10 Spatial Patterns

LEARNING OBJECTIVES

10.1 Introduction

One assumption of regression analysis as applied to spatial data is that the residuals are independent and thus are not spatially autocorrelated – that is, there is no spatial pattern to the errors. Residuals that are not independent can affect estimates of the variances of the coefficients, and hence make it difficult to judge their significance.

We have also seen in other, previous chapters, that lack of independence among observations can affect the outcome of *t*-tests, ANOVA, correlation, and regression, often leading one to find significant results where none in fact exists. One reason for learning more about spatial patterns and their detection, then, is an indirect one – we seek to assess spatial dependence so that we may ultimately correct our statistical analyses based upon dependent spatial data.

In addition to a desire to remove the complicating effects of spatially dependent observations, spatial analysts also seek to learn whether geographic phenomena cluster in space. Here they have a direct interest in the phenomenon and/or process itself. For example, crime analysts wish to know if clusters of criminal activity exist. Health officials seek to learn about disease clusters and their determinants.

In this chapter we will investigate statistical methods aimed at detecting spatial patterns and assessing their significance. We will in particular focus upon statistical tests of the null hypothesis that a spatial pattern is random. The structure of the chapter follows from the fact that data are typically in the form of either point locations (where exact locations of, e.g., disease or crime are available), or in the form of aggregated areal information (where, e.g., information is available only on regional rates).

10.2 The Analysis of Point Patterns

Carry out the following experiment.

Draw a rectangle that is 6 inches by 5 inches on a sheet of paper. Locate 30 dots at random within the rectangle. This means that each dot should be located independently of the other dots. Also, for each point you locate, every subregion of a given size should have an equal likelihood of receiving the dot.

Then draw a six-by-five grid of 30 square cells on top of your rectangle. You can do this by making little tick marks at one-inch intervals along the sides of your rectangle. Connecting the tick marks will divide your original rectangle into 30 squares, each having a side of length one inch.

Give your results a score, as follows. Each cell containing no dots receives one point. Each cell containing one dot receives 0 points. Each cell containing two dots receives 1 point. Cells containing three dots receive 4 points, cells containing four dots receive 9 points, cells containing 5 dots receive 16 points, cells containing 6 dots receive 25 points, and cells containing 7 dots receive 36 points. Find your total score by adding up the points you have received in all 30 cells.

DO NOT READ ON UNTIL YOU HAVE COMPLETED THE INSTRUCTIONS ABOVE!

Classify your pattern as follows.

If your score is 16 or less, your pattern is significantly more uniform or regular than random.

If your score is between 17 and 45, your pattern is characterized as random.

If your score is greater than 45, your pattern exhibits significant clustering.

On average, a set of 30 randomly placed points will receive a score of 29, 95% of the time, a set of randomly placed points will receive a score between 17 and 45. The majority of people who try this experiment produce patterns that are more uniform or regular than random, and hence their scores are less than 29. Their point patterns are more spread out than a truly random pattern. When individuals see an empty space on their diagram, there is an almost overwhelming urge to fill it in by placing a dot there! Consequently, the locations of dots placed on a map by individuals are not independent of the locations of previous dots, and hence an assumption of randomness is violated.

Consider next Figures 10.1 and 10.2, and suppose you are a crime analyst looking at the spatial distribution of recent crimes. Make a photocopy of the page, and indicate in pencil where you think the clusters of crime are. Do this by simply encircling the clusters (you may define more than one on each diagram).

DO NOT READ THE NEXT PARAGRAPH UNTIL YOU HAVE COMPLETED THIS EXERCISE!

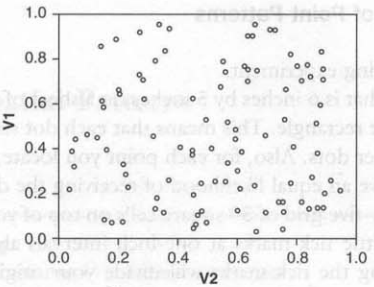


FIGURE 10.1 **Spatial pattern of crime**

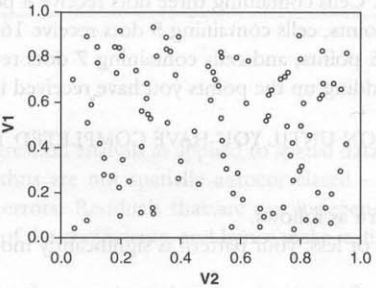


FIGURE 10.2 **Spatial pattern of crime**

How many clusters did you find? It turns out that both diagrams were generated by locating points at random within the rectangle! In addition to having trouble drawing random patterns, individuals also have a tendency to "see" clusters where none exist. This results from the mind's strong desire to organize spatial information.

Both of these exercises point to the need for objective, quantitative measures of spatial pattern – it is simply not sufficient to rely on one's visual interpretation of a map. Crime analysts cannot necessarily pick out true clusters of crime just by looking at a map, nor can health officials always pick out significant clusters of disease from map inspection.

10.2.1 Quadrat Analysis

The experiment involving the scoring of points falling within the "6 × 5" rectangle is an example of *quadrat analysis*, developed primarily by ecologists in the first half of the twentieth century. In quadrat analysis, a grid of square cells of equal size is

used as an overlay, on top of a map of incidents. One then counts the number of incidents in each cell. In a random pattern, the mean number of points per cell will be roughly equal to the variance of the number of points per cell.

If there is a large amount of variability in the number of points from cell to cell (some cells have many points; some have none, etc.), this implies a tendency toward *clustering*. If there is very little variability in the number of points from cell to cell (e.g., when all or almost all of the cells have about the same number of points), this implies a tendency toward a pattern that is termed *regular*, *uniform*, or *dispersed* (where the number of points per cell is about the same in all cells). The statistical test used to evaluate the null hypothesis of spatial randomness makes use of a chi-square statistic involving the variance-mean ratio:

$$\chi^2 = \frac{(m-1)\sigma^2}{\bar{x}}, \tag{10.1}$$

where m is the number of quadrats, and $\bar{x}$ and σ^2 are the mean and variance of the number of points per quadrat, respectively. This value is then compared with a critical value from a chi-square table, with $m - 1$ degrees of freedom.

Quadrat analysis is easy to employ, and it has been a mainstay in the spatial analyst's toolkit of pattern detectors over several decades. One important issue is the size of the quadrat; if the cell size is too small, there will be many empty cells, and if clustering exists on any but the smallest spatial scales, it may be missed. If the cell size is too large, one may miss patterns that occur *within* cells. One may find patterns on some spatial scales and not at others, and thus the choice of quadrat size can seriously influence the results. Curtiss and McIntosh (1950) suggest an "optimal" quadrat size of two points per quadrat. Bailey and Gatrell (1995) suggest that the mean number of points per quadrat should be about 1.6.

10.2.1.1 Summary of the Quadrat Method

1. Divide a study region into m cells of equal size.
2. Find the mean number of points per cell ($\bar{x}$). This is equal to the total number of points divided by the number of cells (m).
3. Find the variance of the number of points per cell, s^2, as follows:

$$s^2 = \frac{\sum_{i=1}^{i=m} (x_i - \bar{x})^2}{m-1} \tag{10.2}$$

   where x_i is the number of points in cell i.
4. Calculate the variance-mean ratio (VMR):

$$\text{VMR} = \frac{s^2}{\bar{x}} \tag{10.3}$$

5. Interpret the results as follows:

   (a) If $s^2/\bar{x} < 1$, the variance of the number of points is less than the mean. In the extreme case where the ratio approaches zero, there is very little variation in the number of points from cell to cell. This characterizes situations where the distribution of points is spread out, or uniform, across the study area.

   (b) If $s^2/\bar{x} > 1$, there is a good deal of variation in the number of points per cell – some cells have substantially more points than expected (i.e., $x_i > \bar{x}$ for some cells i), and some cells have substantially fewer than expected (i.e., $x_i < \bar{x}$). This characterizes situations where the point pattern is more clustered than random. A value of $s^2/\bar{x}$ near one indicates that the points are close to randomly distributed across the study area.

6. Hypothesis testing.

   (a) Multiply the VMR by $m - 1$; the quantity $\chi^2 = (m - 1)\text{VMR}$ has a chi-square distribution, with $m - 1$ degrees of freedom, when H_0 is true. This fact allows us to obtain critical values, χ^2_L and χ^2_H, from a chi-square table. In particular, we will reject H_0 if either $\chi^2 < \chi^2_L$ or $\chi^2 > \chi^2_H$. If the number of cells (m) is greater than about 30, $(m - 1)\text{VMR}$ will, when H_0 is true, have a normal distribution with mean $m - 1$ and variance equal to $2(m - 1)$. This means that we can treat the quantity

$$z = \frac{(m-1)\text{VMR} - (m-1)}{\sqrt{2(m-1)}} = \sqrt{(m-1)/2}\,(\text{VMR} - 1) \qquad (10.4)$$

   as a normal random variable with mean 0 and variance 1. With $\alpha = 0.05$, the critical values are $z_L = -1.96$ and $z_H = +1.96$. The null hypothesis of no pattern is rejected if $z < z_L$ (implying clustering) or if $z > z_H$ (implying uniformity).

EXAMPLE

We wish to know whether the pattern observed in Figure 10.3 is consistent with the null hypothesis that the points were placed at random. We first calculate the VMR. There are 100 points on the 10 x 10 grid, implying a mean of one point per cell. There are six cells with three points, 20 cells with two points, 42 cells with one point, and 32 cells with no points. The variance is

$$\frac{\{6(3-1)^2 + 20(2-1)^2 + 42(1-1)^2 + 32(0-1)^2\}}{99} = \frac{76}{99}$$
$$= 0.77, \qquad (10.5)$$

(Continued)

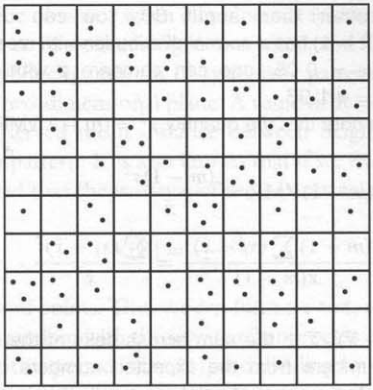


FIGURE 10.3 **A spatial point pattern**

and, since the mean is equal to one, this is also our observed VMR. Since VMR < 1, there is a tendency toward a uniform pattern. How unlikely is a value 0.77 if the null hypothesis is true – is it unlikely enough that we should reject the null hypothesis?

Since the number of degrees of freedom (df) is large, the sampling distribution of $\chi^2 = (m - 1)$ VMR begins to approach the shape of a normal distribution. In particular, we can use Equation 10.4, so that in our example, we have

$$z = \frac{99(0.77) - 99}{\sqrt{2(99)}} = \sqrt{99/2}(0.77 - 1) = -1.618 \qquad (10.6)$$

This falls within the critical values of z and hence we do not have strong enough evidence to reject the null hypothesis.

If cells of a different size had been used, the results, and possibly the conclusions, would have been different. By aggregating the cells in Figure 10.3 to a 5×5 grid of 25 cells, the VMR declines to 0.687 (based on a variance of 1.658^2 and a mean of four points per cell). The χ^2 value is $24(.687) = 16.5$. Since the degrees of freedom are less than 30, we will use the chi-square table (Table A.5) to assess significance. With 24 degrees of freedom, and using interpolation to find the critical values at $p = 0.025$ and $p = 0.975$ yields $\chi^2_L = 12.73$ and $\chi^2_U = 40.5$. Since our observed value of 16.5 falls between these limits, we again fail to reject the hypothesis of randomness.

To summarize, after finding VMR in steps 1–4 above, calculate $\chi^2 = (m - 1)$ VMR, and compare it with the critical values found in a chi-square table, using df $= m - 1$.

(Continued)

(Continued)

If $m - 1$ is greater than about 30, you can use the fact that $z = \sqrt{(m-1)/2}(\text{VMR} - 1)$ has a normal distribution with mean 0 and variance 1, implying that, for $\alpha = 0.05$, one can compare z with the critical values $z_L = -1.96$ and $z_H = +1.96$.

It is interesting to note that the quantity $\chi^2 = (m-1)\text{VMR}$ may be written as:

$$\chi^2 = (m-1)\text{VMR} = \frac{(m-1)s^2}{\bar{x}}$$

$$= \frac{(m-1)\sum (x_i - \bar{x})^2}{\bar{x}(m-1)} = \frac{\sum (x_i - \bar{x})^2}{\bar{x}} \qquad (10.7)$$

The quantity $\sum (x_i - \bar{x})^2/\bar{x}$ is the sum across cells of the squared deviations of the observed numbers from the expected numbers of points in a cell, divided by the expected number of points in a cell. This is commonly known as the chi-square goodness-of-fit test.

10.2.2 Nearest Neighbor Analysis

Clark and Evans (1954) developed nearest neighbor analysis to analyze the spatial distribution of plant species. They developed a method for comparing the observed average distance between points and their nearest neighbors with the distance that would be expected between nearest neighbors in a random pattern.

We begin by defining R_0 to be the observed average distance between points and their nearest neighbors. Let R_e be the expected distance between points and their nearest neighbors when points are distributed randomly. Intuitively, if R_0 is small relative to R_e, the pattern will be clustered; if R_0 is large relative to R_e, the pattern will be more dispersed than random.

R_0 may be calculated as $\sum_{i=1}^{n} d_i/n$ where n is the number of points in the study area, and where d_i is the distance from point i to its nearest neighbor. Note that nearest neighbors may be reflexive – that is, they may be nearest neighbors of each other.

R_e is calculated as one over twice the square root of the density of points:

$$R_e = \frac{1}{2\sqrt{\rho}} = \frac{1}{2\sqrt{n/A}} \qquad (10.8)$$

where ρ is the density of points, and A is the size of the study area.

The nearest neighbor statistic, R, is defined as the ratio between the observed and expected values:

$$R = \frac{R_0}{R_e} = \frac{\bar{d}}{1/(2\sqrt{\rho})} = 2\bar{d}\sqrt{\rho}. \tag{10.9}$$

R varies from 0 (a value obtained when all points are in one location, and the distance from each point to its nearest neighbor is zero), and a theoretical maximum of about 2.14, for a perfectly uniform or systematic pattern of points spread out on an infinitely large two-dimensional plane. A value of $R = 1$ indicates a random pattern, since the observed mean distance between neighbors is equal to that expected in a random pattern. It is also known that if we examined many random patterns, we would find that the variance of the nearest neighbor statistic, R, is

$$\mathrm{V}[R] = \frac{4 - \pi}{2\pi n} \tag{10.10}$$

where n is the number of points. Thus we can form a z-test, to test the null hypothesis that the pattern is random with the now familiar process of starting with the statistic (R), subtracting its expected value (1), and dividing by its standard deviation:

$$z = \frac{(R - 1)}{\sqrt{\mathrm{V}[R]}} = \frac{\sqrt{\pi n}(R_0 - 1)}{\sqrt{4 - \pi}} \approx 1.913(R - 1)\sqrt{n} \tag{10.11}$$

The quantity z has a normal distribution with mean 0 and variance 1, and hence tables of the standard normal distribution may be used to assess significance. A value of $z > 1.96$ implies that the pattern has significant uniformity, and a value of $z < -1.96$ implies that there is a significant tendency toward clustering.

The strength of this approach lies in its ease of calculation and comprehension. Several cautions should be noted in the interpretation of the nearest neighbor statistic. The statistic, and its associated test of significance, may be affected by the shape of the region. Long, narrow, rectangular shapes may have relatively low values of R simply because of the constraints imposed by the region's shape. Points in long, narrow rectangles are *necessarily* close to one another. Boundaries can also make a difference in the analysis. One solution to the boundary problem is to place a buffer area around the study area. The nearest neighbors are found for all points within the study area (but not for the points in the buffer area). Points inside of the study area (such as point A in Figure 10.4) may have nearest neighbors that fall into the buffer area, and these distances (rather than distances to those points that are nearest within the study area) should be used in the analysis.

Another potential difficulty with the statistic is that, since only nearest neighbor distances are used, clustering is only detected on a relatively small spatial scale. To overcome this it is possible to extend the approach to second- and higher-order nearest neighbors.

Most importantly, it is often of interest to ask not only whether clustering exists, but whether clustering exists over and above some background factor (such as

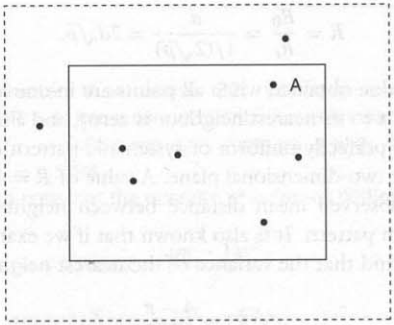


FIGURE 10.4 **Boundary effects in nearest neighbor analysis**

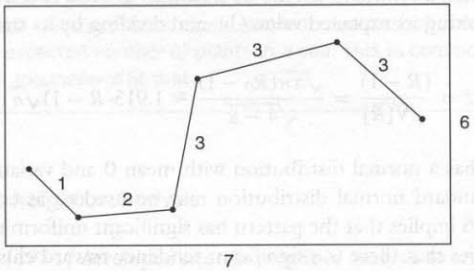


FIGURE 10.5 **Nearest neighbor distances**

population). Nearest neighbor methods are not particularly useful in these situations because they only relate to spatial location, and not to other attributes. The approaches to the study of pattern that are described in the next section do not have this limitation.

10.2.2.1 Illustration For the point pattern in Figure 10.5, distances are given along the lines connecting the points. The mean distance between nearest neighbors is $R_0 = (1 + 2 + 3 + 1 + 3 + 3)/6 = 13/6 = 2.167$. The expected mean distance between nearest neighbors in a pattern of six points placed randomly in a study region with area $7 \times 6 = 42$ is

$$R_e = 1/(2\sqrt{\lambda}) = 1/(2\sqrt{6/42}) = 1.323. \qquad (10.12)$$

The nearest neighbor statistic is $R = 2.167/1.323 = 1.638$, which means that the pattern displays a tendency toward uniformity. To assess significance, we can calculate

the z-statistic from (10.10) as $1.913(1.638 - 1)\sqrt{6} = 2.99$, which is much greater than the critical value of 1.96, which in turn implies rejection of the null hypothesis of a random pattern. However, we have neglected boundary effects, and these have a significant effect on the results. As an alternative way to test the null hypothesis, we can randomly choose six points by choosing random x-coordinates in the range (0, 7) and random y-coordinates in the range (0, 6). Then we compute the mean distance from each of the six points to their nearest neighbors, and repeat the whole process many times. Simulating the random placement of six points in the 7×6 study region 10,000 times led to a mean distance between nearest neighbors of 1.62. This is greater than the expected distance of $R_e = 1.323$ noted above. This greater-than-expected distance can be attributed directly to the fact that points near the border of the study region are relatively farther from other points in the study region than they presumably would have been to points just outside of the study region. Ordering the 10,000 mean distances to nearest neighbors reveals that the 9500th highest one is 2.29. This implies that only 5% of the time would we expect a mean distance greater than 2.29. Our observed distance of 2.167 is less than 2.29, and so we, having accounted for boundary effects through our Monte Carlo simulation, accept the null hypothesis.

10.3 Geographic Patterns in Areal Data

10.3.1 An Example Using a Chi-Squared Test

In a regression of housing prices on housing characteristics, suppose that we have data on 51 houses that are located in three neighborhoods. How might we tell whether there is a tendency for positive or negative residuals to cluster in one or more neighborhoods? One idea is to note whether each residual is positive or negative, and then to tabulate the residuals by neighborhood (see the hypothetical data in Table 10.1).

We can use a chi-square test to determine whether there is any neighborhood-specific tendency for residuals to be positive or negative. Under the null hypothesis of no spatial pattern (i.e., no interaction between the rows and columns of the table), the expected values are equal to the product of the row and column totals, divided by the overall total. For example, we expect $23(16)/51 = 7.22$ positive residuals in neighborhood 1. These expected values are given in parentheses in Table 10.2.

TABLE 10.1 **Hypothetical residuals**

		Neighborhood			
		1	**2**	**3**	**Total**
	+	10	6	7	23
	−	6	15	7	28
Total		16	21	14	51

TABLE 10.2 **Observed and expected frequencies of residuals**

	Neighborhood			
	1	**2**	**3**	**Total**
+	10	6	7	23
	(7.22)	(9.47)	(6.31)	
−	6	15	7	28
	(8.78)	(11.53)	(7.69)	
Total	16	21	14	51

The chi-square statistic is

$$\chi^2 = \sum_{i=1}^{n} \frac{(O - E)^2}{E}, \tag{10.13}$$

where O is the observed frequency, and E is the expected frequency. When the null hypothesis is true, this statistic has a χ^2 distribution with degrees of freedom equal to the number of rows minus one, times the number of columns minus one.

In this example, the value of chi-square is 4.40, which is less than the critical value of 5.99, using $\alpha = 0.05$, and 2 degrees of freedom. Therefore the null hypothesis of no pattern is not rejected.

The observed chi-square statistic for the data in Table 10.2 is:

$$\chi^2 = \frac{(10 - 7.22)^2}{7.22} + \frac{(6 - 9.47)^2}{9.47} + \frac{(7 - 6.31)^2}{6.31} + \frac{(6 - 8.78)^2}{8.78} + \frac{(15 - 11.53)^2}{11.53} + \frac{(7 - 7.69)^2}{7.69} = 4.40$$

Now that spatial autocorrelation in the residuals has been detected, what can be done about it? One possibility is to include a new, location-specific dummy variable. This will serve to capture the importance of an observation's location in a particular neighborhood. In our present housing price example, we could add two variables, one for two of the three neighborhoods (following the usual practice of omitting one category). You should also note that if there are k neighborhoods, it is *not* necessary to have $k - 1$ dummy variables; rather, you might choose to have only one or two dummy variables for those neighborhoods having large deviations between the observed and predicted values.

10.3.2 Moran's I

Moran's I statistic (1948, 1950) is one of the classic (as well as one of the most common) ways of measuring the degree of spatial autocorrelation in areal data. Moran's I is calculated as follows:

$$I = \frac{n \sum_i^n \sum_j^n w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{(\sum_i^n \sum_j^n w_{ij}) \sum_i^n (y_i - \bar{y})^2}, \tag{10.14}$$

where there are n regions and w_{ij} is a measure of the spatial proximity between regions i and j. It is interpreted much like a correlation coefficient. Values near +1 indicate a strong spatial pattern (high values tend to be located near one another, and low values tend to be located near one another). Values near −1 indicate strong negative spatial autocorrelation; high values tend to be located near low values. (Spatial patterns with negative autocorrelation are either extremely rare or nonexistent!) Finally, values near 0 indicate an absence of spatial pattern.

Though Equation 10.14 is perhaps daunting at first glance, it is helpful to realize that if the variable of interest is first transformed into a z-score $\{z = (x - \bar{x})/s\}$, a much simpler expression for I results:

$$I = \frac{n \sum_i \sum_j w_{ij} z_i z_j}{(n-1) \sum_i \sum_j w_{ij}} \tag{10.15}$$

The conceptually important part of the formula is the numerator, which sums the products of z-scores in nearby regions. Pairs of regions where *both* regions exhibit above-average scores (or below-average scores) will contribute positive terms to the numerator, and these pairs will therefore contribute toward positive spatial autocorrelation. Pairs where one region is above average and the other is below average will contribute negatively to the numerator, and hence to negative spatial autocorrelation.

The weights $\{w_{ij}\}$ can be defined in a number of ways. Perhaps the most common definition is one of *binary connectivity*: $w_{ij} = 1$ if regions i and j are contiguous, and $w_{ij} = 0$ otherwise. "Contiguous" can in turn be defined as requiring regions to share at least a common point (termed *queen's case* contiguity), or, more restrictively, a common boundary of nonzero length (termed *rook's case* contiguity). Sometimes the w_{ij} defined in this way are then standardized to define new w_{ij}^* by dividing by the number of regions i is connected to; i.e., $w_{ij}^* = w_{ij}/\sum_j w_{ij}$. In this case all regions i are characterized by a set of weights linking i to other regions that sum to one; i.e., $\sum_j w_{ij} = 1$.

Alternatively, $\{w_{ij}\}$ may be defined as a function of the distance between i and j (e.g., $w_{ij} = d_{ij}^{-\beta}$ or $w_{ij} = \exp[-\beta d_{ij}]$), where the distance between i and j could be measured along the line connecting the centroids of the two regions. It is conventional to use $w_{ii} = 0$. It is also common, though not necessary, to use symmetric weights, so that $w_{ij} = w_{ji}$.

It is important to recognize that the value of I is very dependent upon the definition of the $\{w_{ij}\}$. Using a simple binary connectivity definition for the map in Figure 10.6 gives us

$$\mathbf{W} = \{w_{ij}\} = \begin{matrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{matrix} \tag{10.16}$$

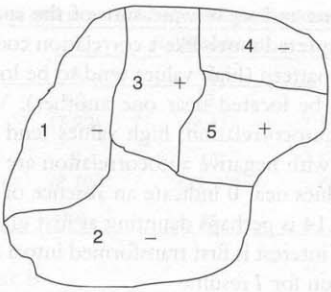


FIGURE 10.6 **Positive and negative residuals in a five-region system**

In this instance, the definition of $\{w_{ij}\}$ causes the neighborhood around region 1 to be much smaller than the neighborhood around region 2 or 3. This is not necessarily "wrong," but suppose that we were interested in the spatial autocorrelation of a disease that was characterized by rates that were strongly associated over small distances, but not correlated over large distances. If we expect disease rates in regions 1 and 2 to be highly correlated while we expect those in regions 4 and 5 to be less highly correlated due to their large spatial separation, our observed value of I would be a combined measure of strong association between close pairs and weak association between distant pairs. For this example, it might be more appropriate to use a distance-based definition of $\{w_{ij}\}$.

10.3.2.1 Illustration Consider the six-region system in Figure 10.7. Using a binary connectivity definition of the weights leads to:

$$W = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \tag{10.17}$$

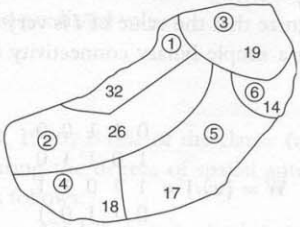


FIGURE 10.7 **Hypothetical six-region system**

where an entry in row i and column j is denoted by w_{ij}. The double summation in the numerator of I (see Equation 10.14) is found by taking the product of the deviations from the mean, for all pairs of adjacent regions:

$$\begin{aligned}(32-21)(26-21)+(32-21)(19-21)+(26-21)(32-21)\\+(26-21)(19-21)+(26-21)(18-21)\\+(26-21)(17-21)+(19-21)(32-21)\\+(19-21)(26-21)+(19-21)(17-21)\\+(19-21)(14-21)+(18-21)(26-21\\+(18-21)(17-21)+(17-21)(19-21)\\+(17-21)(26-21)+(17-21)(18-21)\\+(17-21)(14-21)+(14-21)(19-21)\\+(14-21)(17-21)=100 \qquad (10.18)\end{aligned}$$

Since the sum of the weights in (10.17) is 18, and since the variance of the regional values is 224/5, based on (10.14) Moran's I is equal to

$$I=\frac{6(100)}{18(224)}=.1488. \qquad (10.19)$$

In addition to this descriptive interpretation, there is a statistical framework that allows one to decide whether any given pattern deviates significantly from a random pattern. If the number of regions is large, the sampling distribution of I, under the hypothesis of no spatial pattern, approaches a normal distribution, and the mean and variance of I can be used to create a Z-statistic in the usual way:

$$Z=\frac{I-\mathrm{E}[I]}{\sqrt{\mathrm{V}[I]}} \qquad (10.20)$$

The value is then compared with the critical value found in the normal table (e.g., $\alpha=0.05$ would imply critical values of -1.96 and $+1.96$).

The mean and variance are equal to

$$\begin{aligned}E[I]&=\frac{-1}{n-1}\\ \mathrm{V}[I]&=\frac{n^2(n-1)S_1-n(n-1)S_2+2(n-2)S_0^2}{(n+1)(n-1)^2S_0^2}, \qquad (10.21)\end{aligned}$$

where

$$S_0 = \sum_i^n \sum_{j \neq i}^n w_{ij}$$

$$S_1 = 0.5 \sum_i^n \sum_{j \neq i}^n (w_{ij} + w_{ji})^2$$

$$S_2 = \sum_k^n \left(\sum_j^n w_{kj} + \sum_i^n w_{ik} \right)^2 \qquad (10.22)$$

Computation is not complicated, but it is tedious enough to not want to do it by hand! Unfortunately, few software packages that calculate the coefficient and its significance are available. Exceptions include Anselin's *Spacestat* (1992) and *Geoda* (2003).

Fortunately, there are also simplifications and approximations that facilitate the use of Moran's *I*. An alternative way of finding Moran's *I* is to simply take the ratio of two regression slope coefficients (see Griffith 1996). The numerator of *I* is equal to the regression slope obtained when the quantity $a_i = \sum_{j=1}^{n} w_{ij} z_j$ is regressed on z_i, and the denominator of *I* is equal to the regression slope obtained when the quantity $b_i = \sum_{j=1}^{n} w_{ij}$ is regressed on $c_i = 1$. The z's represent the z-scores of the original variables, and the slope coefficients are found using no-intercept regression (i.e., constraining the result of the regression so that the intercept is equal to zero).

In addition, Griffith gives $2/\sum\sum w_{ij}$ as an approximation for the variance of the Moran coefficient. This expression, though it works best only when the number of regions is sufficiently large (about 20 or more), is clearly easier to compute than the one given in Equations 10.21 and 10.22! Alternatively, when observational units are on a square grid, and connectivity is indicated by the four adjacent cells, the variance may be approximated by $1/(2n)$, where n is the number of cells. Based on either a grid of hexagonal cells or a map displaying "average" connectivity with other regions, the variance may be approximated by $1/(3n)$. An example is given Section 10.5.

The use of the normal distribution to test the null hypothesis of randomness relies upon one of two assumptions:

1. Normality. It can be assumed that regional values are generated from identically distributed normal random variables (i.e., the variables in each region arise from normal distributions that have the same mean and same variance in each region).
2. Randomization. It can be assumed that all possible permutations (i.e., regional rearrangements) of the regional values are equally likely.

The formulae given above (Equations 10.21 and 10.22) for the variance assume that the normality assumption holds. The variance formula for the randomization assumption is algebraically more complex, and gives values that are only slightly different than those given above in (see, e.g., Griffith 1987).

If either of the two assumptions above holds, the sampling distribution of I would have a normal distribution if the null hypothesis of no pattern is true. One of the two assumptions must hold to generate the sampling distribution of I so that critical values of the test statistic may be established. For example, if the first assumption was used to generate regional values, I could be computed; this could then be repeated many times, and a histogram of the results could be produced. The histogram would have the shape of a normal distribution, a mean of $E[I]$, and a variance of $V[I]$. Similarly, the observed regional values on a map could be randomly rearranged many times, and the value of I computed each time. Again, a histogram could be produced; it would again have the shape of a normal distribution with mean $E[I]$ and a variance slightly different than $V[I]$. If we can rely on one of these two assumptions, we do not need to perform these experiments to generate histograms, since we know beforehand that they will produce normal distributions with known mean and variance.

Unfortunately, there are many circumstances in geographical applications that lead the analyst to question the validity of both assumptions. For example, maps of counties by township are often characterized by high population densities in the townships corresponding to or adjacent to the central city, and by low population densities in outlying townships. Rates of crime or disease, though they may have equal means across townships, are unlikely to have equal variances. This is because the outlying towns are characterized by greater uncertainty – they are more likely to experience atypically high or low rates simply because of the chance fluctuations associated with a relatively smaller population base. Thus assumption 1 is not satisfied, since all regional values do not come from identical distributions – some regional values, namely the outlying regions, are characterized by higher variances. Likewise, not all permutations of regional values are equally likely – permutations with atypically high or low values out in the periphery are more likely than permutations with atypically high or low values near the center.

How can we test the null hypothesis of no spatial pattern in this instance? One approach is to use Monte Carlo simulation. Since the z-test described above by Equation (10.20) is no longer valid, we need an alternative way to come up with critical values. The idea is to first assume that the null hypothesis of no spatial pattern is true. Suppose we have data on the number of diseased individuals (n_i) and the population (p_i) in each region. For each individual, assign disease to that the individual with probability $\sum_i n_i / \sum_i p_i$, which is the overall disease rate in the population. Then calculate Moran's I. This is repeated many times, and the resulting values of Moran's I may be used to create a histogram depicting the relative frequencies of I when the null hypothesis is true. Furthermore, the values can be arranged from lowest to highest, and this list can be used to find critical values of I. For example, if the simulations are carried out 1000 times, and critical values are desired for a test using $\alpha = 0.05$, they can be found from the ordered list of I values. The lower critical value would be the 25^{th} item on the list, and the upper critical value would be the 975^{th} item on the list.

Illustration of the Monte Carlo method

Dominik Hasek, the goalie for the gold-medal Czech ice hockey team in the 1998 Olympics, saves 92.4% of all shots he faces when he plays professionally for the Buffalo Sabres of the National Hockey League (NHL). The average save percentage of other goalies in the NHL is 90%. Hasek tends to face about 31 shots per game, while the Sabres manage just 25 shots per game on the opposing goalie. To evaluate how much Hasek means to the Sabres, compare the outcomes of 1000 games using Hasek's statistics with the outcomes of 1000 games assuming the Sabres had an "average" goalie, who stops 90% of the shots against him.

Solution

Take 31 random numbers between 0 and 1. Count those greater than 0.924 as goals against the Sabres with Hasek. Take 25 numbers from a uniform distribution between 0 and 1, and count those greater than 0.9 as goals for the Sabres. Record the outcome (win, loss, or tie). Repeat this 1000 times (preferably using a computer!), and tally the outcomes. Finally, repeat the entire experiment using random numbers greater than 0.9 (instead of 0.924) to generate goals against the Sabres without Hasek. Each time the experiment is performed, a different outcome will be obtained. In one comparison, the results were as follows:

	Wins	Losses	Ties
Scenario 1 (with Hasek)	434	378	188
Scenario 2 (without Hasek)	318	515	167

To evaluate Hasek's value to the team over the course of an 82-game season, the outcomes above may first be converted to percentages, multiplied by 82, and then rounded to integers yielding:

	Wins	Losses	Ties
Scenario 1	36	31	15
Scenario 2	26	42	14

Thus Hasek is "worth" about 10 wins; that is, they win about 10 games a year that they would have lost if they had an "average" goalie.

10.4 Local statistics

10.4.1 Introduction

Besag and Newell (1991) classify the search for clusters into three primary areas. First are "general" tests, designed to provide a single measure of overall pattern for a map consisting of point locations. These general tests are intended to provide a test of the

null hypothesis that there is no underlying pattern, or deviation from randomness, among the set of points. Examples include the nearest neighbor test, the quadrat method, and the Moran statistic, all outlined above. In other situations, the researcher wishes to know whether there is a cluster of events around a single or small number of prespecified foci. For example, we may wish to know whether disease clusters around a toxic waste site, or we may wish to know whether crime clusters around a set of liquor establishments. Finally, Besag and Newell describe "tests for the detection of clustering." Here there is no *a priori* idea of where the clusters may be; the methods are aimed at searching the data and uncovering the size and location of any possible clusters.

General tests are carried out with what are called "global" statistics; again, a single summary value characterizes any deviation from a random pattern. "Local" statistics are used to evaluate whether clustering occurs around particular points, and hence are employed for both focused tests and tests for the detection of clustering. Local statistics have been used in both a confirmatory manner, to test hypotheses, and in an exploratory manner, where the intent is more to suggest, rather than confirm, hypotheses.

Local statistics may be used to detect clusters, either when the location is prespecified (focused tests) or when there is no *a priori* idea of cluster location. When a global test finds no significant deviation from randomness, local tests may be useful in uncovering isolated hotspots of increased incidence. When a global test does indicate a significant degree of clustering, local statistics can be useful in deciding whether (a) the study area is relatively homogeneous in the sense that local statistics are quite similar throughout the area, or (b) there are local outliers that contribute to a significant global statisitc. Anselin (1995) discusses local tests in more detail.

10.4.2 Local Moran Statistic

The local Moran statistic is

$$I_i = n(y_i - \bar{y}) \sum_j w_{ij}(y_j - \bar{y}) \tag{10.23}$$

The sum of local Moran's is equal to, up to a constant of proportionality, the global Moran; i.e, $\Sigma I_i = I$. For example, the local Moran statistic for region 1 in Figure 10.7 is

$$I_1 = (32 - 21)[(26 - 21) + (19 - 21)] = 33. \tag{10.24}$$

The expected value of the local Moran statistic is

$$E[I_i] = \frac{-\sum_j w_{ij}(y_j - \bar{y})}{n - 1} \tag{10.25}$$

and the expression for its variance is more complicated. Anselin gives the variance of I_i, and assesses the adequacy of the assumption that the test statistic has a normal distribution under the null hypothesis.

10.4.3 Getis's G_i Statistic

To test whether a particular location i and its surrounding regions have higher than average values on a variable (x) of interest, Ord and Getis (1995) have used the statistic

$$G_i^* = \frac{\sum_j w_{ij}(d)x_j - W_i^*\bar{x}}{s\{[nS_{1i}^* - W_i^{*2}]/(n-1)\}^{1/2}}, \tag{10.26}$$

where s is the sample standard deviation of the x values, and $w_{ij}(d)$ is equal to 1 if region j is within a distance of d from region i, and 0 otherwise. Also,

$$W_i^* = \sum_j w_{ij}(d)$$

$$S_{1i}^* = \sum_j w_{ij}^2 \tag{10.27}$$

One can see that the numerator of Equation (10.26) represents, for region i, the difference between the weighted value of x in the neighborhood of i and the value that would be expected if the neighborhood was "average" in its x characteristics. Ord and Getis note that, when the underlying variable has a normal distribution, so does the test statistic. Furthermore, the distribution is asymptotically normal even when the underlying distribution of the x-variables is not normal, if the distance d is sufficiently large. Since the statistic (10.26) is written in standardized form, it can be taken as a standard normal random variable, with mean 0 and variance 1.

For region 1 in Figure 10.7, we will use weights equal to 1 for regions 1, 2, and 3, and weights equal to 0 for other regions. The G_i statistic is

$$G_1^* = \frac{87 - 3(21)}{6.69\sqrt{\frac{6(3)-9}{5}}} = 1.543. \tag{10.28}$$

Since this variable has a normal distribution with mean 0 and variance 1 under the null hypothesis that region 1 is not located in a region of particularly high values, we can use a one-sided test with $\alpha = 0.05$ and $z = 1.645$. We therefore fail to reject the null hypothesis.

10.5 Finding Moran's *I* Using *SPSS for Windows 12.0*

Consider the six-region system in Figure 10.7. With connectivity defined by a binary 0–1 weight for adjacent regions, we have the weight matrix given by Equation 10.17.

To compute the value of Moran's I in *SPSS*, we first convert the six regional values to z-scores. These may be found by using Analyze, Descriptives, and Explore, and then clicking on Save standardized scores as variables. For the six regions, the z-scores are 1.64, .747, −.299, −.448, −.598, and −1.046. Then the quantities $a_i = \sum_j w_{ij} z_j$ are found. These are simply weighted sums of the z-scores, where the regions that i is connected to are those z-scores that are summed. For example, region 1 is connected to region 2 and 3. For region 1, $a_1 = .747 - .299 = .448$. The six a_i scores are .448, .299, .747, .149, −1.046, and −.896. Now perform a regression, using the a's as the dependent variable and the z's as the independent variable. In *SPSS*, click on Analyze, Regression, Linear, and then define the dependent and independent variables. Then, under Options, make sure the box labeled "Include constant in equation" is NOT checked. This yields a regression coefficient of .446 for the numerator.

For the denominator, we again use no-intercept regression to regress six y-values on six x-values. The six "y-values" are the sum of the weights in each row (2, 4, 4, 2, 4, and 2 for rows 1–6, respectively). The six x-values are 1, 1, 1, 1, 1, and 1 (this will always be a set of n ones, where n is the number of regions). After again making sure that a constant is NOT included in the regression equation, one finds the regression coefficient is 3.0. Moran's I is simply the ratio of these two coefficients: $.446/3 = .1487$.

The variance of I in this example may be found from Equation 10.21:

$$V[I] = \frac{2(36)(5)(18) - 4(6)(5)(60) + 2(4)(18)^2}{7(5)^2(18)^2} = .033. \qquad (10.29)$$

The z-value associated with a test of the null hypothesis of no spatial autocorrelation is $(.1487 - (-0.2)/\sqrt{.033} = 1.92$. This would exceed the critical value of 1.645 under a one-sided test (which we would use for example if our initial alternative hypothesis was that positive autocorrelation existed), and would be slightly less than the critical value of 1.96 in a two-sided test. We note, however, that we are on shaky ground in assuming that this test statistic has a normal distribution, since the number of regions is small. We also note that, in this case, the approximation of $1/(3n)$ described in Section 10.3.3 for the variance of I would have yielded a variance of $1/18 = .0555$, which is not too far from that found above using Equation 10.21. The approximation of two divided by the sum of the weights, also described in Section 10.3.3, would have yielded $2/18 = .1111$. This approximation works better for systems with a greater number of regions.

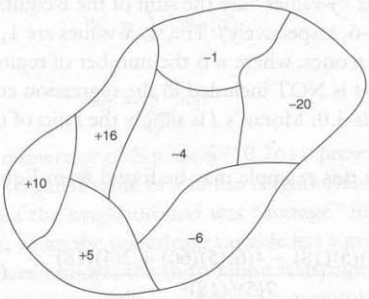
EXERCISES

1. The following residuals are observed in a regression of wheat yields on precipitation and temperature over a six-county area:

County:	1	2	3	4	5	6	
+	7	10	12	9	14	15	Number of Positive Residuals
−	12	8	19	10	10	10	Number of Negative Residuals

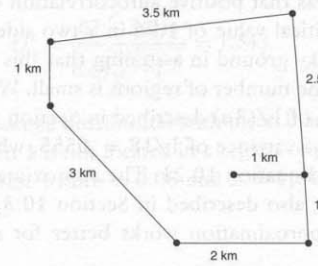
Use the chi-square test to determine whether there is any interaction between location and the tendency of residuals to be positive or negative. If you reject the null hypothesis of no pattern, then describe how you might proceed in the regression analysis.

2. A regression of sales on income and education leaves the following residuals:



Use Moran's I to determine whether there is a spatial pattern to the residuals. If you reject the null hypothesis, describe how you would proceed with the regression analysis.

3. (a) Find the nearest neighbor statistic for the following pattern:



(b) Test the null hypothesis that the pattern is random by finding the z-statistic:

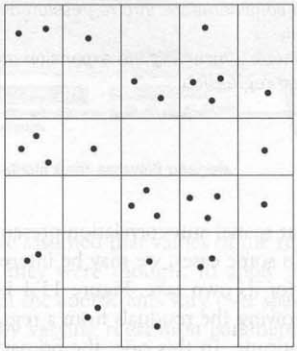
$$z = 1.913(R-1)\sqrt{n}.$$

(c) Find the chi-square statistic, $\chi^2 = (m-1)\sigma^2/\bar{x}$ for a set of 81 quadrats, where 1/3 of the quadrats have 0 points, 1/3 of the quadrats have 1 point, and 1/3 of the quadrats have 2 points. Then find the z-value to test the hypothesis of randomness, where

$$z = \frac{\chi^2 - (m-1)}{\sqrt{2(m-1)}},$$

where m is the number of cells. Compare it with a critical value of $z = -1.96$ and $z = +1.96$.

4. Vacant land parcels are found at the following locations:



Find the variance and mean of the number of vacant parcels per cell, and use the variance-mean ratio to test the hypothesis that parcels are distributed randomly (against the two-tailed hypothesis that they are not).

5. Find the nearest neighbor statistic (the ratio of observed to expected mean distances to nearest neighbors) when n points are equidistant from one another on the circumference of a circle with radius r, and there is one additional point located at the center of the circle. Hints: The area of a circle is πr^2 and the circumference of a circle is $2\pi r$.
6. For the nearest neighbor test, prove that the following two z-scores are equivalent:

$$\frac{R-1}{\sigma_R} = \frac{r_o - r_e}{\sigma_r},$$

where

$$\sigma_R = .52/\sqrt{n}; \sigma_r = \frac{.26}{\sqrt{n\rho}} R = r_o/r_e$$

Thus there are two, equivalent ways of carrying out the nearest-neighbor test.

7. Find the nearest neighbor statistic for four points located at the vertices of a rectangle of length 5 and width 4.