen by chance in 5 out of 100 cases (\( \alpha = 0.5 \)). With this information, we can conclude that none of the patterns in Figure 5.5 has a statistically significant negative spatial autocorrelation or a dispersion pattern.

### 5.5.2 Randomization Sampling

In randomization sampling, the probability of a polygon’s being black or white will depend on the total number of black polygons and the total number of white polygons in the polygon pattern being studied. Using Figure 5.5 as an example again, each of the three possible arrangements has exactly three black polygons and four white polygons. The probability of a polygon’s being black is 3/7, and the probability of its being white is 4/7.

The equations needed to estimate the expected number of joints of all types in nonfree sampling are

\[
E_{BB} = \frac{1}{2} W[n_b(n_b - 1)/n(n - 1)]
\]

\[
E_{WW} = \frac{1}{2} W[n_w(n_w - 1)/n(n - 1)]
\]

\[
E_{BW} = W[n_b(n_b - n_w)/n(n - 1)]
\]

with standard deviations equal to

\[
\sigma_{BB} = \sqrt{\frac{1}{4} \left[ \frac{S_1(n_b(n_b - 1)}{n(n - 1)} + \frac{(S_2 - 2S_1)n_b}{n} \right] - [E_{BB}]^2}
\]

\[
\sigma_{BW} = \sqrt{\frac{1}{4} \left[ \frac{2S_1n_b}{n(n - 1)} + \frac{(S_2 - 2S_1)n_b(n_b + n_w - 2)}{n} \right] - [E_{BW}]^2}
\]

if the stochastic row-standardized matrix is used. \( \sigma_{WW} \) is similar to \( \sigma_{BB} \) defined above, except that \( n_w \) is used instead of \( n_b \). Most of these terms were defined in Section 5.4. If the binary connectivity matrix is used instead, the expected value for black-white joints and the corresponding standard deviation are

\[
E_{BW} = \frac{2Jn_bn_w}{n(n - 1)}
\]

\[
\sigma_{BW} = \sqrt{E_{BW} + \frac{\sum L(L - 1)n_b}{n(n - 1)} + \frac{4(J(J - 1) - \sum L(L - 1))n_b^2}{n^4}}
\]

\[
\sqrt{-E_{BW}^2}
\]

The values of \( J, \sum L(L - 1) \) are as same as those for the free sampling case.

For the polygon patterns in Figure 5.5a, the values of \( E_{BW} \) and \( \sigma_{BW} \) are

\[
E_{BW} = \frac{2 \times 11 \times 3 \times 4}{7 \times 6} = \frac{264}{42} = 6.326
\]

and

\[
\sigma_{BW} = \sqrt{\frac{6.326}{52 \times 3 \times 4} + \frac{1}{7 \times 6} + \frac{4(1110 - 52) \times 3 \times 2 \times 4 \times 3}{7 \times 6 \times 5 \times 4} - 6.326^2}
\]

\[
= \sqrt{6.326 + 14.857 + 19.886 - 39.514} = \sqrt{1.515} = 1.23
\]

Similarly, we can use the Z-score, \( Z = (O_{BW} - E_{BW})/\sigma_{BW} \), to see how each of our sample polygon patterns compares to a random pattern under the condition of three black and four white polygons:

\[
Z = \frac{4 - 6.326}{1.23} = -1.85 \text{ for Figure 5.5a,}
\]

\[
Z = \frac{6 - 6.326}{1.23} = -0.23 \text{ for Figure 5.5b,}
\]

and

\[
Z = \frac{8 - 6.326}{1.23} = 1.39 \text{ for Figure 5.5c.}
\]

Given these Z-scores and the threshold value of ±1.96, we can conclude that none of the polygon patterns in Figure 5.5a has a significant negative spatial autocorrelation at \( \alpha = 0.05 \), but Figure 5.5a exhibits a negative spatial autocorrelation...
lower than a random pattern at $\alpha = 0.1$ (threshold value = ±1.645) level of significance.

5.6 MORAN AND GEARY INDICES

Joint count statistics are useful global measures of spatial autocorrelation for variables with only two outcomes. This situation is quite restrictive, as most real-world cases deal with variables at interval or ratio measurement scales. In these cases, Moran’s I and Geary Ratio C can be used.

**Moran’s I** and **Geary’s Ratio** have some common characteristics, but their statistical properties are different. Most analysts favor Moran’s I mainly because its distribution characteristics are more desirable (Cliff and Ord, 1973, 1981). Still, both statistics are based on a comparison of the values of neighboring areal units. If neighboring areal units over the entire study area have similar values, then the statistics should indicate a strong positive spatial autocorrelation. If neighboring areal units have very dissimilar values, then the statistics should show a strong negative spatial autocorrelation. The two statistics, however, use different approaches to compare neighboring values.

### 5.6.1 Moran’s I

*Moran’s I* can be defined simply as

$$I = \frac{n \sum w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{W \sum(x_i - \bar{x})^2},$$

where $x_i$ is the value of the interval or ratio variable in areal unit $i$. Other terms have been defined previously. The value of Moran’s I ranges from $-1$ for negative spatial autocorrelation to $1$ for positive spatial autocorrelation.

If no spatial autocorrelation exists, the expected value of Moran’s I is

$$E_I = \frac{1}{(n-1)}.$$

When calculating Moran’s I, the spatial weights matrices most commonly used are the binary and stochastic matrices. If a binary matrix is used, $W$ in the denominator is basically twice the number of shared boundaries in the entire study region, or $2J$. However, it is possible to use other types of weights matrices. For our purpose, let us assume that a binary matrix is used.

In the numerator of Moran’s I, if $i$ and $j$ are neighbors, then $w_{ij}$ will be $1$. Therefore, if $i$ and $j$ are not neighbors, the expression will be $0$ for that pair of $i$ and $j$. If they are neighbors, the values of $i$ and $j$ are first compared with the mean of that variable. Their deviations from the mean are then multiplied. The products of the deviations from the mean are then summed for all pairs of areal units as long as they are neighbors. If both neighboring values are above the mean, the product is a large positive number. So is the product if both neighboring values are below the mean (product of two negative numbers).

These situations reflect the presence of positive spatial autocorrelation (i.e., similar values are next to each other). But if the value of one areal unit is above the mean and the value of the neighboring unit is below the mean, the product of the two mean deviations will be negative, indicating the presence of negative spatial autocorrelation. Therefore, over the entire study region, if similar values (can be high-high or low-low) are more likely than dissimilar values between neighbors, Moran’s I tends to be positive, and vice versa.

The numerator of Moran’s I is based upon the covariance, $(x_i - \bar{x})(x_j - \bar{x})$, which is a cross-product. This covariance structure is also the basis of the Pearson product-moment correlation coefficient, which is defined as

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n\delta_x\delta_y},$$

which measures how closely the distributions of the two variables, $x$ and $y$, resemble each other. For a given observation, $i$, if both the $x$ and $y$ values are above their means, the product will be large and positive. Similar result will occur when both the $x$ and $y$ values are below their means. Only if one of the two variables
TABLE 5.6a  Mean Deviations and Squared Values of the Mean Deviations

<table>
<thead>
<tr>
<th>Name</th>
<th>Med Inc</th>
<th>(x - \bar{x})</th>
<th>(x - \bar{x})^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geauga</td>
<td>41,113</td>
<td>10,131</td>
<td>102,637,161</td>
</tr>
<tr>
<td>Cuyahoga</td>
<td>28,996</td>
<td>-2,387</td>
<td>5,697,769</td>
</tr>
<tr>
<td>Trumbull</td>
<td>28,996</td>
<td>-2,387</td>
<td>5,697,769</td>
</tr>
<tr>
<td>Summit</td>
<td>28,996</td>
<td>1,986</td>
<td>3,944,196</td>
</tr>
<tr>
<td>Portage</td>
<td>30,253</td>
<td>-729</td>
<td>531,441</td>
</tr>
<tr>
<td>Ashtabula</td>
<td>24,126</td>
<td>-6,856</td>
<td>47,004,736</td>
</tr>
<tr>
<td>Lake</td>
<td>35,605</td>
<td>-4,623</td>
<td>21,372,129</td>
</tr>
<tr>
<td>Total</td>
<td>216,874</td>
<td></td>
<td>189,005,048</td>
</tr>
<tr>
<td>Mean</td>
<td>30,982</td>
<td></td>
<td>158</td>
</tr>
</tbody>
</table>

TABLE 5.6b  Weights Multiplied by the Cross-Products of Mean Deviations

<table>
<thead>
<tr>
<th>Geauga</th>
<th>Cuyahoga</th>
<th>Trumbull</th>
<th>Summit</th>
<th>Portage</th>
<th>Ashtabula</th>
<th>Lake</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,131</td>
<td>28,996</td>
<td>28,996</td>
<td>28,996</td>
<td>30,253</td>
<td>24,126</td>
<td>35,605</td>
<td>216,874</td>
</tr>
<tr>
<td>-2,387</td>
<td>-2,387</td>
<td>-2,387</td>
<td>-1,986</td>
<td>-729</td>
<td>-6,856</td>
<td>-4,623</td>
<td>4,623</td>
</tr>
<tr>
<td>Geauga</td>
<td>10,131</td>
<td>-2,387</td>
<td>-2,387</td>
<td>-2,387</td>
<td>-1,986</td>
<td>-6,856</td>
<td>4,623</td>
</tr>
<tr>
<td>Cuyahoga</td>
<td>-2,387</td>
<td>10,131</td>
<td>10,131</td>
<td>10,131</td>
<td>10,131</td>
<td>10,131</td>
<td>10,131</td>
</tr>
<tr>
<td>Trumbull</td>
<td>-2,387</td>
<td>-2,387</td>
<td>10,131</td>
<td>10,131</td>
<td>10,131</td>
<td>10,131</td>
<td>10,131</td>
</tr>
<tr>
<td>Summit</td>
<td>-1,986</td>
<td>-2,387</td>
<td>-2,387</td>
<td>10,131</td>
<td>10,131</td>
<td>10,131</td>
<td>10,131</td>
</tr>
<tr>
<td>Portage</td>
<td>-729</td>
<td>-2,387</td>
<td>-2,387</td>
<td>-1,986</td>
<td>-6,856</td>
<td>-4,623</td>
<td>4,623</td>
</tr>
<tr>
<td>Ashtabula</td>
<td>-6,856</td>
<td>-2,387</td>
<td>-2,387</td>
<td>-1,986</td>
<td>-6,856</td>
<td>4,623</td>
<td>4,623</td>
</tr>
<tr>
<td>Lake</td>
<td>4,623</td>
<td>-2,387</td>
<td>-2,387</td>
<td>10,131</td>
<td>-6,856</td>
<td>-4,623</td>
<td>4,623</td>
</tr>
</tbody>
</table>
| Grand Total = -232,462,782
and \( j \). The binary weight is reported in Table 5.1. In Table 5.6b, the mean deviations are also listed together with the corresponding counties. As long as the pair of counties has a 1 in the binary matrix, their corresponding mean deviations will be used to form a product reported in Table 5.6b. These cross-products are then summed to give a grand total of \(-232,462.782\). Therefore, \[
I = \frac{7 \times (-232,462,782)}{26 \times 189,005,048} = -0.3311.
\]

The calculated value for Moran’s I seems to indicate a negative spatial autocorrelation. But we have to compare the calculated Moran’s I with the expected value. In this example, the expected value is \[
E(I) = \frac{-1}{(7 - 1)} = -0.1667.
\]

Since we do not know if this difference is statistically significant, we have to determine if this difference between the calculated and expected Moran’s I values occurred by chance or was very unlikely to happen. To test for the significance of Moran’s I, we adopt the same method used to test the significance of joint count statistics.

The difference between the calculated (observed) and expected values of Moran’s I is scaled by the standard error of Moran’s I in order to derive the z-value. As mentioned before, the expected value of Moran’s I is \( E_I = \frac{1}{(n - 1)} \), i.e., the value that occurs if there is no spatial autocorrelation. The above definition of expected value for Moran’s I applied to any sampling assumptions adopted, but the estimations of variance and standard error vary according to the sampling assumption. Similar to the situation in joint count statistics analysis, there are two sampling assumptions for Moran’s I: normality or randomization.

**Normality** sampling assumes that the attribute values of \( x_i \) are independently drawn from a normal distribution and are not limited by the current spatial pattern. Thus, the distribution properties of Moran’s I are derived from repeated sampling of a set of values from the normal distribution. Different sets of values and their corresponding means are therefore different. The variance of Moran’s I under the normality assumption is

\[
\sigma^2(I) = \frac{n^2S_1 - nS_2 + 3(W)^2}{(W)^2(n^2 - 1)},
\]

where all the terms were defined in previous sections.

Under the randomization assumption, however, the set of values is fixed. What is not fixed is the location associated with each value. In other words, there are many ways to distribute the set of values in the spatial system. The one we observe is one of many possible spatial patterns given the set of values. The configuration that yields no significant spatial autocorrelation is the one generated by distributing the set of values independently and randomly to areal units. Therefore, the variance is based upon the number of possible permutations of the \( n \) data values over the \( n \) locations and is defined as

\[
\sigma^2(I) = \frac{n[(n^2 - 3n + 3)S_1 - nS_2 + 3W^2]}{(n - 1)(n - 2)(n - 3)(W^2)}.
\]

where all these terms are defined as before.

Using the seven Ohio counties data and the binary connectivity matrix, \( n = 7 \), \( S_1 = 52, S_2 = 416 \), and \( W = 26 \). Therefore, under the normality assumption,

\[
\sigma^2(I) = \frac{7^2 \times 52 - (7)(416) + 3(26)^2}{(26)^2(7^2 - 1)} = 0.0513.
\]

If a randomization assumption is used, \( \sigma^2(I) = 0.04945 \). The corresponding z-scores are then

\[
z_n(I) = \frac{-0.3311 - (-0.1667)}{\sqrt{0.0513}} = -0.7258 \quad \text{for normality}
\]

and

\[
z_r(I) = \frac{-0.3311 - (-0.1667)}{\sqrt{0.0495}} = -0.7396 \quad \text{for randomization}.
\]

Because both z-scores are larger than \(-1.96\) using the traditional criterion (the 0.05 confidence level), both tests indicate that Moran’s I is negative. Regardless of which assumption we adopt, we cannot conclude that there is a significant negative spatial autocorrelation in median household income among the seven Ohio counties. The negative Moran’s I may be generated by chance; it is not due to a systematic process.

### 5.6.2 Geary’s Ratio

Similar to the Moran’s I method of measuring spatial autocorrelation, Geary’s Ratio also adopts a cross-product term (Getis, 1991). Geary’s Ratio is formally defined as

\[
c = \frac{(n - 1) \sum \sum w_{ij}(x_i - x_j)^2}{2W \sum (x_i - \bar{x})^2}.
\]

As in Moran’s I, Geary’s Ratio can accommodate any type of spatial weights matrix, although the most popular types are the binary and stochastic matrices. When this formula is compared with the one for Moran’s I, it is apparent that the most significant difference between them is the cross-product term in the numerator. In Moran’s I, the cross-product term is based upon the deviations from the mean of the two neighboring values. In Geary’s Ratio, instead of comparing the neighboring values with the mean, the two neighboring values are compared with each other directly. To a large degree, we are not concerned about whether \( x_i \) is
larger than $x_j$ or vice versa, but we are concerned about how dissimilar the two neighboring values are. Therefore, the differences between neighboring values are squared to remove the directional aspect of the differences. The Geary Ratio ranges from 0 to 2, with 0 indicating a perfect positive spatial autocorrelation (i.e., all neighboring values are the same; thus, the cross-product term becomes 0) and 2 indicating a perfect negative spatial autocorrelation. In contrast to Moran’s I, the expected value of the Geary Ratio is not affected by sample size $n$ but is always 1.

Using the Ohio counties example, we have all the parameters to calculate the Geary Ratio from the Moran’s I calculation except for the sum of the weights times the square of the difference. Table 5.7 shows the derivation of this term. Using the binary connectivity matrix in Table 5.1 again, only if the corresponding cell in Table 5.1 is 1, the corresponding values of the pair of neighbors are compared in Table 5.7. The difference of the value has to be squared too. The sum of the weights times the sum of the squared difference is $2,227,237,238$. Then the Geary Ratio is

$$c = \frac{6 \times 2,227,237,238}{2 \times 26 \times 189,005,048} = 1.3697,$$

indicating a slightly negative situation, which is consistent with the result from Moran’s I.

As in using Moran’s I, we also have to test if the observed Geary Ratio is statistically significant. To derive the $z$-score, we need to know the expected $c$ and its variance. We know that the expected value of $c$ is 1. For variance estimations under the normality assumption

$$\sigma^2(c) = \frac{(2S_1 + S_2)(n - 1) - 4W^2}{2(n + 1)W^2},$$

while under the randomization assumption

$$(n - 1)S_1 \left[ n^2 - 3n + 3 - (n - 1)(m_4/m_2^2) \right] - \frac{1}{4}(n - 1)S_2 \left[ n^2 + 3n - 6 - (n^2 - n + 2)(m_4/m_2^2) \right]$$

$$\sigma^2(c) = \frac{W^2 \left[ n^2 - 3(n - 1)^2(m_4/m_2^2) \right]}{n(n - 2)(n - 1)W^2}.$$

In our Ohio example, the variances are 0.0385 and 0.0288 for the normality and randomization assumptions, respectively. Therefore, the respective $z$-scores, based upon the observed minus expected values, are 1.8341 for normality and 2.119 for randomization. Please note that because a 0 in the Geary Ratio means perfect positive spatial autocorrelation and 1 (expected value) means no spatial autocorrelation, a negative $z$-score means positive spatial autocorrelation and a positive $z$-score indicates negative spatial autocorrelation. Because the $z$-score under the randomization assumption and the observed Geary Ratio is

| TABLE 5.7 Deriving the Numerator of the Geary Ratio (Weights Multiplied by the Cross-Products of Squared Differences) |
|-------------|---------|-----------|
| Gender      | Cuyahoga | Summit    |
| Geary       | 41.113  | 41.113    |
| Cuyahoga    | 41.113  | 41.113    |
| Summit      | 41.113  | 41.113    |
| Total       | 41.113  | 41.113    |
| Cotuban     | 24.224  | 24.224    |
| Cuyahoga    | 24.224  | 24.224    |
| Cotuban     | 24.224  | 24.224    |
| Total       | 24.224  | 24.224    |
| Lake        | 35.668  | 35.668    |
| Cotuban     | 35.668  | 35.668    |
| Total       | 35.668  | 35.668    |
larger than 1 (negative spatial autocorrelation), we can conclude that there is a significant negative spatial autocorrelation in median household income among the seven Ohio counties under the randomization assumption. Please note that in this example, the two spatial autocorrelation measures (Moran’s I and Geary’s Ratio) do not yield consistent conclusions. Even using only one spatial autocorrelation measure (the Geary Ratio) but basing it upon different sampling assumptions gives different results.

ArcView Notes In the project file for this chapter, under the new menu Spatial Autocorrelation, there is a menu item for calculating Moran’s I and Geary’s Ratio. Similar to the procedure for joint count statistics, a spatial weights matrix (binary, stochastic, or distance-based) has to be constructed first. The procedure will ask for the ID field and the variable to be analyzed for spatial autocorrelation. Be sure that the variable you have chosen is an interval or ratio variable. After selecting the weights matrix created previously, the procedure will ask the user to choose a sampling assumption: normality or randomization. If the selected weights matrix is distance-based, the user will be asked to use either a power function or a proportional distance function as the weight. A report window will be created displaying all related statistics, including the observed and expected statistics, variances, and the z-values. This report can be saved as a text file for future use.

5.7 GENERAL G-STATISTIC

Moran’s I and Geary’s Ratio have well-established statistical properties to described spatial autocorrelation globally. They are, however, not effective in identifying different types of clustering spatial patterns. These patterns are sometimes described as “hot spots” and “cold spots.” For instance, if high values are close to each other, Moran’s I and Geary’s Ratio will indicate relatively high positive spatial autocorrelations. This cluster of high values may be labeled as a hot spot. But the high positive spatial autocorrelation indicated by both Moran’s I and Geary’s Ratio could also be created by low values close to each other. This type of cluster can be described as a cold spot. Moran’s I and Geary’s Ratio cannot distinguish these two types of spatial autocorrelation. The general G-statistic (Getis and Ord, 1992) has the advantage over Moran’s I and Geary’s Ratio of detecting the presence of hot spots or cold spots over the entire study area. These hot spots or cold spots can be thought of as spatial concentrations.

Similar to Moran’s I and Geary’s Ratio, the general G-statistic is also based on the cross-product statistics. The cross-product is often labeled as a measure of spatial association. Formally, the general G-statistic is defined as

\[ G(d) = \frac{\sum \sum w_{ij}(d)x_i x_j}{\sum \sum x_i x_j}. \]

for \( i \neq j \). The G-statistic is defined by a distance, \( d \), within which areal units can be regarded as neighbors of \( i \). The weight \( w_{ij}(d) \), is 1 if areal unit \( j \) is within \( d \) and is 0 otherwise. Thus, the weights matrix is essentially a binary symmetrical matrix, but the neighboring relationship is defined by distance, \( d \). The sum of this weights matrix is

\[ W = \sum_i \sum_j w_{ij}(d). \]

where \( j \neq i \). Because of this nature of the weight, some of the \( x_i, x_j \) pairs will not be included in the numerator if \( i \) and \( j \) are more than \( d \) away from each other. On the other hand, the denominator includes all \( x_i, x_j \) pairs, regardless of how far apart \( i \) and \( j \) are. Apparently, the denominator is always larger than or equal (in the most extreme case using a very large \( d \)) to the numerator. Basically, the numerator, which dictates the magnitude of \( G(d) \) statistics, will be large if neighboring values are large and small if neighboring values are small. This is a distinctive property of the general G-statistic. A moderate level of \( G(d) \) reflects spatial association of high and moderate values, and a low level of \( G(d) \) indicates spatial association of low and below-average values.

Before calculating the general G-statistic, one has to define a distance, \( d \), within which areal units will be regarded as neighbors. In the seven Ohio counties, we choose 30 miles for demonstration purposes. Referring to Table 5.8, which shows the centroid distances among these seven counties, 30 miles is large enough for each county to include at least one other county as its neighbor but is not large enough to include all counties for any given county. Based upon the 30-mile criterion used to define neighbors, a binary matrix is derived in Table 5.8. Using median household income as the variable again, the general G-statistic is

\[ G(d) = \frac{22,300,327,504}{40,126,136,560} = 0.5557. \]

But the more detailed interpretation of the general G-statistic has to rely on its expected value and the standardized score (z-score).

To derive the z-score and to test for the significance of the general G-statistic, we have to know the expected value of \( G(d) \) and its variance. The expected value of \( G(d) \) is

\[ E(G) = \frac{W}{n(n - 1)}. \]

The expected value of \( G(d) \) indicates the value of \( G(d) \) if there is no significant spatial association or if the level of \( G(d) \) is average. In the Ohio case,
TABLE 5.8 Converting a Distance Matrix into a Binary Matrix Using 30 Miles as the Threshold

<p>| Distance Matrix Based upon Centroid Distance |</p>
<table>
<thead>
<tr>
<th>ID</th>
<th>Geauga</th>
<th>Cuyahoga</th>
<th>Trumbull</th>
<th>Summit</th>
<th>Portage</th>
<th>Ashtabula</th>
<th>Lake</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geauga</td>
<td>0</td>
<td>25.1508</td>
<td>26.7057</td>
<td>32.7059</td>
<td>25.0389</td>
<td>26.5898</td>
<td>12.6265</td>
</tr>
<tr>
<td>Cuyahoga</td>
<td>25.1508</td>
<td>0</td>
<td>47.8151</td>
<td>23.4834</td>
<td>31.6155</td>
<td>50.8064</td>
<td>28.2214</td>
</tr>
<tr>
<td>Trumbull</td>
<td>26.7057</td>
<td>47.8151</td>
<td>0</td>
<td>41.8561</td>
<td>24.4759</td>
<td>29.5633</td>
<td>36.7535</td>
</tr>
<tr>
<td>Summit</td>
<td>32.7059</td>
<td>23.4834</td>
<td>41.8561</td>
<td>0</td>
<td>17.8031</td>
<td>58.0869</td>
<td>43.7375</td>
</tr>
<tr>
<td>Portage</td>
<td>25.0389</td>
<td>31.6155</td>
<td>24.4759</td>
<td>17.8031</td>
<td>0</td>
<td>45.3341</td>
<td>37.4962</td>
</tr>
<tr>
<td>Ashtabula</td>
<td>26.5898</td>
<td>50.8064</td>
<td>29.5633</td>
<td>58.0869</td>
<td>45.3341</td>
<td>0</td>
<td>24.7400</td>
</tr>
<tr>
<td>Lake</td>
<td>12.6265</td>
<td>28.2214</td>
<td>36.7535</td>
<td>42.7375</td>
<td>37.4962</td>
<td>24.7400</td>
<td>0</td>
</tr>
</tbody>
</table>

Based upon 30 Miles as the Threshold, Distance Matrix Is Converted into Binary Matrix

<table>
<thead>
<tr>
<th>ID</th>
<th>Geauga</th>
<th>Cuyahoga</th>
<th>Trumbull</th>
<th>Summit</th>
<th>Portage</th>
<th>Ashtabula</th>
<th>Lake</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geauga</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cuyahoga</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Trumbull</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Summit</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Portage</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ashtabula</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Lake</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ E(G) = \frac{22}{7 \times 6} = 0.5238. \]

Intuitively, because the observed \(G(d)\) is slightly higher than the expected \(G(d)\), we may say that the observed pattern exhibits some positive spatial association. However, we cannot conclude that this level is significance until we test it. Then we have to derive the z-score of the observed statistic based upon the variance. According to Getis and Ord (1992), the variance of \(G(d)\) is

\[ \text{Var}(G) = E(G^2) - (E(G))^2, \]

where

\[ E(G^2) = \frac{1}{(m_1^n - m_2^n)^2 n^{(4)}} \left[ B_0 m_2^4 + B_1 m_4 + B_2 m_2^2 m_2 + B_3 m_1 m_3 + B_4 m_1^4 \right], \]

where \(m_1\) and \(n^{(x)}\) were defined in Section 5.4. The other coefficients are as follows:

\[ B_0 = (n^2 - 3n + 3)S_1 - nS_2 + 3W^2, \]
\[ B_1 = -(n^2 - n)S_1 - 2nS_2 + 3W^2, \]
\[ B_2 = -2nS_1 - (n + 3)S_2 + 6W^2, \]
\[ B_3 = 4(n - 1)S_1 - 2(n + 1)S_2 + 8W^2, \]
and

\[ B_4 = S_1 - S_2 + W^2, \]

where \(S_1\) and \(S_2\) were defined in Section 5.4.

The median household income data of the seven Ohio counties give \(E(G^2) = 0.2829\). Therefore, the value of \(G(d)\) is

\[ \text{Var}(G) = 0.2829 - (0.5238)^2 = 0.0085 \]

and the standardized score is

\[ Z(G) = \frac{0.5557 - 0.5238}{\sqrt{0.0085}} = 0.3463, \]

which is smaller than 1.96, our standard marker indicating the 0.05 level of significance. In other words, the calculated \(G(d)\) has a mild level of spatial association, and the z-score indicates that the counties with high median household income are close to (within 30 miles of) counties with moderate income; this relationship is not statistically significant. That is, the pattern is probably created by chance rather than by some systematic process.

5.8 LOCAL SPATIAL AUTOCORRELATION STATISTICS

All of the spatial autocorrelation statistics discussed so far share a common characteristic: they are global statistics because they are summary values for the entire study region. It is reasonable to suspect that the magnitude of spatial autocorrela-
tion does not have to be uniform over the region (spatial homogeneity), but rather varies according to the location. In other words, it is likely that the magnitude of spatial autocorrelation is high in some subregions but low in other subregions within the study area. It may even be possible to find positive autocorrelation in one part of the region and negative autocorrelation in another part. This phenomenon is called spatial heterogeneity.

In order to capture the spatial heterogeneity of spatial autocorrelation, we have to rely on another set of measures. All these measures are based upon their global counterparts discussed above (Moran's I, the Geary Ratio, and the general G-statistic) but are modified to detect spatial autocorrelation at a local scale.

### 5.8.1 Local Indicators of Spatial Association (LISA)

Local Indicators of Spatial Association refer to the local version of Moran's I and Geary Ratio (Anselin, 1995). In order to indicate the level of spatial autocorrelation at the local scale, a value of spatial autocorrelation has to be derived for each areal unit. The local Moran statistic for areal unit $i$ is defined as

$$I_i = z_i \sum_{j \neq i} w_{ij} z_j,$$

where $z_i$ and $z_j$ are in deviations from the mean or

$$z_i = (x_i - \bar{x}) / \delta$$

and $\delta$ is the standard deviation of $x_i$. Similar to the interpretation of Moran’s $I$, a high value of local Moran means a clustering of similar values (can be high or low), while a low value of local Moran indicates a clustering of dissimilar values. In general, $w_{ij}$ can be the row-standardized matrix, but other spatial weights matrices are also appropriate. If the weight is in row-standardized form, the local Moran for areal unit $i$ is basically the mean deviation of $i$ multiplied by the sum of the products of the mean deviations for all $j$ values and the spatial weights defining the spatial relationship between $i$ and $j$.

Table 5.9 is similar to Table 5.6. From Figure 5.1 and Table 5.1, we can see that Geauga County has six neighboring counties. If the row-standardized stochastic weights are adopted, the weight for each of Geauga County’s neighbors is $1/6$. The row-standardized weights are reported in Table 5.9a. The number of deviations from the mean value of median household income for all seven counties is reported in Table 5.9b. Among the seven Ohio counties, let us focus on Geauga County. Because Geauga County has six neighbors, the weight for each neighbor is $1/6$. Applying this weight to the mean deviations for all the neighbors of Geauga County defined by the 30-mile criterion, $w_{ij} \cdot z_j$ for each neighbor is reported. The same step is applied to other counties. In Table 5.9b, the products of the weight and mean deviation for all neighboring counties are summed and then multiplied by the mean deviation of the county of concern. That gives us the local
Moran for each of the seven counties. The major difference among the counties is their weights, depending on the number of neighbors each of them has.

As with other statistics, just deriving the local Moran values for each county is not very meaningful. High or low local Moran values may occur just by chance. These values have to be compared with their expected values and interpreted with their standardized scores. According to Anselin (1995), the expected value under the randomization hypothesis is

\[ E[I_i] = -w_i/(n-1), \]

and

\[ \text{Var}[I_i] = w_i^2 \left( \frac{n - m_i/m_j^2}{(n-1)} \right) + 2w_i(kh) \left( \frac{2m_i/m_j^2 - n}{(n-1)(n-2)} \right) - \frac{w_i^2}{(n-1)^2}, \]

where

\[ w_i^2 = \left( \sum_j w_{ij} \right)^2 \]

and

\[ w_i^{(2)} = \sum_j w_{ij}^2; \quad i \neq j. \]

The term

\[ 2w_i(kh) = \sum_{k \neq i} \sum_{h \neq i} w_{ik} w_{ih}. \]

Table 5.9c reports the expected values, variances, and z-scores for all seven counties. Note that because each county has its own local Moran, each local Moran has its associated expected value and variance. This is an advantage of the local Moran. A value is derived for each areal unit, and therefore the results can be mapped. Figure 5.7 consists of two maps: the local Moran values and the z-scores for all seven counties. The local Moran reflects how neighboring values are associated with each other. From Figure 5.6, we can see that the median household income of Geauga County is highest and that one of its neighboring counties, Ashtabula County, has the lowest income. Therefore, the local Moran for Geauga County is rather low and in fact negative. But Ashtabula County has the lowest local Moran because it is very different from its surrounding counties, including Geauga County. On the other hand, Cuyahoga and Summit counties are similar in their relatively low income levels, and therefore their local Moran values are moderately high (similar low values are next to each other). Lake and Portage Counties, and to some extent Trumbull County, surrounding Geauga County, have relatively high or moderate income levels. Therefore, these three counties have very high local Moran values, indicating that neighboring units have rather similar values. Still, none of the standardized local Moran scores exceeds the +1.96 range. Thus, the pattern we observe may be the outcome of a random process.

A local version of the Geary Ratio is also available. Formally, it is defined as

\[ c_i \sum_j w_{ij}(z_i - z_j)^2. \]

Unfortunately, the distribution properties of the local Geary are not as desirable as local Moran. Still, mathematically, we can interpret the local Geary the same way as the global Geary Ratio. Clustering of similar values will create a relatively low local Geary, while clustering of dissimilar values will yield a relatively high local Geary. Similar to the Moran, a local Geary value is computed for each areal unit; therefore, the results can be mapped.
5.8.2 Local G-Statistics

Another local measure of spatial autocorrelation is the local version of the general G-statistic (Getis and Ord, 1992). The local G-statistic is derived for each areal unit to indicate how the value of the areal unit of concern is associated with the values of surrounding areal units defined by a distance threshold, \( d \). Formally, the local G-statistic is defined as

\[
G_i(d) = \frac{\sum_j w_{ij}(d)x_j}{\sum_j x_j}; \quad j \neq i.
\]

All other terms were defined previously in the discussion of the general G-statistic. It is best to interpret the statistic in the context of the standardized score. To obtain the standardized score, we need to know the expected value and the variance of the statistic. The expected value is defined as

\[
E(G_i) = W_i/(n - 1),
\]

where

\[
W_i = \sum_j w_{ij}(d).
\]

The definition of the variance is similar to the definition of the general G-statistic. It is defined as

\[
\text{Var}(G_i) = E(G_i^2) - [E(G_i)]^2
\]

and

\[
E(G_i^2) = \frac{1}{(\sum_j x_j)^2} \left[ \frac{W_i(n - 1 - W_i) \sum_j x_j^2}{(n - 1)(n - 2)} \right] + \frac{W_i(W_i - 1)}{(n - 1)(n - 2)}.
\]

where \( j \neq i \). Given the standardized score of \( G_i(d) \) using the above expected value and variance, a high score appears when the spatial clustering is formed by similar but high values. If the spatial clustering is formed by low values, the \( z \)-score will tend to be highly negative. A \( z \)-score around 0 indicates no apparent spatial association pattern. A related statistic is labeled \( G_i^2(d) \). This statistic is almost identical to \( G_i(d) \), except that it includes cases where \( j = i \). Because these two statistics are so similar, we will focus on \( G_i(d) \). Readers who are interested in the other statistic can refer to Getis and Ord (1992).

Given the interpretation of the local G-statistic, we should expect that, using median household income as the variable, Geauga County will have a high local G-statistic because its two neighbors, Lake and Portage counties, have relatively high values. As shown in the upper map in Figure 5.8 and Table 5.10, the local G-statistic for Geauga County is the highest. Summit County has the lowest local G-statistic because both of its neighbors (defined by the 30-mile distance selected...
a priori) have relatively low levels of income. It is interesting that for Lake County, the three neighbors within 30 miles (Cuyahoga, Geauga, and Ashtabula counties) have very different income levels. As a result, the local G-statistic for Lake County is moderately negative, indicating moderate negative spatial autocorrelation.

### Table 5.10 Local G-Statistics and Related Statistics

<table>
<thead>
<tr>
<th>County</th>
<th>Gi</th>
<th>E(Gi)</th>
<th>VAR(Gi)</th>
<th>Z(Gi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geauga</td>
<td>0.83503</td>
<td>0.83333</td>
<td>0.0609</td>
<td>0.02169</td>
</tr>
<tr>
<td>Cuyahoga</td>
<td>0.56148</td>
<td>0.50000</td>
<td>0.05198</td>
<td>0.26965</td>
</tr>
<tr>
<td>Trumbull</td>
<td>0.50086</td>
<td>0.50000</td>
<td>0.0487</td>
<td>0.02597</td>
</tr>
<tr>
<td>Summit</td>
<td>0.31322</td>
<td>0.33333</td>
<td>0.08899</td>
<td>-0.06741</td>
</tr>
<tr>
<td>Portage</td>
<td>0.52671</td>
<td>0.50000</td>
<td>0.05326</td>
<td>0.11574</td>
</tr>
<tr>
<td>Ashtabula</td>
<td>0.54425</td>
<td>0.50000</td>
<td>0.05229</td>
<td>0.19252</td>
</tr>
<tr>
<td>Lake</td>
<td>0.51765</td>
<td>0.50000</td>
<td>0.05528</td>
<td>0.08507</td>
</tr>
</tbody>
</table>

5.9 MORAN SCATTERPLOT

The development of local spatial autocorrelation statistics acknowledges the fact that spatial processes, including spatial autocorrelation, can be heterogeneous. Apparently, we can adopt one or more of the previously described local measures of spatial autocorrelation to evaluate the entire study area if there is any local instability in spatial autocorrelation. From the statistical visualization and spatial exploratory analysis perspectives, it will be informative and useful if the analyst can identify areas with unusual levels of spatial autocorrelation. Those areas can be regarded as the outliers. A very effective visual diagnostic tool is the Moran scatterplot based upon a regression framework and Moran's I statistic (Anselin, 1995). Assuming that \( z \) is a vector of \( x_i \), the deviation from the mean \( (x_i - \bar{x}) \) and \( W \) is the row-standardized spatial weights matrix, we may form a regression of \( Wx \) on \( x \), while the slope of this regression indicates how the neighboring values are related to each other. In other words, the regression is

\[
x = a + I W x,
\]

where \( a \) is a vector of the constant intercept term and \( I \) is the regression coefficient representing the slope. The slope is therefore also the Moran's I global statistic.

Moran's I reflects the level of spatial autocorrelation, and the statistic is a global summary statistic. Different observations within the study region, however, may show different levels of spatial autocorrelation with neighbors. By plotting \( Wx \) on \( x \) superimposing the regression line, the scatterplot can potentially indicate outliers in terms of magnitude of spatial autocorrelation. If all observations have a similar level of spatial autocorrelation, the scatterplot will show observations lying close to the regression line. If certain observations show unusually high or low levels of spatial autocorrelation locally in reference to their neighbors, those observations will be plotted far above or below the regression line. This regression line reflects the general trend of spatial autocorrelation in the entire region, and the slope parameter is equivalent to Moran's I. In other words, those observations deviating from the general trend of spatial autocorrelation have spatial correlations that are very different from the overall level. Thus, the Moran scatterplot is useful in identifying unusual observations in regard to relationships with neighbors.

In the Ohio case, the median household income values among the seven counties have moderately negative spatial autocorrelation. High-value counties are close to low-value counties. This finding is the result of using a few counties for demonstration. The scatterplot in Figure 5.9 shows that the income (\( x \) is in-

---

**Figure 5.9** Moran scatterplot for median household income, seven Ohio counties.
verse related to the product of the spatial weights and income (Wx). The slope is \(-0.3058\), which is moderately negative and not too different from the Moran's I using the binary weights matrix \((-0.3311)\). Please note that the value of the slope parameter is identical to the Moran's I of median household income when a row-standardized weights matrix is used. The R-square turns out to be quite high, mainly because of a few observations. The extremes are Geauga County, as shown by the point in the lower right corner, and Ashtabula County, as indicated by the point in the upper left corner. The scatterplot basically shows that the high income level of Geauga County (x) is negatively associated with the low income levels of surrounding counties. By contrast, the low level income of Ashtabula County contrasts to the high income levels of its surrounding counties. Because only seven observations are involved, the scatterplot does not reveal any obvious outliers. Figure 5.10 provides another example using 49 states in the continental United States and their median house values. The R-square is lower than in the Ohio case \((0.4732)\), but the slope of Moran's I is 0.5526, indicating moderately positive spatial autocorrelation. In this case, the median house value of a state is positively and moderately associated with the house value levels of the surrounding states. An outlier, however, is California, the state close to the r-axis at the extreme right. The deviation of California from the upward-sloping pattern indicates extremely high house value, in contrast to the values of the surrounding states.

Under the new menu Spatial Autocorrelation, there is a menu item for creating a Moran scatterplot. The scatterplot requires a row-standardized stochastic weights matrix. The procedure will ask for the matrix and the variable to be analyzed. The regression coefficients (the intercept and the slope) will be displayed in a window. The deviations from the mean (x) and the weight times the mean (Wx) will be written to the attribute table in ArcView for plotting a chart. The procedure will continue to ask the user if a scatterplot should be created. If the answer is yes, a Moran scatterplot will be created in the Chart window together with the regression parameters using values in the two new fields written to the attribute table. Please note that if the identification tool in ArcView is chosen, the user can click on the observation (a point) in the plot to find out, which observation the point represents.

5.10 APPLICATION EXAMPLES

The methods discussed in this chapter are used to calculate spatial autocorrelation in various forms. The point count statistic procedure is used with nominal or binary data. Moran's I, Geary's Ratio, and G-statistics are for overall evaluation of spatial patterns, while local indicators and local G-statistics measure local variation in spatial autocorrelation. Finally, the Moran scatterplot is a tool to visually identify unusual levels of spatial autocorrelation among areal units.

In working with these methods for this example, we will use the U.S. county database. Data themes have been extracted for the states of Ohio, Indiana, Illinois, and Kentucky. We will examine the spatial patterns of one variable, the median housing value, in these four states.

There is a median housing value for each county in each of the four states. First, we calculate the average of all median housing values in each state as a cutoff point to create a binary category. Those counties with median housing values less than the average in their state will be assigned a value of 1. Those counties with median housing value greater than the state's average will be assigned a value of 0.

Below we describe the steps for calculating each index of spatial autocorrelation. The general structure for processing each data theme is as follows:

1. Copy the shapefiles from the companion website to this book to your working directory. This is needed so that we can make changes in these files.
2. Calculate descriptive statistics of the chosen variable, Median_val. We will use the average of all median housing values as the cutoff point, setting those counties with higher median housing values equal to 0 and those counties with lower median housing values equal to 1.
3. Edit the attribute table to add a new field for the new binary values of 1s and 0s as defined above.
4. Calculate weight matrices and save them in the working directory.
5. Calculate the joint count statistic using the newly added binary field.
6. Calculate Moran’s I and Geary’s C.
7. Calculate the global G-statistic.
8. Calculate the local indicators for each county.
9. Calculate the local G-statistics for each county.
10. Construct a Moran scatterplot.
11. Repeat Steps 1 to 10 for each state.

When the procedure is completed, we will be able to examine all the results and compare them for the four states. At this point, it is necessary to point out that the more counties (polygons) we have in a data theme, the more time will be needed to complete the calculation. The time required depends on the computer’s speed as well as the distance measures selected.

5.10.1 Preparation
To use the various spatial autocorrelation indices, we need to prepare the data theme and its attribute table. To do so, follow these steps:

1. Start the ArcView and open the project file, Ch5.apr, downloaded from the website.
2. Use View/Add Theme from the View document’s menu to add the oh.shp downloaded from the website.
3. Because we are going to make changes on the themes, if you want to keep the original copy of the theme, you must create a duplicate. Use the Theme/Convert to Shape File menu item from the View document. Save oh.shp in your working directory. When asked if the theme is to be added to the View document, click the Yes button. The original oh.shp can be deleted from the Table of Contents in the View document by clicking it (to make it active) and then using Edit/Delete Theme to remove it from the View document.

Next, we will add a binary attribute field, using the variable being analyzed. First, we will calculate descriptive statistics for this variable and then use its mean to create a new attribute field, hprice.

1. Use Theme/Table from the View document menu to open the associated attribute table.
2. In the Table document, Attributes of Oh.shp, scroll to the right until the field of median housing value (Median_val) appears.
3. Click the title button of Median_val to highlight this field. Use the Field/Statistics... menu item to calculate the descriptive statistics of this field. We see that the average of all 88 counties’ median housing value is $54,884. Figure 5.11 shows the descriptive statistics derived from the median housing values.
4. In the Table document, Attributes of Oh.shp (make the table active), use Table/Start Editing from the menu to begin editing. Figure 5.12 shows the screen when setting the Table for editing.
5. Use Edit/Add Field from the Table document’s menu to invoke the Field Definition window. In it, change the name of the field to be hprice, type to be number, width to be 2, and decimal places to be 0. Notice that a new field is created and highlighted.
6. Use the Table/Query menu item to open the query window named Attributes of Oh.shp. From the list of fields, scroll down and double click

Figure 5.11 Descriptive statistics for median housing values.
Figure 5.12  Edit table document.

[Median.val] to add it to the query. Next, click the "<" button to add it to the query. Finally, type $54,884 to complete the query. Now click the New Set button to select all counties whose median housing value is less than $54,884.

7. Use the Field/Calculate... menu item to open the Field Calculator window. In the [hprice] box, type 1 to set the value of hprice for all selected records to be 1. Click OK to proceed.

8. To assign another value to the unselected records, use the Edit/Switch Selection menu item to select the counties that have not yet been assigned values for hprice. When these are selected, use the Field Calculator to assign them a value of 0 by repeating Step 7 but typing 0 in the [hprice] box.

9. Unselect all records so that the subsequent procedures are applied to all records. To do so, use the Edit/Select None menu item.

10. Use Table/Stop Editing to finish the procedure of adding the binary field. When asked whether you wish to save edits, click the Yes button to save the changes.

With the data theme prepared, the next task is to calculate the weight matrices.

5.10.2 Weight Matrices

Adding the binary field allows the calculation of joint count statistics. In calculating weight matrices, the binary field is directly related to the weight matrix.

The procedure for calculating various weight matrices is as follows:

1. From the View document's menu, use the Spatial Autocorrelation/Creating Weight Matrices menu item to start the process of creating a binary weight matrix. Figure 5.13 shows the added user interface in the project file, Ch5.apr.

2. In the Info window, click OK to confirm that no feature was selected at this point.

3. In the Selection window, click the Yes button to agree to select all records.

4. In the ID Selection window, choose Cnty.fips as the ID field and then click OK to proceed.

5. In the Distance Definition window, choose Binary Connectivity as the distance measure and then click OK to proceed.

Figure 5.13  Menu items for spatial autocorrelation statistics.
6. When prompted, navigate to the working directory. Give ohbinary.dbf as the name of the file. Click OK to proceed. In the subsequent Info window, click OK to acknowledge the creation of the distance matrix.

7. Repeat the steps to create other matrix files by selecting different distance measures in the Distance Definition window. Save the output files using the names ohstoch.dbf for Stochastic Weight, ohnparts.dbf for Distance between Nearest Parts, and ohdist.dbf for Centroid Distance. Store all of these weight matrices in the working directory.

5.10.3 Joint Count Statistics

Joint count statistics use binary variables and either a binary connectivity matrix or a stochastic matrix. To calculate joint count statistic, follow these steps:

1. Use the Spatial Autocorrelation/Joint Count Statistics menu item to start the process. In the Check for Input window, click Yes to indicate that the matrix has been created.

2. In the Get Input window, choose Cnty.fips as the ID field. Click OK to proceed.

3. In the next Get Input window, scroll down the list and select hprice as the variable. Click OK to continue.

4. When prompted, navigate to the directory and select ohbinary.dbf as the weight matrix file if the binary matrix is used.

5. When the calculation is completed, in the Sampling window, choose Normality as the sampling scheme. Click OK to proceed.

6. From the Report window, we see that the z-values for AA and AB joints are all statistically significant (> 1.96 or < -1.96) while the z value for BB joints is not statistically significant. Figure 5.14 shows the results of running joint count statistics.

7. In the Write txt file window, click the Yes button to save the results. Give ohjcs.txt as the text file name.

As the results show, the distribution of median housing values is far from random. In fact, we can conclude that median house value has a significant positive spatial autocorrelation, because the z-value of positive autocorrelation (AA) is larger than 1.96, which is affirmed by the very low negative z-value for AB. But this is measured only at the nominal scale. We shall use other indices to examine the spatial pattern of median housing values in this state.

5.10.4 Moran-Geary Indices

While they are structured differently, both Moran's I and Geary's Ratio are useful measurements of spatial autocorrelation. To calculate their index values, follow these steps:

1. From the View document's menu, use the Spatial Autocorrelation/Moran-Geary menu item to start.

2. In the Check for Input window, click the Yes button to confirm the existence of a spatial weight matrix.

3. In the Get Input window, select Cnty.fips as the ID field.

4. In the next Get Input window, scroll down to select Median_val as the variable to be analyzed.

5. In the Choose Spatial Weight Matrix File window, navigate to the directory and select ohstoch.dbf. At this point, the selection of the spatial weighting matrix file is not limited to the use of ohstoch.dbf or ohbinary.dbf. Users may choose to use ohdist.dbf instead. However, the selection of the matrix file should also be reflected in the selection of the next window.

6. In the Distance Definition window, select Stochastic Weight as the distance measure used to create the matrix file. Again, if you are using a weight matrix file other than ohstoch.dbf, the selection of the distance measure should reflect the matrix file used.

7. From the Report window, we see that Moran's I is 0.449089, with a z-value of 6.94788 under normality sampling, and Geary's C is 0.56978, with a z-value of -6.42476 under Normality sampling. Figure 5.15 shows the results from Moran-Geary menu item.

8. Save the result as ohmg.txt in the directory.
The values of the two indices indicate strong spatial autocorrelation. They confirm the conclusion from joint count statistics analysis that the pattern is not random.

### 5.10.5 Global G-Statistics

Other measures that can be used to measure the global pattern of spatial autocorrelation are the G-statistics. In applying them to the Ohio county data theme, the procedures are as follows:

1. From the View document’s menu, use **Spatial Autocorrelation/Global G-Statistics** to start.
2. In the **Distance Entry** window, enter 50 (miles) as the search distance for defining neighbors.
3. In the **Get Input** window, select **Cnty_fips** as the ID field.
4. In the next **Get Input** window, select **Median_val** as the variable to be analyzed.
5. When prompted, choose **ohcdist.dbf** as the centroid distance matrix file from the directory.
6. Upon completion of the calculation, the statistics are reported in the **Result** window. We can see that the calculated value is 0.144867, with a z-value of 1.7808, not statistically significant at the 95% level. Figure 5.16 shows the output from running G-statistics.
7. When prompted, save the result in the same directory under the name **ohgg.txt**.

### 5.10.6 Local Indicators

To examine how a spatial pattern changes within itself, we will apply the local indicators. To do so, follow these steps:

1. From the View document’s menu, use **Spatial Autocorrelation/Local Indicator** to begin.
2. Click the **Yes** button in the **Check for input window** to confirm the creation of a weight matrix file.
3. Choose **Median_val** as the variable to be analyzed.
4. Choose **ohstoch.dbf** as the stochastic weight matrix file.
5. When the calculation is done, click the Table document, **Attributes of Oh.shp**, to make it active. We see that four attributes have been created: **I_i, Exp_I_i, Val_I_i, Z_I_i, and C_i**. The fields give the values of local indicators, their expected values, variances, z-values, and so on. If desired, the user can double click the **Oh.shp** theme in the View document to invoke the Legend Editor. Specify **Graduated Color** to be the Legend Type and **I_i** to be the Classification Field. Click the **Apply** button to display the spatial pattern of the calculated local indicators. Figure 5.17 gives an example for mapping the local indicators.

It may be necessary to adjust the classification in the Legend Editor or the color scheme to see how these local indicators display the changes.

### 5.10.7 Local G-Statistics

Other local indices are the local G-statistics. To calculate them for the counties, follow these steps:
1. From the View document’s menu, use **Spatial Autocorrelation/Local G-Statistics** to continue.

2. In the **Distance Entry** window, type 50 as the distance in miles.

3. In the **Get Input** window, choose **Cnty_fips** as the ID field.

4. In the next **Get Input** window, choose **Median.val** as the variable to be analyzed.

5. When prompted, choose **ohcdist.dbf** as the centroid distance matrix file.

6. When the calculation is complete, click the Table document to make it active. Again, scroll to the right to see the newly created fields: GI, Ex.GI, Var.GI, and 2.GI, similar to the output from local indicators. These fields give values of local G-statistics, their expected values, their variances, and their z-values.

If desired, the Legend Editor can be used to map the distribution of local G-statistics.

### 5.10.8 Moran Scatterplot

One way to see the index values graphically is to construct a scatterplot to display them. We can use the Moran scatterplot menu to do so:

1. From the View document’s menu, use the **Spatial Autocorrelation/Moran ScatterPlot** menu item to begin.
2. Confirm the creation of a stochastic weight matrix by clicking the **Yes** button in the next window. A stochastic matrix is needed for Moran scatterplot.
3. Choose **Cnty_fips** as the ID field.
4. Choose **ohcdist.dbf** as the weight matrix file.
5. In the **Bivariate Regression Results** window, we see that the **R-squared** value is 0.4145, with the intercept of the regression line being 0.0258799 and the slope being 0.449148. Click **OK** to proceed. Note that the slope is the same as the Moran’s I reported earlier.
6. In the **Bivariate Regression** window, click the **Yes** button to create a Moran scatterplot.
7. When the plot is created, expand the **Chart** window horizontally to give the scatterplot more space for the x-axis. Figure 5.18 shows an example of the results.

In Moran scatterplots, the x-axis is the original value (deviation from the mean) and Wx is the predicted x value based on neighboring values. The scatterplots show regression lines between the two axes. The slope of the regression line is identical to that of Moran’s I using the stochastic weight matrix. If the slope runs from lower left to upper right, the spatial autocorrelation is positive. The spatial autocorrelation is negative if the slope runs from upper left to lower right.

Because the magnitude of spatial autocorrelation over a region is not uniform, to determine the spatial heterogeneity of spatial autocorrelation, the scatterplot can show how closely each area unit aligns with the regression line. The model
is summarized by the $R$-square value that indicates how well the regression represents the distribution. If all areal units have a similar magnitude of spatial autocorrelation, they should lie very close to a straight line. However, outliers can be easily identified.

### 5.10.9 Comparing States

Once the procedures for calculating various spatial autocorrelation statistics are completed, they can be repeated for Indiana, Illinois, and Kentucky. The results of these calculations are shown in Table 5.11.

From the average of median housing values in the counties of each state, we can see that Ohio has the highest housing values, followed by Indiana, Illinois, and Kentucky. The $Z$-values for joint count statistics among the four states show various degrees of departure from a random pattern. For example, Ohio has more AA joins but fewer AB joins than a random pattern. For Indiana, only AB joins are different from a random pattern with statistical significance. Illinois and Kentucky have patterns similar to that of Ohio.

With Moran’s I and Geary’s Ratio, all four states show spatial patterns with statistical significance. Illinois has the strongest spatial pattern. With the largest Moran’s I and the smallest Geary’s Ratio, Illinois has the most clustered pattern, in which adjacent polygons show similar median housing values. This finding is further verified by the global G-statistics. Furthermore, Illinois has the highest $R$-square value in the Moran scatterplot, with the steepest slope for the regression line.

<table>
<thead>
<tr>
<th>Statistical Measure</th>
<th>Ohio</th>
<th>Indiana</th>
<th>Illinois</th>
<th>Kentucky</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average of median</td>
<td>$54,884</td>
<td>$48,715</td>
<td>$47,098</td>
<td>$41,393</td>
</tr>
<tr>
<td>housing values</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA joins</td>
<td>74</td>
<td>61</td>
<td>53</td>
<td>73</td>
</tr>
<tr>
<td>Z(AA) value</td>
<td>2.4289</td>
<td>1.1005</td>
<td>2.4540</td>
<td>1.7234</td>
</tr>
<tr>
<td>BB joins</td>
<td>89</td>
<td>91</td>
<td>141</td>
<td>130</td>
</tr>
<tr>
<td>Z(BB) value</td>
<td>1.1959</td>
<td>1.1650</td>
<td>1.2084</td>
<td>1.2473</td>
</tr>
<tr>
<td>AB joins</td>
<td>68</td>
<td>88</td>
<td>75</td>
<td>106</td>
</tr>
<tr>
<td>Z(AB) value</td>
<td>-2.9935</td>
<td>1.9537</td>
<td>-2.5123</td>
<td>-2.3975</td>
</tr>
<tr>
<td>Moran Index</td>
<td>0.4491</td>
<td>0.3144</td>
<td>0.7471</td>
<td>0.5000</td>
</tr>
<tr>
<td>Z(Moran)</td>
<td>6.9479</td>
<td>4.9828</td>
<td>12.2387</td>
<td>8.7378</td>
</tr>
<tr>
<td>Geary’s Ratio</td>
<td>0.5698</td>
<td>0.6667</td>
<td>0.1986</td>
<td>0.4739</td>
</tr>
<tr>
<td>Z (Geary)</td>
<td>1.7808</td>
<td>1.9137</td>
<td>2.8359</td>
<td>4.1025</td>
</tr>
</tbody>
</table>

### 5.11 SUMMARY

In real-world GIS applications, spatial data are often treated as data without a spatial dimension. Classical statistical techniques, which often assume that the observations are independent of each other, are used indiscriminately on spatial data as if they are ordinary data. The unique characteristics of spatial data are ignored, and the analytical results may be biased. In this chapter, we have argued that a unique characteristic of spatial data is (spatial) dependency, and evaluating spatial dependency, or autocorrelation, is an important step in analyzing spatial data. If the assumption of independence is violated, classical statistical techniques for drawing inferences will be inappropriate.

Different types of spatial autocorrelation and measures have been discussed in this chapter. Global measures are summary measures for the entire region, while local measures depict the situation for each areal unit. Some measures are effective in identifying spatial trends, while others are efficient in distinguishing hot spots from cold spots. All of these measures and tools are descriptive and exploratory. To model spatial autocorrelation, more advanced techniques and models are needed.

### REFERENCES


