

General Issues in Inferential Spatial Statistics

13.1 Types of Spatial Pattern

13.2 The Concept of Autocorrelation

13.3 Who Is My Neighbor?

Geographers often examine spatial patterns on the earth's surface that are produced by physical or cultural processes. These patterns represent the spatial distribution of a variable across a study area. Sometimes geographic variables are displayed as point patterns with dot maps. In chapter 4, descriptive spatial statistics (or geostatistics), such as the mean center, standard distance, and standard deviational ellipse were introduced to summarize point patterns. In other instances, explicit spatial patterns representing data summarized for a series of subareas within a larger study region can be displayed effectively using choropleth maps.

Whether data are presented as points or areas, geographers often want to describe and explain an existing pattern. In this chapter, we focus on several general issues associated with inferential spatial statistics. A recurring theme is the analysis of a sample point pattern or sample area pattern to determine if the overall or "global" arrangement is random or nonrandom. When testing a point or area pattern for randomness, the question is whether the population pattern from the sample was generated by a spatially random process. A secondary theme has to do with "local" inferential spatial statistics that identify and map clusters ("hot spots" and "cold spots") of points or areas within a larger area that have a concentration of very high (or very low) data values.

13.1 TYPES OF SPATIAL PATTERN

Geographers often want to compare an existing spatial pattern to a particular theoretical pattern. Spatial patterns may appear clustered, dispersed, or random (fig. 13.1). In

case 1, both the point and area patterns have a **clustered** appearance. On the point pattern map, the density of points appears to vary significantly from one part of the study area to another, with many points concentrated in the northwest

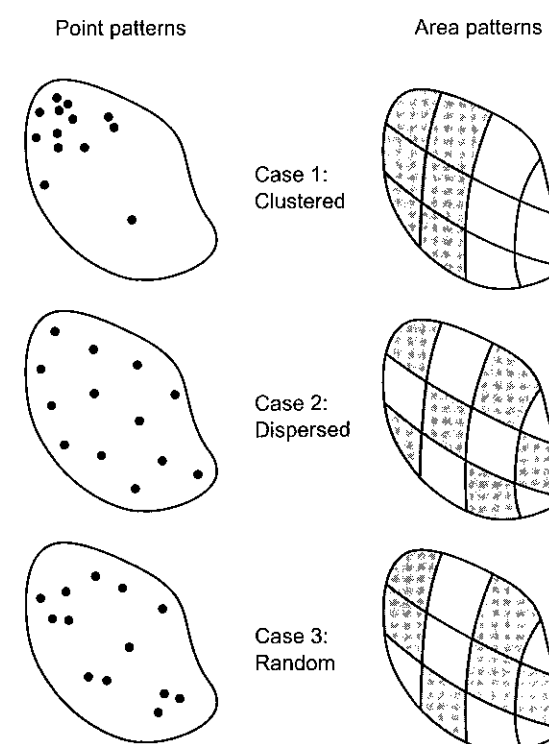


FIGURE 13.1
Types of Point and Area Patterns

portion of the area. Perhaps the points represent sites of tertiary economic activity (retail and service functions), which often cluster around a location with high accessibility and high profit potential, such as a highway interchange. On the clustered area pattern map, the shaded subareas could represent political precincts where a majority of registered voters are Democrat, with Republican majority precincts not shaded. Such a clustered pattern would likely occur if there is a distinctly nonrandom spatial distribution of voters by income, race, or ethnicity in the region.

Other spatial patterns seem evenly **dispersed** or **regular**. The set of points in case 2 (fig. 13.1) appears uniformly distributed across the study area, perhaps suggesting that a systematic spatial process produced the locational pattern. One hypothesis in classical central place theory, for example, is that settlements are uniformly distributed across the landscape to best serve the needs of a dispersed rural population. The area pattern in case 2 exhibits a regular or alternating type of spatial arrangement. This pattern could represent county populations in the same region where the central place distribution of settlements is hypothesized. Shaded counties could have above average populations while counties not shaded are below average.

The spatial patterns in case 3 (fig. 13.1) appear **random** in nature, with no dominant trend toward either clustering or dispersion. A random point or area pattern logically suggests that a spatially random (Poisson) process is operating to produce the pattern. In each of these examples we use the term *appear*, based on our subjective opinion of the distribution of the data. However, the New Mexico wildfire distribution example from chapter 5 illustrates a process that can be statistically evaluated to determine if the wildfire pattern is random or clustered.

In most geographic problems, a point or area pattern will not provide a totally clear indication that the pattern is clustered, dispersed, or random. Rather, many real-world patterns show a combination of these arrangements with tendencies from “purely” random toward either clustered or dispersed.

13.2 THE CONCEPT OF AUTOCORRELATION

In 1970, geographer Waldo Tobler stated that “Everything is related to everything else, but near things are more related than distant things.” This quote became known as Tobler’s Law, but, more importantly, it illustrates the phenomenon of **spatial autocorrelation**. Spatial autocorrelation simply measures the degree to which a geographic variable is correlated with itself through space. Like traditional correlation discussed in chapter 16, spatial correlation is measured as either being positive, negative, or non-existent.

Positive spatial autocorrelation means that geographic objects that are near one another tend to be similar. That

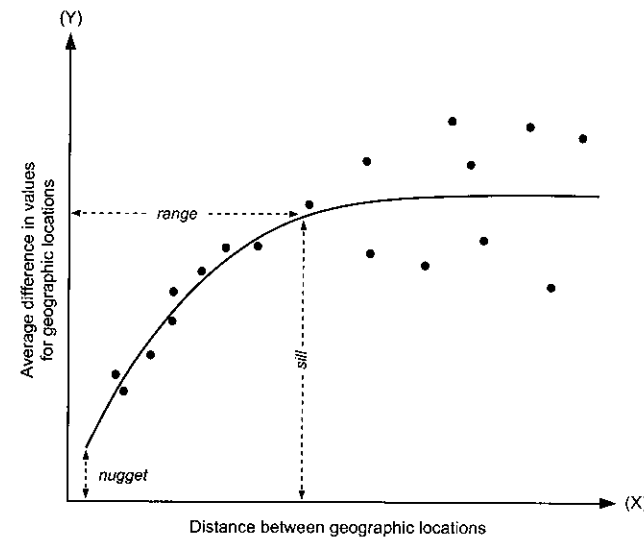


FIGURE 13.2
An Example of a Variogram

is, high value features tend to be located near other high value features, medium value features near other medium value features, and so on. Negative spatial autocorrelation means that geographic objects near one another tend to have sharply contrasting values (a high value located adjacent to a low value, for example). Most geographic phenomena have some measure of positive spatial autocorrelation. Nearby cities tend to get similar amounts of rainfall; expensive homes tend to be located near other expensive homes.

One can visualize spatial autocorrelation through use of a **variogram**. A variogram is a type of scatterplot that displays the differences in values between geographic locations against the differences in the distance between the geographic locations. The Y axis records the average variance in values among a set of geographic objects, but for reasons beyond our discussion, the actual measure used is *half* the variance (termed *semi-variance*). The X axis shows the average distance among the objects, which is termed the **range**. For simplicity, the distances are aggregated into more easily understood groups like 100 miles, 200 miles, etc. By plotting a variogram you can determine the average difference in values among locations that are 200 miles apart, 500 miles apart, and any other logical distance that you would like. A generalized example of a variogram is shown in figure 13.2.

In this example, the Y axis represents the average difference in *values* between geographic locations, while the X axis represents the difference in *distances* between geographic objects. You can see that geographic locations near one another tend to have smaller differences, while geographic locations with larger differences are generally further apart. Rather than showing the individual points, a curved line is used to represent the nature of the rela-

tionship. While beyond the scope of our discussion, in advanced geostatistics the mathematical properties of the best-fitting curve are used to estimate values at unknown locations. The first part of the variogram (left portion of the X axis) illustrates the change in average differences in data values as the distances become greater, while the flat portion of the variogram (right portion of the X axis) shows there is no change in the average differences in data values when the locations are very far apart. In other words, the flat portion of the graph suggests no relationship between the average difference in values for pairs of geographic objects and their distance from one another, while the earlier portion of the graph shows that the values become less similar as the distances among the locations increase.

The **range** represents the distance at which the differences in values of geographic locations are no longer correlated, while the **sill** represents the average difference in value where there is no relationship between location and value. Typically, if two objects are in the same location, they should have the same value and thus no difference. However, whenever samples are taken, some degree of uncertainty always occurs due to measurement error, the inherent properties of the measuring device, or some other reason. The **nugget** represents the degree of uncertainty when measuring values for geographic locations that are very close to one another.

For our purposes, we do not address the mathematical derivation of the variogram (you can learn more about variograms in some of the more advanced readings listed at the end of this chapter). Rather than going into statistical detail, we illustrate the variogram concept with a short example. In chapter 1, an example shows the Last Spring Frost (LSF) date for selected weather stations in

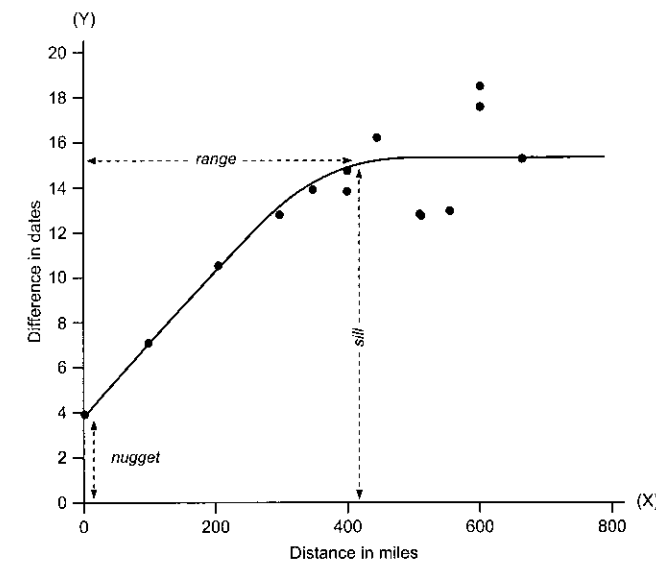


FIGURE 13.3
Variogram of Last Spring Frost (LSF) Date

the Southeast United States. Naturally, we would expect the LSF date for two nearby weather stations to be similar, while stations further separated from one another would have less similarity.

This relationship is illustrated perfectly by the set of observations and curve that “best fits” through those observations (fig. 13.3). Between 0 and 400 miles it appears that if the distance separating a pair of weather stations is large, then their dates of last spring frost are likely to be quite different as well. Furthermore, this relationship between distance and differences in LSF date appears linear. Notice, however, what happens with pairs of weather stations located more than 400 miles apart: the scatterplot of points disperses around the best fit curve and appears random. This means there is no relationship between the LSF dates of weather stations when they are more than 400 miles apart. Therefore, the **range** is said to represent the distance at which the LSF dates are no longer spatially autocorrelated—in this example, about 400 miles.

The concept of spatial autocorrelation is critical in analyzing geographic phenomena for a number of reasons. First, by just pushing a button, GIS software allows users to analyze the relationship between points no matter how far apart they are. However, in our example we have concluded that there is no relationship in LSF values between weather station pairs that are more than 400 miles apart. Therefore, using a GIS or statistical process that assumes a relationship among the data may have limited use if the points are more than 400 miles apart. In short, to make intelligent observations about spatial data, we must understand the spatial structure of the data.

Second, and more importantly, the presence of spatial autocorrelation presents a real challenge when performing statistical analysis. Most inferential statistics assume that observations are independent of one another. Unfortunately, we have just demonstrated that LSF dates are actually spatially correlated (and thus dependent) with one another for weather stations less than 400 miles apart. Why is this important? Simply stated, confidence intervals calculated from spatially autocorrelated data will likely be different than confidence intervals calculated from data that are not spatially autocorrelated.

Recall from chapter 8 that sample size is an important component in the computation of standard error; the greater the sample size, the better our estimates. In the case of spatially autocorrelated data, imagine a geographer sampling 30 objects, with five of those objects very close to one another. Essentially, selecting five nearby observations (assuming spatial autocorrelation within that distance) is almost like selecting the same observation five times. While you may believe you have 30 independent observations, in reality (since five of them are spatially autocorrelated), it is probably more practical to recognize that you essentially have only 25 independent observations. Therefore, the value of *n* in the confidence interval

equation is too high (you will be erroneously making n equal to 30 rather than 25), thus giving you a smaller standard error than warranted. Also, if the five spatially autocorrelated values are similar, then the standard deviation for the entire dataset will also most likely be lower, yielding an even lower standard error. Therefore, you must take extreme care when interpreting results for inferential analysis when the data exhibits spatial autocorrelation.

A final characteristic of autocorrelation worth mentioning here is that it can be measured either globally or locally. For instance, a social geographer may be interested in determining if the distributions of various racial and ethnic groups in a city have different levels of spatial autocorrelation. Perhaps it is hypothesized that one racial or ethnic group has a higher positive autocorrelation across the entire area of the city than another. That is, spatial autocorrelation is being used as a comparative **global** measure of segregation or integration. It is “global” in the sense that a single spatial autocorrelation value is derived for each group across the entire study area, telling us the degree to which the overall spatial pattern of each group is clustered, random, or dispersed. In chapter 15, we apply a measure of global autocorrelation to several racial and ethnic groups in Cleveland, Ohio.

By contrast, **local** autocorrelation methods compare each geographic object with its surrounding neighbors, thereby measuring whether the values in the immediate neighborhood are clustered, random, or dispersed. Many global methods for measuring spatial autocorrelation have been modified to provide local measures. Also in chapter 15, we evaluate the spatial pattern of obesity levels for Pennsylvania counties and identify particular counties having significant local spatial autocorrelation.

For the last half century, geographers and statisticians have continued to study methods to identify and account for the amount of spatial autocorrelation among variables. It is not unlike correlation over time. For instance, during a long recession, a company's sales for one year may be very similar to the sales in a previous year. Economists apply advanced statistical formulas to account for this correlation over time. Geographers are now applying similar principles to account for correlation over space.

13.3 WHO IS MY NEIGHBOR?

In the famous parable of the Good Samaritan, a man asked “who is my neighbor?” Geographers studying spatial autocorrelation have recently been asking the same question. Geographers have different ways of defining the term “neighbor.” One approach is to define **neighbor** in a binary way—that is, two geographic objects are either neighbors or they are not. One acceptable criterion for this binary relationship can be represented as a form of geographic association called **adjacency**. For example,

Pennsylvania is adjacent to New York because they share a common boundary. However, neither state is adjacent to California. Another approach is to define “neighbor” using a **distance threshold** (or cut-off distance) measure. For example, if we establish a distance threshold of 300 miles, New York City and Boston would be considered neighbors, while Boston and Los Angeles would not.

“Neighbor” may also be defined as a continuous variable, by measuring the strength of “neighborliness” between two objects as a function of the distance separating them. This logic is comparable to spatial interaction models. For example, a simple **inverse-distance** weight ($1/\text{distance}$) could establish a measure of proximity. Using this model, the inverse distance weight of New York City and Boston is $1/189$ miles, or .005, while the inverse distance weight of Boston and Los Angeles is $1/2588$ miles, or .0004. In this case, the “neighborliness” or interaction measure is 12 times stronger for New York City and Boston than it is for Boston and Los Angeles. Yet another measure of neighborly interaction is **inverse-distance squared** ($1/\text{distance}^2$).

However neighbors are defined, geographers typically represent them in the form of a weight, where w_{ij} is the weight between geographic objects i and j . The weight can be either 0 or 1 for a binary representation or a ratio such as $1/\text{distance}$ for a continuous function.

KEY TERMS

autocorrelation (spatial autocorrelation), 206
 clustered, random and dispersed patterns, 205, 206
 global and local measures of autocorrelation, 208
 neighbor definitions: adjacency, distance threshold,
 inverse-distance, inverse-distance squared, 208
 range, sill, and nugget in a variogram, 207
 variogram, 206

MAJOR GOALS AND OBJECTIVES

If you have mastered the material in this chapter, you should now be able to:

1. Understand the concepts of clustering, dispersion, and randomness in point and area patterns.
2. Recognize situations in which point and area patterns appear clustered, dispersed, or random.
3. Explain why it is important to measure spatial autocorrelation.
4. Explain the basic purpose of a variogram and understand the terminology associated with a variogram.
5. Identify some ways in which the term “neighbor” is defined in spatial statistics.
6. Explain the basic difference between “global” and “local” measures of spatial association.

REFERENCES AND ADDITIONAL READING

Included below are a number of references that refer to geostatistics (spatial statistics) in general. Many of these sources also cover the point pattern and area pattern techniques discussed in chapters 14 and 15.

- Anselin, Luc. “Local Indicators of Spatial Association—LISA.” *Geographical Analysis*, Vol. 27 No. 2 (1995): 93–115.
- Bailey, T. C. and A. C. Gatrell. *Interactive Spatial Data Analysis*. London: Longman, 1995.
- Barnes, R. *Variogram Tutorial*. Golden Software Inc. <http://www.goldensoftware.com/variogramTutorial.pdf>
- Cliff, A. and J. Ord. *Spatial Processes: Models and Applications*. London: Pion, 1981.
- Ebdon, D. *Statistics in Geography. A Practical Approach*. Oxford: Basil Blackwell, 1985.
- Goodchild, M. *Spatial Autocorrelation*, CATMOG No. 47. Norwich, England: Geo Books, 1988.
- Griffith, D. A. *Spatial Autocorrelation: A Primer*. Washington, DC: Association of American Geographers, 1988.
- Griffith, D. and C. Amrhein. *Statistical Analysis for Geographers*. Englewood Cliffs, NJ: Prentice Hall, 1991.
- Isaaks, E. H. and R. M. Srivastava. *An Introduction to Applied Geostatistics*. Oxford: Oxford University Press, 1989.
- Lee, J. and D. W. S. Wong. *Statistical Analysis with ArcView GIS*. New York: Wiley, 2001.
- Mitchell, A. *The ESRI Guide to GIS Analysis. Vol. 2: Spatial Measurements and Statistics*. Redlands, CA: ESRI Press, 2005.
- Odland, J. *Spatial Autocorrelation*. Vol. 9, Scientific Geography Series. Beverly Hills, CA: Sage, 1988.
- Ripley, B. *Spatial Statistics*. New York: Wiley, 1981.

Point Pattern Analysis

14.1 Nearest Neighbor Analysis

14.2 Quadrat Analysis

This chapter presents methods for analyzing point patterns that are distributed across an entire study region (that is, they are *global* methods of analysis). Such spatial patterns are characterized by observations appropriately displayed as a series of points or dots on a map and are very common in geographic studies. For example, urban geographers plot the location of settlements as points on a map. Economic geographers study the spatial pattern of retail activities by mapping store locations as dots on the map. Physical geographers show glacial features like drumlins as a series of points, while crime analysts may plot the location of burglaries as a series of dots.

In some problems, the *spacing of individual points* within the overall point pattern is important, in which case a statistical procedure known as "nearest neighbor analysis" is applied. For example, an urban geographer might analyze the existing configuration of fire stations in a city to determine whether the pattern is random or more dispersed than random. Suppose one of the goals in the provision of fire service is to provide an equitable or dispersed distribution of fire stations throughout the city so that no one lives or works extremely far from the nearest fire station. The geographer might suggest the siting of new stations or relocation of existing facilities to meet this goal. The proposed new configuration of fire stations could then be analyzed to determine if it were more dispersed than the existing pattern.

In other instances, the *nature of the overall point pattern* is important, and a statistical test known as "quadrat analysis" is often appropriate. For example, a biogeographer might select a random sample of trees in a national forest to determine which trees are diseased, plot the location of each diseased tree as a point on a map, and

analyze the points. The study would determine whether the trees were randomly distributed throughout the study area or clustered in certain portions of the forest. The result of this analysis would provide some guidance concerning the most appropriate type of treatment (e.g., widespread aerial spraying versus concentrated treatment from the ground).

For many geographic studies that begin with analysis of a variable having locations on a dot map, a primary objective is to determine the form of the pattern of points. The nature of the point pattern might reveal information about the underlying processes that produced the distribution. In addition, a series of point patterns of the same variable recorded at different times can help determine temporal changes in the locational process. In short, point pattern analysis offers quantitative tools for examining a spatial arrangement of point locations on the landscape as represented by a conventional dot map.

14.1 NEAREST NEIGHBOR ANALYSIS

Nearest neighbor analysis (NNA) is a widely used procedure to determine the spatial arrangement of a pattern of points within a study area. The distance of each point to its "nearest neighbor" is measured and the mean nearest neighbor distance for all points is determined. The spacing within a point pattern can be analyzed by comparing this observed mean distance to some expected mean distance, such as that for a random (Poisson) distribution.

The nearest neighbor technique was originally developed by biologists interested in studying the spacing of plant species within a region. They measured the dis-

tance separating each plant from its nearest neighbor of the same species and determined whether this arrangement was organized in some non-random manner or was the result of a random process. Geographers have applied the technique in numerous research problems, including the study of settlements in central place theory, economic functions within an urban region, and the distribution of earthquake epicenters in an active seismic region. In all applications, the objective of nearest neighbor analysis is to describe the pattern of points within a study area and make inferences about the underlying process.

The nearest neighbor methodology is illustrated using the example of seven points shown in figure 14.1. A coordinate system is created and the horizontal (*X*) and vertical (*Y*) positions of the points recorded (table 14.1). For each of the points, the nearest neighbor (*NN*) is determined as the point closest in straight-line (Euclidean) distance. The distances to each nearest neighbor (*NND*) are then calculated from the coordinates or measured from the map. From the set of nearest neighbor distances, the *mean* nearest neighbor distance ($\overline{NND}$) is determined using the basic formula for the mean:

$$\overline{NND} = \frac{\sum NND}{n} \quad (14.1)$$

where *n* = number of points.

The mean nearest neighbor distance of the seven-point pattern is 2.67 distance units. This mean nearest neighbor distance provides an index of spacing for the set of points. However, the usefulness of this descriptive

index comes from comparing the index value for an observed pattern to the results produced from certain distinct or theoretical point distributions. This objective is analogous to the situation discussed in section 12.1, where an observed frequency distribution is compared to a perfectly normal frequency distribution.

In NNA, mean nearest neighbor distances can be calculated for three distinct theoretical point arrangements: perfectly random, perfectly dispersed, and perfectly clustered. In each case, the spacing index is determined for an area containing a certain density of points. For example, if points are arranged in a random spatial pattern, the mean nearest neighbor distance ($\overline{NND}_R$) would be determined as follows:

$$\overline{NND}_R = \frac{1}{2\sqrt{\text{Density}}} \quad (14.2)$$

where $\overline{NND}_R$ = mean nearest neighbor distance in a random pattern

Density = number of points (*n*)/Area

Since the study area displayed in figure 14.1 has a dimension of 10 by 10 units, the area represented is 100 square units. The corresponding density of points is therefore 7/100 or .07. For an area with this point density, a random arrangement of points within the study area should produce mean nearest neighbor distance:

$$\overline{NND}_R = \frac{1}{2\sqrt{.07}} = 1.89$$

TABLE 14.1

Coordinates and Nearest Neighbor Information for Example*

Point	X	Y	NN	NND
A	1.3	0.9	E	3.94
B	3.2	4.4	C	2.00
C	3.3	6.4	B	2.00
D	5.6	3.8	E	1.36
E	4.8	2.7	D	1.36
F	8.1	7.4	G	4.21
G	9.4	3.4	D	3.82
SUM				18.69

where *NN* = nearest neighbor
NND = nearest neighbor distance

$$\overline{NND} = \frac{\sum NND}{n} = \frac{18.69}{7} = 2.67$$

* See figure 14.1 for graph of point locations.

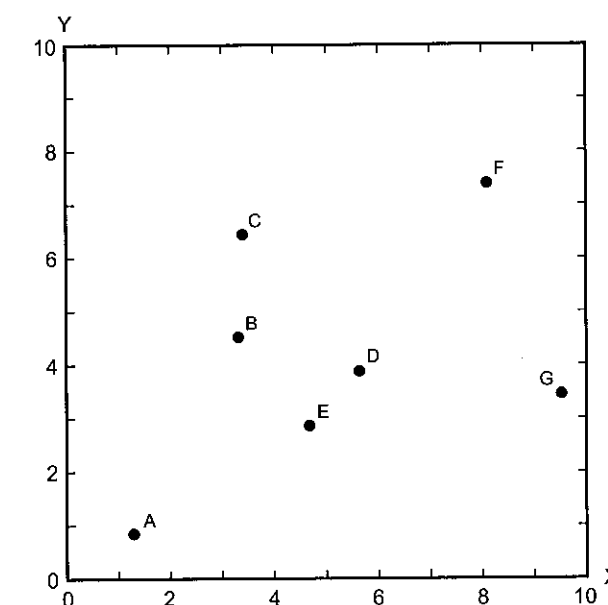


FIGURE 14.1
Location of Points for Nearest Neighbor Problem

If the arrangement of points shows maximum dispersion or a perfectly uniform pattern, the mean nearest neighbor distance ($\overline{NND}_D$) would be determined as:

$$\overline{NND}_D = \frac{1.07453}{\sqrt{\text{Density}}} \quad (14.3)$$

Since this pattern represents the most dispersed or separated arrangement of points, it also serves as the maximum value for the nearest neighbor index. The mean nearest neighbor distance for a dispersed pattern whose point density matches that in figure 14.1 is:

$$\overline{NND}_D = \frac{1.07453}{\sqrt{.07}} = 4.06$$

The pattern of spacing most distinct from dispersion or uniformity is "clustering." When all points lie at the same position, the pattern shows maximum clustering of points, and each nearest neighbor distance is zero. Therefore, in a perfectly clustered pattern, the mean nearest neighbor distance $\overline{NND}_C$ is also zero, representing the lowest possible value for the index:

$$\overline{NND}_C = 0 \quad (14.4)$$

Since perfectly clustered and perfectly dispersed patterns provide the extreme spacing arrangements for a set of points, the nearest neighbor index offers a useful method to measure the spacing of locations within an observed point pattern. However, the mean nearest neighbor distance is an absolute (as opposed to relative)

Nearest Neighbor Analysis	
Primary Objective:	Determine whether a random (Poisson) process has generated a point pattern
Requirements and Assumptions:	1. Random sample of points from a population 2. Sample points are independently selected
Hypotheses:	$H_0: \overline{NND} = \overline{NND}_R$ (point pattern is random) $H_A: \overline{NND} \neq \overline{NND}_R$ (point pattern is not random) $H_A: \overline{NND} > \overline{NND}_R$ (point pattern is more dispersed than random) $H_A: \overline{NND} < \overline{NND}_R$ (point pattern is more clustered than random)
Test Statistic:	$Z_n = \frac{\overline{NND} - \overline{NND}_R}{\sigma_{\overline{NND}}}$

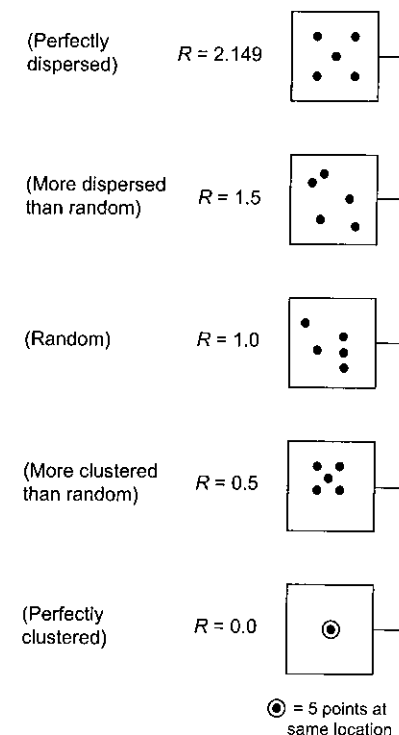


FIGURE 14.2
Continuum of R Values in Nearest Neighbor Analysis

index, dependent upon the units used to measure distance. Therefore, direct comparison of results from different problems or different regions is difficult. Although the minimum value of the nearest neighbor index is always zero (a perfectly clustered pattern), the maximum value corresponding to a perfectly dispersed pattern is not constant, but a function of the point density. To overcome this complication, a **standardized nearest neighbor index (R)** is often used. This index is found by dividing the mean nearest neighbor distance ($\overline{NND}$) by the corresponding value for a random distribution with the same point density:

$$R = \frac{\overline{NND}}{\overline{NND}_R} \quad (14.5)$$

With the standardized index, a perfectly clustered pattern produces an R value of 0.0, a perfectly random distribution 1.0, and a perfectly dispersed arrangement generates the maximum R value of 2.149 (fig. 14.2). Thus, an actual point pattern can be measured for relative spacing along a continuous scale from perfect clustering to perfect dispersal. For the set of points from figure 14.1, the standardized nearest neighbor index is:

$$R = \frac{2.67}{1.89} = 1.41$$

This spacing pattern is moderately dispersed, since it lies between a perfectly dispersed distribution and a random one.

In addition to its use as a descriptive index of point spacing, the nearest neighbor methodology can also be used to infer results from a sample of points to the population from which the sample was drawn. A difference test can be used to determine if the observed nearest neighbor index ($\overline{NND}$) differs significantly from the theoretical norm ($\overline{NND}_R$), which would occur if the points were randomly arranged. The expectation is that a Poisson process operates over space to produce a random pattern of points. The null hypothesis is that no difference exists between observed and random nearest neighbor values. The test statistic (Z_n) follows a format similar to other difference tests discussed in chapters 9 through 12:

$$Z_n = \frac{\overline{NND} - \overline{NND}_R}{\sigma_{\overline{NND}}} \quad (14.6)$$

Example: Nearest Neighbor Analysis—Community Services in Toronto, Canada

The nearest neighbor methodology is now applied to point patterns representing community service sites within Toronto, Ontario. Some public services are best located in a highly dispersed pattern to provide relatively equal service distance for all parts of the region. Such a concern for spatial equity is especially important for emergency services such as police and fire. Other community activities may be sited to offer the region a high degree of efficiency or cost-effectiveness, with less concern for equal spacing of services within the region. Many nonemergency services exhibit more clustered arrangements in their point patterns, possibly reflecting financial rather than spatial constraints.

Four community services in Toronto are selected for a nearest neighbor analysis of their spacing. Two of the services, police and fire, provide emergency protection from sites of their facilities to various locations within the region. The other two services, elementary education and voting locations, provide nonemergency services for persons who travel to these locations. The facility sites for the four services are shown in figure 14.3.

A nearest neighbor analysis is applied to each pattern to measure the spacing of service sites. The calculation procedure for analyzing the spacing of police facilities is summarized in table 14.2. The nearest neighbor distances (in kilometers) for each of the 20 sites are shown in figure 14.4. The mean nearest neighbor distance for police stations in Toronto is 3.63 kilometers. If the stations had been distributed randomly across the city, the mean spacing would have been 2.83 kilometers (table 14.2). To hold the influence of point density constant and allow more useful comparisons of nearest neighbor results, a standardized nearest neighbor index (R) is calculated. This ratio of $\overline{NND}$ to $\overline{NND}_R$ for police stations in Toronto ($R = 1.28$) suggests that the actual spacing of points is more dispersed than random. Although dispersal of sites is expected for an emergency public service, does the R value of

where $\sigma_{\overline{NND}}$ = standard error of the mean nearest neighbor distances

The standard error for the nearest neighbor test can be estimated with the following formula:

$$\sigma_{\overline{NND}} = \frac{.26136}{\sqrt{n(\text{Density})}} \quad (14.7)$$

where: n = number of points

Either a directional (one-tailed) or non-directional (two-tailed) approach can be applied, depending on the form of the alternate hypothesis (H_A). If the problem suggests a clear rationale for the actual point pattern being either dispersed or clustered (as opposed to the null hypothesis of randomness), a one-tailed approach is warranted. However, without an underlying reason or theoretical expectation that the null hypothesis should be rejected on one side as opposed to the other, a two-tailed non-directional approach is best.

TABLE 14.2

Worktable for Nearest Neighbor Analysis: Police Service Facilities in Toronto, Ontario

$H_0: \overline{NND} = \overline{NND}_R$ (point pattern is random)
 $H_A: \overline{NND} \neq \overline{NND}_R$ (point pattern is not random)

Calculate mean nearest neighbor distance:

$$\overline{NND} = \frac{\sum NND}{n} = \frac{72.68}{20} = 3.63$$

where: n = number of points

Calculate random nearest neighbor distance:

$$\overline{NND}_R = \frac{1}{2\sqrt{\text{Density}}}$$

where: $\text{Density} = n/\text{Area}$

$$\overline{NND}_R = \frac{1}{2\sqrt{\frac{20}{641.45}}} = \frac{1}{2\sqrt{0.0312}} = \frac{1}{0.1766} = 2.83$$

Calculate standardized nearest neighbor distance:

$$R = \frac{\overline{NND}}{\overline{NND}_R} = \frac{3.63}{2.83} = 1.28$$

Calculate test statistic:

$$Z_n = \frac{\overline{NND} - \overline{NND}_R}{\sigma_{\overline{NND}}}$$

where: $\text{Density} = n/\text{Area}$

$$\sigma_{\overline{NND}} = \frac{0.26136}{\sqrt{20(0.0312)}} = \frac{0.26136}{0.7899} = 0.3309$$

$$Z_n = \frac{3.63 - 2.83}{0.3309} = \frac{0.80}{0.3309} = 2.42 \quad (p = 0.0078)$$

(continued)

1.28 differ significantly from 1.0, the result for a random pattern of points?

Equations 14.6 and 14.7 are used to test for a significant difference between an observed nearest neighbor index and the corresponding index value for a random spacing of points. The null hypothesis is that the observed and random nearest neighbor distances are equal. In this problem, a one-tailed (directional) test is logical, since police facilities are hypothesized to have a dispersed locational pattern. Therefore, the alternate hypothesis is that the observed nearest neighbor index is larger than the corresponding random value. The resulting test statistic for police locations in Toronto ($Z_n = 2.42$;

$p = .0078$) indicates that you can reject the null hypothesis with more than 99% certainty and conclude that the spacing of police stations is more dispersed than random.

Inferential interpretations in this problem should be made only with extreme care. It could be argued that the existing pattern of locations is a "natural" sample from the many possible locations where police stations could have been sited. If this argument is not accepted, the inferential assumption of an independently selected random sample of points is not met. If this is the case, you might want to focus primarily on a descriptive comparison of nearest neighbor indices (R -values) for the police stations and the other community services.

TABLE 14.3						
Nearest Neighbor Values for Selected Public Facilities in Toronto, Ontario						
Public Facility	$\overline{NND}$	Density	$\overline{NND}_R$	R	Z_n	p
Fire	2.06	.1278	1.40	1.47	8.21	.0000
Police	3.63	.0312	2.83	1.28	2.42	.0153
Elementary schools	0.59	.7358	0.58	1.02	0.95	.3439
Voting locations	0.27	2.4351	0.32	0.84	-12.23	.0000

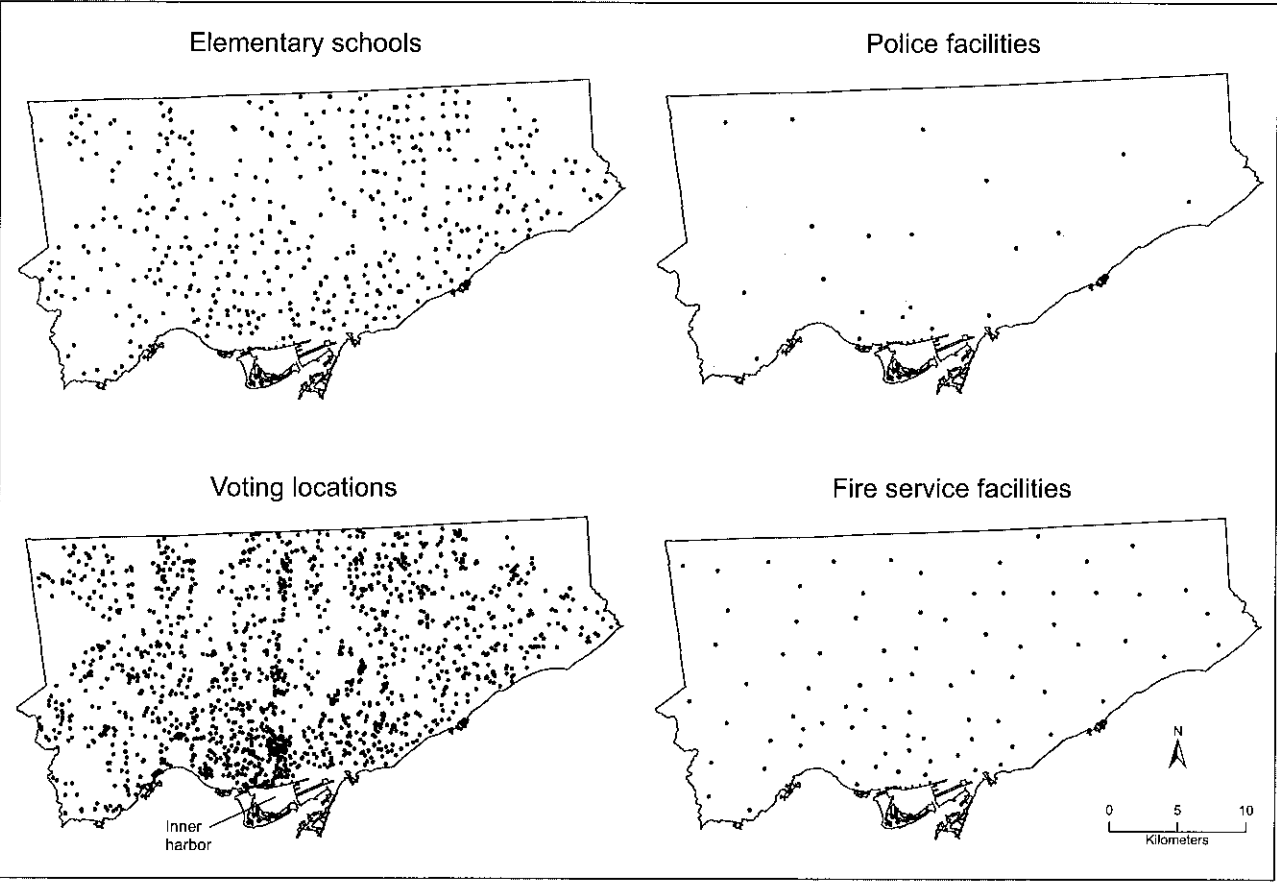


FIGURE 14.3
Location of Selected Public Facilities in Toronto, Ontario
Source: City of Toronto Planning Division

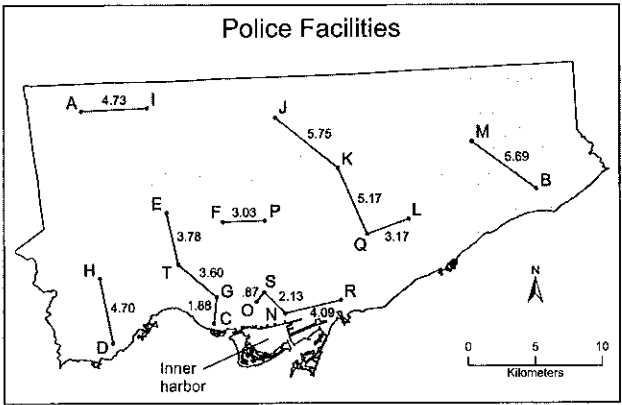


FIGURE 14.4
Nearest Neighbor Distances for Police Service Facilities in Toronto, Ontario

The corresponding nearest neighbor values and test statistic results for each of the four service patterns are shown in table 14.3. The public services are ranked in decreasing size of the R index (from more dispersed to more clustered). The pattern of fire stations in Toronto is the most dispersed ($R = 1.47$) followed by police stations ($R = 1.28$), thus supporting the assumption of an even distribution of emergency services to meet an equity requirement. The third most dispersed service is elementary education facilities ($R = 1.02$). However, statistical evidence ($p = 0.34$) indicates that this dispersion is not significantly different from a random pattern.

The most clustered spacing pattern among the four services is voting locations ($R = 0.84$), the only service where the random nearest neighbor distance exceeds the observed average nearest neighbor distance. Unlike emergency services which are located to provide relatively equal response times for all parts of a region, voters must utilize a specific voting location based on their place of residence. This suggests that areas with higher population densities would require additional voting locations producing a more clustered spatial pattern. This clustered pattern is reflected in a lower R index and is evident in the areas just north of Toronto's inner harbor.

These results are clearly influenced by the way in which the problem is structured. A critical issue when using the nearest neighbor procedure is the delimitation of the study area boundary. In many research problems (like the analysis of public services in Toronto), a political boundary defines the study area. In other problems, where no formal or logical boundary exists for enclosing the point locations, you must designate an operational boundary or territorial limit for the region under study. The position of the boundary does not directly affect the distance between nearest neighbors or the mean nearest neighbor index. However, boundary position does affect both the area of the study region and the point density, factors that determine the random nearest neighbor distance. Therefore, specification of the study area bound-

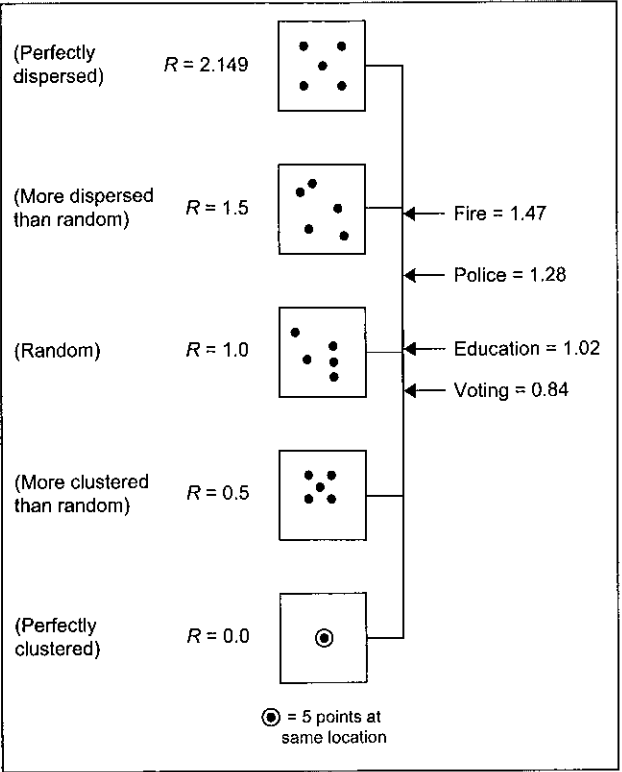


FIGURE 14.5
Continuum of R Values in Nearest Neighbor Analysis: Toronto, Ontario

Spacing of facilities for the four selected services in Toronto confirms the expected arrangements of emergency and non-emergency services (fig. 14.5). As hypothesized, the pattern of emergency fire services exhibits the most dispersed spacing, followed closely by the pattern of police facilities. Also as expected, voting locations in Toronto are more clustered than random. Although education services visually appear more dispersed than random, statistical evidence suggests that the spacing does not differ significantly from a random pattern.

ary influences the outcome of a point pattern analysis. Ideally, a functional study area boundary should be defined that is consistent with the type of points being analyzed. Some researchers suggest that the boundary should be defined just beyond the outermost points within the study area.

Another issue related to study area boundary is the specification of nearest neighbors at the edge of a region. In some actual situations, the nearest neighbor for a point located near the boundary may actually lie outside the study area. Perhaps the nearest fire station or school lies across the border at a shorter distance than the nearest neighbor lying inside the study area. Therefore, results are influenced by how nearest neighbor distances are handled near the study area boundary. Administra-

tive boundaries do not necessarily offer the best approach to delimiting the study area in point pattern problems.

To adapt to various situations, geographers have expanded the NNA concept to include not just the single nearest neighbor, but n nearest neighbors (where n represents all nearby neighbors, and the definition of “nearby” varies depending on the specific situation). If the complete distribution of all point-pair distances in a spatial pattern is considered, then we have a classical k nearest neighbor (often called a *Ripley’s K* function). In evaluating more than one nearest neighbor, you simply determine the mean distance from each point to the n closest points in the distribution. For example, we can determine the average distance of the closest four police stations to any other police station in Toronto.

The k function is a true second order statistic that evaluates the distribution of points over the full set of distances in the point set. That is, for each point the k function determines how many points are within distance d . Because distance is a continuous variable, the k function typically breaks the data up into distance intervals. For example, we might determine how many police stations are within 1km, 2km, 3km, . . . and 10km of each other and compare that to an expected number of police stations from a theoretical distribution of a completely random set of points. The purpose is to determine if a set of points (services or facilities) is clustered at multiple distances.

14.2 QUADRAT ANALYSIS

An alternate methodology for studying the spatial arrangement of point locations is **quadrat analysis**. Rather than focusing on the spacing of points within a study area, quadrat analysis examines the frequency of points occurring in various parts of the area. A set of quadrats (usually square cells) is superimposed on the study area, and the number of points in each cell is determined. By analyzing the distribution of cell frequencies, the point pattern arrangement within the study area can be described.

Whereas nearest neighbor analysis uses the average spacing of the closest points, quadrat analysis considers the variability in the number of points per cell (fig. 14.6). If each of the quadrats contains the same number of points (case 1), the pattern would show no variability in frequencies from cell to cell and would be perfectly dispersed. By contrast, if a wide disparity exists in the number of points per cell for the set of quadrats examined (case 2), the variability of the cell frequencies would be large, and the pattern would display a clustered arrangement. In a third alternative (case 3), the variability of cell frequencies is moderate, and the pattern of points would reflect a random or near random spatial arrangement.

However, the absolute variability of the cell frequencies cannot be used as an effective descriptive measure of the point pattern because it is influenced by the density

of points—that is, the mean number of points per cell. This relationship is directly analogous to the influence of the mean on the standard deviation of a variable. Recall from chapter 3 that the coefficient of variation is used for meaningful comparisons of relative variability between distributions. In quadrat analysis, an index known as the **variance-mean ratio (VMR)** standardizes the degree of variability in cell frequencies relative to the mean cell frequency:

$$VMR = \frac{VAR}{MEAN} \tag{14.8}$$

where: VMR = variance-mean ratio
 VAR = variance of the cell frequencies
 $MEAN$ = mean cell frequency

The mean and variance are the basic descriptive statistics that summarize the central tendency and variability of a

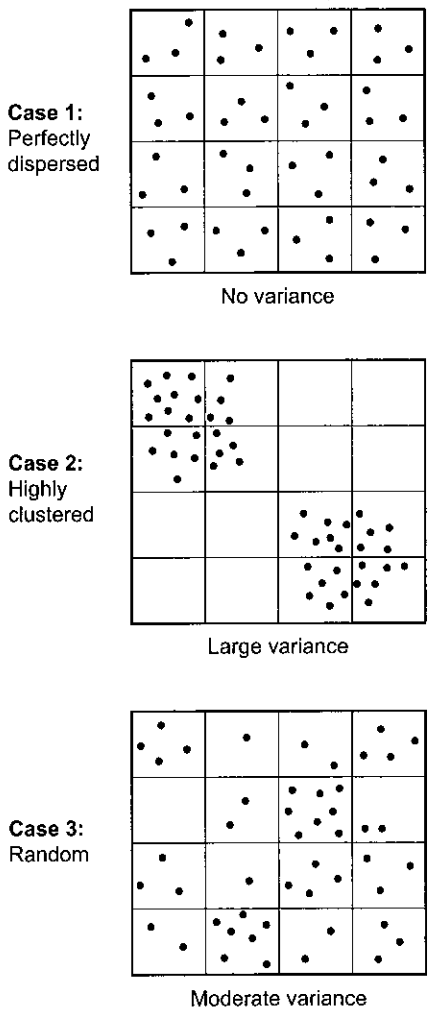


FIGURE 14.6
Quadrat Analysis: The Relationship of Cell Frequency Variability and Point Pattern

variable. Since the data from quadrat analysis are usually summarized as frequency counts (number of points in each quadrat), formulas for the weighted mean and weighted variance (weighted standard deviation squared) are usually applied (see tables 3.4 and 3.8).

Interpretation of the variance-mean ratio offers insight into the spatial pattern of points within the study area. In a dispersed distribution of points, for example, the cell frequencies will be similar, and variance of cell frequency counts will be very low. In an extreme case (such as case 1 in fig. 14.6), if each quadrat contains exactly the same number of points, the variance will be zero, making the variance-mean ratio zero. Conversely, if a point pattern is highly clustered with most cells containing no points and a few cells having many points, the variance of cell frequencies will be large relative to the mean cell frequency. This type of spatial pattern will produce a large value for the variance-mean ratio.

If the set of points is randomly arranged across the cells of the study area, an intermediate value of variance

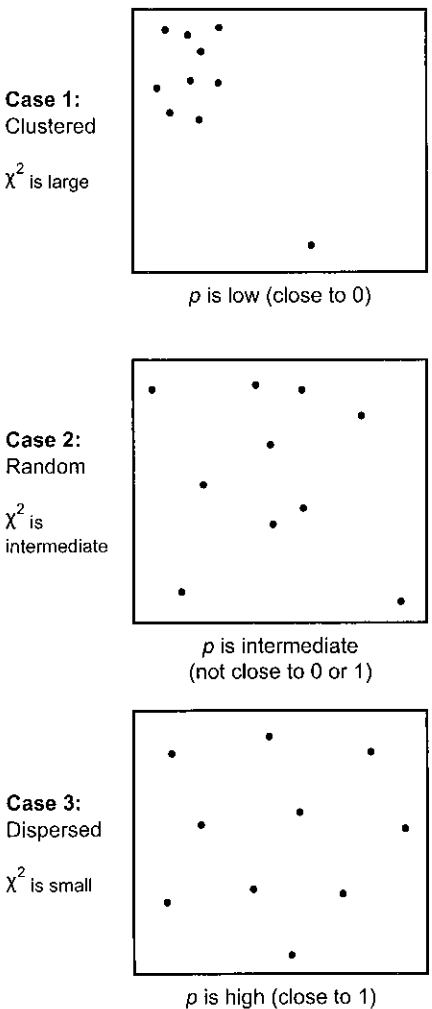


FIGURE 14.8
Quadrat Analysis: The Relationship of Chi-square Value and p

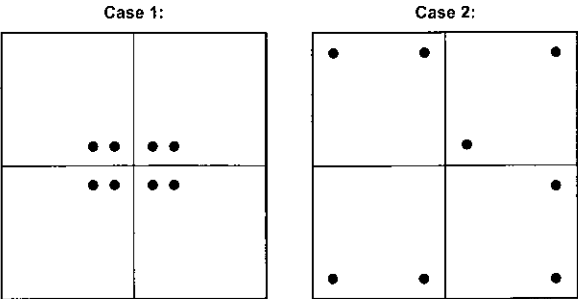


FIGURE 14.7
Two Point Patterns with the Same Variance-mean Ratio, but Different Visual Impressions

will occur. For a perfectly random point pattern, the variance of the cell frequency is equal to the mean cell frequency. To understand this result, recall from section 5.4 that the Poisson distribution is used to describe the frequency of values for a randomly generated spatial or temporal pattern. In the Poisson distribution, the mean frequency equals the variance of the frequencies. Therefore, when using the variance-mean ratio to describe spatial point patterns, a result close to one (variance equals mean) suggests that the distribution has a random arrangement.

Since quadrat analysis measures dispersion related to the frequencies of points in the cell boundaries and not the explicit location of points in relationship to one another, you must be careful when interpreting the VMR. For example, both case 1 and case 2 in figure 14.7 have the same VMR. There are eight points and four quadrats. However, the actual distribution of the points in relationship to one another in the large study area (ignoring the four internal quadrats) would suggest that case 1 has a

Quadrat Analysis

Primary Objective: Determine whether a random (Poisson) process has generated a point pattern

Requirements and Assumptions:

- 1. Random sample of points from a population
- 2. Sample points are independently selected

Hypotheses:

- $H_0 : VMR = 1$ (point pattern is random)
- $H_A : VMR \neq 1$ (point pattern is not random)
- $H_A : VMR > 1$ (point pattern is more dispersed than random)
- $H_A : VMR < 1$ (point pattern is more clustered than random)

Test Statistic:

$$\chi^2 = VMR(m - 1)$$

clustered pattern of points while the distribution of points in case 2 is dispersed.

In addition to its use as a descriptive index, the variance-mean ratio can also be applied inferentially to test a distribution for randomness. The test statistic is chi-square, defined as a function of both the *VMR* and number of cells (*m*):

$$\chi^2 = VMR(m - 1) \quad (14.9)$$

The null hypothesis for this test is expressed as no difference between the observed distribution of points and a distribution of points resulting from a random process (that is, *VMR* = 1). Either a directional (one-tailed) or non-directional (two-tailed) approach can be used, depending on the test objective of the alternate hypothesis.

Rejection of the null hypothesis can occur if a point pattern is more clustered than random (*VMR* > 1) or more dispersed than random (*VMR* < 1). The difference test

determines whether an observed *VMR* value differs significantly from a theoretical random value (*VMR* = 1.0). The interpretation of the chi-square test statistic and corresponding *p*-value reflects these two possibilities. A large *VMR* generally produces a larger chi-square value (with a small *p*-value), suggesting greater variability of the cell frequencies and a clustered arrangement of points. Conversely, a small *VMR* produces a smaller chi-square value, indicating lower variability of cell frequencies and a more dispersed distribution of points. In such instances, the corresponding *p*-value is larger. Intermediate chi-square values result from variance-mean ratios closer to one, suggesting a lower probability of rejecting the null hypothesis and therefore more likelihood that the distribution of points is actually random. Thus, a continuum of *p*-values from 0 to 1 indicates transition from clustered to random to dispersed (fig. 14.8). Either very high or very low *p*-values suggest rejection of the null hypothesis, while intermediate values do not support rejection.

Example: Quadrat Analysis—Distribution of Wildfires in New Mexico

The example of wildfire distribution (section 5.4) is re-examined here using quadrat analysis. The objective now is to describe the pattern of points within the study area by analyzing the distribution of cell frequencies. The resulting frequency pattern indicates whether the wildfire starting points have a dispersed, random, or clustered arrangement. The chi-square difference test is then applied to determine whether the observed pattern differs significantly from a theoretical random pattern.

In this analysis, wildfire locations are assumed to be independent random events. This assumption is justified if the sample is considered a "natural" sample over one time period. If a natural sampling process is assumed, this particular pattern of wildfire starting points represents a random sample from an infinite number of sample patterns that could have occurred during a given time period.

As discussed in section 5.4, 122 cells covering New Mexico have at least half of their area in the state. This set of 122 quadrats excludes only 116 of the over 16,000 recorded wildfire starting points for the state over the entire 30 year time period. One of our objectives is to determine if the distributions of wildfires in New Mexico are different over the last three decades: 1980s, 1990s, and 2000s.

In addition to an analysis of wildfire distributions for all three decades, we also focus attention on a random set of 2,500 points. Using GIS processing, a random set of 2,500 points is generated, and yields a mean frequency of 20.9 wildfires per cell—2,500 wildfires in 120 cells (fig. 14.9). The variability in cell frequencies around these means largely determines the nature of the point pattern. For this set of quadrats, the variance of the random points is 18.9 per cell. The two summary statistics produce a variance-mean ratio of .904, suggesting a random distribution (table 14.4).

A difference test is applied to determine whether the *VMR* of .904 of our computer-generated random set of points does

not differ significantly from that of a theoretical random pattern (*VMR* = 1). The null hypothesis assumes no difference between the observed and expected *VMRs*. Selection of a directional (one-tailed) or non-directional (two-tailed) test depends on whether *a priori* evidence exists to suggest a point pattern that is more clustered than random or more

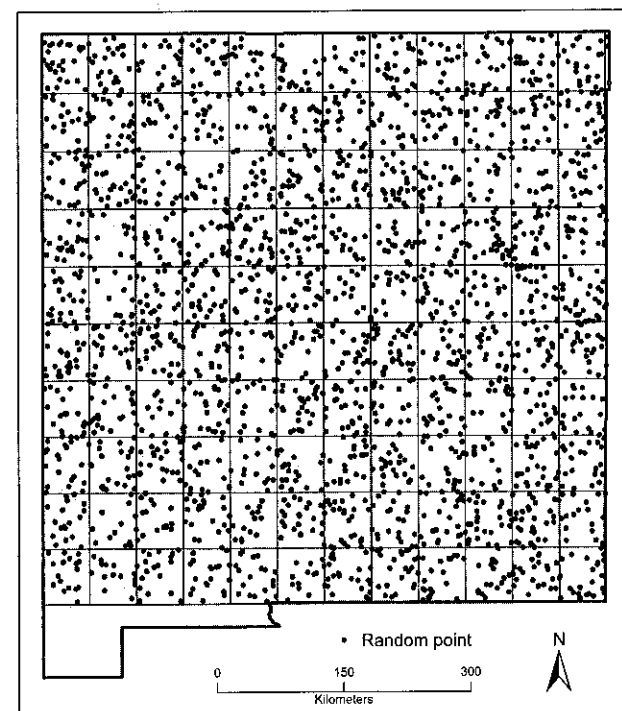


FIGURE 14.9

GIS-generated Random Set of 2,500 Points Superimposed on 120 Cells in New Mexico

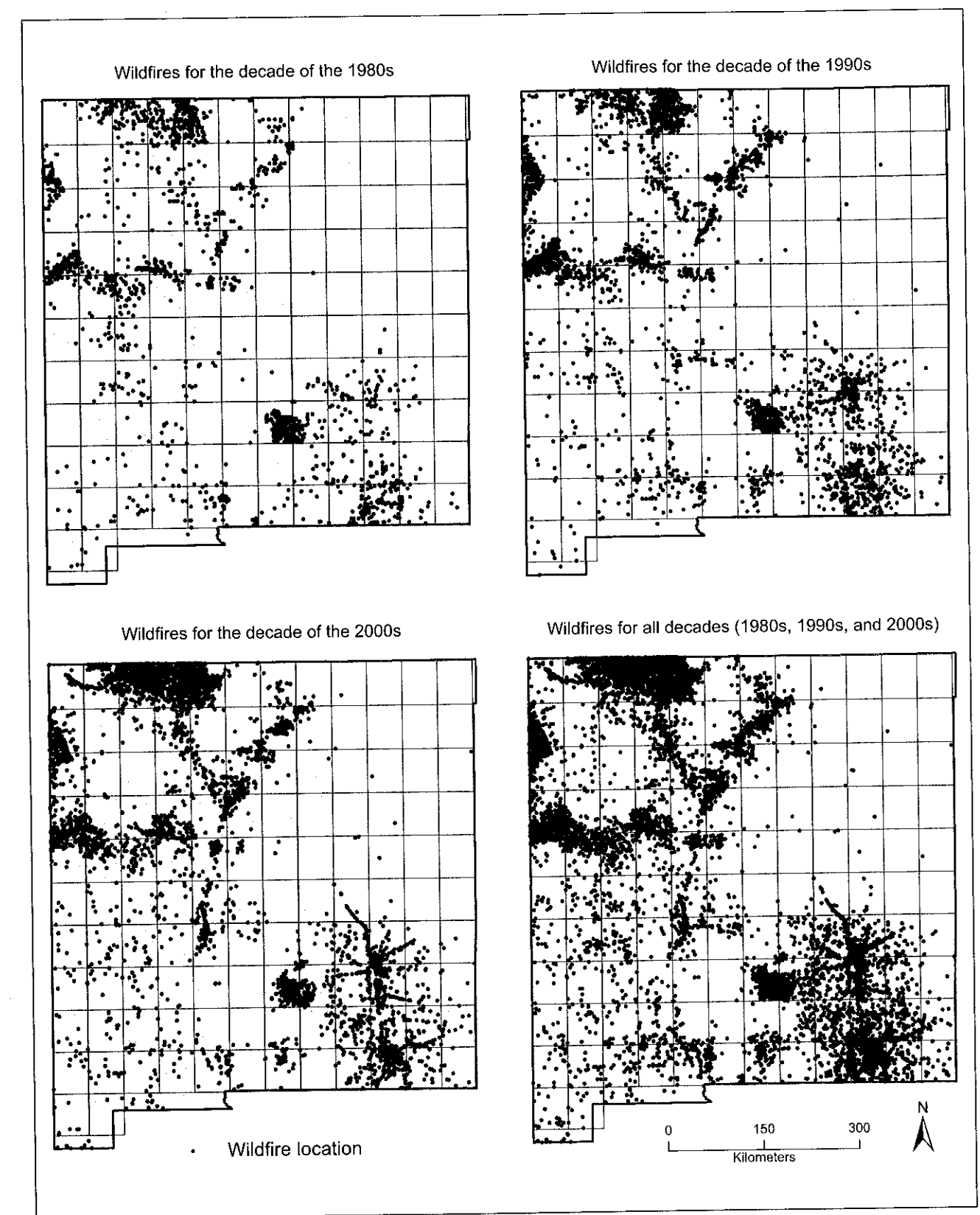


FIGURE 14.10

Point Patterns for Wildfires in New Mexico: 1980s, 1990s, and 2000s
Source: U.S. Geological Survey, Federal Wildland Fire Occurrence Website

(continued)

dispersed than random. Because we generated a fictitious set of random points, there is no *a priori* evidence suggesting the pattern should be clustered (or dispersed), therefore a non-directional, two-tailed difference test is appropriate. Conversely, an alternate hypothesis that assumes, *a priori*, a clustered or dispersed point pattern would require a directional, one-tailed difference test.

As could be expected, the resulting chi-square test statistic of 107 produces a rather large *p*-value of 0.77 (table 14.4). Thus, we have very strong confidence that the null hypothesis should be accepted and that our computer generated pattern is, in fact, random. Figure 14.10 shows the point pattern distribution for wildfires in New Mexico for the decades of 1980–1989, 1990–1999, and 2000–2009.

A cursory view of the maps seems to show that the distribution of wildfire starting points is very highly clustered over all three decades. This makes intuitive sense, as many areas in New Mexico are desert, and only a few areas are heavily forested. In cases like this, a *VMR* will almost certainly describe a highly clustered distribution. However, even though the “natural pattern” of forested areas would logically result in a clustered distribution of wildfires, there is still merit in applying the *VMR* technique to compare multiple decades (and detect any changes in the overall clustering pattern of wildfires). Table 14.5 shows the *VMR* results for each decade.

You can see that each decade exhibits a highly clustered distribution, but as we said, that is not surprising given the distribution of forests, deserts, and human settlements in the state. However, the decades of 1990–99 and 2000–09 are dramatically more clustered than the decade of 1980–89. The increasing clustering of wildfire starting points could possibly be explained by the fact that expansion of “the human foot-

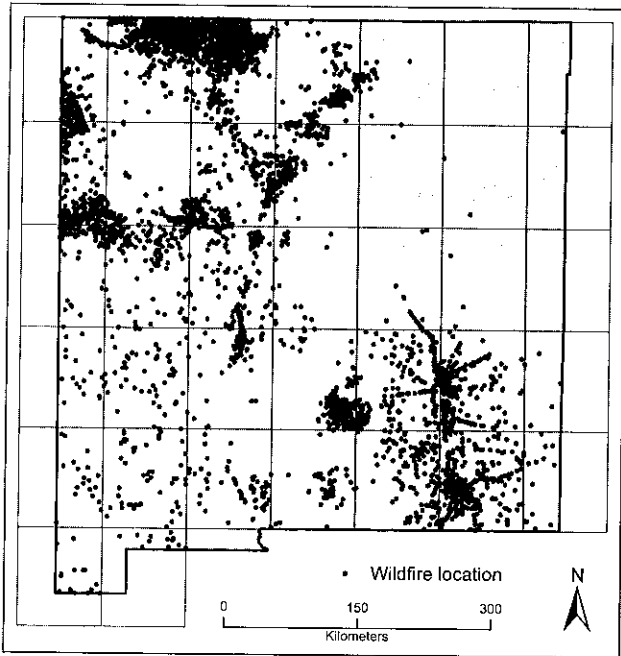


FIGURE 14.11
New Mexico Wildfires from 2000 to 2009 Overlain with a Larger Quadrat Size

print” across New Mexico has relegated forested areas to an ever-smaller portion of the state. Quite simply, there are fewer places where a wildfire could start, and these places are increasingly clustered on public lands such as national parks, national forests, and Bureau of Land Management lands.

TABLE 14.4

Worktable for Quadrat Analysis: New Mexico Wildfires (2,500 Random Points)

H_0 : *VMR* = 1 (point pattern is random)
 H_A : *VMR* <> 1 (point pattern is not random)

Calculate mean cell frequency:

Mean cell frequency = n/m
where: n = number of points
 m = number of cells

$$MEAN = \frac{2,500}{120} = 20.9$$

Calculate variance of cell frequencies:

Number of points per cell (<i>x</i>)	Number of cells (<i>f</i>)	Total points	fx^2
11	1	11	121
12	2	24	288
13	2	26	338
14	4	56	784
15	3	45	675
16	4	64	1,024
17	9	153	2,601
18	8	144	2,592
19	15	285	5,415
20	10	200	4,000
21	10	210	4,410
22	11	242	5,324
23	9	207	4,761
24	13	312	7,488
25	3	75	1,875
26	2	52	1,352
27	3	81	2,187
28	8	224	6,272
32	1	32	1,024
33	2	66	2,178
SUM	120	2,509	54,709

$$VAR = \frac{\sum f_i X_i^2 - \frac{(\sum f_i X_i)^2}{m}}{m - 1}$$

where: f_i = frequency of cells with i wildfires
 X_i = number of wildfires per cell

$$VAR = \frac{54,709 - \frac{(2,509)^2}{120}}{120 - 1} = \frac{54,709 - 52,459}{119} = 18.9$$

Calculate variance-mean ratio:

$$VMR = \frac{VAR}{MEAN} = \frac{18.9}{20.9} = 0.904$$

Calculate test statistic:

$$\chi^2 = VMR(m - 1) = 0.9(119) = 107 \quad (p = 0.77)$$

TABLE 14.5

Variance-mean Ratios (VMRs) for Wildfire Distribution in New Mexico

Distribution	Variance-Mean Ratio
Computer generated random sample	.904
1980–1989	20,381
1990–1999	32,645
2000–2009	30,818

Does the size of quadrats used to cover the study area influence results from quadrat analysis? To examine this question, the 2000–09 wildfire distribution is evaluated using a cell design with quadrat areas four times as large. This cell structure

Although quadrat analysis offers a useful approach to studying the spatial arrangement of points, several procedural issues must be considered. Different cell sizes produce varying levels of mean point frequency and variance per cell. Moreover, studies have shown that if the point pattern is held constant but the quadrat size is decreased (i.e., the number of quadrats is increased), the variance of the point frequencies usually declines faster than the mean. However, this is not always the case as we can see from our example. Thus, different cell sizes in a single problem can generally be expected to produce different values of the variance-mean ratio.

Although more research needs to be conducted on optimal quadrat size, two guidelines can be offered. When visual examination of a point pattern reveals distinct clusters, the size of quadrats should match that of the clusters. This cell size allows the results from quadrat analysis to reflect more accurately the visual impression of the pattern. Other researchers have suggested a general “rule of thumb” that cell size should be equal to twice the average area per point. In other words, the mean frequency of points should be close to 2.0 per quadrat. This guideline approximates the cell structure used with the 122 cells covering the New Mexico forest fire example. Taylor (1977) presents a more detailed discussion of cell size in quadrat problems.

Another issue in the use of quadrat analysis also requires some attention. Like nearest neighbor analysis, delimitation of the study area can influence the results of a study. In the use of quadrat analysis, a way to handle cells around the study area boundary needs to be determined. Cells on the boundary will inevitably contain area both inside and outside the study region. In the New Mexico wildfire problem, cells with more than half of their area outside the study region were eliminated. Alternative decision rules can be used, and these may change the results obtained from the analysis. No matter what decision rule is applied in a particular problem, the geographer needs to be consistent in handling the boundary issue.

produces 37 larger quadrats covering 7,177 wildfires for the decade (fig. 14.11). To be consistent, boundary cells whose area is mostly (more than half) outside New Mexico are eliminated.

The larger cell network produces fewer quadrats with higher frequencies per quadrat and a larger mean cell frequency of 211 wildfires per quadrat. The variance of the cell frequencies about this mean is also higher at 81,203; since the variance remains much larger than the mean, the pattern once again displays a highly clustered arrangement. Standardizing this variance to the mean produces a variance-mean ratio of 15,070, much lower than the value found with the smaller quadrat size (*VMR* = 30,818), although still highly significant. You might want to speculate about why the *VMR* for the larger quadrats is considerably smaller than the *VMR* for the smaller quadrats.

KEY TERMS

nearest neighbor analysis, 210
point pattern analysis, 210
quadrat analysis, 216
standardized nearest neighbor index, 212
variance-mean ratio (*VMR*), 216

MAJOR GOALS AND OBJECTIVES

If you have mastered the material in this chapter, you should now be able to:

1. Identify the types of geographic problems or situations for which nearest neighbor analysis is appropriate.
2. Identify the types of geographic problems or situations for which quadrat analysis is appropriate.
3. Explain the logic and procedures used in nearest neighbor analysis and explain how they differ from those used in quadrat analysis.

REFERENCES AND ADDITIONAL READING

Boots, B. N. and A. Getis. *Point Pattern Analysis*. Newbury Park, CA: Sage, 1988.
Getis, A. “Temporal Analysis of Land Use Patterns with Nearest Neighbor and Quadrat Methods.” *Annals, Association of American Geographers*. Vol. 54 (1964): 391–399.
Taylor, P. J. *Quantitative Methods in Geography*. Houghton Mifflin: Boston, 1977.
Thomas, R. *An Introduction to Quadrat Analysis*. CATMOG No. 12. Norwich, England: Geo Books, 1977.

For those interested in looking further at the distribution of public services in Toronto, Canada, see the City of Toronto Planning Division website: www.toronto.ca/planning. For those interested in the spatial distribution of wildfire occurrences, see the Department of the Interior, U.S. Geological Survey, Federal Fire Occurrence Website: <http://wildfire.cr.usgs.gov/firehistory/data.html>

Area Pattern Analysis

15.1 Join Count Analysis

15.2 Moran's *I* Index (Global)

15.3 Moran's *I* Index (Local)

This chapter presents several methods for analyzing area patterns. Some of these methods are global in nature, encompassing an entire study area, while one method (a variation of Moran's *I*) analyzes a local region. Two statistical procedures, the join count statistic and Moran's *I* Index, are discussed in more detail. The join count statistic basically examines the structure of all subarea connections (joins) within the larger study area. Each particular subarea is either "joined" or "not joined" to a number of adjacent subareas, and the overall pattern of this subarea joining structure is analyzed. The Moran's Index analyzes ordinal or interval/ratio area patterns to determine if they are significantly autocorrelated.

For example, an economic development planner might be interested in evaluating changing spatial patterns of poverty in Appalachia. Through area pattern analysis of a chronological series of county-level choropleth maps showing Appalachian poverty, the spatial distribution of poverty can be evaluated over time to determine whether it is becoming more dispersed or clustered. In another example, an urban geographer could analyze a map pattern depicting the number of existing home sales by census tract to determine the degree to which sales are spatially concentrated in certain portions of the city. If a nonrandom clustering of high turnover rates is found in certain segments of the city, attention could be focused on why the sales exhibit such spatial patterns.

The goals and objectives in area pattern analysis are similar to those in point pattern analysis. Just as nearest neighbor and quadrat analysis examine the random or nonrandom nature of point patterns, descriptive and inferential procedures are available to analyze area pat-

terns. A choropleth or area pattern map is considered clustered if adjacent or nearby areas tend to have highly similar values or scores. Alternatively, if the values of adjacent or nearby areas tend to be dissimilar, then the spatial pattern is considered more dispersed than random (refer back to fig. 13.1).

Area pattern analysis is appropriate for studying many practical problems in geography. It is particularly valuable when examining the way in which area patterns of a specific variable change over time. For example, suppose a medical geographer is concerned with the diffusion of influenza in a metropolitan area. If the number of cases is reported by census tract for several time periods, the influenza morbidity rate could be displayed on a series of choropleth maps and the degree of non-randomness in the patterns analyzed. If the incidence of the disease becomes more dispersed over successive time periods, researchers would be very concerned because the dispersion indicates that the influenza is spreading more widely throughout the metro area. It could also mean a significant number of cases were appearing in other areas not close to existing areas of high incidence. Conversely, if the morbidity rate patterns were becoming more clustered over successive time periods, then a closer examination of those areas with higher incidence of influenza would be appropriate. Such knowledge could lead to the implementation of effective strategies for disease control.

Suppose a political geographer wants to study the spatial pattern of voter registration rates across a number of precincts within a community. If registration rates have been mapped by precinct before each of the last sev-

eral elections, a useful comparative analysis of area patterns is possible. If the pattern is becoming more clustered over successive time periods, a different registration strategy may be necessary. An analysis of area pattern trends might also provide insights into demographic or economic variables that seem to be related to the voter registration rate.

Join count analysis is used when areas on a map are represented in binary form, such as yes/no, presence/absence, or above/below a particular threshold. For example, each census tract in an influenza morbidity rate map could be classified as having either an above average or below average incidence of influenza. Moran's *I* is used when areas on a map are represented in ordinal or interval/ratio form, with the actual values or ranks of values. For example, the census tracts might indicate the actual influenza morbidity rate, or a ranking of census tracts from the highest to lowest morbidity rate.

15.1 JOIN COUNT ANALYSIS

The **join count** is the basic organizational statistic for the analysis of area patterns. A "join" is operationally defined as two areas sharing a common edge or boundary. The procedures involved in calculating a join count statistic are relatively straightforward. The fundamental building block is the number of joins in the pattern and the nature of the join structure in the study area. Figure 15.1 illustrates the simplest situation: a binary classification of data and a two-category choropleth map. Many geographic variables are binary in nature, or can be downgraded effectively to binary without sacrificing important locational information. This figure shows the join structure for the subareas in the study region. For example, area *A* is joined to two other areas, *B* and *D*. The number of joins associated with each area is listed adjacent to the join structure map.

To comply with accepted practice, each area can be identified as either "black" or "white." It is then possible to refer to the *observed number* of "black-white," "black-black," and "white-white" joins in the pattern. In case 1 of figure 15.1, the area pattern appears clustered, with relatively few black-white (dissimilar) joins and a rather large number of black-black and white-white (similar)

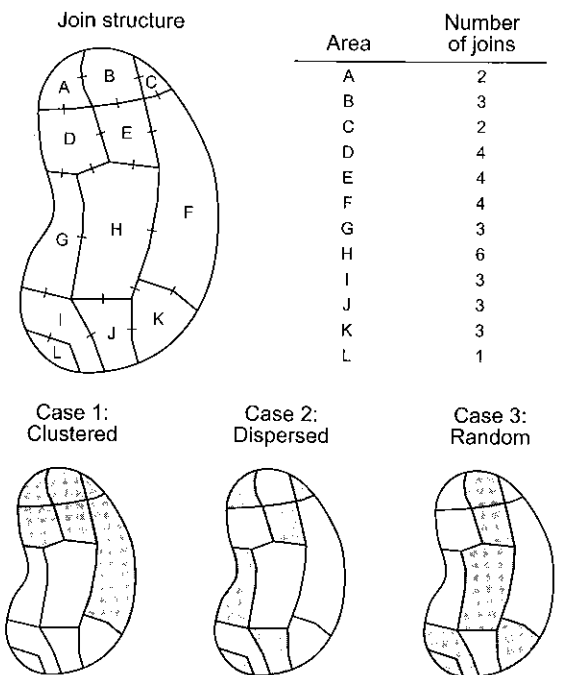


FIGURE 15.1
Join Structure and Examples of Area Patterns Appearing Clustered, Dispersed, and Random

joins. In case 2, a dispersed (checkerboard-like) pattern has more black-white joins and fewer joins of similar-category areas. With random patterning (case 3), the result is an intermediate number of both similar and dissimilar joins. These situations are all reflected in table 15.1.

To determine if a join count distribution has been generated by a random process, the number of black-white joins that would be expected from a theoretical random arrangement of binary observations must be calculated. That is, the *expected number* of black-white joins in a purely random pattern is compared with the one actually observed. If the observed join count is significantly different from the expected join count, the observed pattern can be described as nonrandom.

How do you determine the expected number of black-white joins? The locational context in which the overall study area pattern is being evaluated suggests two

TABLE 15.1

Join Counts Associated with Area Patterns

Case	Total joins *	Dissimilar areas joined	Similar areas joined		Total
		Black-white joins	Black-black joins	White-white joins	
1 (clustered)	19	5	7	7	14
2 (dispersed)	19	15	0	4	4
3 (random)	19	12	4	3	7

* See figure 15.1 for map of join structure and area patterns.

possible approaches for generating expected join counts. Some area pattern analyses may be conducted under the hypothesis of **free sampling**. Free sampling should be used if the researcher can determine the probability of an area being either black or white based on some theoretical idea or by referring to a larger study area. Once the probability of an area being black or white has been determined, the expected number of black-white joins in a random pattern having those probabilities can also be determined. For example, in the influenza morbidity rate problem, suppose each census tract throughout the metropolitan area is classified as having either above average incidence of the disease (black) or below average incidence (white). If the pattern of disease is being analyzed in only one portion of the metropolitan area, logic would suggest that the probability of above- or below-average incidence of influenza should be typical or representative of that found throughout the metro area, and the free sampling hypothesis would be appropriate.

The **non-free sampling** hypothesis is used when no reference to a larger study area or general theory is appropriate. In this case, only the study region itself is considered in the analysis, and the expected number of black-white joins is estimated from a random patterning of only those subareas in the study region. For instance, a political geographer studying the spatial pattern of community voter registration rates by precinct may not have any reason to believe this pattern is typical or representative of any larger area (such as the state in which the community is located). Therefore, if it is hypothesized that the area pattern has resulted from the unique characteristics found within the study area, then a non-free sampling hypothesis should be applied.

In many geographic problems, you will find it difficult to conclude with any degree of confidence that a subarea pattern is typical of a larger area. If this is the case, use of the expected number of black-white joins from a non-free sampling hypothesis frees you from having to make any restrictive assumption. For the sake of brevity, we will only show the slightly simpler calculation procedure for area pattern analysis in a non-free sampling context.

In a non-free sampling context the observed number of black-white joins is compared with the number of such joins expected in a purely random arrangement. This will help you determine if the area pattern has been generated by a random process. The null hypothesis is that the observed number of black-white joins is equal to the expected number of black-white joins in a purely random area pattern. If reason exists to hypothesize that the observed area pattern is non-random in a certain direction, then a one-tailed alternate hypothesis is appropriate. For instance, an area pattern could be hypothesized either as more clustered than random or more dispersed than random. Without an *a priori* reason to hypothesize a clustered or dispersed pattern, a two-tailed alternate

hypothesis is selected, and the conclusion will be that an area pattern is random or nonrandom.

The test statistic (Z_b) for non-free sampling is:

$$Z_b = \frac{O_{BW} - E_{BW}}{\sigma_{BW}} \tag{15.1}$$

where: O_{BW} = observed number of black-white joins
 E_{BW} = expected number of black-white joins
 σ_{BW} = standard error of the expected number of black-white joins

In non-free sampling, the expected number of black-white joins is calculated by incorporating the observed number of black areas and white areas directly into the equation:

$$E_{BW} = \frac{2JBW}{N(N-1)} \tag{15.2}$$

where: E_{BW} = expected number of black-white joins
 J = total number of joins
 B = number of black areas
 W = number of white areas
 N = total number of areas (black plus white)
 $= B + W$

The expected number of black-white joins in the study region is determined by asking how many black-white joins would occur from a theoretical random pattern containing the same number of black and white areas as the observed area pattern. That is, the question becomes how many black-white joins would be generated from a random arrangement of the observed number of black and white areas in the area pattern.

Certain generalizations can be made concerning the expected number of black-white joins. Given a particular total number of areas ($N = B + W$), the value of the denominator in equation 15.2 is constant. If most of the areas in the observed pattern are either black or white, then their product (BW) in the numerator will be a smaller number, producing a smaller expected number of black-white joins. Conversely, if about the same number of black and white areas is found in the observed pattern, then their product will be a larger number, producing a larger expected number of black-white joins. These results seem logical, for a pattern with a large number of black areas and very few white areas (or vice versa) would not be expected to have a large number of black-white joins. The expected number of black-white joins is maximized when an equal number of black and white areas occur in the overall pattern.

Not all randomly generated patterns will contain exactly the same number of black-white joins, so a mea-

sure of the amount of variability expected as a result of sampling must be included. The standard error of expected black-white joins is:

$$\sigma_{BW} = \sqrt{E_{BW} + \frac{\sum L(L-1)BW}{N(N-1)} + \frac{4[J(J-1) - \sum L(L-1)]B(B-1)W(W-1)}{N(N-1)(N-2)(N-3)} - E_{BW}^2} \tag{15.3}$$

where: σ_{BW} = standard error of the expected number of black-white joins
 J = total number of joins
 $\sum L$ = total number of links $= 2J$
 B = number of black areas
 W = number of white areas
 N = total number of areas (black plus white)
 $= B + W$

Note: $\sum L = 2J$. If areas A and B are joined, then A is "linked" to B and B is "linked" to A, making the sum of all links twice the number of joins.

The area pattern may now be analyzed with the non-free sampling hypothesis (equation 15.1 with supplementary equations 15.2 and 15.3). If the observed number of black-white joins is larger than the expected number of black-white joins, then Z_b will be a positive value, indicating a pattern more dispersed than random. A relatively large number of black-white (dissimilar) joins will occur in a dispersed pattern (fig. 15.1). Conversely, if the observed number of black-white joins is smaller than the

Join Count Analysis (Binary Categories)

Primary Objective: Determine whether a random process has generated a binary (two-category) area pattern

- Requirements and Assumptions:**
1. Each area is assigned to one of two categories
 2. Each pair of areas must be defined as either adjacent ("joined") or nonadjacent (not "joined") in a consistent manner

- Hypotheses:**
- $H_0 : O_{BW} = E_{BW}$ (area pattern is random)
 - $H_A : O_{BW} \neq E_{BW}$ (area pattern is not random)
 - $H_A : O_{BW} > E_{BW}$ (area pattern is more dispersed than random)
 - $H_A : O_{BW} < E_{BW}$ (area pattern is more clustered than random)

Test Statistic:

$$Z_b = \frac{O_{BW} - E_{BW}}{\sigma_{BW}}$$

expected number of such joins, then Z_b will be negative, and the pattern will appear more clustered than random. The greater the magnitude of Z_b (either positive or negative), the greater the likelihood that the area pattern being analyzed is not random. Just like other Z-score difference tests, this Z_b test statistic for area pattern analysis indicates the number of standard deviations separating the expected random black-white join count from the observed value.

Example: Join Count Analysis—Obesity Distribution Patterns across the United States: 1995, 2002, 2009

Returning to the discussion of obesity, we have shown previously that the obesity rate in the United States has increased each year. However, while the nation suffers a high obesity rate, the geographic distribution of highly obese states may not be evenly distributed. That is, only certain portions of the country may exhibit high levels of obesity. A geographer might be interested in determining if the areas of high obesity in the United States are clustered, random, or dispersed in order to tailor a specific approach for combating the epidemic. You also might want to know if the overall spatial pattern of obesity is becoming more dispersed, random, or clustered over time. Figure 15.2 shows binary maps of obesity in the conterminous United States (excluding Alaska and Hawaii, since they are obviously not "joined" to any other state) for three different years: 1995, 2002, and 2009. Black areas represent the states that are above the national mean obesity level and white areas represent states below this level. As shown in table 15.2 the number of states with obesity levels above the national mean is relatively similar for each year.

Across the conterminous United States there are 212 links and 106 joins (recall from equation 15.3 that $\sum L = 2J$); table 15.3 shows the state linkage pattern needed for the area pattern analysis. This general state linkage pattern serves all three dates of analysis (1995, 2002, and 2009) although the detailed calculation procedure which follows is shown only for the 1995 pattern.

TABLE 15.2
Number of States with Obesity Levels Above the National Average

Year	States with obesity above the national average	Number of black-white joins
1995	26	21
2002	27	32
2009	25	25

(continued)

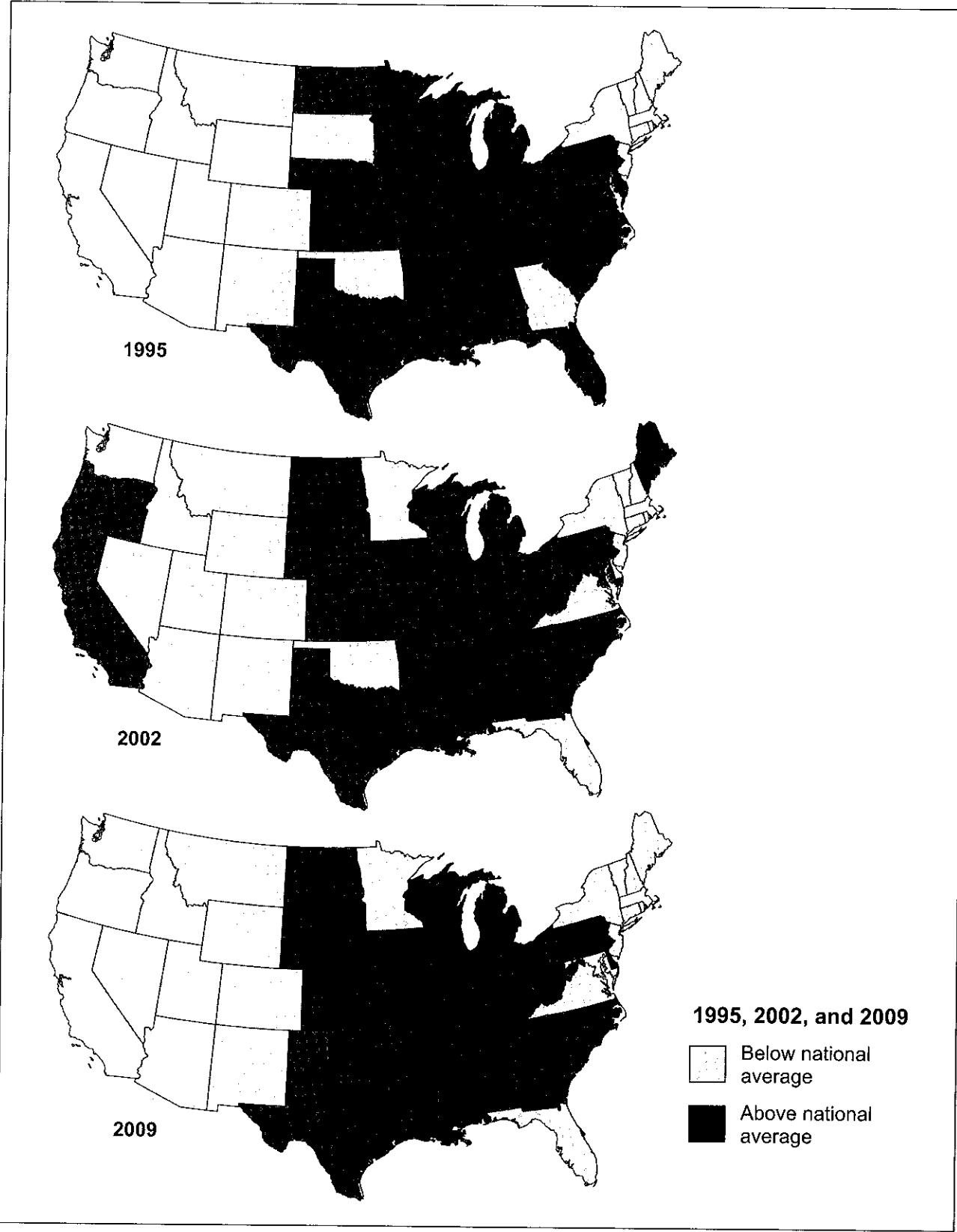


FIGURE 15.2
State Obesity Levels in the United States: 1995, 2002, and 2009
Source: Centers for Disease Control and Prevention (CDC)

In this example, a “black-white” or dissimilar join occurs when a state with an obesity rate above the national mean shares a common boundary with a state having an obesity rate below the national mean. Conversely, a similar join

TABLE 15.3			
State Linkage Pattern			
State Name	Number of links (L)	L - 1	L (L - 1)
Alabama	4	3	12
Arizona	5	4	20
Arkansas	6	5	30
California	3	2	6
Colorado	7	6	42
Connecticut	3	2	6
Delaware	3	2	6
Florida	2	1	2
Georgia	5	4	20
Idaho	6	5	30
Illinois	5	4	20
Indiana	4	3	12
Iowa	6	5	30
Kansas	4	3	12
Kentucky	7	6	42
Louisiana	3	2	6
Maine	1	0	0
Maryland	4	3	12
Massachusetts	5	4	20
Michigan	2	1	2
Minnesota	4	3	12
Mississippi	4	3	12
Missouri	8	7	56
Montana	4	3	12
Nebraska	6	5	30
Nevada	5	4	20
New Hampshire	3	2	6
New Jersey	3	2	6
New Mexico	5	4	20
New York	5	4	20
North Carolina	4	3	12
North Dakota	3	2	6
Ohio	5	4	20
Oklahoma	6	5	30
Oregon	4	3	12
Pennsylvania	6	5	30
Rhode Island	2	1	2
South Carolina	2	1	2
South Dakota	6	5	30
Tennessee	8	7	56
Texas	4	3	12
Utah	6	5	30
Vermont	3	2	6
Virginia	5	4	20
Washington	2	1	2
West Virginia	5	4	20
Wisconsin	4	3	12
Wyoming	6	5	30
TOTAL	212		856

occurs when either two states with obesity rates above (or below) the national mean share a common boundary. On the 1995 map, 21 of the 106 joins are dissimilar, while the other 85 joins are similar. Our research task is to determine if the observed number of dissimilar joins is significantly different from the number of dissimilar joins expected in a random obesity level pattern.

The worktable for the 1995 join count analysis is presented in table 15.4. The expected number of dissimilar joins in a purely random pattern is calculated by incorporating the observed number of states with obesity levels above and below the national mean directly into the equation. The result indicates that nearly 54 (53.75) dissimilar joins are expected in a random pattern.

To determine whether this contrast between the number of observed dissimilar joins (21) and the number of expected dissimilar joins (53.75) is statistically significant, the standard error of expected dissimilar joins (σ_{BW}) must be incorporated into the test statistic. The standard error of expected dissimilar joins is 4.92 (table 15.4). The resultant test statistic value (Z_b) is -6.66, with an associated p -value of .000. These statistics indicate a highly significant tendency toward clustering of states having obesity rates above the national mean.

Using a similar statistical test procedure, analyses of the 2002 and 2009 obesity patterns (fig. 15.2) have also been generated. The results of these analyses are summarized in table 15.5, and an interesting trend emerges across the United States. In all three years, the distribution of states with obesity levels above the national mean is very highly clustered. However, 1995 appeared to have the most clustering over the 15-year period.

The knowledgeable geographer should be able to offer some tentative hypotheses about why the obesity rate in the United States was initially most clustered in 1995, became a bit less clustered in 2002, and once again become more clustered in 2009. First, the overall patterns are all highly clustered (examine fig. 15.2 and table 15.5). Generally speaking, for all three dates, similar observable regional patterns are evident. Most states in the New England, Rocky Mountain, and Pacific census divisions have below national mean obesity rates most of the time. Conversely, most states in the South and Midwest census regions have above national mean obesity rates most of the time. These dominant regional patterns would explain the continuing high levels of obesity clustering. Perhaps we are over-emphasizing the importance of these changes. After all, the obesity patterns are all extremely clustered.

When there are exceptions to this overall pattern, measurable differences in the level of clustering seem to emerge. The most clustered pattern is that of 1995 ($Z_p = -6.66$) with only 21 black-white or dissimilar joins. All of the New England, Rocky Mountain, and Pacific states have below national average obesity rates. Georgia is one of only a few notable outliers on the 1995 pattern. If Georgia's obesity rate was just a little higher, it would have been above the national mean, and the national pattern would have been even more clustered. Perhaps some additional insights might emerge if we took a closer look at obesity pattern changes within Georgia. Maybe the state obesity level was below the national mean back in

(continued)

1995 because of the rapid growth in the Atlanta area (where obesity levels are generally lower than in the rest of Georgia). As the rapid growth in Atlanta has slowed a bit in recent years, perhaps this demographic change allowed Georgia to slip back above the national obesity level.

The least clustered of the three patterns (although still very highly clustered) is 2002 ($Z_p = -4.33$). On the 2002 map, there are more notable outliers (Florida, Maine, Virginia, Minnesota, California, and Oregon) which produces more (32) dissimilar joins (table 15.2).

Are these "cluster shifts" significant in any practical policy context? That is, do these country-wide join count analyses direct us to any particular new obesity reduction plan? The answer is probably no, at least not directly. But the newly emerging outlier states that don't fit the general regional pattern are worth exploring further. In addition to the Georgia situation just discussed, we might hypothesize why the entire

Rocky Mountain and Pacific census divisions are now less obese than the national average. What (if anything) distinguishes the people of the West from much of the rest of the country? Will these regional differences become stronger as time goes by?

TABLE 15.5

Results from the Join Count Analysis—Trends in Obesity Level Clustering: 1995, 2002, and 2009

Year	Z	p-value
1995	-6.66	0.000
2002	-4.33	0.000
2009	-5.89	0.000

TABLE 15.4

Worktable for Join Count Analysis: Obesity in the Conterminous United States, 1995

$H_0: O_{BW} = E_{BW}$ (area pattern is random)

$H_A: O_{BW} \neq E_{BW}$ (area pattern is not random)

number of "black" states = 26

number of "white" states = 22

$O_{BW} = 21$

$$E_{BW} = \frac{2JBW}{N(N-1)} = \frac{2(106)(26)(22)}{48(47)} = \frac{121,264}{2,256} = 53.75$$

$$\begin{aligned} \sigma_{BW} &= \sqrt{E_{BW} + \frac{\sum L(L-1)BW}{N(N-1)} + \frac{4[J(J-1) - \sum L(L-1)]B(B-1)W(W-1)}{N(N-1)(N-2)(N-3)} - E_{BW}^2} \\ &= \sqrt{53.75 + \frac{(856)(26)(22)}{48(47)} + \frac{40[(106)(105) - 856]26(25)(22)(21)}{48(47)(46)(45)} - 53.75^2} \\ &= \sqrt{53.75 + \frac{494,780}{2,256} + \frac{40[11,130 - 856](300)(300)}{4,669,920} - 2,889.06} \\ &= \sqrt{53.75 + 217.03 + 2,642.68 - 2,889.06} \\ &= \sqrt{24.4} \\ &= 4.92 \end{aligned}$$

$$Z_D = \frac{O_{BW} - E_{BW}}{\sigma_{BW}} = \frac{21 - 53.75}{4.92} = \frac{-32.75}{4.92} = -6.66 \quad (p = 0.000)$$

15.2 MORAN'S I INDEX (GLOBAL)

The Moran's Index (I), developed by Australian statistician Patrick Moran in 1948, is a popular technique for quantifying the level of spatial autocorrelation in a set of geographic areas, with a value assigned to each area. In contrast to nearest neighbor analysis which examines patterns in the *location* of points, or the join count analysis that only evaluates a binary areal representation of the data, Moran's Index takes into account the geographic locations (points or areas) in addition to the actual *attribute values* to determine if areas are clustered, randomly located, or dispersed across the overall study area. The index is positive when nearby locations have similar attribute values (clustered), negative when nearby geographic locations have dissimilar attribute values (dispersed) and approximately zero when attribute values are randomly dispersed throughout the study area.

Geographers have applied the Moran's Index in numerous research problems, such as exploring the spatial distribution of income, rates of disease, and the racial and ethnic composition of urban neighborhoods. In all applications, the objective of Moran's technique is to identify significant spatial patterns within a study area. This technique is appropriate for both point and polygon (area) features, and attribute values must be measured on an ordinal or interval/ratio scale.

Moran's Index proceeds by identifying each unique pair of neighboring locations in the study area. The rela-

tionship between neighboring locations is shown for a simplified study region comprising four areas as shown in figure 15.3. Each area is identified by a letter and the magnitude of the area's attribute value appears in parentheses.

The general form of Moran's Index for areas is shown in equation 15.4, and the mathematical equivalent, as presented in Ebdon (1988), appears in equation 15.5:

$$I = \frac{\left(\text{number of areas} \right) \left(\text{sum of cross-products for all contiguous pairs } (i, j) \right)}{\left(\text{number of joins} \right) \left(\text{variance of the area attribute values} \right)} \quad (15.4)$$

$$I = \frac{n \sum (x_i - \bar{x})(x_j - \bar{x})}{J \sum (x - \bar{x})^2} \quad (15.5)$$

where:

n = the number of areas

x = an area attribute value

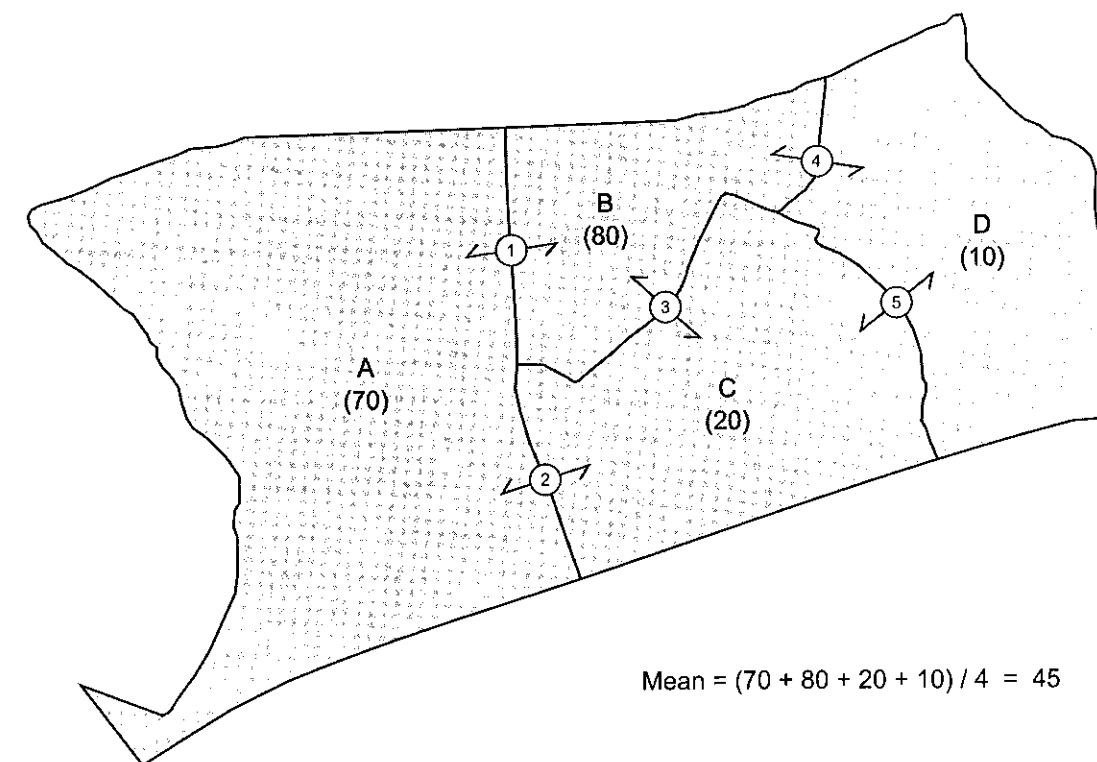
$\bar{x}$ = the mean of all area attribute values

x_i and x_j = the values of contiguous pairs

$\sum (x_i - \bar{x})(x_j - \bar{x})$ = the sum of all contiguous pairs

J = the number of joins

$\sum (x - \bar{x})^2$ = the variance of the attribute values.



$$\text{Mean} = (70 + 80 + 20 + 10) / 4 = 45$$

FIGURE 15.3
Areas and Attribute Values for Moran's I

The deviation values for each pair of neighboring locations or contiguous pairs (i and j) are multiplied together and summed $\sum (x_i - \bar{x})(x_j - \bar{x})$ to create a weighted cross-product. When a neighboring pair of areas both have larger values than the study area mean, their cross-product will be positive (this is the case, for example, with pair AB in fig. 15.3). The cross-product of a neighboring pair will also be positive if both have smaller values than the study area mean (which is the case with pair CD). Positive cross-products therefore indicate that areas with similar attribute values (either high or low) are clustered. In addition, the larger the deviations from the study area mean, the larger the absolute magnitude of the cross-product.

Conversely, when one area of a neighboring pair is larger than the study area mean, while the other area is smaller, the cross-product will be negative (as with pair BD). A negative cross-product therefore indicates dispersion. Once again, the larger the deviations from the study area mean, the larger the absolute magnitude of the cross-product.

It is important to note that the Moran's Index of autocorrelation explicitly takes into account the locations of areas relative to one another. In the procedure as discussed here, the key factor is whether two areas are joined (making them a neighboring pair) or not. We are using *simple adjacency weighting or contiguity* to determine the degree of autocorrelation—only those areas sharing a common boundary are included in the calculations. For instance, areas A and C share a boundary and are included in the calculations, while areas A and D do not share a boundary and are not included.

Other spatial weighting schemes are available to calculate Moran's Index, but these are not shown here. One alternative method is *inverse distance*. With this option, all pairs of geographic areas are included in the analysis, but nearby areas carry more weight in the calculation. If this method were used in our analysis, the cross-product for areas B and C would heavily influence the Moran's Index value (by virtue of their closeness or proximity). By contrast, since areas A and D are relatively far apart, their cross-product value would be given a lighter weight.

To summarize, the numerator of equations 15.3 and 15.4 represents the sum of all contiguous or adjacent area cross-products multiplied by the number of areas in the study region. A positive sum indicates that areas with similar values (either high or low) tend to be located nearby each other, which suggests spatial clustering among the subareas. A negative sum indicates that areas with high attribute values tend to be located near features with low attribute values, which indicates a dispersed pattern. Sometimes the positive cross-products are balanced by the negative cross-products and their overall sum will be close to zero, which indicates a spatially random pattern. The weighted cross-product numerator is divided by the variance and the sum of the spatial

weights (represented as either “joins” or a distance-based weight). This rescales Moran’s Index so that it ranges from -1.00 to 1.00 . A Moran’s Index value less than zero indicates a dispersed pattern of area attribute values, an Index near zero indicates a random pattern, and an Index greater than zero indicates a clustered pattern. Mathematically, the Moran’s Index is similar to a correlation coefficient, as we shall see in the next chapter.

While Moran's technique provides an index for the level of spatial autocorrelation for a set of areas, the question becomes how different from zero (randomness) must the Index be in order to indicate significant dispersion or clustering? To answer this question, the Index is usually converted to a Z-score using equation 15.6. A *p*-value is then calculated which allows you to determine the significance of the difference between the observed Moran's Index and the corresponding expected Index value for a purely random distribution of values across the same number of areal units.

The null hypothesis is that the area values are arranged in a completely random spatial pattern, and that spatial autocorrelation is not present within the study area. The Z score is computed as:

$$Z = \frac{I - E_I}{\sqrt{Var_I}} \quad (15.6)$$

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

Moran's Index

Primary Objective: Identify significant spatial patterns within a study area

Requirements and Assumptions:

1. Minimum of (30) geographic features
2. Attribute values measured on an ordinal or interval/ratio scale

Hypotheses:

H_0 : Attribute values are randomly distributed across features in the study area

H_A : Attribute values are not randomly distributed across features in the study area

Test Statistic:

$$I = \frac{n \sum (x_i - \bar{x})(x_j - \bar{x})}{J \sum (x - \bar{x})^2}$$

Interpretation:

Assuming a significant p -value:

- $I < 0$ (observed pattern is dispersed)
- $I = 0$ (observed pattern is random)
- $I > 0$ (observed pattern is clustered)

Variance under the *randomization* assumption is computed as:

$$Var_i =$$

$$\frac{n[J(n^2+3-3n)+3J^2-n\sum L^2]-k[J(n^2-n)+6J^2-2n\sum L^2]}{J^2(n-1)(n-2)(n-3)}$$

where: k represents the kurtosis for the variable x

If the p -value is not significant, then you should not reject the null hypothesis, as we cannot confidently say that the observed pattern is different from complete spatial randomness. This conclusion is equivalent to a statement that the study area does not exhibit significant spatial autocorrelation. However, if the p -value is significant and the Z -score is positive, we can confidently reject the null hypothesis and say that similar area attribute values (either high or low) are clustered in the study area. If the p -value is significant and the Z -score is negative, we

TABLE 15.6

Worktable for Moran's I: Five-area Example Pattern

H_0 : no spatial autocorrelation in the data

H_A : spatial autocorrelation exists

Joins			Area values		
Area	L	L^2	x	$(x - \bar{x})$	$(x - \bar{x})^2$
A	2	4	70	25	625
B	3	9	80	35	1,225
C	3	9	20	-25	625
D	2	4	10	-35	1,225
Joins(J) = 5		$\sum L^2 = 26$	$\bar{x} = \frac{180}{4} = 45$	$\sum (x - \bar{x})^2 = 3,700$	

Join number	x_i	$(x_i - \bar{x})$	x_j	$(x_j - \bar{x})$	$(x_i - \bar{x})(x_j - \bar{x})$
1	70	25	80	35	875
2	70	25	20	-25	-625
3	20	-25	80	35	-875
4	80	35	10	-35	-1,225
5	10	-35	20	-25	875
SUM					-975

$$E_t = \frac{-1}{(n-1)} = -\frac{1}{4-3} = -0.33$$

$$I = \frac{n \sum (x_i - \bar{x})(x_j - \bar{x})}{I \sum (x - \bar{x})^2} = \frac{4(-975)}{5(3,700)} = -0.21$$

$$Var_i = \frac{n^2J + 3J^2 - n\sum L^2}{J^2(n^2 - 1)} = \frac{(4^2 \times 5) + (3 \times 5^2) - (4 \times 26)}{5^2(4^2 - 1)} = \frac{51}{375} = 0.136$$

$$Z = \frac{I - E_I}{\sqrt{\text{Var}_I}} = \frac{-0.21 - (-0.33)}{\sqrt{0.136}} = 0.326$$

$$p = 0.74$$

can confidently reject the null hypothesis and say that the area attribute values are more dispersed across the study area than what we would expect by random chance.

The Moran's *I* computation procedure is presented in table 15.6 using the simplified study area having four sub-areas and five joins, as shown in figure 15.3. Calculations

are only shown using the normalization assumption. In this example, a Moran's *I* of $-.21$ and a *p* value of $.74$ indicates that we do not have enough evidence to reject the null hypothesis and should therefore conclude that the study area does not exhibit significant spatial autocorrelation.

**Example: Moran's *I* Index (Global)—
Distribution of Racial and Ethnic Groups in Cleveland, Ohio Census Block Groups**

The Moran's Index is applied at the census block group level to explore the spatial distribution of racial and ethnic groups within Cleveland, Ohio (fig. 15.4). The objective is to determine the extent to which various racial and ethnic populations are dispersed, randomly located or clustered. Such a concern is important because while progress has been made toward providing equal access to housing and other impor-

tant public services for all racial and ethnic groups, historical patterns of segregation still exist, particularly in older, industrial American cities. Among other concerns, urban geographers have demonstrated a strong link between patterns of segregation and higher rates of crime and poverty.

Three racial groups and one ethnic group in Cleveland, Ohio are selected to assess the magnitude of spatial autocorre-

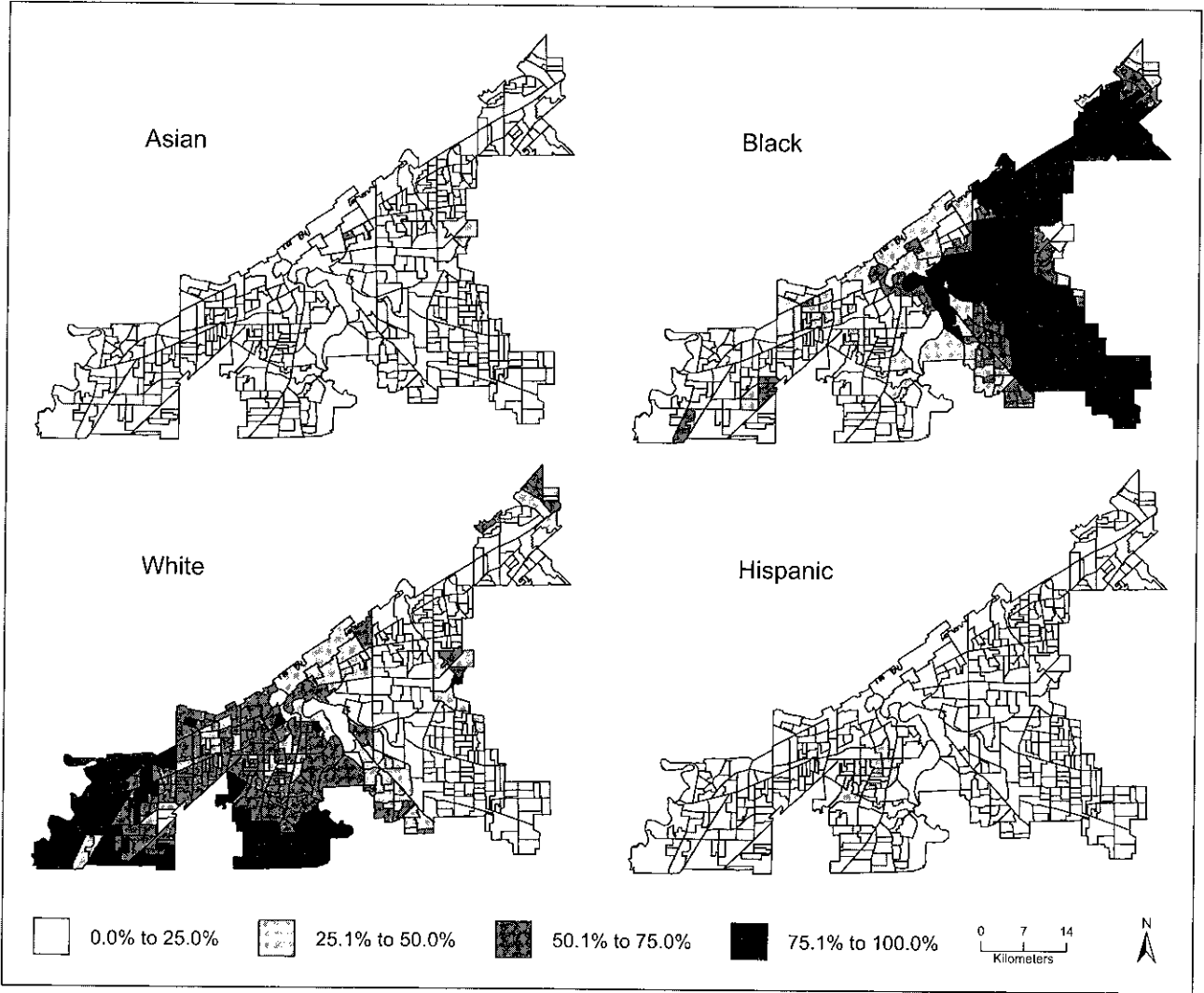


FIGURE 15.4
Distribution of Various Racial and Ethnic Groups by Census Block Group in Cleveland, Ohio
Source: United States Bureau of the Census

TABLE 15.7
Moran's *I* Values and Test Statistics for Selected Racial and Ethnic Groups in Cleveland, Ohio Census Block Groups

Group	Moran's Index	Z	p-value
Asian	0.35	18.39	0.000
Black	0.58	64.6	0.000
White	0.64	65.4	0.000
Hispanic	0.81	65.7	0.000

lation using Moran's Index. As reported in the 2010 Census, Cleveland is racially diverse with 53.3% of the population reporting their race as black, 33.4% white (non-Hispanic), 1.8% Asian, and 10.0% reporting their ethnicity as Hispanic/Latino.

A global Moran's analysis is applied to each group to determine the degree to which areas with similar (either high or low) racial/ethnic composition cluster together across the city. Simple adjacency or contiguity is the method used to determine the magnitude of spatial autocorrelation. The null hypothesis in each case is that the values representing the percentage of each racial/ethnic group are arranged in a completely random spatial pattern. The Moran's indices and test statistic results for each of the four racial/ethnic groups are shown in table 15.7.

In this table, the groups are ranked in increasing magnitude on the Moran's index (that is, increasing levels of cluster-

ing since all of the Z values are positive). The distribution of Asians is the least clustered. With an index value of 0.35, we can ask whether that value differs significantly from the expected value of -0.002 for a purely random distribution across the same number of census block groups. The resulting test statistics for percent Asian population ($Z = 18.39, p = .000$) indicate that you can reject the null hypothesis with near 100% certainty and conclude that the spatial pattern of Asians in Cleveland is non-random. Indeed, with large positive test statistics and *p*-values of .000, we can conclude that all groups are significantly more clustered than random. In summary, we can conclude that a high level of residential segregation (and spatial autocorrelation) currently exists in Cleveland.

Strongly stated inferential conclusions in this type of problem should be made with extreme care. The underlying null hypothesis of complete spatial randomness is very restrictive. Arguably, almost any spatial distribution that would be of interest to a geographer will show some measurable level of spatial autocorrelation. A more reasonable approach to the interpretation of these results is to consider them descriptive summaries of the patterns within the study area. Rather than a strict inferential interpretation, it might be better to focus on a relative comparison of the magnitude of the Moran's indices across the four racial/ethnic groups. Such a comparison of Moran's values across groups is valid in this instance because all analyses are conducted for the same study area, using similar variables (all related to racial/ethnic composition), and making the same assumptions regarding adjacency and contiguity when creating hypotheses.

15.3 MORAN'S *I* INDEX (LOCAL)

In addition to looking at global measures of spatial autocorrelation, geographers may also be interested in assessing whether spatial autocorrelation exists at a local level. As a global indicator of spatial autocorrelation, Moran's *I* may miss certain subtleties within a geographic dataset. That is, while the overall spatial arrangement might indicate that no global spatial autocorrelation exists, there may in fact be pockets (typically called hotspots) of spatial autocorrelation not picked up by the global measure. To resolve this issue, geographers have adapted global methods of spatial autocorrelation like Moran's *I* to include a local component.

This concept, described by geographer Luc Anselin (1995) and others, is termed *Local Indicators of Spatial Association (LISA)*. LISA statistics quantify the similarity of each geographic observation with an identified group of geographic neighbors, based on one of the geographic weighting schemes discussed earlier. It is often used to identify local clusters (geographic locations where adjacent or nearby areas have similar values) or spatial outliers (geographic locations that are different

from adjacent or nearby areas). Therefore, rather than a single measure of spatial autocorrelation, the LISA methods determine individual measures for each geographic entity as follows:

$$I = \frac{(x_i - \bar{X})}{S_i^2} \sum_{j=1, j \neq i} w_{ij} (x_j - \bar{X}) \quad (15.7)$$
$$S_i^2 = \sum_{j=1, j \neq i} \frac{(x_j - \bar{X})^2}{n-1} - \bar{X}^2$$

where: x_i = the value for a particular geographic entity
 x_j = the value for the neighboring geographic entity
 $\bar{X}$ = the average of all attributes
 w_{ij} = the spatial weights.

The overall computation of the local Moran index is beyond the scope of this book, but a short example will help illustrate its usefulness.

Example: Moran's I Index (Local)—Distribution of Obesity in Pennsylvania Counties

Consider the pattern of obesity levels in Pennsylvania counties shown in figure 15.5. You can definitely see a pattern of high obesity in the southwest corner of the Commonwealth, while the southeastern corner has lower obesity values. However, due to the overall configuration, the Moran's I statistic for the Commonwealth is a single measure, computed as .69 with a p-value of .25, indicating no spatial autocorrelation at the global (statewide) level. However, computation of the local Moran's I value provides a p-value for each county.

Figure 15.6 shows the counties with a Z-score above 2.0 or less than -2.0 ($p = 0.00$). Those counties with high positive Z-scores indicate places whose obesity level is very similar to neighboring counties, suggesting local clustering. It is not all

that surprising to see the metropolitan Philadelphia counties highlighted as an urban region with a cluster of low obesity level counties. A pair of adjacent counties east of Pittsburgh (Cambria County and Blair County) with smaller urban areas (Johnstown and Altoona, respectively) is also identified as very similar to neighboring counties in a region of high obesity levels. Finally, Fayette County in southwest Pennsylvania has a large negative Z-score indicating that it is an outlier in obesity level when compared to its neighboring counties. This pattern of local spatial autocorrelation values should raise other basic geographic questions about why the pattern has the specific distribution and variation characteristics shown on the two maps.

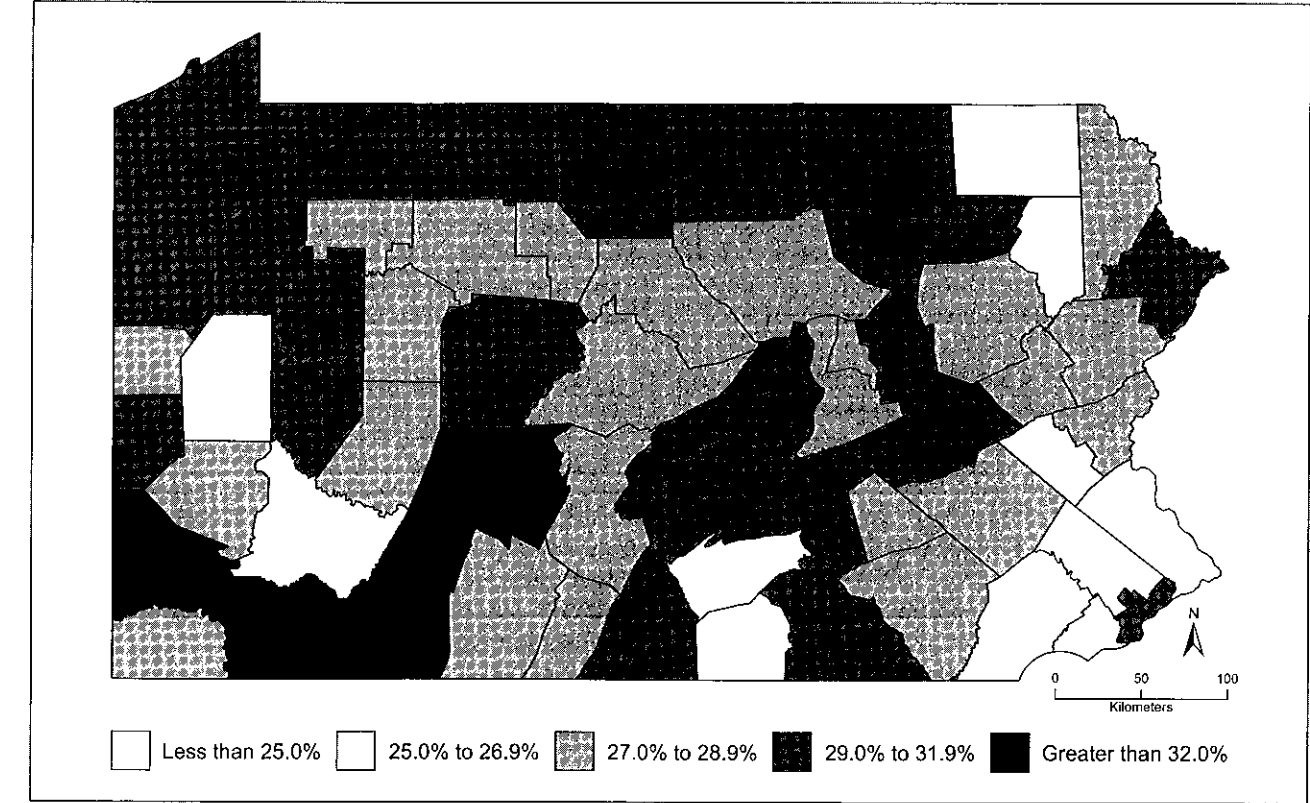


FIGURE 15.5
Distribution of Obesity Levels in Pennsylvania Counties, 2010
Source: Centers for Disease Control and Prevention (CDC)

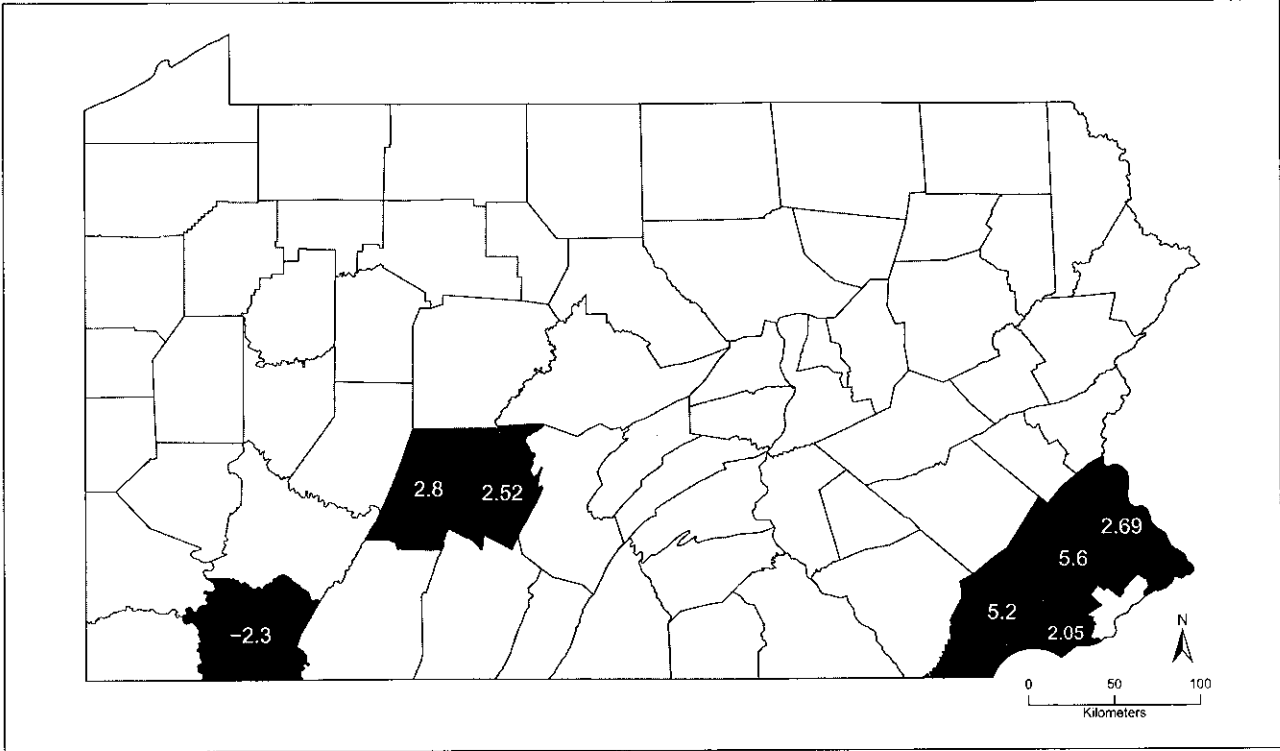


FIGURE 15.6
Pennsylvania Counties with Significant Local Spatial Autocorrelation

KEY TERMS

- area pattern analysis, 222
- free and non-free sampling, 231
- join count analysis, 223
- Moran's I index (global and local), 229, 233

MAJOR GOALS AND OBJECTIVES

- If you have mastered the material in this chapter, you should now be able to:
1. Understand the purposes of area pattern analysis.
 2. Define and explain join structure and the joint count statistic in area pattern analysis.
 3. Distinguish between the contexts or problem settings in which the hypotheses of free sampling and non-free sampling are best used.
 4. Explain the basic objectives of Moran's index of autocorrelation.
 5. Understand the different goals of global and local autocorrelation.

REFERENCES AND ADDITIONAL READING

Anselin, Luc. "Local Indicators of Spatial Association—LISA." *Geographical Analysis*, Vol. 27, No. 2 (1995): 93–115.

Bailey, T. C. and A. C. Gatrell. *Interactive Spatial Data Analysis*. London: Longman, 1995.

Ebdon, D. *Statistics in Geography. A Practical Approach*. Oxford: Basil Blackwell, 1985.

Griffith, D. and C. Amrhein. *Statistical Analysis for Geographers*. Englewood Cliffs, NJ: Prentice Hall, 1991.

Lee, J. and D. W. S. Wong. *Statistical Analysis with ArcView GIS*. New York: Wiley, 2001.

Mitchell, A. *The ESRI Guide to GIS Analysis. Vol. 2: Spatial Measurements and Statistics*. Redlands, CA: ESRI Press, 2005.

Ripley, B. *Spatial Statistics*. New York: Wiley, 1981.

For state-level obesity data and county-level obesity data (for Pennsylvania or any other state), see the Centers for Disease Control and Prevention www.cdc.gov. For spatial distributions of racial and ethnic groups by census block group or by census tract for Cleveland, Ohio or any other metropolitan area, see the U.S. Bureau of the Census www.census.gov.