# CHAPTER 3

# PATTERN DETECTORS

The descriptive spatial statistics introduced in the previous chapter are useful in summarizing point distributions and in making comparisons between point distributions with similar attributes. But the use of these descriptive statistics is only the first step in geographic analysis. The next step is to analyze a point distribution to see if there is any recognizable pattern. This step requires additional tools, such as those to be discussed in this chapter.

Every point distribution is the result of some processes at a given point in time and space. To fully understand the various states and the dynamics of particular geographic phenomena, analysts need to be able to detect spatial patterns from the point distributions. This is because successful formulation or structuring of spatial processes often depends to a great extent on the ability to detect changes between point patterns at different times or changes between point patterns with similar characteristics. The recognition and measurement of patterns from point distributions is therefore a very important step in analyzing geographic information.

In considering how cities distribute over a region, one can easily find situations in which cities distribute unevenly over space. This is because the landscape, transportation network, natural resources, and possibly many other factors might have influenced the decision to choose different locations for the settlements to start with and the different growth rates these settlements might have had afterward.

At a global or continental scale, cities are often represented as points on a map. At a local scale, incidences of disease or crime in a city or incidences of fire in a forest may be plotted similarly. When studying point distributions such as these, analysts may try to relate them to particular patterns based on their experience or knowledge generated from previous studies, particularly studies that have developed theories and models. One example is to examine how cities distribute in a region, resembling the theoretical pattern (hexagonal pattern) of city distribution under the Central Place Theory. To do so, analysts would need tools that do more than summarize the statistical properties of point distributions.

The techniques that we discuss in this chapter, although having limitations, are useful in detecting spatial patterns in point distributions (Getis and Boots, 1988). We will first introduce Quadrat Analysis, which allows analysts to determine if a point distribution is similar to a random pattern. Next, the nearest neighbor analysis compares the average distance between nearest neighbors in a point distribution to that of a theoretical pattern (quite often a random pattern). Finally, the spatial autocorrelation coefficient measures how similar or dissimilar an attribute of neighboring points is.

## 3.1 SCALE, EXTENT, AND PROJECTION

The methods to be discussed in this chapter are used mainly to detect or measure spatial patterns for point distributions. It is necessary to pay special attention to three critical issues when using these methods.

First, we need to choose a proper geographic scale to work with when using points to represent some geographic objects. This is because geographic objects may be represented differently at different scales, depending on how they are treated. Whether to hold the scale constant in a study of certain geographic objects or to allow the scale to be changed is a rather important issue to consider when working with sets of points scattering over space.

As pointed out earlier, cities are often represented by points at a global or continental scale. The City of Cleveland, for example, appears to be only a point when a map shows it with other major cities in the United States. The same city, however, becomes a polygonal object that occupies an entire sheet of map when a larger-scale map shows the city with all its streets, rivers, and other details. Similarly, a river may be represented in a small-scale map as a linear feature, but it is an ecosystem if an analyst's focus is on its water, riverbeds, banks, and all of its biological aspects.

The second issue is the extent of geographic areas in the study. Analysts often need to determine to what extent the areas surrounding the geographic objects of interest are to be included in their analysis. Let's assume that we are working on a study that examines intercity activities including Akron, Cincinnati, Cleveland, Columbus, and Dayton in Ohio. When only the geographic extent of the state of Ohio is used, the five cities seem to scatter quite far apart from each other, as shown in Figure 3.1. However, these five cities would seem to be very closely clustered if we define the entire United States to be the study area with respect to them. To increase this difference further, the five cities essentially cluster nearly at one location if they are considered from the perspective of the entire world.

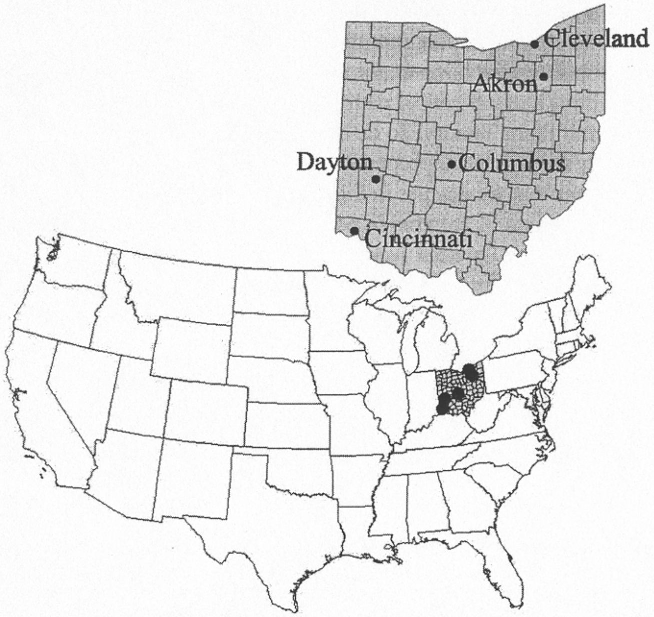


**Figure 3.1** Comparative clusterness.

Delimiting the study area properly for a project is never easy. There are cases in which political boundaries are appropriate choices, but there are also cases in which they are not. While no single, simple solution can be offered for all cases, analysts are urged to consider this issue carefully.

The last issue is the projection used in maps when displaying the distribution of geographic objects. Four possible distortions may be caused by different projections: *area*, *shape*, *direction* and *distance*. Among the more than 20 different projections that can be applied in mapping, there is no projection that can perfectly transform geographic locations from their locations on the globe (a three-dimensional space) to a map (a two-dimensional plane). Therefore, we will need to be sensitive to the needs of the project or study at hand. Careful consideration of the purposes, scales, and accuracy of data is needed for a successful study.

In detecting point patterns, it is important to consider the impact of different projections. This is because both area and distance are used intensively in the analysis of point patterns. In Quadrat Analysis, the size of the study area affects the density of points. In the nearest neighbor analysis and the spatial autocorrelation coefficient, distances between points play a critical role in the calculation of these two indices.

Not surprisingly, the larger the study area, the more significant the impact of different projections will be. If the study area is small, such as a residential neighborhood or a small village/city the different areal and distance measurements by different projections may not matter. However, studies that encompass the entire

United States or the world have to pay special attention to selecting proper projections to work with.

**ArcView Notes**

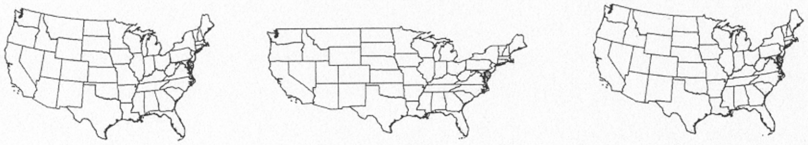
Two types of projections are most popular when working with data concerning the United States. They are known as *State Plane* projections and *Universal Mercator Projections* (*UTM*). When working with projects at local scales, e.g., one or several counties, analysts often use the State Plane projection to display their data in **View** documents or in maps created in **Layout** documents. For projects dealing with an entire state, UTM is often a better choice. In projects dealing with data at a global or continental scale, the choice of projections will depend on where the areas are located and the purpose of the analysis.

As an example, one can see that the shape of the mainland United States changes between projections in Figure 3.2a. Consequently, these different shapes of the United States will affect the measurements of area and distance between places. To use proper projections in a study, factors such as areal extent, the purpose of the study, and the location of the study area need to be assessed carefully.

Compared to the entire United States, the changes caused by changing projections in the case of the State of Ohio are much less significant. The projections showed in Figure 3.2b display few changes compared to the extent of the area of concern.

Although ArcView can be specified to manage projected data, we recommend adding unprojected data (i.e., the data are in the latitude-longitude geo-referencing system) as themes to View documents if possible. This is because ArcView provides more flexibility in allowing its users to change a View document's projection if the themes are added as unprojected data. To do this,

1. access the menu item **View** and the **Properties** from the drop-down menu list;
2. specify the proper data unit and map unit based on your data;
3. use the **Projection** button to access the dialog box for changing projections;



Equal-Area Projection Un-Projected Equal-Distance Projection

**Figure 3.2a** Examples of distortions caused by changing projections.

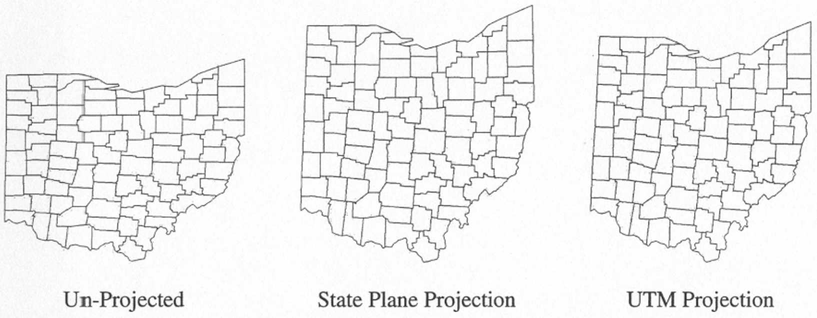


**Figure 3.2b** Examples of distortions caused by changing projections.

4. choose the projection category based on the geographic extent of your data and then, as required, choose the projection zone for your study area.

## 3.2 QUADRAT ANALYSIS

The first method for detecting spatial patterns from a point distribution is *Quadrat Analysis*. This method evaluates a point distribution by examining how its density changes over space. The density, as measured by Quadrat Analysis, is then compared with usually, but not necessarily a theoretically constructed random pattern to see if the point distribution in question is more clustered or more dispersed than the random pattern.

The concept and procedures of Quadrat Analysis are relatively straightforward. First, the study area is overlain with a regular square grid, and the number of points falling in each of the squares are counted. Some squares will contain none of the points, but other squares will contain one, two, or more points. With all squares counted, a frequency distribution of the number squares with given number of points can be constructed. Quadrat Analysis compares this frequency distribution with that of a known pattern, such as a theoretically random pattern.

The squares are referred to as *quadrats*, but quadrats do not always need to be squares. Analysts can use other geometric forms, such as circles or hexagons, as appropriate for the geographic phenomenon being studied. The selection among various forms of quadrats should be based on previous successful experience or the characteristics of the phenomenon in question. In addition, within each analysis, the shape and size of the quadrats have to be constant.

In considering an extremely clustered point pattern, one would expect all or most of the points to fall inside one or a very few squares. On the other hand, in an extremely dispersed pattern, sometimes referred to as a *uniform* pattern, one would expect all squares to contain relatively similar numbers of points. As a result, analysts can determine if the point distribution under study is closer to a clustered, random, or dispersed pattern. As examples of clustered and dispersed patterns, Figure 3.3 shows the 164 cities in the state of Ohio, along with hypothetical clustered and dispersed patterns with the same number of points.

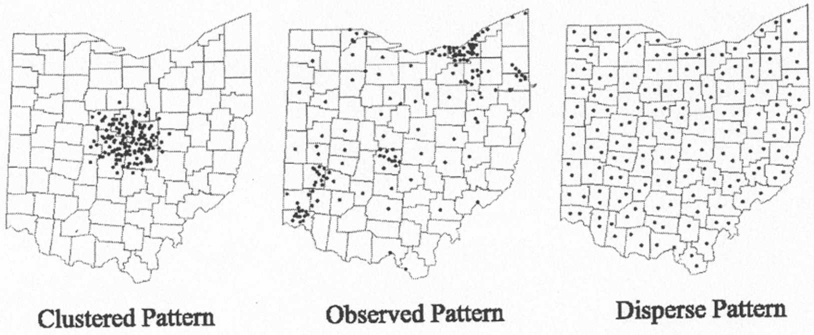


**Figure 3.3** Ohio cities with hypothetical clustered and cispersed patterns.

Overlaying the study area with a regular grid partitions the study area in a systematic manner to avoid over- or undersampling of the points anywhere. Since Quadrat Analysis evaluates changes in density over space, it is important to keep the sampling interval uniform across the study area. There is, however, another way to achieve the same effect. This involves randomly placing quadrats of a fixed size over the study area (Figure 3.4). Statistically, Quadrat Analysis will achieve a fair evaluation of the density across the study area if it applies a large enough number of randomly generated quadrats.

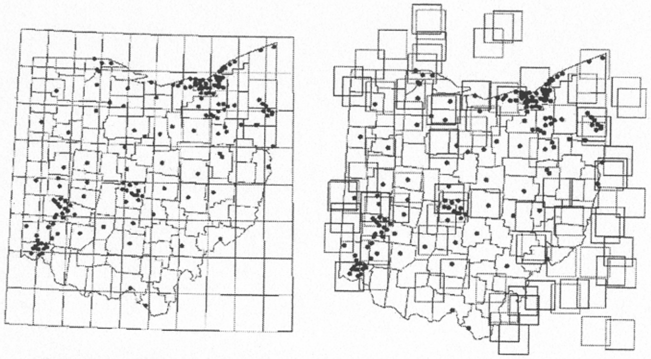
**ArcView Notes**

In the accompanying project file, `Ch3.apr`, three types of quadrats are used. Users can choose to use squares, hexagons, or circles. In addition, users can choose between placing a regular grid over the study area or throwing randomly generated quadrats over it.

Similar to `Ch1.apr` and `Ch2.apr`, `Ch3.apr` can be found on the companion website to this book. Use ArcView to open this project file before adding data themes to the View document.

The last issue that needs careful consideration when applying Quadrat Analysis is the size of quadrats. According to the Greig-Smith experiment (Greig-Smith, 1952) and the subsequent discussion by Taylor (1977, pp. 146–147) and Griffith and Amrhein (1991, p. 131), an optimal quadrat size can be calculated as follows:

$$\text{Quadrat size} = \frac{2 \cdot A}{n},$$



Complete Coverage with Regular Grid

Random Quadrats

**Figure 3.4** Quadrat analysis: complete coverage by grids and random quadrats.

where $A$ is the area of the study area and $n$ is the number of points in the distribution. This suggests that an appropriate square size has a width of $\sqrt{2 \cdot A/n}$.

Once the quadrat size for a point distribution is determined, Quadrat Analysis can proceed to establish the frequency distribution of the number of points in each quadrat, using either complete coverage (with a regular square grid or a hexagon grid) or randomly generated quadrats. This frequency distribution needs to be compared to a frequency distribution describing a random point pattern.

A statistical test known as the *Kolmogorov-Simirnov test* (or simply the *K-S test*), can be used to test the difference statistically between an observed frequency distribution and a theoretical frequency distribution. This test is a simple, straightforward test, both conceptually and computationally.

As an example, let's take the 164 Ohio cities and use 80 squares to construct the frequency distribution for Quadrat Analysis. The frequency distribution of the cities falling into squares is listed in Table 3.1. In this table, the left-hand column lists the number of cities in each square. The second column shows that there are 36 squares with no city at all, 17 squares with only one city, 10 squares with two cities, and so on. For an example of a uniform/dispersed pattern, the third column lists frequencies that are made up to approximate an even distribution of cities across all squares. The right-hand column indicates that all cities are located within one square.

By observing the differences among the three frequency distribution columns, it is clear that the observed pattern is more clustered than the dispersed pattern. But it is not as clustered as that of the right-hand column. While the differences among the columns can be visually estimated, we need some way of measuring the difference quantitatively. At this stage of our analysis, we can apply the K-S test.

**TABLE 3.1 Frequency Distribution: 164 Ohio Cities with Three Hypothetical Frequency Distributions**

| Number of Cities in Each Square | Observed Pattern | Uniform/Dispersed | Clustered |
|---|---|---|---|
| 0 | 36 | 0 | 79 |
| 1 | 17 | 26 | 0 |
| 2 | 10 | 26 | 0 |
| 3 | 3 | 26 | 0 |
| 4 | 2 | 2 | 0 |
| 5 | 2 | 0 | 0 |
| 6 | 1 | 0 | 0 |
| 7 | 1 | 0 | 0 |
| 8 | 1 | 0 | 0 |
| 9 | 1 | 0 | 0 |
| 10 | 1 | 0 | 0 |
| 11 | 1 | 0 | 0 |
| 12 | 1 | 0 | 0 |
| 13 | 1 | 0 | 0 |
| 14 | 1 | 0 | 0 |
| 28 | 1 | 0 | 0 |
| 164 | 0 | 0 | 1 |

The K-S test allows us to test a pair of frequency distributions at a time. Let's take the observed pattern and the dispersed pattern to start with. First, we assume that the two frequency distributions are similar enough that we cannot detect differences between them that are statistically significant. This concept is slightly confusing for those who have not had much experience with statistical analysis, but it is simply making an assumption that allows a small degree of difference to be acceptable. If the difference between two frequency distributions is indeed very small, then the difference might have happened simply by chance. The larger the difference, the less likely that it occurred by chance.

The test is as follows:

1. Assume that there is no statistically significant difference between two frequency distributions.
2. Decide on a level of statistical significance—for example, allowing only 5 out of 100 times ($\alpha = 0.05$).
3. Convert all frequencies to cumulative proportions in both distributions.
4. Calculate the $D$ statistic for the K-S test:

$$D = \max |O_i - E_i|,$$

where $O_i$ and $E_i$ are cumulative proportions of the $i$th category in the two distributions. The max || term indicates that we do not care which one is larger than the other; we are concerned only with their difference. $D$ is then the maximum of the absolute differences among all pairwise comparisons.

5. Calculate a critical value as the basis for comparison:

$$D_{\alpha=0.05} = \frac{1.36}{\sqrt{m}},$$

where $m$ is the number of quadrats (or observations).

In a 2-sample case,

$$D_{\alpha=0.05} = 1.36\sqrt{\frac{m_1+m_2}{m_1 m_2}}$$

where $m_1$ and $m_2$ are numbers of quadrats in the 2 groups.

6. If the calculated $D$ is greater than the critical value of $D_{\alpha=0.05}$, we will conclude that the two distributions are significantly different in a statistical sense.

Taking the example of the 164 Ohio cities, Table 3.2 lists the frequencies and their converted proportions. The right-hand column lists the absolute differences between the two columns of cumulative proportions. The largest absolute difference, in this case, is 0.45. Therefore, $D = \max |O_i - E_i| = 0.45$.

Because this is a 2-sample case, the critical value $D(\alpha = 0.05)$ can be calculated as follows:

$$D(\alpha = 0.05) = 1.36\sqrt{\frac{80+80}{80 * 80}} = 0.215.$$

**TABLE 3.2** ***D* Statistics for K-S Test**

| Number of Cities in Each Square | Observed Pattern | Cumulative Observed Proportions | Dispersed Pattern | Cumulative Pattern Proportions | Absolute Difference |
|---|---|---|---|---|---|
| 0 | 36 | 0.45 | 0 | 0 | 0.45 |
| 1 | 17 | 0.66 | 26 | 0.325 | 0.34 |
| 2 | 10 | 0.79 | 26 | 0.65 | 0.14 |
| 3 | 3 | 0.83 | 26 | 0.975 | 0.15 |
| 4 | 2 | 0.85 | 2 | 1 | 0.15 |
| 5 | 2 | 0.88 | 0 | 1 | 0.13 |
| 6 | 1 | 0.89 | 0 | 1 | 0.11 |
| 7 | 1 | 0.90 | 0 | 1 | 0.10 |
| 8 | 1 | 0.91 | 0 | 1 | 0.09 |
| 9 | 1 | 0.93 | 0 | 1 | 0.08 |
| 10 | 1 | 0.94 | 0 | 1 | 0.06 |
| 11 | 1 | 0.95 | 0 | 1 | 0.05 |
| 12 | 1 | 0.96 | 0 | 1 | 0.04 |
| 13 | 1 | 0.98 | 0 | 1 | 0.03 |
| 14 | 1 | 0.99 | 0 | 1 | 0.01 |
| 28 | 1 | 1 | 0 | 1 | 0 |

The critical value of 0.215 is apparently far smaller than 0.45, indicating that the difference between two frequency distributions is statistically significant at the $\alpha = 0.05$ level. With this, we can easily reject our initial hypothesis that there is no significant difference between a dispersed pattern of 164 points and the distribution formed by the 164 Ohio cities. In other words, the 164 Ohio cities do not distribute in a dispersed manner.

In the above example, we examined the difference between an observed point pattern and a dispersed pattern. However, it is more common to compare an observed point pattern to a point pattern generated by a random process. A well-documented process for generating a random point pattern is the Poisson process. The Poisson random process is appropriate to generate count data or frequency distributions. Quadrat Analysis often compares an observed point pattern frequency distribution to a frequency distribution generated by the Poisson random process.

A Poisson distribution is strongly determined by the average number of occurrences, $\lambda$. In the context of Quadrat Analysis, $\lambda$ is defined as the average number of points in a quadrat. Assume that we have $m$ quadrats and $n$ points in the entire study area, $\lambda = n/m$, the average number of points in a quadrat. Let $x$ be the number of points in a quadrat. Using the Poisson distribution, the probability of having $x$ points in a quadrat is defined as

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where $e$ is the Euler's 2.71828 constant and $x!$ is the factorial of $x$, which can be defined as $x(x-1)(x-2)\ldots(1)$ and 0!, by definition, is 1. Using the probabilities for various values of $x$ based upon the Poisson distribution, we can generate a frequency distribution in the same format as those shown in Table 3.1 but for a random point distribution.

Generating a probability value from a Poisson distribution is rather simple. But if a set of probabilities for a range of $x$ values is required, it becomes quite tedious, as the factorial and the $e$ function have to be applied every time. Fortunately, there is a shortcut that can be used to generate a set of probabilities based upon the Poisson distribution. We know that if $x = 0$, the Poisson probability is reduced to

$$p(0) = e^{-\lambda}.$$

Other probabilities can be derived based upon $p(0)$. In general,

$$p(x) = p(x-1) * \frac{\lambda}{x}.$$

If $x$ is 1, then $p(x-1) = p(0)$. Using this shortcut formula, it would be efficient to derive an entire set of Poisson probabilities.

In the Ohio example, there are 164 points (cities) and 80 quadrats were used in the previous example. Therefore, $\lambda = 164/80 = 2.05$. Using this $\lambda$ value, a set of

**TABLE 3.3 Comparing the 164 Ohio Cities to a Random Pattern Generated by a Poisson Process**

| Number of Cities in Each Square | Observed Pattern | Observed Probability | Cumulative Observed Probability | Poisson Probability | Cumulative Poisson Probability | Difference |
|---|---|---|---|---|---|---|
| 0 | 36 | 0.45 | 0.4500 | 0.1287 | 0.1287 | 0.3213 |
| 1 | 17 | 0.2125 | 0.5514 | 0.2639 | 0.3926 | 0.2699 |
| 2 | 10 | 0.1125 | 0.7875 | 0.2705 | 0.6631 | 0.1244 |
| 3 | 3 | 0.0375 | 0.8250 | 0.1848 | 0.8480 | 0.0230 |
| 4 | 2 | 0.025 | 0.8500 | 0.0947 | 0.9427 | 0.0927 |
| 5 | 2 | 0.025 | 0.8750 | 0.0388 | 0.9816 | 0.1066 |
| 6 | 1 | 0.0125 | 0.8875 | 0.0133 | 0.9948 | 0.1073 |
| 7 | 1 | 0.0125 | 0.9000 | 0.0039 | 0.9987 | 0.0987 |
| 8 | 1 | 0.0125 | 0.9125 | 0.0010 | 0.9997 | 0.0872 |
| 9 | 1 | 0.0125 | 0.9241 | 0.0002 | 0.9999 | 0.0759 |
| 10 | 1 | 0.0125 | 0.9375 | 0.0000 | 1 | 0.0625 |
| 11 | 1 | 0.0125 | 0.9500 | 0.0000 | 1 | 0.0500 |
| 12 | 1 | 0.0125 | 0.9625 | 0.0000 | 1 | 0.0375 |
| 13 | 1 | 0.0125 | 0.9750 | 0.0000 | 1 | 0.0250 |
| 14 | 1 | 0.0125 | 0.9875 | 0.0000 | 1 | 0.0125 |
| 28 | 1 | 0.0125 | 1 | 0.0000 | 1 | 0.0000 |

Poisson probabilities can be derived using the shortcut formula. Table 3.3 shows the derivations of the probabilities. The first two columns in the table are identical to those in Table 3.1. The third and fourth columns are the observed probabilities and the cumulative probabilities based upon the observed pattern. In the fifth column, a Poisson distribution was generated based upon $\lambda = 2.05$. This probability distribution indicates the probability that a quadrat may receive different numbers of points. The cumulative probabilities of the Poisson distribution are also derived in the sixth column. The last column reports the absolute differences between the two sets of cumulative probabilities. The largest of these differences is the K-S $D$ statistic, which is 0.3213, much greater than the critical value of 0.1520 using the 0.05 level of significance.

If the observed pattern is compared to a random pattern generated by a Poisson process, we can exploit a statistical property of the Poisson distribution to test the difference between the observed pattern and the random pattern in addition to the K-S statistic. This additional test is based upon the variance and mean statistics of a Poisson distribution.

We know that $\lambda$ is the mean of a Poisson distribution. A very interesting and useful property of a Poisson distribution is that the variance is also $\lambda$. In other words, if a distribution potentially is generated by a random process like the Poisson process, the distribution should have the same mean and variance. If the mean and variance form a ratio, the variance-mean ratio, the ratio should be very close to 1. Therefore, given an observed point pattern and the frequency distribution of points by quadrats, we can compare the observed variance-mean ratio to 1 to see if they are significantly different. Given the information similar to that in Table 3.1, the mean of the observed distribution is $\lambda = 2.05$. The variance is basically a variance for grouped data. The mean $\lambda$ has to be compared with the number of points in each quadrat. The difference is then squared and multiplied by the quadrat frequencies. The sum of these products divided by the number of quadrats produces the observed variance. Using the observed variance and $\lambda$, we can form a variance-mean ratio. This ratio is then compared to 1, and the difference has to be standardized by the standard error in order to determine if the standardized score of the difference is larger than the critical value (quite often 1.96 at the 0.05 level of significance).

The K-S test and the variance-mean ratio may yield inconsistent results. But because the K-S test is based upon weak-ordered data (Taylor, 1977) while variance-mean ratio is based upon an interval scale, the variance-mean ratio tends to be a stronger test. However, variance-mean ratio test can be used only if a Poisson process is expected.

Quadrat Analysis is useful in comparing an observed point pattern with a random pattern. Theoretically, we can compare the observed pattern with any pattern of known characteristics. For instance, after we compare the observed pattern with a random pattern and the result indicates that they are significantly different, the next logical step is to test if the observed pattern is similar to a clustering pattern or a dispersed pattern. Quite often, through visual inspection, the analyst can hypothesize what pattern the observed pattern resembles. Using other statistical distributions, such as the negative gamma or the negative binomial, we can generate point patterns with specific distribution properties. These patterns can then be compared with the observed pattern to see if they are different. Quadrat Analysis, however, suffers from certain limitations. The analysis captures information on points within the quadrats, but no information on points between quadrats is used in the analysis. As a result, Quadrat Analysis may be insufficient to distinguish between certain point patterns. Figure 3.5 is an example.

In Figures 3.5a and 3.5b, both spatial configurations have eight points with four quadrats. Visually, the two point patterns are different. Figure 3.5a is a more dispersed pattern, while Figure 3.5b is definitely a cluster pattern. Using quadrat

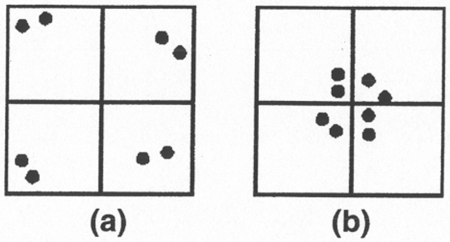


**Figure 3.5** Local clusters with regional dispersion.

analysis, however, the two patterns yield the same result. In order to distinguish patterns depicted in Figure 3.5, we have to use Nearest Neighbor Analysis.

**ArcView Notes**

In `Ch3.apr`, Quadrat Analysis is implemented as a menu item accessible from the View document. Similar to other project files on the companion website to this book, `Ch3.apr` needs to be opened before adding data themes for analysis.

Using the 164 Ohio cities as an example, `Ch3.apr` can be used to determine if that distribution is more clustered or more dispersed:

1. Download `Ch3.apr` from the companion website, and open it in ArcView.
2. In the View document, add two themes for Chapter 3 obtained from the website. These two themes are named `ohcities.shp` and `ohcounty.shp`. After they are added, change their display order, if necessary, so that the `ohcities.shp` is at the top of the Table of Contents in the View document.
3. Access the **View/Properties** menu item from the View document's menu. In the **View Properties** window, click the **Projection** button. In the **Projection Properties** window, change the projection **Category** to `UTM-1983` and the projection **Type** to `Zone 17`. Click **OK** to return to the **View Properties** window. Click **OK** again to return to the View document.
4. There are 88 counties in the state of Ohio. Let's use approximately this many quadrats in our analysis since these counties are fairly similar in size. To start Quadrat Analysis, choose **Quadrat Analysis** from the drop-down menu of **Point Patterns**.
5. In the **Coverage Choice** window, choose **complete** and then click the **OK** button. Notice that the other choice is **random**, which is used when Quadrat Analysis is carried out using randomly generated quadrats.
6. In the **Shape Choice** window, choose **square** and then click the **OK** button. Similar to the previous step, two other options available. They are **hexagons** and **circles** (only for random coverage) in the drop-down list.
7. In the next **Input** window, we see that the number of points is 164. Enter **81** as the preferred number of quadrats.
8. Click the **OK** button for the next three windows as ArcView informs you of the length of the side of the quadrat (47,690.7 meters in our case), 9 vertical steps, and 9 horizontal steps. This is because the 81 quadrats form a 9 by 9 grid.
9. In the **Theme source FileName window**, navigate to your working directory and save the quadrat arrangement.
10. The first statistic is the **Lambda**, which is the average number of points per quadrat (in our case, 2.02469).
11. Next is the frequency, which is followed by a measurement of variance and then the **variance-mean ratio**.
12. In the **K-S D Statistic** window, we have, in our case, 0.361868.
13. For the level of **Alpha** (statistical significance), choose 0.05 and then click the **OK** button.
14. The calculated critical value is given as 0.1511 in our case. When this value is compared to 0.361868, the difference is statistically significant.

The resulting **View** document can be seen in Figure 3.6. The procedures we just went through placed a 9 by 9 grid over the study area.

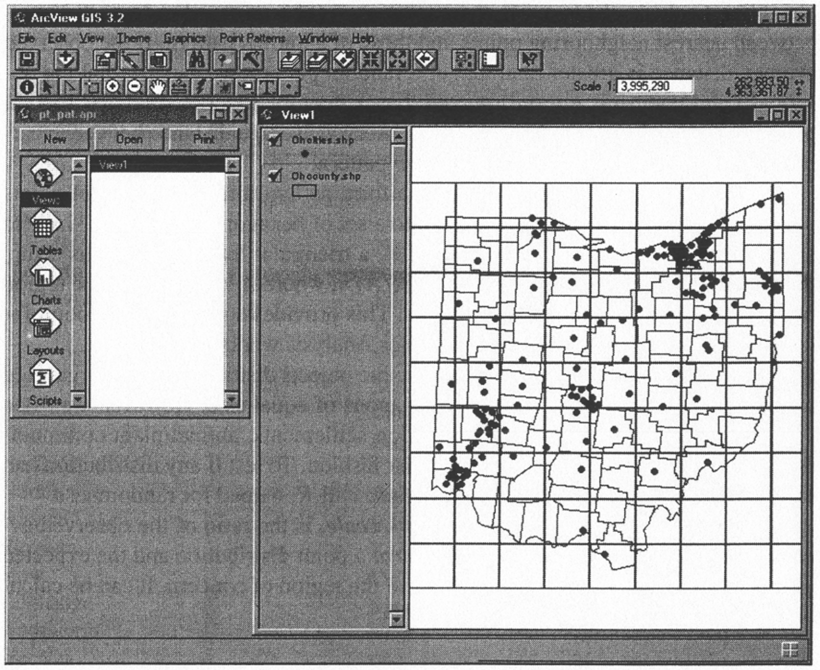


**Figure 3.6** Quadrat analysis of 164 Ohio cities with a 9 by 9 grid.

As an alternative, it is possible to run the same project file for random quadrats. To do so, choose **random** in the **Coverage Choice** window. When asked for the number of quadrats preferred, give **81**. For the width of the square, give 25 (miles) since most counties are about this size.

Finally, the number of quadrats may be increased to ensure that the resulting frequency distribution in Quadrat Analysis approaches a normal frequency distribution. There is no number that is optimal in all cases, but the number of quadrats should be chosen based on the size of the study area, density of points, and computational time and resources.

## 3.3 NEAREST NEIGHBOR ANALYSIS

Quadrat Analysis tests a point distribution with the *points per area* (*density*) concept. The method to be discussed in this section, the *Nearest Neighbor Analysis*, uses the opposite concept of *area per point* (*spacing*). Quadrat Analysis examines how densities in a point distribution change over space so that the point pattern can be compared with a theoretically constructed random pattern. For the Nearest Neighbor Analysis, the test is based on comparing the observed average distances between nearest neighboring points and those of a known pattern. If the observed average distance is greater than that of a random pattern, we can say that the observed point pattern is more *dispersed* than a random pattern. Similarly, a point pattern is said to be more *clustered* if its observed average distance between nearest neighbors is less than that of a *random* pattern.

In a homogeneous region, the most uniform pattern formed by a set of points occurs when this region is partitioned into a set of hexagons of identical size and each hexagon has a point at its center (i.e., a triangular lattice). With this setup, the distance between points will be $1.075\sqrt{A/n}$, where $A$ is the area of the region of concern and $n$ is the number of points. This provides a good starting point for us to understand how the Nearest Neighbor Analysis works.

In the real world, we rarely see geographic objects distributing in an organized manner such as being partitioned by hexagons of equal size. However, we often see geographic objects, such as population settlements, animal/plant communities, or others distribute in a more irregular fashion. To test if any distribution had any recognizable patterns, let's use a statistic call $R$, named for randomness.

The $R$ statistic, sometimes called the $R$ *scale*, is the ratio of the observed average distance between nearest neighbors of a point distribution and the expected distance of the average nearest neighbor of the region of concern. It can be calculated as follows:

$$\mathrm{R} = \frac{r_{\text{obs}}}{r_{\text{exp}}},$$

where $r_{\text{obs}}$ is the observed average distance between nearest neighbors and $r_{\text{exp}}$ is the expected average distance between nearest neighbors as determined by the theoretical pattern being tested.

To measure the observed average distance between nearest neighbors, we can calculate the distance between each point and all of its neighbor points. The shortest distance among these neighbors will be associated with the nearest point. When this process is repeated for all points, a table such as Table 3.4 can be calculated. The points in Table 3.4 are based on the distribution shown in Figure 3.7. For this set of 17 cities, the observed average distance between pairs of nearest neighbors is $r_{\text{obs}} = 6.35$ miles.

For the theoretical random pattern, let's use the following equation to calculate the expected average distance between nearest neighbors:

$$r_{\text{exp}} = \frac{1}{2\sqrt{n/A}},$$

where the $n$ is the number of points in the distribution and the $A$ is the area of the space of concern. In our example, the area of the five counties is 3,728 square miles. Therefore, the expected average distance is

$$r_{\text{exp}} = \frac{1}{2\sqrt{17/3728}} = 7.40.$$

**TABLE 3.4 Observed Distances Between Nearest Neighbor Cities in the Five-County Area of Northeastern Ohio**

| City Name | Nearest City | Nearest Distance |
|---|---|---|
| Akron | Tallmadge | 5.37 |
| Alliance | North Canton | 14.96 |
| Barberton | Norton | 2.39 |
| Brunswick | Medina | 7.93 |
| Canton | North Canton | 4.40 |
| Cuyahoga Falls | Stow | 4.52 |
| Kent | Stow | 4.36 |
| Massillon | Canton | 7.90 |
| Medina | Brunswick | 7.93 |
| North Canton | Canton | 4.40 |
| Norton | Barberton | 2.39 |
| Portage Lakes | Barberton | 4.00 |
| Ravenna | Kent | 6.32 |
| Stow | Cuyahoga Falls | 4.52 |
| Tallmadge | Kent | 4.38 |
| Wadsworth | Norton | 4.52 |
| Wooster | Wadsworth | 17.64 |
| Average nearest distance | | 6.35 (miles) |

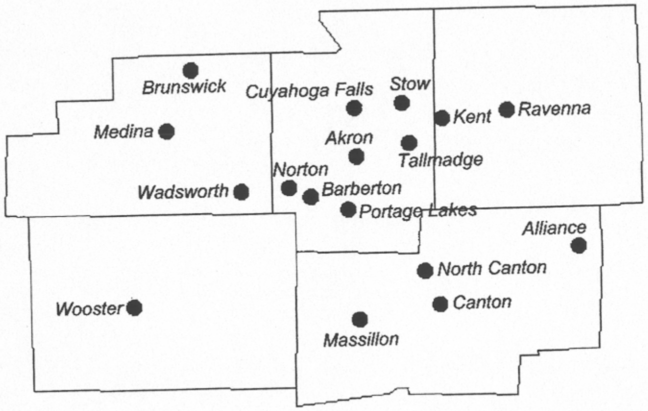


**Figure 3.7** Cities in the five-county area of northeastern Ohio region.

With both distances calculated, we can now compute the $R$ statistic:

$$\mathrm{R} = \frac{r_{\text{obs}}}{r_{\text{exp}}} = \frac{6.35}{7.40} = 0.8581.$$

With this $R$ scale, we know that the point pattern formed by the 17 cities is more clustered than a random pattern.

Now we know how the $R$ statistic is calculated and how to determine if a pattern is more clustered or more dispersed than a random pattern. Many conclusions can be drawn from this calculation with regard to how the 17 cities relate to each other. But we are still not sure to what degree this pattern is more clustered than other. Is it much more clustered or just slightly more clustered? To appreciate the implications of various values of the $R$ statistic, Figure 3.8 shows a series of hypothetical distributions and their associated $R$ values.

Figure 3.8 shows that the more clustered patterns are associated with smaller $R$ values ($r_{\text{obs}} < r_{\text{exp}}$) while the more dispersed patterns are associated with larger

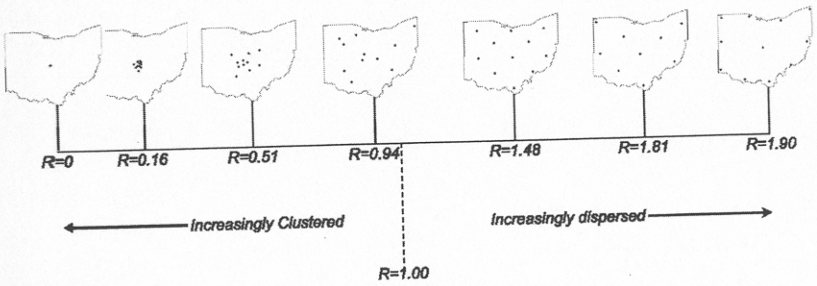


**Figure 3.8** The scale of $R$ statistics.

$R$ values ($r_{\text{obs}} > r_{\text{exp}}$). Figure 3.8 is useful for establishing a general concept of how $R$ values relate to various patterns. It is, however, not sufficient to measure quantitatively the difference between an observed pattern and a random pattern.

The $R$ scale ranges from $R = 0$ (completely clustered) to $R = 1$ (random) to $R = 2.149$ (completely dispersed). When $R = 0$, all distances between points are zero, indicating that all points are found at the same location. When $R = 1$, $r_{\text{obs}} = r_{\text{exp}}$; the pattern being tested should therefore be a random pattern. When $R$ approximates values or 2 or more, the pattern displays various degrees of dispersion.

When using Nearest Neighbor Analysis, one way to measure the extent to which the observed average distance differs from the expected average distance is to compare this difference with its *standard error* ($SE_r$). The standard error describes the likelihood that any differences occur purely by chance. If the calculated difference is relatively small when compared to its standard error, we say that this difference is not statistically significant. By contrast, when we have a difference that is relatively large with respect to its standard error, we claim that the difference is statistically significant; that is, it does not occur by chance.

The concept of standard error is rooted in classical statistical theories. In a normal distribution, there is about a 68% chance that some differences between one negative standard error and one positive standard error will occur by chance when in fact there should not be any difference between two populations being compared. Described in equation, this means that:

$$\text{Probability}(< 68\%) = (-1SE_r, +1SE_r).$$

Following this, we can define a calculated difference to be statistically significant only when it is smaller than $-1SE_r$ or greater than $+1SEE_r$. Or, if we want to be more rigid, we would call a difference statistically significant only if it is smaller than $-1.96SEE_r$ or greater than $1.96SEE_r$. This is because the probability of having a difference of that magnitude is 5 out of 100 times or less:

$$\text{Probability}(< 95\%) = (-1.96SE_r, +1.96SE_r).$$

To calculate the standard error for the observed distances, we can use the following equation:

$$SE_r = \frac{0.26136}{\sqrt{n^2/A}},$$

where $n$ and $A$ are as defined previously. With this standard error, we can now see how the difference is compared with it by calculating a standardized $Z$ score:

$$Z_R = \frac{r_{\text{obs}} - r_{\text{exp}}}{SE_r}.$$

As mentioned earlier, if $Z_R > 1.96$ or $Z_R < -1.96$, we conclude that the calculated difference between the observed pattern and the random pattern is statistically significant. Alternatively, if $-1.96 < Z_R < 1.96$, we conclude that the observed pattern, although it may look somewhat clustered or somewhat dispersed visually, is not significantly different from a random pattern.

In our example of 17 Ohio cities, we have

$$SE_r = \frac{0.26136}{\sqrt{17^2/3728}} = 0.9387.$$

And the $Z_R$ score is

$$Z_R = \frac{6.35 - 7.4}{0.9387} = \frac{-1.05}{0.9387} = -1.12,$$

meaning that it is not statistically different from a random pattern.

As demonstrated above, a point pattern may seem clustered or dispersed by visual inspection or even described by calculating its $R$ value. However, we will not be able to reach a conclusion unless a statistical test confirms or rejects this conclusion. In other words, the calculated $R$ value should be confirmed by $Z_R$ scoring to ensure its statistical significance.

Please note that in the Ohio example, $Z_R$ is negative, indicating that the nearest neighbor distance of the observed pattern is smaller than expected but insignificant. The sign of the $z$-score, however, indicates that the observed pattern has a clustering tendency. In other words, if the $z$-score indicates that the difference between the observed and expected nearest neighbor distances is statistically significant, the sign of the statistic can show if the observed pattern is probably clustered or dispersed. Following the logic of hypothesis testing, we can conduct a one-tailed test to see if the $z$-score is really negative (smaller than $-1.645$ at the 0.05 significance level) or really positive (greater than 1.645). These tests ultimately can provide a conclusion if the observed pattern is significantly different from a clustering pattern or a dispersed pattern.

With its ability to detect patterns in a point distribution, Nearest Neighbor Analysis has been extended to accommodate second, third, and higher order neighborhood definitions. When two points are not immediate nearest neighbors but are the second nearest neighbors, the way distances between them are computed will need to be adjusted accordingly. The extension is straightforward in concept and has been used on special occasions when this relationship is important.

For instance, using the first order nearest neighbor statistic, we cannot distinguish the two point patterns shown in Figure 3.5a and Figure 3.5b because the nearest neighbors of each point in both patterns are very close. But if the second order nearest neighbors are used in the analysis, the result will show that Figure 3.5a has a dispersed pattern because the second nearest neighbors are all far away in other quadrats. On the other hand, the result for Figure 3.5b will indicate a clustering pattern because all second nearest neighbors are still quite close. By combining the results of the first order and second order nearest neighbor analyses, we can conclude that Figure 3.5a has a local clustering but regional dispersed pattern, while Figure 3.5b has a clustering pattern on both local and regional scales. To a large extent, using different orders of nearest neighbor statistics can detect spatially heterogeneous processes at different spatial scales.

Finally, it should be noted that Nearest Neighbor Analysis has certain problems. When it is used to examine a point distribution, the results are highly sensitive to the geographic scale and the delineation of the study area. A set of cities may be considered very disperse if they are examined at a local scale. These same cities may seem extremely clustered if they are viewed at a continental or global scale. The 17 Ohio cities may seem quite dispersed if we use only the five counties as the study area. However, they are considered to be very clustered if they are plotted on a map that shows the entire United States.

Understanding the limitations of Nearest Neighbor Analysis, we should always be careful to choose an appropriate geographic scale to properly display the geographic objects we study. Furthermore, the delineation of study areas should be justified by meaningful criteria. In many cases political boundaries may make sense, but other situations may require boundaries defined by natural barriers such as coastlines, rivers, or mountain ranges.

**ArcView Notes** Nearest Neighbor Analysis is implemented in the ArcView project file `Ch3.apr` found on the companion website to this book. This project file has a customized menu item called **Point Patterns.** From **Point Patterns,** a drop-down menu provides access to all tools for detecting point patterns, as discussed in this chapter. To use Nearest Neighbor Analysis, bring in data layers by Add Theme and then choose **Point Patterns/Ordered Neighbor Statistics.**

As an example, we will use the 164 cities in Ohio as our data set to describe the procedures needed for calculating the $R$ statistic and its standardized $Z_R$ score:

1. Start ArcView and use **File/Open Project** to open the `Ch3.apr` from `\AVStat\Chapter3\Scripts\`.
2. Click the **Add Theme** button and then navigate to the directory with `Ohcities.shp` and `Ohcounty.shp`. Then click the **OK** button.
3. Change the order of the two theme so that `Ohcities.shp` is at the top.
4. Click `Ohcities.shp` to make it active.
5. Choose **Point Patterns/Ordered Neighbor Statistics** from the menu bar in the **View** document.

6. In the *Set Properties* window, click the **Yes** button to continue.
7. In the *Nearest* window, click **OK** to continue. The default selection of **Nearest** is the first order neighbor relationship. Higher order neighborhoods can be accessed by clicking the appropriate entry from the drop-down list in this window.
8. Note that the calculated Observed Neighbor Distance is 8.51217. Click **OK** to continue.
9. The Expected Neighbor Distance shown in the next window is 11.6421. Click **OK** to continue.
10. The *R* value is shown in the **Nearest Neighbor Statistic:** window to be 0.731156. Click **OK** to continue.
11. The $Z_R$ score is 6.58691, as shown in the next window. Click **OK** to continue.

As a result of this run, we found that the standardized $Z_R$ score is very high, much higher than the 1.96 threshold we discussed earlier. We can conclude that the 164 cities are definitely clustered.

## 3.4 SPATIAL AUTOCORRELATION

In detecting spatial patterns of a point distribution, both Quadrat Analysis and Nearest Neighbor Analysis treat all points in the distribution as if they are all the same. These two methods analyze only the locations of points; they do not distinguish points by their attributes.

In this section, we will discuss a method for detecting spatial patterns of a point distribution by considering both the locations of the points and their attributes. This method uses a measure known as the *spatial autocorrelation coefficient* to measure and test how clustered/dispersed points are in space with respect to their attribute values. This measure is considered to be more powerful and more useful than the two methods discussed previously in certain ways. Different geographic locations rarely have identical characteristics, making it necessary to consider the characteristics of points in addition to their locations. Not only do locations matter; the conditions of these locations or activities happening there are also of great importance.

Spatial autocorrelation of a set of points is concerned with the degree to which points or things happening at these points are similar to other points or phenomena happening there. If significantly positive spatial autocorrelation exists in a point distribution, points with similar characteristics tend to be near each other. Alternatively, if spatial autocorrelation is weak or nonexistent, adjacent points in a distribution tend to have different characteristics. This concept corresponds to what was once called the *first law of geography* (Tobler, 1970): *everything is related to everything else, but near things are more related than distant things* (also cited and discussed in Gould, 1970, pp. 443–444; Cliff and Ord, 1981, p. 8; and Goodchild, 1986, p. 3).

With the spatial autocorrelation coefficient, we can measure

1. The proximity of locations and
2. The similarity of the characteristics of these locations.

For proximity of locations, we calculate the distance between points. For similarity of the characteristics of these locations, we calculate the difference in the attributes of spatially adjacent points.

There are two popular indices for measuring spatial autocorrelation in a point distribution: *Geary's Ratio* and *Moran's I*. Both indices measure spatial autocorrelation for interval or ratio attribute data. Following the notation used in Goodchild (1986, p. 13), we have

$c_{ij}$ representing the similarity of point $i$'s and point $j$'s attributes,

$w_{ij}$ representing the proximity of point $i$'s and point $j$'s locations, with $w_{ii} = 0$ for all points,

$x_i$ represents the value of the attribute of interest for point $i$, and

$n$ represents the number of points in the point distribution.

For measuring spatial autocorrelation, both Geary's Ratio and Moran's I combine the two measures for attribute similarity and location proximity into a single index of $\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}w_{ij}$. It is used as the basis for formulating both indices. In both cases, the spatial autocorrelation coefficient ($SAC$) is proportional to the weighted similarity of attributes of points. Specifically, the equation for spatial autocorrelation coefficient takes the general form

$$SAC \approx \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}w_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}.$$

In the case of Geary's Ratio for spatial autocorrelation, the similarity of attribute values between two points is calculated as

$$c_{ij} = (x_i - x_j)^2.$$

The difference in attribute values for point $i$ and point $j$ is calculated as $(x_i - x_j)$. These differences for all pairs of $i$ and $j$ are then squared before being summed so that positive differences will not be offset by negative differences. Specifically, Geary's Ratio is calculated as follows:

$$C = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}w_{ij}}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\sigma^2} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\sigma^2},$$

where $\sigma^2$ is the variance of the attribute $x$ values with a mean of $\overline{x}$ or

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{(n-1)}.$$

In the case of Moran's I, the similarity of attribute values is defined as the difference between each value and the mean of all attribute values in question. Specifically, for Moran's I,

$$c_{ij} = (x_i - \overline{x})(x_j - \overline{x})$$

and the index can be calculated as

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}c_{ij}}{s^2 \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{s^2 \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}},$$

where $s^2$ is the sample variance or

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}.$$

In Geary's Ratio and Moran's I, all terms can be calculated directly from the attribute values of the points. The only item not yet defined is $w_{ij}$, which is the proximity of locations between point $i$ and point $j$. We often use the inverse of the distance between point $i$ and point $j$. This assumes that attribute values of points follow the first law of geography. With the inverse of the distance, we give smaller weights to points that are far apart and larger weights to points that are closer together. For example, $w_{ij}$ can be defined as $1/d_{ij}$, where $d_{ij}$ is the distance between point $i$ and point $j$.

The two indices are similar in format. The difference between them is whether the differences in attribute values ($x_i$ and $x_j$) are calculated directly ($x_i - x_j$) or via their mean $(x_i - \overline{x})(x_j - \overline{x})$. As a result, the two indices yield different numeric ranges, as shown in Table 3.5. In Table 3.5, possible values for both indices are listed with respect to three possible spatial patterns: clustered, random, and dispersed. Note that neither index's scale corresponds to our conventional impression of correlation coefficient of $(-1, 1)$ scale.

**TABLE 3.5 Numeric Scales for Geary's Index and Moran's Index**

| Spatial Patterns | Geary's C | Moran's I |
|---|---|---|
| Clustered pattern in which adjacent points show similar characteristics | $0 < C < 1$ | $I > E(I)$ |
| Random pattern in which points do not show particular patterns of similarity | $C \sim 1$ | $I \sim= E(I)$ |
| Dispersed/uniform pattern in which adjacent points show different characteristics | $1 < C < 2$ | $I < E(I)$ |

$E(I) = (-1)/(n-1)$, with $n$ denoting the number of points in the distribution.

For Geary's Ratio, a value of 1 approximates the random pattern, whereas values greater than 1 suggest a dispersed or uniform pattern that has adjacent points displaying different characteristics. For Geary's Ratio, a value of less than 1 suggests a clustered pattern in which adjacent points show similar attribute values.

The numeric scale of Moran's I is anchored at the expected value of $E(I) = -1/n - 1$ for a random pattern. Index values that are less than $E(I)$ are typically associated with a uniform/dispersed pattern. At the other end of the scale, index values greater than $E(I)$ typically indicate clustered patterns where adjacent points tend to have similar characteristics.

When analyzing a point distribution, if we assume that the way the attribute values are assigned to the points is only one of the many possible arrangements using the same set of values, we adopt the assumption known as *randomization*, or *nonfree sampling*. Alternatively, we may assume that the attribute values in a set of points are only one of an infinite number of possibilities; each value is independent of others in the set of points. This assumption is sometimes called the *normality* or *free sampling* assumption. The difference between these two assumptions affects the way the variances of Geary's Ratio and Moran's I are estimated.

For both indices, we can calculate variances under free sampling and nonfree sampling assumptions. Free sampling allows replacements of observations in sampling possible outcomes, while nonfree sampling does not allow replacement.

Let's use $R$ for the nonfree sampling assumption (randomization) and $N$ for the free sampling assumption (normality). Following Goodchild (1986), we can estimate the expected values for a random pattern and the variances for Geary's C by

$$E_N(C) = 1$$

$$E_R(C) = 1$$

$$VAR_N(C) = \frac{[(2S_1 + S_2)(n-1) - 4W^2]}{2(n+1)W^2}$$

$$VAR_R(C) = \frac{(n-1)S_1[n^2 - 3n + 3 - (n-1)k]}{n(n-2)(n-3)W^2} - \frac{(n-1)S_2[n^2 + 3n - 6 - (n^2 - n + 2)k]}{4n(n-2)(n-3)W^2} + \frac{W^2[n^2 - 3 - (n-1)^2 k]}{n(n-2)(n-3)W^2},$$

where

$$W = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$$

$$S_1 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij} + w_{ji})^2}{2}$$

$$S_2 = \sum_{i=1}^{n}(w_{i\cdot} + w_{\cdot i})^2$$

$$k = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2}.$$

For Moran's I, the expected index value for a random pattern and the variances are

$$E_N(I) = E_R(I) = \frac{-1}{n-1}$$

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

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

with $W$, $S_1$, $S_2$, and $k$ similarly defined.

Once the expected values and their variances are calculated, the standardized Z scores can be calculated as

$$Z = \frac{I - E(I)}{VAR(I)}$$

or

$$Z = \frac{C - E(C)}{VAR(C)}.$$

Note that the same critical values of $-1.96 < Z < 1.96$ can be applied with a statistical significance level of 5%, or 0.05.

While the calculation of the spatial autocorrelation coefficient is straightforward, the definition of the similarity of locations can have some variations from those defined here. For example, $w_{ij}$ can take a binary form of 1 or 0, depending

on whether point $i$ and point $j$ are spatially adjacent. If two points are spatially adjacent, $w_{ij} = 1$; otherwise, $w_{ij} = 0$. If we use the concept of nodal region in the geography literature, each point in a distribution can be seen as the centroid of a region surrounding it. If the two adjacent regions share a common boundary, the two centroids of their corresponding regions can be defined as spatially adjacent.

Aside from various ways of defining $w_{ij}$, it is also possible to vary how the distance between points is used. For example, rather than defining $w_{ij} = 1/d_{ij}$, one can use $w_{ij} = 1/d_{ij}^b$ where $b$ may take any appropriate value based on specific characteristics or empirical evidence associated with the geographic phenomena in question. This is because the distances measured by driving a car between two places can have quite different meaning from the distances measured by flying between two places or making phone calls between them. Many empirical studies indicated that $b = 2$ is widely applicable to many geographic phenomena.

Finally, it should be noted that spatial autocorrelation coefficients discussed here are also used for calculating similarity among polygon objects. We will discuss those uses in more detail in Chapter 5.

**ArcView Notes**

The use of Geary's Ratio and Moran's I for a set of points in ArcView involves intensive computational time and a large amount of storage space on computer disks. This is because of the need to first calculate the distances between all pairs of points in the point theme. Depending on the hardware configuration and the amount of data, the computation will vary in the time it takes to finish calculating the distances. When working with a large data set, users should expect longer computational time than when working with a small data set.

The use of the spatial autocorrelation coefficient here is also included in the project file, `Ch3.apr`. Using the data set of the 17 northeast Ohio cities, the procedure for using `Ch3.apr` to calculate spatial autocorrelation is as follows:

1. Open `Ch3.apr` with ArcView and then add `5Ccities.shp` and `5counties.shp`.
2. Click the `5Ccities.shp` theme to make it active.
3. From the **Point Patterns** menu item, choose **Create Distance Matrix** to start the process for computing the distances between points.
4. In the ID Selection window, select `City_fips` as the ID and then click **OK** to continue.
5. In the Output FileName window, navigate to your working directory and then click **OK** to save the file `distmatrix.dbf`.
6. Click **OK** in the next window to finish the processes of calculating distances.

The output file, distmatrix.dbf, contains the distances between all pairs of points in the point theme. If you wish, it can be added into the ArcView project by

1. highlighting the **Tables** icon in the project window,
2. clicking the **Add** button to invoke the Add Table window,
3. choosing the distmatrix.dbf file, where you stored it, and then clicking **OK** to add it to the current ArcView project, and
4. opening it as a regular ArcView Tables document.

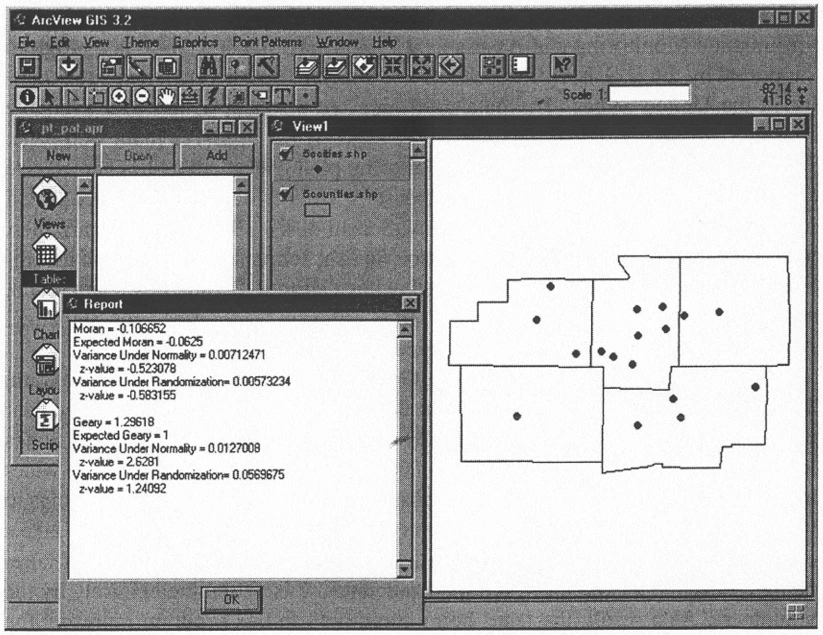
With distance matrix computed, we are now ready to calculate index values for spatial autocorrelation by following these steps:

1. Choose the **Point Patterns** item from the menu bar.
2. In the drop-down submenu, choose **Moran-Geary** to proceed.
3. In the *Check for Input* window, answer the question "Have you created a spatial weight matrix?" by clicking **Yes** to proceed.
4. As previously specified, in the *Get Input* window, choose City_fips as the ID field. Click **OK** to continue.
5. In the next *Get Input* window, select Pop1990 as the variable for this calculation. Click **OK** to continue.
6. For the next window, navigate to the directory where you have saved the distmatrix.dbf file. Choose it and then click **OK** to proceed.
7. In the *Weight* window, choose **Inverse to Distance** as the weight option for this calculation. Click **OK** to continue.
8. When the calculation is finished, the values for both indices are listed in the *Report* window (Figure 3.9) along with their expected values and variances under nonfree sampling (randomization) and free sampling (normality) assumptions. Note that the standardized *Z*-scores are also calculated for testing the statistical significance of the resulting index values.

In this example, we can see that the Moran's Index value is not statistically significant under either assumption. Geary's Index value, on the other hand, is statistically significant under the free sampling (normality) assumption.

## 3.5 APPLICATION EXAMPLES

The methods discussed in this chapter are used to detect spatial patterns of point distributions. Quadrat Analysis is concerned with how densities of points change over space, and it is a spatial sampling approach. Quadrats of consistent size and shape are overlaid on points in the study area. The frequency distribution of the number of points in each quadrat is constructed and compared with the frequency distribution of a theoretical random pattern. Nearest Neighbor Analysis, on the other hand, exploits the spacing between neighbors. The distances between nearest neighboring points are measured. The average of the distances from all possible pairs of nearest neighbors is compared to that of a theoretical random pattern.



```
Moran = -0.106652
Expected Moran = -0.0625
Variance Under Normality = 0.00712471
  z-value = -0.523078
Variance Under Randomization= 0.00573234
  z-value = -0.583155

Geary = 1.29618
Expected Geary = 1
Variance Under Normality = 0.0127008
  z-value = 2.6281
Variance Under Randomization= 0.0569675
  z-value = 1.24092
```

**Figure 3.9** Reporting Moran's I and Geary's C Index values.

Both Quadrat Analysis and Nearest Neighbor Analysis are useful in detecting spatial patterns and comparing them with other known patterns. However, only the locations of the points are considered. These two methods do not take into account that different points in a distribution may be different in some ways or may represent different activities or events. Therefore, the use of these methods is limited.

Nevertheless, spatial autocorrelation coefficients consider the similarity of point locations as well as the similarity of attributes of the points. These coefficients calculate how attribute values change over space with respect to the locations of the points. Both Moran's I and Geary's C have been discussed in this chapter. They are equally useful but differ in their numeric scales.

In this section, we will look at how these methods can help us understand how point data distribute and, to a great extent, how we can use them to detect if the data distribute in any distinguishable pattern. We will use two sets of point-based data that represent the water transparency of monitored lakes.

The first data set concerns the transparency of lakes, as monitored by the Environmental Monitoring and Assessment Program (EMAP). EMAP is a program supported by the U.S. EPA that collects and analyzes data on environmental quality in the United States. The data used here are a subset of the data available from the program's website, which can be reached via the EPA's main web page. EMAP has developed an intensive monitoring plan that attempts to characterize fully the lake water quality of the northeastern United States. The data collected include transparency (as measured by the Secchi disk), water chemical variables, watershed characteristics, and information on fish, zooplanton, and diatom assemblages. The sampled lakes were selected using a stratified probabilistic approach, which randomly selected lakes based on a criterion that defined the statistical population of lakes in northeastern United States (Larsen et al., 1994).

Another data set is taken from the Great American Secchi Dip-In program. The North American Lake Management Society supports this program. Each year during the week of July 4th, thousands of volunteers across the United States and Canada dip their Secchi disks into lakes of their choice to measure the water transparency (Jenerette et al., 1998). These volunteers then report their findings to the program's office, along with their answers to other questions about the lake's environmental conditions and how the lake is being used. The selection of lakes being monitored has no prestructured framework. It is determined entirely by the volunteers. As a result, the lakes being monitored in this program represent the lakes that are being used, that volunteers care about, and consequently, the ones that need our attention.

One of the issues discussed recently is the sampling process used in the two lake monitoring programs. EMAP, through great efforts, selected lakes using what it considered to be a random pattern based on a stratified probabilistic approach. The Dip-In program, on the other hand, lets volunteers made the selections. The philosophical and theoretical approaches behind the two programs are entirely different, and it will be interesting to examine how the outcomes differ. To provide a visual impression of how the monitored lakes distribute, Figure 3.10 shows the locations of the EMAP and Dip-In lakes.

On the issue of how to better sample lakes to be monitored, we can use the methods discussed in this chapter to examine how the two data sets differ. We will measure to what degree the lakes being monitored by the two programs deviate from a random pattern to indicate indirectly how the sampling outcomes of the two programs differ.

Now that the ArcView project file, `Ch3.apr`, is available to us, it is just a matter of running the script for the two data sets. To examine the data in more detail, we can divide each data set by state boundaries to create subsets of data in both cases. This is done so that we can see how the spatial patterns change between scales. When the entire data set is used in the analysis, we test the spatial pattern at a multistate scale. When testing the subsets, we are examining the spatial patterns at a more detailed local scale.

Both data sets contain water transparency data (as an attribute in their ArcView shapefiles). Each of the data sets is divided into eight subsets for the following eight states: Connecticut (CT), Delaware (DE), Massachusetts (MA), Maine (ME), New Hampshire (NH), New York (NY), Rhode Island (RI), and Vermont (VT). Including the entire northeastern United States, each program has nine data sets in this analysis.

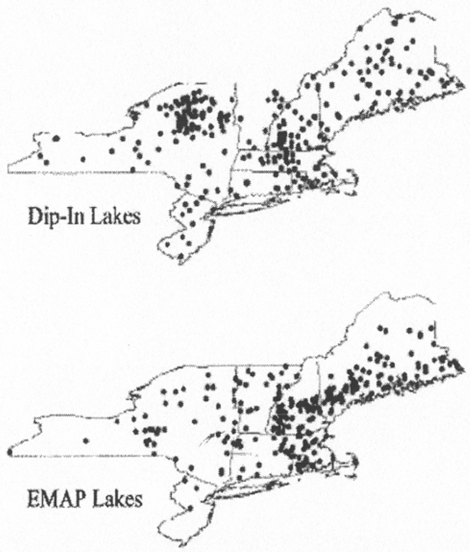


**Figure 3.10** Spatial distribution of Dip-In lakes and EMAP lakes.

Table 3.6 lists the results of running `Ch3.apr` on each of the data sets and subsets. It gives the statistics and their $Z$ scores for Quadrat Analysis and Nearest Neighbor Analysis. For Quadrat Analysis, 100 quadrats were used in each run.

In Table 3.6, the number of data points in each data set is shown in parentheses with the program's name. For example, EMAP has 350 lakes for the entire northeastern United States, and Dip-In has 303 lakes. Some subsets are excluded because one of the programs has fewer than five data points. They are being displayed with a gray screen. To identify significant results easily, those $z$-scores that are greater than 1.96 or less than $-1.96$ are highlighted in boldface italics because these $z$-scores indicate statistical significance at the $\alpha = 0.05$ (or 5%) level.

For the entire northeastern United States, neither program shows any spatial pattern when examined and tested by Quadrat Analysis. However, they are both considered nonrandom patterns by Nearest Neighbor Analysis. They deviate from the random pattern with a statistical significance at the $\alpha = 0.05$ level. When the data are partitioned into subsets for individual states, monitored lakes in both the EMAP and Dip-In programs show statistically significant dispersed patterns in Massachusetts, Maine, New Hampshire, and New York. For Vermont, EMAP's lakes show much more dispersion than the Dip-In lakes.

**TABLE 3.6 Quadrat Analysis and Nearest Neighbor Analysis of EMAP Data and Dip-In Data**

| | | Quadrat Analysis | | Nearest Neighbor Analysis | |
|---|---|---|---|---|---|
| | | *D* Statistics | *Z* score | NNA Statistics | *Z* score |
| NE | EMAP (350) | 0.7505 | 0.0727 | 0.3227 | ***24.2433*** |
| USA | Dip-In (303) | 0.4308 | 0.0781 | 0.4128 | ***19.5540*** |
| CT | EMAP (14) | *6.0000* | *0.3490* | *0.3103* | *4.9375* |
| | Dip-In (4) | *23.5008* | *0.6240* | *0.8198* | *0.6896* |
| DE | EMAP (13) | *6.3846* | *0.3610* | *0.7529* | *1.7041* |
| | Dip-In (4) | *23.0000* | *0.6240* | *0.4931* | *1.9397* |
| MA | EMAP (38) | 1.5263 | 0.2206 | 0.4301 | ***6.7213*** |
| | Dip-In (32) | 2.0004 | 0.2200 | 0.6724 | ***3.5449*** |
| ME | EMAP (74) | 0.4553 | 0.1581 | 0.4416 | ***9.1909*** |
| | Dip-In (99) | 0.3456 | 0.1367 | 0.4612 | ***10.2563*** |
| NH | EMAP (45) | 1.1346 | 0.2027 | 0.5979 | ***5.1605*** |
| | Dip-In (78) | 0.4665 | 0.1540 | 0.4497 | ***9.2991*** |
| NY | EMAP (142) | 0.3486 | 0.1141 | 0.3403 | ***15.0407*** |
| | Dip-In (44) | 1.1818 | 0.2050 | 0.7324 | ***3.3958*** |
| RI | EMAP (4) | *24.0008* | *0.6240* | *1.0870* | *0.3328* |
| | Dip-In (18) | *4.5556* | *0.3090* | *0.6070* | *3.1899* |
| VT | EMAP (18) | 4.5000 | 0.3090 | 0.3119 | ***5.5853*** |
| | Dip-In (21) | 3.7156 | 0.2700 | 0.7998 | 1.7556 |

*Notes*:

- Numbers in parentheses are the number of points in each data set.
- Results from data sets with four or fewer points are screened.
- Z scores above 1.96 are highlighted in boldface italic.

Two observations can be made. First, Nearest Neighbor Analysis is a more powerful method than Quadrat Analysis because it detects what the Quadrat Analysis fails to detect. Second, volunteers, without special instructions, selected lakes that show dispersion similar to that of the lakes selected by EMAP's stratified sampling, but to a lesser extent.

When examining the data sets for spatial patterns, we often wonder what spatial pattern each data set will display. The spatial autocorrelation coefficients can be used to assist the detection. Table 3.7 shows the results of running `Ch3.apr` for calculating Geary's C index and Moran's I index. Similar to Table 3.6, states with fewer than five lakes in either the EMAP or Dip-In program are dropped from further analysis. Any $z$-score that is either greater than 1.96 or less than $-1.96$ is highlighted in boldface italic text, as it is statistically significant at the $\alpha = 0.05$ level.

With measures of water transparency data being tested by Geary's Ratio, EMAP shows some degree of a regionalized pattern but not much significance. Geary's Ratio is 0.9268 for the entire EMAP data set, suggesting that the spatial pattern shows smooth changes in water transparency between neighboring lakes.

**TABLE 3.7 Spatial Autocorrelation In EMAP data and Dip-In Data**

| | | Geary's C | *Z* score | Moran's I | *Z* score |
|---|---|---|---|---|---|
| NE | EMAP (350) | 0.9268 | 0.0000 | 0.0279 | ***5.0743*** |
| USA | Dip-In (303) | 0.6814 | −8.8528 | 0.1928 | ***16.3977*** |
| CT | EMAP (14) | 1.4593 | 1.9502 | −0.3569 | −1.4290 |
| | Dip-In (4) | 1.0253 | 0.3296 | −0.3319 | 0.0043 |
| DE | EMAP (13) | 0.9383 | −0.5924 | 0.0910 | 1.5382 |
| | Dip-In (4) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| MA | EMAP (38) | 1.3206 | 0.0000 | 0.1774 | ***−2.8878*** |
| | Dip-In (32) | 0.8213 | −1.0881 | 0.1120 | 1.1451 |
| ME | EMAP (74) | 0.9376 | 0.0000 | −0.0082 | 0.1458 |
| | Dip-In (99) | 0.6876 | −4.5359 | 0.0779 | ***3.0414*** |
| NH | EMAP (45) | 0.8728 | −1.0937 | 0.1139 | 1.7596 |
| | Dip-In (78) | 0.7964 | −3.6065 | 0.1474 | ***5.0601*** |
| NY | EMAP (142) | 0.8806 | 0.0000 | 0.0149 | 1.0342 |
| | Dip-In (44) | 0.9555 | −0.4744 | −0.0301 | −0.1131 |
| RI | EMAP (4) | 1.0040 | 0.0404 | −0.2987 | 0.0993 |
| | Dip-In (18) | 0.7976 | −1.2640 | −0.1070 | −0.3794 |
| VT | EMAP (18) | 1.5770 | 1.9899 | −0.4946 | ***−2.1601*** |
| | Dip-In (21) | 0.9505 | −0.5391 | −0.0184 | 0.3640 |

*Notes*:

- Numbers in parentheses are the number of points in each data set.
- Results from data sets with four or fewer points are screened.
- Z scores above 1.96 are highlighted in boldface italic.
- Z scores are calculated for the free sampling (or normality) assumption.

With the $z$-score approaching 0.0, the lakes monitored by EMAP do not show spatial autocorrelation with enough statistical significance. Also with Geary's Ratio, the lakes in the Dip-In program seem to show a much stronger degree of spatial autocorrelation. This means that the neighboring lakes tend to show similar values of water transparency. As suggested by the $z$-score, this data set may be more appropriate for analysis of a regional trend. When Moran's I is used, the lakes in both programs show a statistically significant departure from a random pattern. This is demonstrated by the high $z$-scores of 5.0786 (EMAP) and 16.3977 (Dip-In).

As for the data on individual states, the Dip-In program has Maine and New Hampshire showing strong regional trends, while Vermont shows contrasting transparency values between neighboring lakes. For the EMAP lakes, none of the states has a strong correlation of a point pattern to be detected by Geary's C index.

When using Moran's I index on data sets of individual states, we see that the EMAP program's lakes in Massachusetts and Vermont show strong dissimilarity between neighboring lakes in terms of water transparency. For Dip-In's data set, Maine and New Hampshire show a strong regional trend, as their transparency values tend to be similar between neighboring lakes.