

Figure 1.11b Steps and results of bivariate regression.

CHAPTER 2

POINT DESCRIPTORS

In the Introduction and in Chapter 1, we discussed that spatial data consist of cartographic data (which describe the locational and geometric characteristics of features) and attributes data (which provide meanings to the geographic features). In Chapter 1, we focused on the analysis of attribute data exclusively, and thus the analysis was aspatial. Starting with this chapter, we will discuss statistical tools that have been implemented in GIS and have been especially designed to analyze only locational information or locational information together with attribute information. In this chapter, the locational information of points will be used in several descriptive geostatistics or centrographic measures to analyze point distribution.

2.1 THE NATURE OF POINT FEATURES

With vector GIS, *geographic features* are represented geometrically by points, lines, or polygons in a two-dimensional space. Geographic features that occupy very little or no areal extent at a given scale (the scale of study) are often represented as *points*. In a vector GIS database, linear features are best described as *lines*, and regions or areal units are often structured as *polygons*. As a whole, we refer to points, lines, and polygons as *geographic objects* since they represent geographic features.

While geographic objects can be conveniently represented as points, lines, and polygons, the relationship between geographic objects and the geographic features they represent is not always fixed. Scales often determine how geographic features are represented. For example, a house on a city map is only a point, but it becomes a polygon when the floor plan of the house is plotted on a map. Similarly, the City of Cleveland, Ohio, is represented by a large polygon that occupies

an entire map sheet. In this way, the details of the city's street network and other facilities are shown on a large-scale map (large because the ratio may be 1:2,000 or larger). Alternatively, Cleveland is often represented only as a point on a small-scale map (small because the ratio may be 1:1,000,000 or smaller) that identifies all of the major cities in the world.

The degree of abstraction also affects how various geographic objects represent geographic features. This is because points can be used not only to represent physical geographic features such as those described above, but also to describe the locations of events and incidences such as disease occurrences or even traffic accidents. In these cases, points do not represent real geographic features, but just locations of the events. Furthermore, for transportation modeling, urban analysis, or location-allocation analysis, areal units with significant spatial extents are often abstractly represented by points (such as centroids) to accommodate specific data structures, as required by the analytic algorithms.

In this chapter, we will focus on point features. Points are defined by coordinates. Depending on the coordinate system and the geographic projections, points on a map can be defined by a pair of latitude and longitude measures, x and y coordinates, easting and northing, and so on. On a small-scale map that covers a large areal extent but shows few geographic details, points can be used to identify locations of cities, towns, tourist attractions, and so on. On a large-scale map, points may represent historical landmarks, street lights, trees, wells, or houses.

While points on any map are all simple geometric objects that are defined by their coordinates, the attributes associated with these points provide specifics to differentiate them according to the characteristics emphasized. Consider a map showing all residential water wells in a city; the points will all look alike except for their locations. If attributes such as owners' names, depths, dates dug, or dates of last water testing were added to the database, more meaningful maps could be created to show spatial variation of the wells according to any of the attributes.

Individually, points may represent the locations of geographic features. The associated attributes help to describe each point's unique characteristics. The description of spatial relationship between individual points, however, requires the application of some of the spatial statistics described in this chapter. Specifically, we will discuss ways to determine where points are concentrated, as described by their locations or weighted by a given attribute. We will also examine how to measure the degree of dispersion in a set of points. This set of tools is also known as centrographic measures (Kellerman, 1981).

The spatial statistical methods to be discussed in this chapter are appropriate for points that represent various types of geographic features in the real world. A word of caution is needed here: the accuracy of locational information and its associated attribute values must be considered carefully. This is because the reliability and usefulness of any inference that results from analyzing the points are often affected by the quality of the data.

Point data obtained from maps may contain cartographic generalizations or locational inaccuracy. On a small-scale map, a point may represent a city whose actual areal extent is a certain number of square miles, while another point may

represent a historical landmark or the location of threatened plant species that occupy only several square inches on the ground. Comparing these points directly, however carefully performed, would be like comparing oranges with apples. Point locations derived from calculating or summarizing other point data can be especially sensitive to the quality of input data because the inaccuracy of input data will be propagated during computation, so that the results are of little value due to the magnitude of the inaccuracy.

For this reason, we urge spatial analysts to be sensitive to the scale of a spatial database and the quality of the data used in the analysis. Statistical methods can be very useful when they are used correctly, but they can be very misleading and deceiving when used inappropriately or carelessly.

2.2 CENTRAL TENDENCY OF POINT DISTRIBUTIONS

In classical statistics, the conventional approach to summarizing a set of values (or numerical observations) is to calculate the measure of central tendency. The central tendency of a set of values gives some indication of the *average* value as their representative. The average family income of a neighborhood can give an out-of-towner a quick impression of the economic status or living style of the neighborhood. If you plan to visit an Asian country for Christmas, it would be useful to know the average temperature in December there so that you know what clothing to take.

When comparing multiple sets of numerical values, the concept of average is particularly useful. Educators can use average scores of state proficiency tests between elementary schools to see how schools compare with one another. Comparing the harvests from farms using a new fertilizer with the harvests from farms without it provides a good basis for judging the effectiveness of the fertilizer. In these and many other similar settings, the central tendency furnishes summary information of a set of values that would otherwise be difficult to comprehend.

Given a set of values, x_i , $i = 1, 2, \dots, n$, measures of central tendency include the mean, weighted mean, and median. The mean, \bar{X} , is simply the arithmetic average of all values as

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

What if the observation values in a data set do not carry the same level of importance? Obviously, the measure of central tendency will not be simply the arithmetic mean. In that case, each value, x_i , in the data set will first be multiplied by its associated weight, w_i . The sum of the weighted values is then divided by the sum of the weights to obtain the *weighted mean*:

$$\bar{X}_w = \frac{\sum_{i=1}^n x_i w_i}{\sum_{i=1}^n w_i}$$

Another measure of central tendency in classical statistics is the *median*. These two measures of central tendency were discussed in Chapter 1.

When dealing with a data set that contains observations distributing over space, one can extend the concept of *average* in classical statistics to the concept of *center*, as a measure of spatial central tendency. Because geographical features have spatial references in a two-dimensional space, the measure of central tendency needs to incorporate coordinates that define the locations of the features or objects. Central tendency in the spatial context will be the mean center, the weighted mean center, or the median center of a spatial point distribution.

There are several ways in which the position of such centers can be calculated; each gives different results based on how the data space is organized. Different definitions of the extent of the study area, distortions caused by different map projection systems, or even different map scales at which data were collected often lead to different results. It is important to realize that there is no one *correct* way of finding the center of a spatial distribution. There are appropriate methods for use in various settings, but there is probably no single correct method suitable for all situations. Therefore, the interpretation of the result of calculating the center of a spatial distribution can only be determined by the nature of the problem.

To describe a point distribution, we will discuss a series of point descriptors in this chapter. For central tendency, mean centers, weighted mean centers, and median centers provide a good summary of how a point set distributes. For the extent of dispersion, standard distances and the standard ellipse measure the spatial variation and orientation of a point distribution.

2.2.1 Mean Center

The *mean center*, or spatial mean, gives the average location of a set of points. Points may represent water wells, houses, power poles in a residential subdivision, or locations where landslides occurred in a region in the past. As long as a location can be defined, even with little or no areal extent, it can be represented as a point in a spatial database. Whatever the points in a spatial database represent, each point, p_i , may be defined operationally by a pair of coordinates, (x_i, y_i) , for its location in a two-dimensional space.

The coordinate system that defines the location of points can be quite arbitrary. Geographers have devised various map projections and their associated coordinate systems, so the locations of points in space can be referred to by their latitude/longitude, easting/northing, or other forms of coordinates. When working with known coordinate systems, the location of a point is relatively easy to define or even to measure from maps. There are, however, many situations requiring the use of coordinate systems with an arbitrary origin as the reference point. Arbitrary coordinate systems are often created for small local studies or for quick estimation projects. In those cases, the coordinate system needs to be carefully structured so that (1) it orients to a proper direction for the project, (2) it situates with a proper origin, and (3) it uses suitable measurement units. For more detailed discussion of these selections, interested readers may refer to the monograph by

TABLE 2.1 Ohio Cities from the Top 125 U.S. Cities and Their Mean Center

City Name	Longitude in Degrees (x)	Latitude in Degrees (y)
Akron	-81.5215	41.0804
Cincinnati	-84.5060	39.1398
Cleveland	-81.6785	41.4797
Columbus	-82.9874	39.9889
Dayton	-84.1974	39.7791
$n = 5$	$\sum x = -414.8908$	$\sum y = 201.4679$
	$\bar{x}_{mc} = \frac{-414.8908}{5} = -82.9782$	$\bar{y}_{mc} = \frac{201.4679}{5} = 40.2936$

Monmonier (1993). All these issues have to be taken into account so that the resulting mean center will approximate its most appropriate location.

With a coordinate system defined, the mean center can be found easily by calculating the mean of the x coordinates (eastings) and the mean of the y coordinates (northings). These two mean coordinates define the location of the mean center as

$$(\bar{x}_{mc}, \bar{y}_{mc}) = \left(\frac{\sum_{i=1}^n x_i}{n}, \frac{\sum_{i=1}^n y_i}{n} \right),$$

where

$\bar{x}_{mc}, \bar{y}_{mc}$ are coordinates of the mean center,
 x_i, y_i are coordinates of point i , and
 n is the number of points.

As an example, Table 2.1 lists the coordinates of 5 cities in Ohio that are among the 125 largest U.S. cities. Their locations are shown in Figure 2.1. The calculation in Table 2.1 shows that the mean center of the five cities is located at -82.9782, 40.2936 or 82.9782W, 40.2936N. The star in Figure 2.1 identifies the location of the calculated mean center. Since this mean center is defined by the mean of x coordinates and the mean of y coordinates, it is located at the geometric center of the five cities, as expected. What it represents is the center of gravity of a spatial distribution formed by the five cities.

2.2.2 Weighted Mean Center

There are situations in which the calculation of mean centers needs to consider more than just the location of points in the spatial distribution. The importance of individual points in a distribution is not always equal. In calculating the spatial mean among a set of cities, the mean center may give a more realistic picture of the central tendency if the mean center is weighted by the population counts



Figure 2.1 Five Ohio largest 125 U.S. cities and their mean center with standard distance.

of these cities. The mean center is pulled closer to a city if the city's population is larger than the populations of the other cities being considered. Similarly, a *weighted mean center* provides a better description of the central tendency than a mean center when points or locations have different frequencies or occurrences of the phenomenon studied. Given points representing the locations of endangered plant species, it makes more sense to calculate their mean center by using the sizes of plant communities at each location as weight of the points.

The weighted mean center of a distribution can be found by multiplying the x and y coordinates of each point by the weights assigned to them. The mean of the weighted x coordinates and the mean of the weighted y coordinates define the position of the weighted mean center.

The equation for the weighted mean center is

$$(\bar{x}_{wmc}, \bar{y}_{wmc}) = \left(\frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}, \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} \right),$$

where

$\bar{x}_{wmc}, \bar{y}_{wmc}$ defines the weighted mean center,

and

w_i is the weight at point p_i .

TABLE 2.2 Ohio Cities from the Top 125 U.S. Cities and Their Weighted Mean Center

City Name	Longitude in Degrees x	Latitude in Degrees y	Population in 1990 p
Akron	-81.5215	41.0804	223,019
Cincinnati	-84.5060	39.1398	364,040
Cleveland	-81.6785	41.4797	505,616
Columbus	-82.9874	39.9889	632,910
Dayton	-84.1974	39.7791	182,044
Sum	$\sum x = -414.8908$	$\sum y = 201.4679$	$\sum p = 1,907,629$
City Name	Longitude \times Population $x \times p$	Latitude \times Population $y \times p$	
Akron	-18,180,843.41	9,161,709.73	
Cincinnati	-30,763,564.24	14,248,452.79	
Cleveland	-41,297,956.46	20,972,800.00	
Columbus	-52,523,555.33	2,530,937.47	
Dayton	-15,327,631.49	7,241,546.48	
Sum	$\sum xp = -158,093,550.90$	$\sum yp = 76,933,883.70$	
$n = 5$	$\sum xp = -158,093,550.9$	$\sum yp = 76,933,883.7$	
	$\bar{x}_{wmc} = \frac{\sum xp}{\sum p} = \frac{-158,093,550.9}{1,907,629}$	$\bar{y}_{wmc} = \frac{\sum yp}{\sum p} = \frac{76,933,883.7}{1,907,629}$	
	$= -82.87$	$= 40.33$	

In the case of the 5 largest 125 U.S. cities in Ohio, the mean center will be shifted toward the Cleveland-Akron metropolitan area if the population sizes are used as weights for the 5 Ohio cities. To calculate this weighted mean center, Table 2.2 lists the coordinates as those of Table 2.1 and population counts in cities in 1990.

The result shown in Table 2.2 gives the location of the weighted mean center as $-82.87, 40.33$, representing a shift toward the northeast direction from the mean center calculated in Section 2.2.1.

ArcView Notes In calculating the mean center, the major inputs are the x and y coordinates for the unweighted case. The additional input will be the weights for the weighted mean. The weights are usually provided as an attribute (such as counts of incidence or the number of people) of that location. The x and y coordinates usually are



not explicitly recorded in ArcView. They must be extracted either by using the **Field Calculator** with **.GetX** and **.GetY** requests issued to the point shape objects or by running the Avenue sample script, `addxycoo.ave`. Either approach will put the x and y coordinates into the attribute table associated with the point theme in ArcView. To use the **Field Calculator**, first add a theme to a View. Open the associated feature table under the Theme menu. Then choose **Start Editing** on the table under the Table menu. Under **Edit**, choose **Add Field** to add a field for the x -coordinate readings and a field for the y -coordinate readings.

After adding the two new fields to the Table, click on the new field for the x -coordinate, then go to **Field/Calculate** menu item. The Field Calculator will appear. It will ask for a formula for the x -coordinate. Double-click on the [Shape] field in the list of fields to include the shape field in the formula. Then type `.GetX`. Be sure to include the period. Then click **OK**. ArcView will extract the x -coordinate of all points and put the values in the new field. Repeat the same process for the y -coordinate but use `.GetY`. Afterward, remember to save the result by choosing the **Table/Stop Editing** menu item.

If one prefers to use the sample script `addxycoo.ave`, first open a new script window by double-clicking on the script icon in the Project window. Then go to **Script/Load Text File** to locate the sample script. Usually, it is under the subdirectory for `ArcView\samples\scripts`. Load the script into the script window. Go to **Script/Compile** to compile it before **Run**.

Similarly, to calculate the mean center in ArcView, we can use either the existing functions of ArcView or Avenue. For the unweighted case, the mean center is just the average of the x and y coordinates for all points. Therefore, using the **Statistics** function under the **Field** menu in **Table** documents, a set of statistics including the mean of the chosen coordinate (x or y) will be provided.

For the weighted case, the x and y coordinates have to be weighted by the appropriate attribute. First, two new columns (one for weighted x and another for weighted y) have to be created (follow the steps above to start editing a table and adding new fields). The two columns are then multiplied by their corresponding weights, as described in the equations above using the **Field Calculator**. Then the weighted mean center is obtained by using the **Statistics** function under the **Field** menu. Please note, however, that the results have to be recorded manually and cannot be displayed graphically.

If you are using Avenue to calculate mean centers, the process of extracting the coordinates can be coded into the Avenue

script. In that case, the coordinates will not need to be stored in the attribute table. Therefore, the attribute table does not need to be modified. This book has included a project file, `Ch2.aprx`, in the accompanying website. This file can be loaded into ArcView as a project (when loading ArcView, use the menu item **File** and then **Open Project** to direct ArcView to the appropriate directory to open this project file). All the geostatistical tools described in this chapter are included in this project file. The layout and the ArcView user interface of that customized project file are shown in Figure 2.2. In contrast to an ordinary ArcView project interface, this project has an additional menu item, **Spatial Analysis**. Under this menu, you can find items for **Spatial Mean**, **Standard Distance**, **Spatial Median**, and **Standard Deviation Ellipse** calculations. This project file, however, is encrypted to prevent

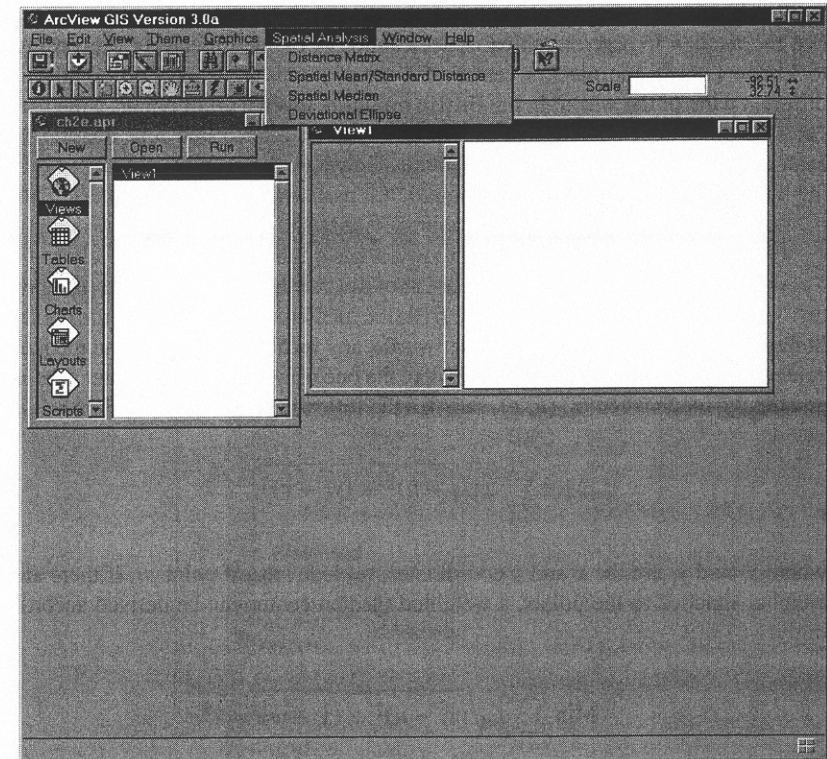


Figure 2.2 Customized ArcView user interface.

novice users from accidentally changing the content of the scripts.

The script for the mean center calculation includes the procedures for extracting the coordinates of point features and the mean center calculation. The script can handle both unweighted and weighted cases, allowing the user to select an attribute from the attribute table of a point theme as the weight. In addition, the script will draw the result (a point) on the associated View document to show the location of the mean center. Finally, if the user wishes to store the location of mean centers for future analysis, there is an option for creating a shapefile of the mean center to be used later.

2.2.3 Median Center

Analogous to classical descriptive statistics of central tendency, the concept of the median of a set of values can be extended to the *median center* of a set of points. But the median in a geographical space cannot be defined precisely. According to Ebdon (1988), the median center of a set of points is defined differently in different parts of the world. In the British tradition, given a set of points, a median center is the center upon which the study region is divided into four quadrants, each containing the same number of points. However, there can be more than one median center dividing the study area into four parts with equal numbers of points if there is sufficient space between points close to the center of the distribution. As this method leaves too much ambiguity, it has not been used extensively.

As used in North America, the concept of median center is the center of minimum travel. That is, the total distance from the median center to each of the points in the region is the minimum. In other words, any location other than the median center will yield a total distance larger than the one using the median center. Mathematically, median center, (u, v) , satisfies the following objective function:

$$\text{Min} \sum_{i=1}^n \sqrt{(x_i - u)^2 + (y_i - v)^2},$$

where x_i and y_i are the x and y coordinates, respectively, of point p_i . If there are weights attached to the points, a weighted median center can be derived accordingly:

$$\text{Min} \sum_{i=1}^n f_i \sqrt{(x_i - u)^2 + (y_i - v)^2}.$$

Please note that the weights, f_i for p_i , can be positive or negative values to reflect the pulling or pushing effects of points to the location of the median center.

To derive the median center, an iterative procedure can be used to explore and to search for the location that satisfies the above objective function. The procedure is as follows:

1. Use the mean center as the initial location in searching for the median center. This is essentially setting (u_0, v_0) equal to (x_{mc}, y_{mc}) .
2. In each iteration, t , find a new location for the median center, (u_t, v_t) , by

$$u_t = \frac{\sum f_i x_i / \sqrt{(x_i - u_{t-1})^2 + (y_i - v_{t-1})^2}}{\sum f_i / \sqrt{(x_i - u_{t-1})^2 + (y_i - v_{t-1})^2}}$$

and

$$v_t = \frac{\sum f_i y_i / \sqrt{(x_i - u_{t-1})^2 + (y_i - v_{t-1})^2}}{\sum f_i / \sqrt{(x_i - u_{t-1})^2 + (y_i - v_{t-1})^2}}.$$

3. Repeat step 2 to derive new locations for the median center until the distance between (u_t, v_t) and (u_{t-1}, v_{t-1}) is less than a threshold defined a priori.

ArcView Notes



The Avenue script for median centers, as incorporated into the project file Ch2.apr, is an extension of the script for the mean center. This is because the mean center is used as the initial location. The script asks the user for a threshold distance to control the termination of the script. Using the mean center as the



Figure 2.3 Spatial median.

initial location, the script goes through the iterative process to search for the median center. Figure 2.3 shows the median center, which is quite far from the spatial mean. Please note that if the coordinates of points are in degrees of longitude and latitude (Map Units, as defined in ArcView), the threshold distance is also defined in that metric scale (degrees).

2.3 DISPERSION OF POINT DISTRIBUTIONS

Similar to using measures such as standard deviations to assist an analyst in understanding a distribution of numeric values, standard distances or standard ellipses have been used to describe how a set of points disperses around a mean center. These are useful tools because they can be used in very intuitive ways. The more dispersed a set of points is around a mean center, the longer the standard distance and the larger the standard ellipse it will have.

Given a set of n data values, x_i , $i = 1, \dots, n$, the *standard deviation*, S , can be computed as

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

where \bar{x} is the mean of all values. The standard deviation is literally the square root of the average squared deviation from the mean.

2.3.1 Standard Distance

Standard distance is the spatial analogy of standard deviation in classical statistics. While standard deviation indicates how observations deviate from the mean, standard distance indicates how points in a distribution deviate from the mean center. Standard deviation is expressed in units of observation values, but standard distance is expressed in distance units, which are a function of the coordinate system or projection adopted.

The standard distance of a point distribution can be calculated by using the following equation:

$$SD = \sqrt{\frac{\sum_{i=1}^n (x_i - x_{mc})^2 + \sum_{i=1}^n (y_i - y_{mc})^2}{n}}$$

where (x_{mc}, y_{mc}) is the mean center of the point distribution.

Since points in a distribution may have attribute values that can be used as weights when calculating their mean center or even their median center, it is also

possible to weight the points with specified attribute values when calculating the standard distance. For *weighted standard distance*, the following equation can be used:

$$SD = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - x_{mc})^2 + \sum_{i=1}^n f_i (y_i - y_{mc})^2}{\sum_{i=1}^n f_i}}$$

where f_i is the weight for point, (x_i, y_i) .

Using the 5 Ohio cities selected from the list of 125 largest U.S. cities, the standard distance is derived and the associated standard distance circle is presented in Figure 2.1. The steps for manually calculating the standard distance and the weighted standard distance are shown in Table 2.3.

TABLE 2.3 Standard Distance and Weighted Standard Distance of 5 Ohio Cities from the Largest 125 U.S. Cities

City Name	Longitude in Degrees x	Latitude in Degrees y	Population in 1990 p
Akron	-81.5215	41.0804	223,019
Cincinnati	-84.5060	39.1398	364,040
Cleveland	-81.6785	41.4797	505,616
Columbus	-82.9874	39.9889	632,910
Dayton	-84.1974	39.7791	182,044
Σ	$\Sigma x = -414.8908$	$\Sigma y = 201.4679$	$\Sigma p = 1,907,629$
	$\bar{x}_{mc} = \frac{-414.8908}{5} = -82.9782,$	$\bar{y}_{mc} = \frac{201.4679}{5} = 40.2936$	
City Name	$x - x_{mc}$	$(x - x_{mc})^2$	$p(x - x_{mc})^2$
Akron	+1.4567	2.1220	473,246.3180
Cincinnati	-1.5278	2.3342	849,742.1680
Cleveland	+1.2997	1.6892	854,086.5472
Columbus	-0.0092	0.0001	63.2910
Dayton	-1.2192	1.4864	270,590.2016
Σ		7.6319	2,450,728.5260
City Name	$y - y_{mc}$	$(y - y_{mc})^2$	$p(y - y_{mc})^2$
Akron	+0.7868	0.6191	138,071.0629
Cincinnati	-1.1538	1.3313	484,646.4520
Cleveland	+1.1861	1.4068	711,300.5888
Columbus	-0.3047	0.0928	58,734.0480
Dayton	-0.5145	0.2647	48,187.0468
Σ		3.7147	1,440,939.1990

(continued)

TABLE 2.3 Continued.

Standard Distance

Step 1:

$$\sum (x - x_{mc})^2 = 7.6319$$

$$\sum (y - y_{mc})^2 = 3.7147$$

Step 2:

$$SD = \sqrt{\frac{\sum_{i=1}^n (x_i - x_{mc})^2 + \sum_{i=1}^n (y_i - y_{mc})^2}{n}}$$

$$= \sqrt{\frac{7.6319 + 3.7147}{5}}$$

$$= \sqrt{\frac{11.3466}{5}}$$

$$= \sqrt{2.2693}$$

$$= 1.5064$$

Weighted Standard Distance

Step 1:

$$\sum p(x - x_{mc})^2 = 2,450,728.526$$

$$\sum p(y - y_{mc})^2 = 1,440,939.199$$

Step 2:

$$SD_w = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - x_{mc})^2 + \sum_{i=1}^n f_i (y_i - y_{mc})^2}{\sum_{i=1}^n f_i}}$$

$$= \sqrt{\frac{2,450,728.526 + 1,440,939.199}{1,907,629}}$$

$$= \sqrt{\frac{3,891,667.725}{1,907,629}}$$

$$= \sqrt{2.0401}$$

$$= 1.4283$$

ArcView Notes

In terms of its application, standard distance is usually used as the radius to draw a circle around the mean center to give the extent of spatial spread of the point distribution it is based on. In the project file Ch2.apr, the added item under Spatial Analysis calculates the mean center before calculating the standard distance. It then uses the standard distance as the radius to draw a circle around the mean center. The script also provides

the options for users to save the mean center with or without the standard distance as point and polygon shapefiles for future use.

Different standard distance circles can be drawn for different types of events or incidences in the same area. The same types of events or incidences can also be drawn in different areas. All these can provide the basis for visual comparison of the extent of spatial spread among different types of events or different areas. Between two neighborhoods with the same number of houses, the neighborhood that has a longer standard distance is obviously spreading over more space geographically than the other neighborhood. Similarly, for all cities in a state, standard distances will be different if their calculation is weighted by different attributes, such as their population sizes.

While similar applications can be structured easily, it is important to understand that comparisons of standard distances between point distributions may or may not be meaningful. For instance, the standard distance of Japan's largest cities weighted by population counts is calculated as 3.27746 decimal degrees, while it is 8.84955 decimal degrees for Brazil's largest cities. If the two standard distances are used alone, they indicate that Brazil has a much more dispersed urban structure than Japan. However, given the very different sizes and territorial shapes of the two countries, the absolute standard distances may not reflect accurately the spatial patterns of how the largest cities spread in these countries.

To adjust for this possible bias, the standard distance may be scaled by the average distance between cities in each country or by the area of each country or region. Alternatively, the standard distances can be standardized or weighted by a factor that accounts for the territorial differences of the two countries. In general, standard distances can be scaled by a variable that is a function of the size of the study areas. In this example, the weighted standard distances are 0.2379 and 0.0272 for Japan and Brazil, respectively, when scaled by their areas, indicating that Japan's largest cities are in fact more dispersed than Brazil's cities.

2.3.2 Standard Deviation Ellipse

The standard distance circle is a very effective tool to show the spatial spread of a set of point locations. Quite often, however, the set of point locations may come from a particular geographic phenomenon that has a directional bias. For instance, accidents along a section of highway will not always form a circular shape represented by a standard distance circle. Instead, they will appear as a linear pattern dictated by the shape of that section of highway. Similarly, occurrences of algae on the surface of a lake will form patterns that are limited by the shape of the lake. Under these circumstances, the standard distance circle will not be able to reveal the directional bias of the process.

A logical extension of the standard distance circle is the *standard deviational ellipse*. It can capture the directional bias in a point distribution. There are three components in describing a standard deviational ellipse: the angle of rotation, the deviation along the major axis (the longer one), and the deviation along the minor axis (the shorter one). If the set of points exhibits certain directional bias, then there will be a direction with the maximum spread of the points. Perpendicular to this direction is the direction with the minimum spread of the points. The two axes can be thought of as the x and y axes in the Cartesian coordinate system but rotated to a particular angle corresponding to the geographic orientation of the point distribution. This angle of rotation is the angle between the north and the y axis rotated clockwise. Please note that the rotated y axis can be either the major or the minor axis. Figure 2.4 illustrates the terms related to the ellipse.

The steps in deriving the standard deviational ellipse are as follows:

1. Calculate the coordinates of the mean center, (x_{mc}, y_{mc}) . This will be used as the origin in the transformed coordinate system.
2. For each point, p_i , in the distribution, transform its coordinate by

$$x'_i = x_i - x_{mc}$$

$$y'_i = y_i - y_{mc}$$

After they have been transformed, all points center at the mean center.

3. Calculate the angle of rotation, θ , such that

$$\tan \theta = \frac{\left(\sum_{i=1}^n x_i'^2 - \sum_{i=1}^n y_i'^2 \right) + \sqrt{\left(\sum_{i=1}^n x_i'^2 - \sum_{i=1}^n y_i'^2 \right)^2 + 4 \left(\sum_{i=1}^n x_i' \sum_{i=1}^n y_i' \right)^2}}{2 \sum_{i=1}^n x_i' \sum_{i=1}^n y_i'}$$

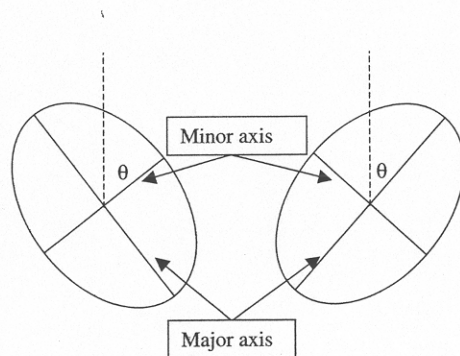


Figure 2.4 Elements defining a standard deviational ellipse.

$\tan \theta$ can be positive or negative. If the tangent is positive, it means that the rotated y axis is the longer or major axis and rotates clockwise from north. If the tangent is negative, it means that the major axis rotates counterclockwise from north. If the tangent is positive, we can simply take the inverse of tangent θ (arctan) to obtain θ for subsequent steps. If tangent is negative, taking the inverse of the tangent of a negative number will yield a negative angle (such as $-x$), i.e., rotating counterclockwise. But angle of rotation is defined as the angle rotating clockwise to the y axis, therefore, 90 degrees have to be added to the negative angle (i.e., $90 - x$) to derive θ . With θ from step 3, we can calculate the deviation along the x and y axes in the following manner:

$$\delta_x = \sqrt{\frac{\sum_{i=1}^n (x'_i \cos \theta - y'_i \sin \theta)^2}{n}}$$

and

$$\delta_y = \sqrt{\frac{\sum_{i=1}^n (x'_i \sin \theta + y'_i \cos \theta)^2}{n}}$$

ArcView Notes



The script for the standard deviational ellipse incorporated in the project file derives all three parameters necessary to fit a standard deviational ellipse to a set of points. While ArcView 3.1 and later versions support the ellipse object, ellipses slightly different in orientation cannot be easily detected visually. Instead, the script uses two perpendicular axes to show the orientation of the points and the spatial spread along the two directions.

2.4 APPLICATION EXAMPLES

We have discussed the concepts and background of a set of centrographic measures for analyzing point data. Although these measures are very useful in analyzing point data, they have not been used as widely as expected. We can still find many examples using these geostatistics in various research publications or applications. After each census in the United States, the Bureau of the Census calculates both the mean center and the median center of the entire U.S. population. Between censuses, the estimated mean center and median are also reported (for example, in the Census, 1996). By plotting the mean centers and median centers for each census year in the past century, it shows that the center of the U.S. population has been moving from the Northeast (Maryland and Delaware) to the West and the Southwest. Today the mean center is somewhere near St. Louis, Missouri.

Thapar et al. (1999) calculated mean population centers of the United States at two different geographical scales: census regions and states. By comparing the mean centers over different censuses, the results help us to depict the gross pattern of population movement at different spatial scales. Thapar et al. (1999) also reviewed several other studies using mean center as a descriptor tool for point data. In another example, Greene (1991) was concerned about the spread of economically disadvantaged groups over time. After deriving the standard distances of these population groups for different cities, Greene created circles based upon the standard distances to compare the location and geographical spread of these groups in several cities over time. In a completely different context, Levine et al. (1995) applied standard deviational ellipses to compare different types of automobile accidents in Hawaii. This study was done to decide if there was any directional bias among the types of accidents. The authors ultimately were able to provide a prescription for the local authority to deal with some "hot spots." By fitting an ellipse to a specific ethnic group, different ellipses are derived for different groups. These ellipses can be laid over each other to indicate the extent of their spatial correspondence. Using this idea, Wong (1999) recently derived a spatial index of segregation. The ellipse-based index was also used to study the spatial integration of different ethnic groups in China (Wong, 2000).

The articles reviewed above are not meant to be an exhaustive list. Readers can identify additional studies using these measures. To illustrate how these measures can be used in ArcView-accompanied Avenue scripts incorporated into the project file (CH2.APR) in this book, a point data set will be analyzed to show how these statistics can be used for real-world data.

2.4.1 Data

The point data set has been selected from the U.S. Environmental Protection Agency Toxic Release Inventory (EPA, 1999). The details of this database and related data quality issues are presented on EPA's website (<http://www.epa.gov>), so they will not be discussed here. Interested readers should carefully review all the data documentation and metadata before using this database in formal analysis and research. In brief, the TRI database is organized by state. Submitted to the EPA by each state government for the TRI database is the information about facilities that produce certain types of substances and chemicals that are known to be hazardous to the environment and the well-being of humans. The TRI database lists the facilities and their characteristics, such as chemicals released to the environment through various channels, together with their locations expressed in longitude and latitude.

We will focus on the TRI facilities in the state of Louisiana. Like most databases, the TRI database also has errors in positional accuracy. After sites with erroneous locational information are eliminated, there are 2,102 records in the Louisiana database. Among these 2,102 records, there are 234 types of chemicals. For the purpose of this exercise, only the longitude, latitude, and chemical information were extracted from the database to be imported into ArcView.

ArcView Notes



The three columns of TRI data (longitude, latitude, and chemical name) have to be converted into **dBase IV** format so that they can be imported into ArcView. Many spreadsheet and database packages can perform this conversion. With a new project opened in ArcView, the dBase table can first be added to the project to create an ArcView Table document. Click the **Table** icon in the project window and then click the **Add** button to bring up the **Add Table** window. Navigate to the directory where the dBase file is stored. Double click the filename to bring it in as an ArcView Table document.

In creating a point theme of the TRI facilities, a new View document was first created. In this View document, the **Add Event Theme** function under the **View** menu item was used to geocode the TRI facilities by choosing the Table document that contains the TRI information as the geocode table. After the point theme was created by the geocode procedure, it was converted into a shapefile (*la.tri.shp*) using the **Convert to Shapefile** function item under the **Theme** menu. Please note that in the table shown in Figure 2.5, two additional columns were added to the original three columns of data. The first new column is the **Shape** field, which is listed as "point." This is a column that ArcView creates and adds to the data set in order to define the geographical object (point, polyline, or polygon) represented in that theme. The last column will be discussed later.

To obtain a list of all chemicals and the number of facilities releasing these chemicals, we can summarize the table shown in Figure 2.5. First, highlight (click) the *chem_name* field. Then go to the **Field** menu to choose the **Summarize** function. In the **Field** window of the dialog box, choose either **Lat** or **Long**. Then in the **Summarize-by** window, choose **Count**. Click the **Add** button so that a *Count_lat* or *Count_long* query can be added to the window on the right. If you want to designate the summary table to be saved in a particular location, make the changes in the **Save-As** window. Part of the summary table is shown in Figure 2.6. The table shown is sorted in descending order according to the numbers of facilities releasing any chemicals. From the table, it is clear that the chemicals released by most facilities are chlorine, ammonia, toluene and methanol.

2.4.2 Analyses: Central Tendency and Dispersion

Although more than 200 types of chemicals are reported in the Louisiana TRI database, it is obvious that some chemicals concentrate in specific locations rather than spreading evenly across the state. The set of statistics described in this chapter can help explore the spatial characteristics of this point data set. Conceptually, we could derive a mean center, a standard distance and the associated circle, and

Shape	Lat	Long	Chem name	NewChem
Point	32.007500	-93.986944	CREOSOTE	0
Point	32.904324	-93.981919	TETRACHLOROETHYLENE	0
Point	32.824400	-93.975600	CREOSOTE	0
Point	32.877222	-93.975278	STYRENE	0
Point	32.346380	-93.973699	ARSENIC COMPOUNDS	0
Point	32.346380	-93.973699	CHROMIUM COMPOUNDS	0
Point	32.346380	-93.973699	COPPER COMPOUNDS	0
Point	32.618056	-93.924167	AMMONIA	0
Point	32.618056	-93.924167	CHLORINE	0
Point	32.618056	-93.924167	ETHYLBENZENE	0
Point	32.618056	-93.924167	HYDROCHLORIC ACID	0
Point	32.618056	-93.924167	MOLYBDENUM TRIOXIDE	0
Point	32.618056	-93.924167	N,N-DIMETHYLFORMAMIDE	0
Point	32.618056	-93.924167	NICKEL COMPOUNDS	0
Point	32.618056	-93.924167	NITRATE COMPOUNDS	0
Point	32.618056	-93.924167	NITRIC ACID	0
Point	32.618056	-93.924167	PHOSPHORIC ACID	0
Point	32.618056	-93.924167	TOLUENE	0
Point	32.433333	-93.908333	1,2,4-TRIMETHYLBENZENE	0
Point	32.433333	-93.908333	CERTAIN GLYCOL ETHERS	0
Point	32.433333	-93.908333	CHROMIUM COMPOUNDS	0
Point	32.433333	-93.908333	COPPER COMPOUNDS	0
Point	32.433333	-93.908333	CRESOL (MIXED ISOMERS)	0
Point	32.433333	-93.908333	MANGANESE COMPOUNDS	0
Point	32.433333	-93.908333	NICKEL COMPOUNDS	0
Point	32.433333	-93.908333	PHENOL	0

Figure 2.5 Portion of the ArcView attribute table showing the TRI data of Louisiana.

a standard deviational ellipse for each chemical. To illustrate the utility of the geostatistics and associated concepts, we choose only two chemicals: copper and ethylene. Using the Avenue script built into the project file CH2.APR, a mean center was created for each chemical. Based upon each mean center, a standard distance was derived for each of the two chemicals. With the standard distances as radii, two standard distance circles were drawn. The results are shown in Figure 2.7.

In Figure 2.7, all the TRI sites are included, with those releasing copper and ethylene represented by different symbols. Also note that some symbols are on top of each other because those facilities released multiple chemicals. As shown in

Chem name	Count	Count Lat
CHLORINE	78	78.000000
AMMONIA	76	76.000000
TOLUENE	76	76.000000
METHANOL	75	75.000000
XYLENE (MIXED ISOMERS)	64	64.000000
PHOSPHORIC ACID	56	56.000000
HYDROCHLORIC ACID	54	54.000000
ZINC COMPOUNDS	52	52.000000
BENZENE	45	45.000000
ETHYLBENZENE	44	44.000000
ETHYLENE	44	44.000000
N-HEXANE	40	40.000000
SULFURIC ACID	38	38.000000
NAPHTHALENE	38	38.000000
METHYLETHYL KETONE	36	36.000000
PROPYLENE	34	34.000000
NITRATE COMPOUNDS	34	34.000000
FORMALDEHYDE	32	32.000000
STYRENE	32	32.000000
CERTAIN GLYCOL ETHERS	30	30.000000
PHENOL	29	29.000000
ETHYLENE GLYCOL	28	28.000000
1,2,4-TRIMETHYLBENZENE	28	28.000000
NICKEL COMPOUNDS	27	27.000000
CYCLOHEXANE	24	24.000000
COPPER COMPOUNDS	22	22.000000

Figure 2.6 Portion of the table summarizing the TRI table by chemicals.

Figure 2.7, most TRIs releasing copper are found in the central and northwestern portions of the states, while most TRIs releasing ethylene are closer to the Gulf of Mexico. A comparison of the mean centers for the two sets of points shows that in general, TRIs releasing copper are mostly located in areas northwest of the TRIs releasing ethylene. But as indicated by the two standard distance circles (1.605 for copper and 1.159 for ethylene as their radii), the standard distance circle for copper TRIs is larger than the one for ethylene TRIs. In other words, TRIs releasing copper are geographically more dispersed in the northwestern part of the state than those releasing ethylene concentrating in the southeast.

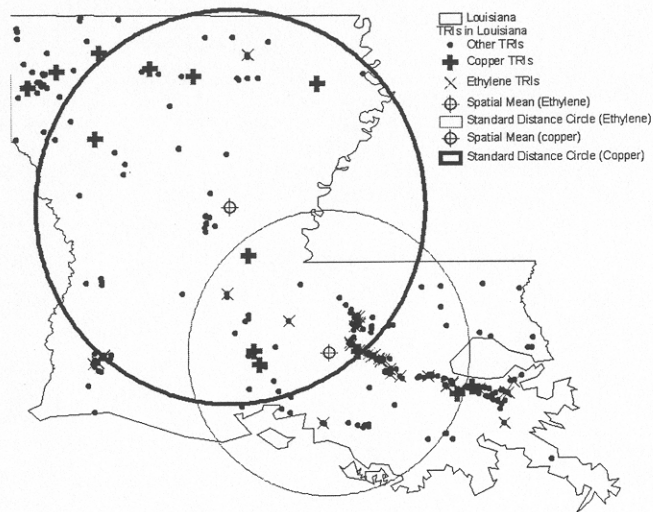


Figure 2.7 Spatial means and standard distances for TRIs in Louisiana releasing copper and ethylene.

In this example, because the two sets of points are from the same geographical region, we do not have to make any adjustment in the geographical scale when we compare the two standard distance circles.

ArcView Notes



The summary table shown in Figure 2.6 indicates that 13 facilities released copper (not shown in the table) and 44 facilities released ethylene. These facilities are then reclassified into three groups by adding a new attribute, *Newchem*, to the attribute table. The three new categories are copper (1), ethylene (2), and others (0). Using the **Query Builder** button in ArcView, we can select facilities according to the values in the *Newchem* field. By selecting *Newchem* = 1 in the **Query Builder** window, we select all TRIs releasing copper. Choosing the menu item **Spatial Mean/Standard Distance** under the **Spatial Analysis** menu will invoke the script to derive the mean center and draw the standard distance circle. The script is designed in such a way that if no points or records are selected, all the points or records will be selected and entered into the calculation of the statistics. After the script is run, a mean center and a standard distance circle will be generated for copper TRI sites. The process can be repeated for ethylene TRI sites.

Note that the calculation of standard distance is based upon the map units defined under **View Properties**. If the map units are in degree-decimal, the resultant standard distance is also in degree-decimal. In general, for large study areas, the data should be projected instead of using latitude-longitude. For small study areas, the distortion due to the unprojected coordinates may be insignificant. Conceptually, we could also derive the median center for each of the two chemicals, but we feel that median center will not provide additional insights into the problem.

2.4.3 Analyses: Dispersion and Orientation

In addition to mean center and standard distance circles, we can analyze the two sets of points using the standard deviational ellipses. Figure 2.8 shows the two ellipses indicated by their axes. The angle of rotation (i.e., the angle from north clockwise to the axis) for the copper ellipse is 44.6 degrees, while for the ethylene ellipse the angle of rotation is only 6.39 degrees. Obviously, the ellipse for ethylene TRIs follows an east-west direction and is relatively small compared to the ellipse for copper TRIs, which is larger and follows a northwest-southeast direction. The difference in orientation of these two sets of points is clearly depicted by the two ellipses.

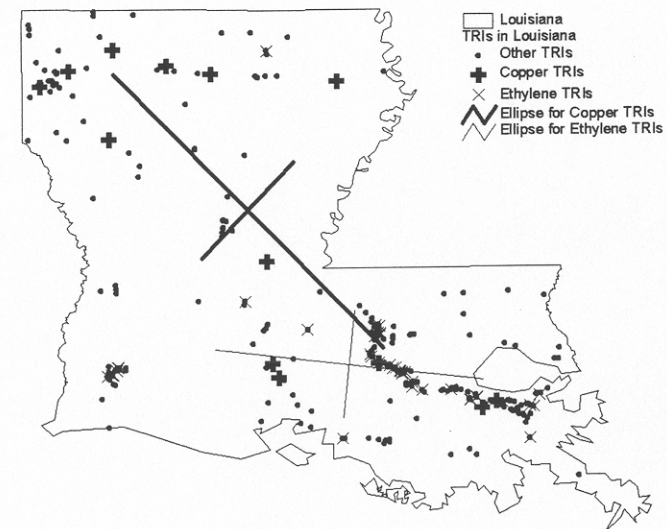



Figure 2.8 The standard deviational ellipses for TRIs releasing copper and ethylene.

ArcView Notes  The setup of the accompanying script to fit the deviational ellipse is similar to that of the other scripts for point descriptors. Users can use the **Query Builder** to select different sets of point to be fitted by the ellipses. This script, however, will produce additional parameters defining the ellipse. These parameters include $\tan \theta$, the lengths of the two axes, and the angle of rotation. As discussed in Section 2.3.2, the angle of rotation has to be adjusted if $\tan \theta$ is negative. The script will make this adjustment automatically when the angle of rotation is reported. In addition, if the script reports a negative $\tan \theta$, it means that the major axis is counterclockwise from the north.

The existing setup of the project file `Ch2.apr` allows users to bring themes into this project for analysis using the built-in functions of the project file. To do this, start ArcView and then use the **Open Project** menu item in the **File** menu to open `Ch2.apr`. When calculating any of the geostatistics discussed in this chapter, use the **Add Theme** menu item in **View** menu to bring in geo-datasets and then choose appropriate menu items from the **Spatial Analysis** menu.

Even though centrophographic measures are very useful in extracting spatial patterns and trends in point sets, there are pitfalls readers should be aware of. First, the statistics may be limited by the boundary effect. In the example here, TRI facilities may not follow state lines between Louisiana and its neighboring states, and the use of the state boundary here may or may not cut off the distribution of TRI facilities. The decision to use this boundary normally requires careful consideration, as the orientation of the boundary may force the point distribution to have spatial patterns that do not exist in reality.

No matter how carefully the analysis is performed, if the data are not properly sampled or properly processed, biased or inaccurate results will cause analyst to make incorrect decisions. This is particularly true in geo-referenced data. For point data, the analyst should explore whether or not the attribute values of any given point are affected by the neighboring points in some way. That is, the analyst must determine if the set of points contains any spatial pattern that is worthy of study.

In later chapters, we will discuss methods for detecting and measuring such spatial patterns. The methods discussed in this chapter are descriptive. They are used to describe and compare point sets as the first step in geographic analysis.

REFERENCES

- Bureau of the Census. (1996). *Statistical Abstract of the United States 1995*. Washington, DC: U.S. Bureau of the Census.
- Environmental Protection Agency (EPA). (1999). <http://www.epa.gov/opptintr/tri/>

Ebdon, D. (1988). *Statistics in Geography*. New York: Basil Blackwell.

Greene, R. (1991). Poverty concentration measures and the urban underclass. *Economic Geography*, 67(3):240–252.

Kellerman, A. (1981). *Centrophographic Measures in Geography. Concepts and Techniques in Modern Geography (CATMOG) No. 32*. Norwich: Geo Book, University of East Angolia.

Levine, N., Kim, K. E., and Nitz, L. H. (1995). Spatial analysis of Honolulu motor vehicle crashes: I. Spatial patterns. *Accident Analysis and Prevention*, 27(5):675–685.

Monmonier, M. (1993). *Mapping it Out*. Chicago: University of Chicago Press.

Thapar, N., Wong, D., and Lee, J. (1999). The changing geography of population centroids in the United States between 1970 and 1990. *The Geographical Bulletin*, 41:45–56.

Wong, D. W. S. (1999). Geostatistics as measures of spatial segregation. *Urban Geography*, 20(7):635–647.

Wong, D. W. S. (2000). Ethnic integration and spatial segregation of the Chinese population. *Asian Ethnicity*, 1:53–72.