3 Analyzing Geographic Distributions and Point Patterns

THEORY

Learning Objectives

This chapter deals with

- Calculating basic statistics for analyzing geographic distributions including mean center, median center, central feature, standard distance and standard deviational ellipse (centrographics)
- Explaining how these metrics can be used to describe spatial arrangements of different sets of point patterns
- Defining locational and spatial outliers
- Introducing the notions of complete spatial randomness, first-order effects and second-order effects
- Analyzing point patterns through average nearest neighbor analysis
- Ripley's *K* function
- Kernel density estimation
- Randomness and the concept of spatial process in creating point patterns

After a thorough study of the theory and lab sections, you will be able to

- Use spatial statistics to describe the distribution of point patterns
- Identify locational and spatial outliers
- Use statistical tools and tests to identify if a spatial point pattern is random, clustered or dispersed
- Use Ripley's *K* and *L* functions to define the appropriate scale of analysis
- Use kernel density functions to produce smooth surfaces of points' intensity over space
- Apply centrographics, conduct point pattern analysis, apply kernel density estimator and trace locational outliers through ArcGIS

3.1 Analyzing Geographic Distributions: Centrography

Centrographic statistics are tools used to analyze geographic distributions by measuring the center, dispersion and directional trend of a spatial arrangement. The centrographic statistics in most common use are the mean center, median center, central feature, standard distance and standard deviational ellipse. Centrographic statistics are calculated based on the location of each feature, which is their major difference with descriptive statistics, which concern only the nonspatial attributes of spatial features.

3.1.1 Mean Center

Definition

Mean center is the geographic center for a set of spatial features. It is a measure of central tendency and is calculated as the average of the x_i and y_i values of the centroids of the spatial features ([3.1]; see Figure 3.1A).

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}, \bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{3.1}$$

n is the number of spatial objects (e.g., points or polygons)
x_i, y_i are the coordinates of the i-th spatial object (centroid in case of polygons)
$\bar{X}, \bar{Y}$ are the coordinates of the mean center

The mean center can be calculated considering weights (3.2). For example, we can calculate the mean center of cities based on their population or income (see Figure 3.1B).

$$\bar{X} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \bar{Y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \tag{3.2}$$

where

w_i is the weight (e.g., income or population) of the i-th spatial object

Why Use

The mean center is used to identify the geographical center of a distribution while the weighted mean center is used to identify the weighted geographical center of a distribution. Mean and weighted centers can be used to compare between the distributions of different types of features (e.g., crime distribution to police station distribution) or the distributions at different time stamps (e.g., crime during the day to crime during the night).

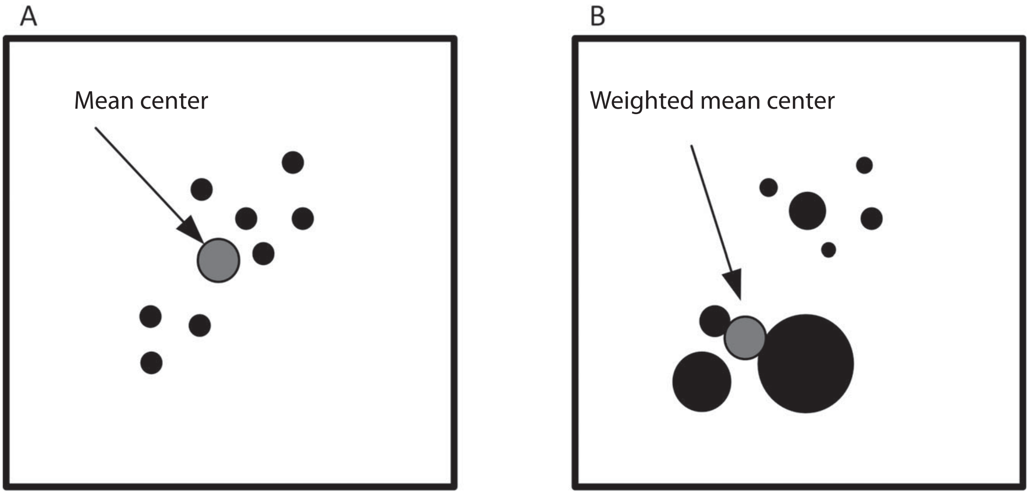

Figure 3.1 (A) Mean center of points' spatial distribution (e.g., cities). The mean center is not an existing point in the dataset. (B) Weighted mean center of a spatial distribution of cities using population as weight. The weighted mean center has been shifted downward relative to the non-weighted mean center (see Figure 3.1A) because the cities in the south are more highly populated.

Interpretation

The mean center offers little information if used on its own for a single geographical distribution. It is used mostly to compare more than one geographic distribution of different phenomena, or in time series data to trace trends in spatial shifts (see Figure 3.2A and B). It should be noted that different geographic distributions might have similar mean centers (see Figure 3.2C). A weighted mean center identifies the location of a weighted attribute. In this respect, it is more descriptive (compared to the mean center), as it indicates where the values of a specific attribute are higher (see Figure 3.1B).

Discussion and Practical Guidelines

The mean center will be greatly distorted where there are locational outliers. Locational outliers should thus be traced before the mean center is calculated. The mean center can be applied to many geographically related problems, such as crime analysis or business geomarketing analysis. Let us consider two examples.

- Crime analysis: Police can identify the central area of high criminality by using a point layer of robberies and calculating the mean center (see Figure 3.1A). This would not be very informative, however, as the mean center would probably lie close to the centroid of the case study area. If

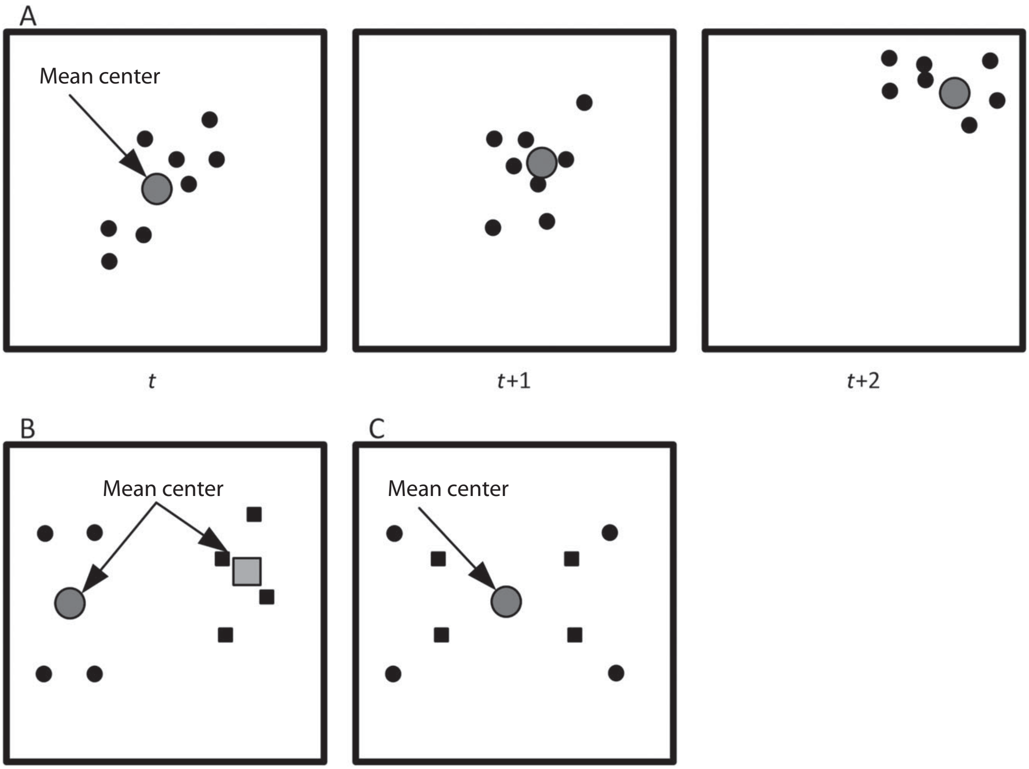

Figure 3.2 (A) Mean center shift for time series data. (B) Different distributions with different mean centers; this is more informative. (C) Different distributions with the same mean center.

data are available for multiple time stamps, the mean center calculation could reveal a shift toward specific directions and thus identify a crime location trend (see Figure 3.2A). Police may further seek to determine why the trend exists and allocate patrols accordingly.

- Geomarketing analysis: Suppose we analyze two different point sets for a city. The first refers to the location of banks (circles) and the second to the location of hotels (squares) (see Figure 3.2B and C). The calculation of the mean center of each point set will probably show different results. We expect that banks will probably be located in the central business district, while hotels closer to the historic center. In this case, the two mean centers will probably lie far apart from each other (see Figure 3.2B). Still, the two mean centers might lie close together if the central business district lies close or engulfs the historic city center (see Figure 3.2C). The mean center is not very informative on its own, but it provides the base upon which more advanced spatial statistics can be built.

3.1.2 Median Center

Definition

Median center is a point that minimizes the travel cost (e.g., distance) from the point itself to all other points (centroids in the case of polygons) in the dataset (see Figure 3.3). It is a measure of central tendency, calculated as shown in (3.3):

$$minimize \sum_{i=1}^{n} d_i \quad (3.3)$$

where

n is the number of spatial objects (e.g., points or polygons)
d_i is the distance of the i-th object to the potential median center

It can also be used with weights (e.g., population, traffic load) calculated as (3.4):

$$minimize \sum_{i=1}^{n} w_i d_i \quad (3.4)$$

where

w_i is the weight of the i-th object. Weights can be positive or negative reflecting the pulling or pushing effects of the events on the location of the median center (Lee & Wong 2001 p. 42).

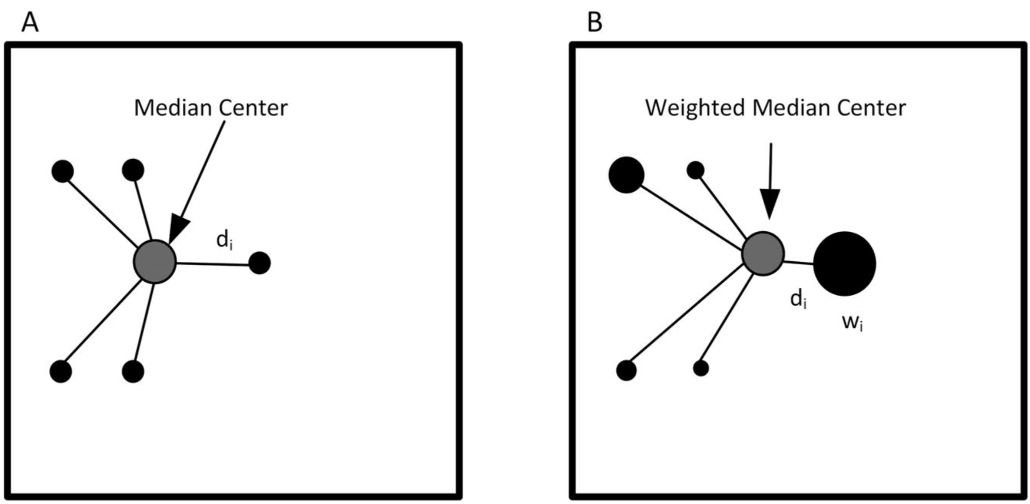

Figure 3.3 (A) Median center. (B) Weighted median center, when using population as weight, shifted to the right.

Why Use

The median center is a measure of the central tendency of spatial data and can be used to find the location that minimizes total travel cost (or weighted cost in the case of a weighted median center).

Interpretation

The median center is a new location and is not necessarily one of the points that exist in the layer. The median center is not as prone to spatial outliers as the mean center is. When there are many spatial objects, the median center is more suitable for spatial analysis. The median center can be also calculated based on weights. For example, to locate a new hospital, we could calculate the median center of postcodes based on their population. The median center will be the location where the total travel cost (total distance) of the population in each postcode to the median center is minimized.

Discussion and Practical Guidelines

There is no direct solution for finding the median center of a spatial dataset. A final solution can be derived only by approximation. The iterative algorithm suggested by Kulin & Kuenne (1962) is a common method used to find the median center (Burt et al. 2009). The algorithm searches for the solution/location that minimizes the total Euclidean distance to all available points. If more than one location minimizes this total cost, the algorithm will calculate only one. Examples related to crime and geomarketing analyses include the following:

- Crime analysis: By calculating the median center of crime, police can locate the point that minimizes police vehicles' total travel cost to the most dangerous (high-crime) areas.
- Geomarketing analysis: Locating a new shop in a way that minimizes the total distance to potential customers in nearby areas.

3.1.3 Central Feature

Definition

Central feature is the object with the minimum total distance to all other features (see Figure 3.4). It is a measure of central tendency. An exhaustive simple algorithm is used to calculate the total distance of a potential central feature to all other features. The one that minimizes the total distance is the central feature (3.5):

$$\textit{find object } j \textit{ the minimizes } \sum\nolimits_{i=1}^{n} d_{ji} \tag{3.5}$$

where

n is the number of spatial objects (e.g., points or polygons)

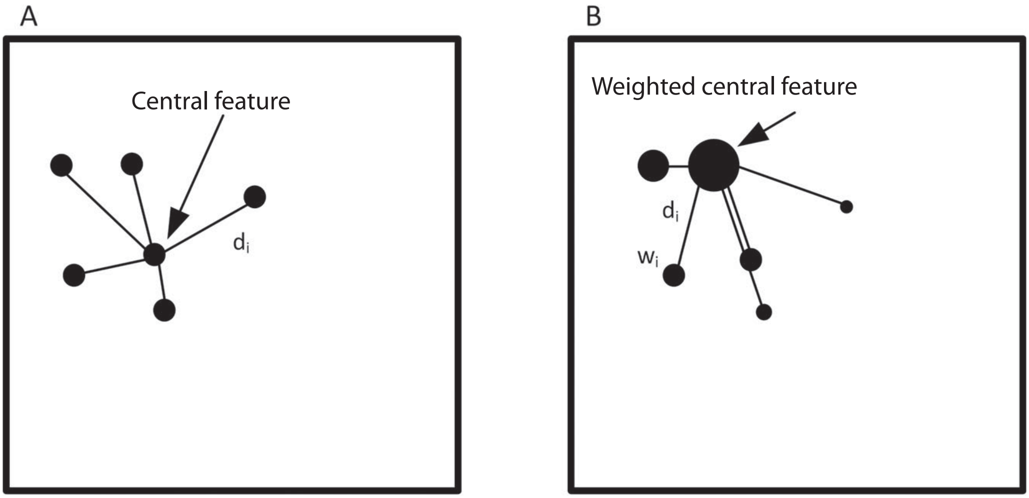

Figure 3.4 (A) The central feature is the existing feature that minimizes total distance. (B) Weighted central feature is the existing feature that minimizes the total distance on some weights (e.g., city population). The points' locations are the same in both graphs.

j is one of the *n* objects selected to be the central feature in an iteration of the algorithm
d_{ij} is the distance of the *i*-th object to the potential central feature *j*. The distance of *j*-th object to itself is zero. Euclidean and Manhattan distance can be used.

It can also be used when weights are selected – e.g., population, traffic load – (3.6):

$$\textit{find object } j \textit{ the minimizes} \sum_{i=1}^{n} w_i d_{ji} \tag{3.6}$$

where

w_{ij} is the weight of the *i*-th object to the potential *j*-th central feature

Why Use

To identify which spatial object from those existing in the dataset is the most centrally located.

Interpretation

The central feature is usually located near the center of the distribution, while the weighted central feature can be located far from the center. In this case, the weighted central feature acts as an attraction pole of a large magnitude.

Discussion and Practical Guidelines

The difference with the median center is that a central feature is an existing object of the database, while the median center is a new point. It can be used if we have to select a specific object from the database rather than a new one – for example, to find which postcode is the central feature in a database. Examples related to crime and geomarketing analyses include the following:

- Crime analysis: By calculating the central feature, police can locate which one of the police stations around the city is most centrally located (i.e., a point layer of police stations).
- Geomarketing analysis: To identify the most centrally located postcode when locating a new shop. Using population as the weight, we could locate the central feature that minimizes the distance between the central postcode and the weighted population.

3.1.4 Standard Distance

Definition

Standard distance is a measure of dispersion (spread) that expresses the compactness of a set of spatial objects. It is represented by a circle the radius of which equals the standard distance, centered on the mean center of the distribution (see Figure 3.5). It is calculated as in (3.7):

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n} + \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n}}. \tag{3.7}$$

n is the total number of spatial objects (e.g., points or polygons)
x_i, y_i are the coordinates of the i-th spatial object
$\bar{X}$, $\bar{Y}$ are the coordinates of the mean center

The weighted standard distance is calculated (Figure 3.5B; [3.8]):

$$SDw = \sqrt{\frac{\sum_{i=1}^{n} w_i(x_i - \bar{X}_w)^2}{\sum_{i=1}^{n} w_i} + \frac{\sum_{i=1}^{n} w_i(y_i - \bar{Y}_w)^2}{\sum_{i=1}^{n} w_i}} \tag{3.8}$$

where

w_i is the weight (the value e.g., income or population) of the i-th spatial object
$\bar{X}_w$, $\bar{Y}_w$ are the coordinates of the weighted mean center

Why Use

To assess the dispersion of features around the mean center. It is analogous to standard deviation in descriptive statistics (Lee & Wong 2001 p. 44).

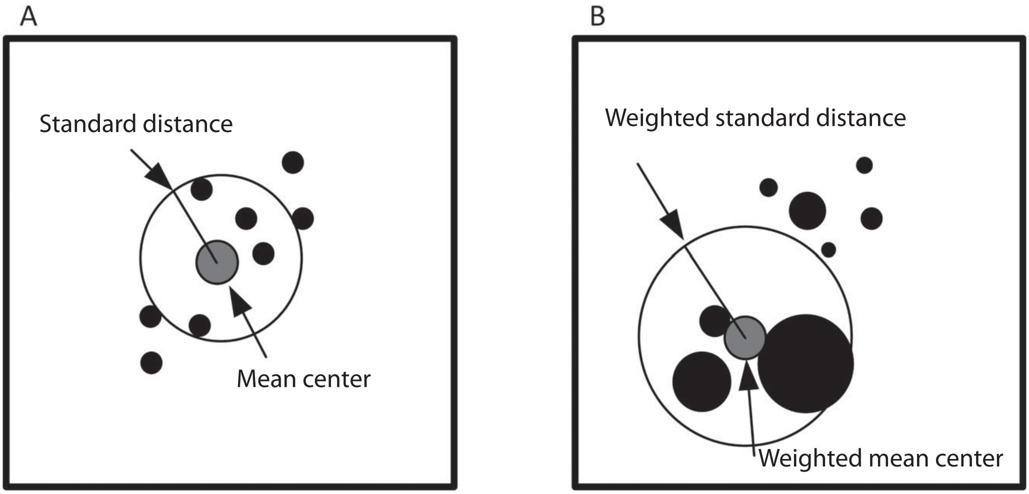

Figure 3.5 (A) Standard distance. (B) Weighted standard distance (weight: city population).

Interpretation

When spatial objects are arranged in such a way that more of them are concentrated in the center with fewer objects scattered towards the periphery (following a Rayleigh distribution) and when we do not use weights, approximately 63% of the spatial objects lie one standard distance from the mean center, and 98% lie within two standard distances from it (Mitchell 2005). The greater the standard distance, the more dispersed the spatial objects are around the mean.

Discussion and Practical Guidelines

Standard distance is a measure of spatial compactness. It is a single measure of spatial objects' distribution around the mean center, similar to statistical standard deviation, which provides a single measure of the values around the statistical mean. Standard distance is more useful when analyzing point layers than when measuring polygons, as polygons are usually human constructed (e.g., administrative boundaries). Examples related to crime and geomarketing analyses include the following:

- Crime analysis: Standard distance can be used to compare crime distributions across several time stamps (e.g., crime during the day relative to crime during the night). If the standard distance during the day is smaller than that at night and is in a different location, this might indicate that police should expand patrols to wider areas during the night (see Figure 3.6A).
- Geomarketing analysis: We could compare the standard distance of different events, such as the spatial distribution of customers of a mall

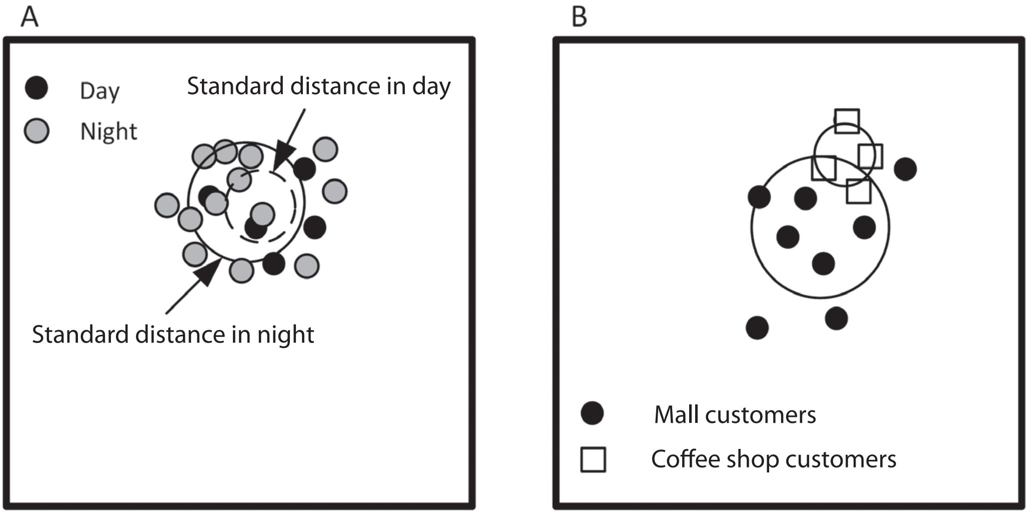

Figure 3.6 (A) Standard distance for the same phenomenon in two different time stamps. (B) Standard distance for two different sets of events. Different standard distance reflects the different dispersion patterns of the two clientele groups.

relative to that of the customers a specific coffee shop. A different standard distance would reveal a more local and less-dispersed pattern for the coffee shop clients compared to those of the mall. An appropriate marketing policy could be initiated based on these findings at the local level (for the coffee shop) and the regional level (for the mall; see Figure 3.6B).

3.1.5 Standard Deviational Ellipse

Definition

Standard deviational ellipse is a measure of dispersion (spread) that calculates standard distance separately in the x and y directions (see Figure 3.7), as in (3.9 and 3.10). The ellipse can also be calculated using the locations influenced by an attribute (weights).

$$SD_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{X})^2}{n}} \tag{3.9}$$

$$SD_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{Y})^2}{n}} \tag{3.10}$$

n is the total number of spatial objects (e.g., points or polygons)
x_i, y_i are the coordinates of the i-th spatial object
$\bar{X}$, $\bar{Y}$ are the coordinates of the mean center

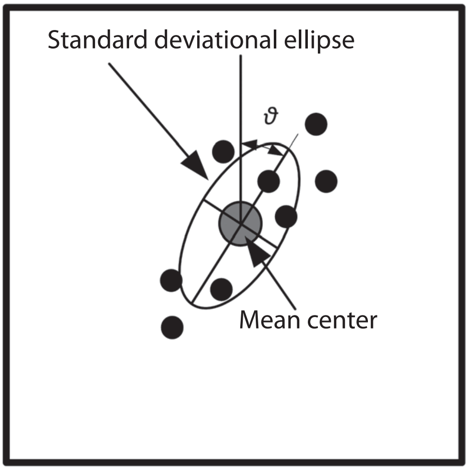

Figure 3.7 Standard deviational ellipse reveals dispersion and directional trend.

These two standard distances are orthogonal to each other and define the standard deviational ellipse. The ellipse is centered on the mean center and is rotated by a particular angle θ form north.

If $x'_i = x_i - \bar{X}$ and $y'_i = y_i - \bar{Y}$ are the deviations of the points form the mean center, the rotation angle is calculated by (3.11):

$$tan\theta = \frac{\left(\sum_i^n x'^2_i - \sum_i^n y'^2_i\right) + \sqrt{\left(\sum_i^n x'^2_i - \sum_i^n y'^2_i\right)^2 + 4\left(\sum_i^n x' \sum_i^n y'\right)^2}}{2\sum_i^n x'_i \sum_i^n y'_i} \tag{3.11}$$

The angle of rotation θ is the angle between north and the major axis. If the sign of the tangent is positive, then the major axis rotates clockwise from the north. If the tangent is negative, then it rotates counterclockwise from north (Lee & Wong 2001 p. 49).

Why Use

The standard deviational ellipse is used to summarize the compactness and the directional trend (bias) of a geographic distribution.

Interpretation

The direction of the ellipse reveals a tendency toward specific directions that can be further analyzed in relation to the problem being studied. Furthermore, we can compare the size and direction of two or more ellipses, reflecting different spatial arrangements. When the underlying spatial arrangement is concentrated in the center with fewer objects lying away from the center, according to a rule of thumb derived from the Rayleigh distribution for two-dimensional data, a one standard deviational ellipse will cover approximately 63% of the spatial features, two standard deviations will contain around 98% percent of the features, and three standard deviations will cover almost all of the features (99.9%; Mitchell 2005). For a nonspatial one-dimensional variable, these percentages are 68%, 95% and 99%, respectively, for the normal distribution and standard deviation statistics.

Discussion and Practical Guidelines

When we calculate a standard deviational ellipse, we obtain (a) an ellipse centered on the mean center of all objects, (b) two standard distances (the length of the long and short axes) and (c) the orientation of the ellipse. Axis rotation is calculated so that the sum of the squares of the distances between the points and axes is minimized. The standard deviational ellipse is more descriptive and is more widely used than standard distance. Examples include the following:

- Crime analysis: A standard deviational ellipse can reveal a direction. If related to other patterns (e.g., bank or shop locations), it could indicate a potential association (e.g., shop burglaries arranged around a specific route). Police can then better plan patrols around these areas (see Figure 3.8A).
- Geomarketing analysis: Plotting the deviational ellipses calculated on the locations of customers who buy specific products allow us to locate the areas where people prefer one product over another. Marketing policies can then be formulated for these locations (see Figure 3.8B).

3.1.6 Locational Outliers and Spatial Outliers

Definition: Locational Outlier

A **locational outlier** is a spatial object that lies far away from its neighbors (see Figure 3.9). As in descriptive statistics, there is no optimal way to define a location outlier. One simple method is to use the same definition as that in descriptive statistics (see Section 2.2.7). Under this definition, an object whose distance to its nearest neighbor exceeds 2.5 deviations from the mean nearest neighbor average (computed for the entire dataset) is considered the locational outlier, as in (3.12):

$$\begin{gathered}\textit{Locational Outlier Distance from nearest neighbor} \geq \\ (\textit{Average Nearesr Neighbor Distance}) + 2.5\ \textit{Standard deviations}\end{gathered} \tag{3.12}$$

Why Use

Locational outliers should be traced, as they tend to distort spatial statistics outputs. For example, if an outlier exists, the mean center will be

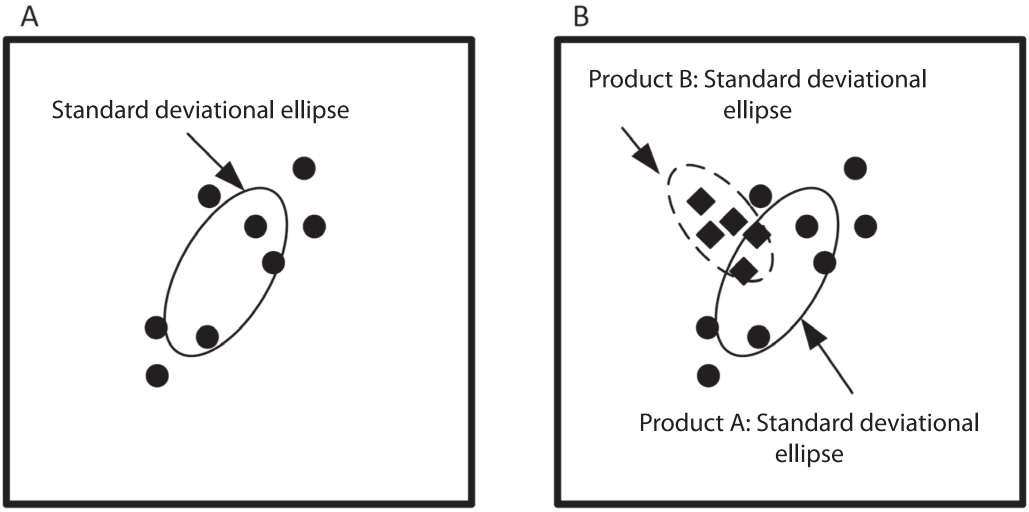

Figure 3.8 (A) Standard deviational ellipse of crimes. A northeast-to-southwest direction is identified. (B) Standard deviational ellipse for two different sets of events. The circles are the locations where customers buy product A, and the squares are the locations where customers buy product B. The directions and dispersion are different, showing that product A penetrates in areas different from those of product B.

significantly different from the mean center when the outlier is not included in the calculations (see Figure 3.9). They may also reveal interesting data patterns (see the Interpretation and Discussion and Practical Guidelines sections that follow).

Interpretation

An outlier typically indicates that we should either remove it from the dataset (at least temporarily) or conduct further research to explain its presence. For example, it may indicate incorrect data entry or a distant location that is abnormal (such as a distant location in which a virus suddenly appears), which should be further studied.

Discussion and Practical Guidelines

Following the definition given earlier, tracing locational outliers requires calculating (a) the nearest neighbor distance of each object (thus creating a distribution of all nearest neighbor distances), (b) the average nearest neighbor distance of all spatial objects, and (c) the standard deviation of the nearest neighbor distances (this is not the standard distance, as discussed in Section 3.1.4).

For polygon features, we can use the area instead of the distance to trace location outliers. We consider locational outliers as those polygons the area of

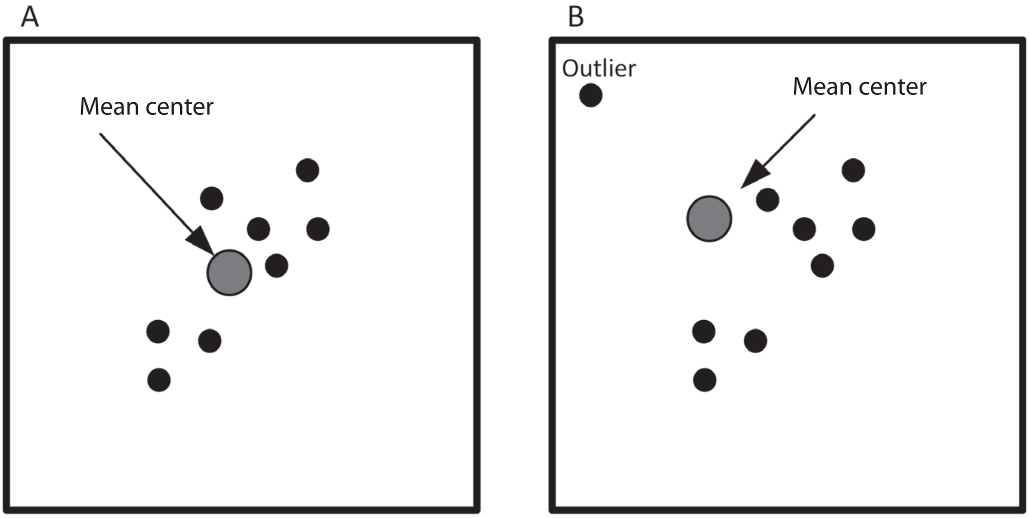

Figure 3.9 (A) Mean center without any locational outlier. (B) Mean center when the outlier is included in the calculation. The mean center is shifted toward the direction of the locational outlier, distorting the real center of the distribution.

which is considerably larger than that of the rest of the objects (again 2.5 or 3 standard deviations away from the mean). In this case, we do not compare nearest neighbors, but we examine the dataset as a whole. Area is related to distance; the larger the area, the more likely it is that the nearest neighbor distance will increase. The calculation of locational outliers using area is not always thorough, as island polygons with small areas may be outliers but will not be classified as such due to their small area value.

Tracing the locational outliers of a polygon shape is essential (especially when calculating spatial autocorrelation; see the next chapter). For example, administrative boundaries, postcodes and school districts are usually human-made. These kinds of spatial data typically have small polygons in the center of a city and larger ones in the outskirts. Thus, the probability of having a locational outlier is high. Spatial statistics tools (in most software packages) estimate locational outliers based on the distances among objects; it is thus better to determine if a polygon is an outlier using distance rather than area. For this reason, polygons should be handled as centroids, and the existence of outliers should be verified based on the distance and distribution of their centroids.

Potential case studies include the following:

- Crime analysis: Identifying if locational outliers in crime incidents exist may reveal abnormal, unexpected behavior that might need additional surveillance. For example, if a credit card that is usually used in specific locations of a city (e.g., shops, restaurants, a house) is also used the same day at a location miles away, that would be a strong indication of fraud.

- Geomarketing analysis: Locational outliers might not be desirable. For example, if a bank wants to locate a new branch, the existence of locational outliers might place the location point in an area that will not be convenient for most of its clientele. The temporary exclusion of the locational outlier will allow for a better branch location. The outlier will then be included in the dataset for other analyses. In other words, when we exclude an outlier, we usually exclude it only for certain spatial statistical procedures, but we include it again because it might be helpful for another type of analysis.

Definition: Spatial Outlier

A **spatial outlier** is a spatial entity whose nonspatial attributes are considerably different from the nonspatial attributes of its neighbors (see Figure 3.10).

Why Use

Spatial outliers reveal anomalies in a spatial dataset that should be further examined.

Interpretation

Spatial outliers can be interpreted differently depending on the problem at hand (see examples and case studies later in this chapter). The basic conclusion we draw when spatial outliers exist is that some locations have attribute values that differ considerably from those of their neighbors. This means that some processes run across space, inferring heterogeneity. Various spatial analysis methods can be used to analyze these processes (see Section 3.2 and Chapter 4).

Discussion and Practical Guidelines

A spatial outlier should not be confused with a locational outlier. To detect locational outliers, we analyze only the distance of a spatial entity to its neighbors. No other attribute analysis takes place. To detect spatial outliers, we study if an attribute value deviates significantly from the attribute values of the neighboring entities. Thus, a spatial outlier does not need to be a locational outlier as well. Additionally, a specific entity may be labeled as a spatial outlier only for a single attribute, while other attribute values might not deviate from the corresponding attribute values of other neighboring entities. Finally, a spatial outlier is not necessarily a global outlier as well, as the spatial outlier is always defined inside a predefined neighborhood.

Let us consider an example. Suppose we study the percentage of flu occurrence within the postcodes of a city (i.e., the percent of people who got the flu in the last month). If we trace a specific postcode with a flu occurrence

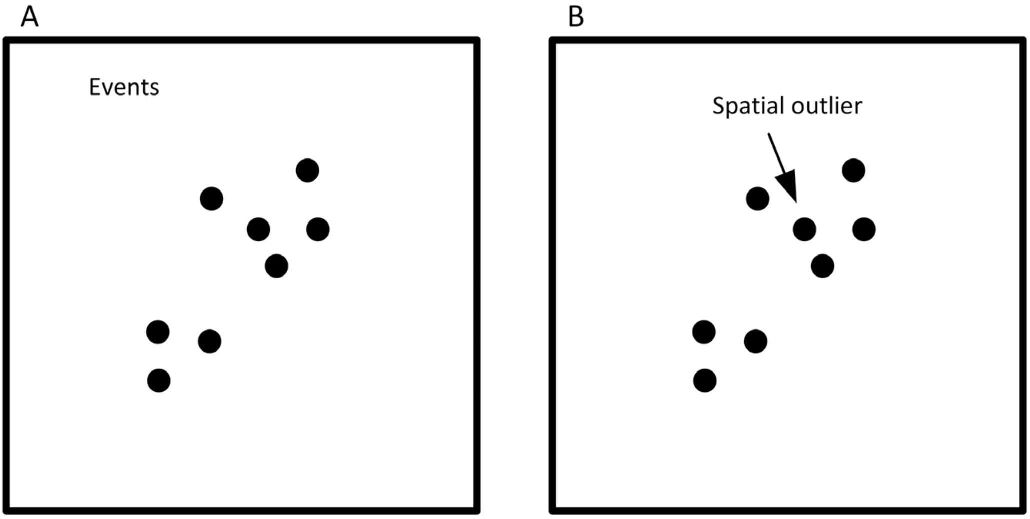

Figure 3.10 (A) Set of events at a set of locations. Each event stands for the location of a house in a neighborhood. (B) If we analyze an attribute (e.g., the size of each house), then a house whose size is considerably different from that of its neighbors is a spatial outlier. For example, the house (event pointed at with an arrow) is a spatial outlier, as its size value is either too large or too small compared to that of its neighbors. As can be observed, a spatial outlier does not have to lie far away from other points.

percentage considerably higher than that of adjacent postcodes, we might label this postcode as a spatial outlier (this might trigger an emergency alert only for this area; Grekousis & Fotis 2012, Grekousis & Liu 2019). This postcode may be centrally located. Then, although it is a spatial outlier, it would not be a locational outlier (neighboring postcodes are in close proximity). If we examine the income attribute of this postcode in relation to the income attributes of the neighboring postcodes, we might find no significant differences in values. This postcode would then be a spatial outlier only for the flu percentage occurrence attribute and not for the income attribute. A spatial outlier does not necessarily mean that the value of the attribute is a global outlier as well. For example, there might be additional postcodes with similarly high flu occurrence percentages but clustered in another area of the city. It is a spatial outlier because it is different (for some attribute) within its neighborhood (so we have to define a neighborhood to trace the spatial outliers). By tracing spatial outliers, we can detect anomalies in many diverse types of data, including environmental, surveillance, health and financial data.

Depicting them in a 3-D graph is a rapid way of detecting spatial outliers. Objects are located on the *X-Y* plain by their geographic coordinates. The value of the attribute for each spatial object is depicted on the Z axis. Those objects that are significantly higher or lower than their neighbors might be spatial outliers. This method is only graphical. To detect outliers based on numerical results, we can use the Local Moran's *I* index (Section 4.4.1) or the optimized outlier analysis (Section 4.4.2), which are explained in the next chapter.

Spatial outlier detection can be used for the following:

- Crime analysis: One could identify if a postcode has a high crime rate while all the neighboring ones have low rates. This might indicate a ghetto.
- Geomarketing analysis: Identifying a specific postcode where people buy a certain product considerably less often than in adjacent ones might indicate where access to the product is not easy or where a better marketing campaign is required.

3.2 Analyzing Spatial Patterns: Point Pattern Analysis

Definitions

Spatial point pattern *S* is a set of locations $S = \{s_1, s_2, s_3, \ldots s_n\}$ in a predefined region, *R*, where *n* events have been recorded (Gatrell et al. 1996; see Figure 3.11). Put simply, a point pattern consists of a set of events at a set of locations, where each event represents a single instance of the phenomenon of interest (O'Sullivan & Unwin 2010 p. 122).

Event is the occurrence of a phenomenon, a state, or an observation at a particular location (O'Sullivan & Unwin 2010 p. 122).

Spatiotemporal point pattern is a spatial pattern of events that evolves over time. In spatiotemporal point patterns, multiple sets of events occur

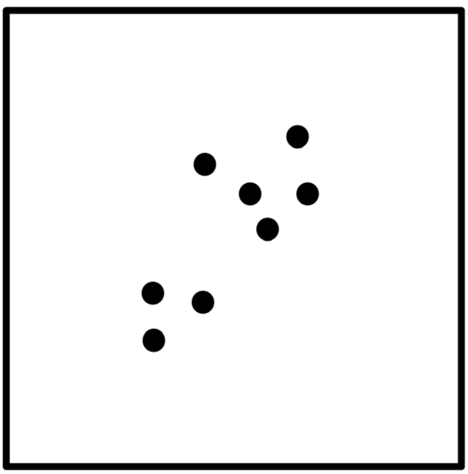

Figure 3.11 Spatial point pattern: Events arranged in a geographic space create a spatial point pattern. A map of events (point pattern) is the realization of a spatial process.

diachronically. These events include several links among places based on physical connections, functional interactions or processes that relate one place to another (Lu & Thill 2003).

Why Use

Point pattern analysis is used mainly to (a) describe the events variation across space and (b) identify relationships among the events by the definition of the spatial process that triggers their arrangement.

Discussion and Practical Guidelines

Sets of point events are widespread in geographical analysis. The locations of shops, crimes, accidents or customers are just a few examples of sets of events that create point patterns in space. Questions can be asked when similar data are available, such as "Where do most of our customers live?" "Where is traffic accident density highest?" and "Are there any crime hot spots in a study area?" Although mapping events provides an initial assessment of their spatial arrangement, more advanced techniques should be used to provide a quantitative evaluation of their geographical distribution. The rationale behind such analysis is deeper. By describing the spatial distribution of events, we attempt to identify the spatial process leading to this formation. The core idea is that there is a spatial process that generates a specific event's arrangement (spatial point pattern) and that this formation is not just the result of a random procedure. In fact, this spatial process is worth further study as the realization of the spatial process in space is the point pattern. Identifying the process allows us to apply the appropriate measures if we want to change the observed pattern (see Figure 3.12).

The term "event" is commonly used in spatial analysis to distinguish the location of an observation from any other arbitrary location within the study area. The term "point" is also often used to describe a point pattern. Events are described by their coordinates $s_i(x_i, y_i)$ and a set of attributes related to the studied phenomenon. The events should be a complete enumeration of the spatial entities being studied and not just a sample (O'Sullivan & Unwin 2010 p. 123). In other words, all available events should be used in point pattern analysis. Although point pattern analysis techniques can be used for samples, the results are very sensitive to missing events. Most point pattern analysis techniques deal only with the location of the events and not with other attributes they might carry (O'Sullivan & Unwin 2010 p. 122). The analysis of attribute value patterns is usually conducted using spatial autocorrelation methods, spatial clustering or spatial regression, as explained in the following chapters (or using geostatistics for field data). The analysis of point patterns may start with centrographic measures, as discussed in Section 3.1, but more sophisticated measures are required, as discussed next.

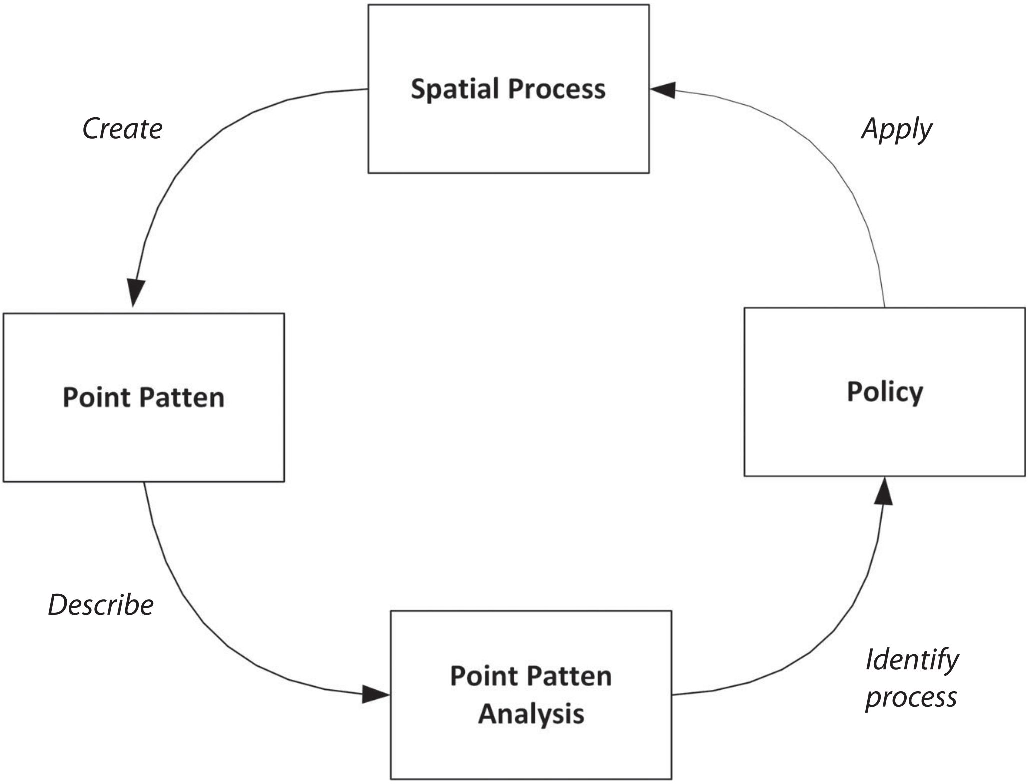

Figure 3.12 A spatial process (not yet known) creates a point pattern that can be described through point pattern analysis. When the spatial process is defined (through the point pattern analysis), measures can be applied to modify the spatial process if needed. For example, the spatial analysis of disease will probably reveal a clustered pattern through an aggregating process. If we locate hot spots, we may apply measures (e.g., vaccination) to specific geographical areas to prevent an expansion of the problem.

3.2.1 Definitions: Spatial Process, Complete Spatial Randomness, First- and Second-Order Effects

Definition

Spatial process is a description of how a spatial pattern can be generated. There are three main types of spatial process (Oyana & Margai 2015 p. 151):

- **Complete spatial randomness process,** also called independent random process (IRP), is a process whereby spatial objects (or their attribute values) are scattered over the geographical space based on two principles (O'Sullivan & Unwin 2010 p. 101):
  - (a) There is an equal probability of event occurrence at any location in the study region (also called first-order stationary).
  - (b) The location of an event is independent of the locations of other events (also called second-order stationary).

- **Competitive process** is a process that leads events to be arranged as far away from each other as possible. The concept is that each event should be located at even, large distance from each neighboring event so that it maximizes its zone of influence/impact. In this respect, events tend to be uniformly distributed.
- **Aggregating process** is a process where events tend to cluster as a result of some pulling action.

There are three main types of spatial arrangement/pattern associated with the above spatial processes (Oyana & Margai 2015 p. 151; see Figure 3.13):

- **Random spatial pattern:** In this type of arrangement, events are randomly scattered all over the study area (see Figure 3.13B). This pattern has a moderated variation and is similar to a Poisson distribution (Oyana & Margai 2015 p. 151). It is the result of an independent random process.
- **Dispersed:** The events are located uniformly around the study area. This is the result of a competitive process, creating a pattern with no or little variation (see Figure 3.13A). Events are located such that they are further away from their neighbors. For instance, the locations of bank branches are more likely to form a dispersed pattern, as there is no reason to have branches of the same bank located near each other.
- **Clustered:** The events create clusters in some parts of the study area, and the pattern has a large variation (see Figure 3.13C). This is the result of an aggregating process. For example, most hotel locations in a city tend to cluster around historical landmarks or major transportation hubs.

First-order spatial variation effect occurs when the values or locations of spatial objects vary from place to place due to a local effect of space (the equal probability assumption of IRP no longer holds; O'Sullivan & Unwin 2003 p. 29). For example, stroke event locations may vary from place to place inside a city

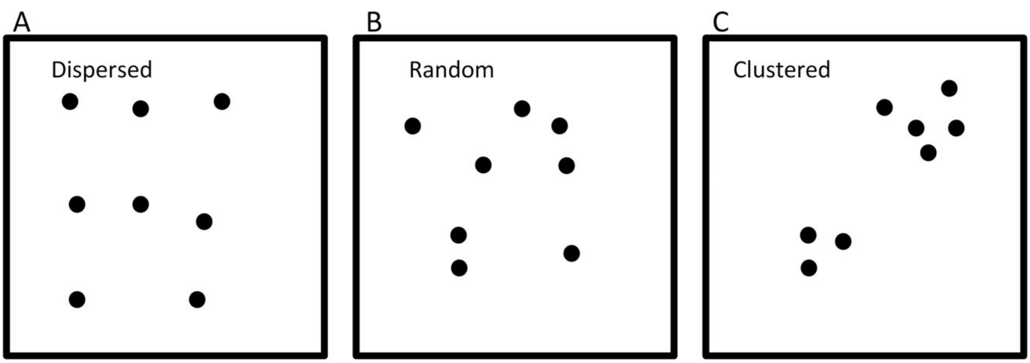

Figure 3.13 (A) In a dispersed pattern, events are scattered in a nearly uniform way. (B) In a random spatial pattern, we cannot identify any clusters or dispersion. (C) In a clustered spatial pattern, clusters will be evident in some parts of the region.

but are expected to be concentrated in places where more of the population is located or in places closer to industrial zones. In other words, events do not occur with the same probability everywhere. This type of spatial effect on the variation of a value or on the location of events is called the first-order effect. First-order effects are mostly associated with density/intensity measures (O'Sullivan & Unwin 2003 p. 112).

Second-order spatial variation effect occurs when there is interaction among nearby locations. Here, the location, or the value of an observation, is highly influenced by a nearby location or the value of a neighboring observation (the independence assumption of IRP no longer holds). For example, immigrants new to a city are more likely to reside in a neighborhood where people of the same ethnicity live, as they would probably feel more comfortable there. In this case, there is a strong local interaction that attracts newcomers; this is a second-order effect of space. Second-order effects are mostly associated with distance measures (O'Sullivan & Unwin 2003 p. 112).

Stationarity is the state in which (a) every location in space has equal probability for an event placement through a spatial process and (b) there is no interaction among events (i.e., independence; O'Sullivan & Unwin 2003 p. 65).

First-order stationary process is a process in which there is no variation in its intensity across space.

Second-order stationary process is a process in which there is no interaction among events.

Intensity of a spatial process is the probability that each small geographical area will receive an event. This probability is the same for every location in a stationary spatial process.

Anisotropic process is a process in which the intensity varies according to specific directions (O'Sullivan & Unwin 2010 p. 108).

Isotropic process, on the other hand, is a spatial process in which directional effects do not exist.

Spatial heterogeneity refers to the non-stationarity of a geographic process (Demšar et al. 2013).

3.2.2 Spatial Process

Each spatial pattern in a particular space and at a specific moment in time is the result of a process occurring within a wider space and time. For a given set of events, it is interesting from the spatial analysis perspective to analyze (a) the observed spatial pattern and (b) the process that created this arrangement (if any). We can summarize the above by posing the following research question: "Can the observed point pattern be the result of a hypothesized spatial process?"

A spatial process can be either (a) deterministic, when inputs and outputs are certain over time, or (b) stochastic, when the outcome is subject to variation and cannot be defined precisely by a mathematical function. Typically, a spatial pattern is the potential realization of some stochastic process. The most common stochastic process is the independent random process (IRP), also known as complete spatial randomness.

First- and second-order effects express the variations in intensity of a spatial process across the study area. A spatial process is first-order stationary if there is no variation in its intensity over space, meaning that there is an equal probability of an event occurring in the study region. A second-order stationary process occurs when there is no interaction among events. Complete spatial randomness is both first- and second-order stationary. Simply put, in complete spatial randomness, there are no first- or second-order effects. When first- and/or second-order effects exist, the chances of an event occurring at some specific location vary (due to dependence and interactions with the neighboring events); thus, the process is no longer stationary (O'Sullivan & Unwin 2003 p. 65). In zero spatial autocorrelation, no first- or second-order effects exist, and location or value variation is random (no patterns are detected). First- and second-order effects are not always easily distinguished, but tracing their existence is extremely important for the accurate modeling of spatial data (O'Sullivan & Unwin 2003 p. 29)

Complete spatial randomness refers to (a) the specific locations in which objects are arranged and (b) the spatial arrangement of the spatial objects' attribute values. In the first case, the techniques used to detect complete spatial randomness only analyze the location of the spatial objects. Most of these techniques are point pattern analysis methods and are used for point features. In the second case, the focus is on how the values of one or more of the variables for fixed locations are spatially arranged. This approach deals mostly with aerial data (polygon features). Polygons are usually artificially created to express certain boundaries. Therefore, instead of analyzing their distribution as polygons, which would offer no valuable results, we study how their attribute values are arranged in space. Spatial autocorrelation methods are used in this case and will be explained in detail in Chapter 4.

However, what is randomness in a spatial context? In nonspatial inferential statistics, we use representative samples to make inferences for the entire population. To do so, we set a null hypothesis and reject it or fail to reject it using a statistical test. In a spatial context, the null hypothesis used is (typically) complete spatial randomness. Under this hypothesis, the observed spatial pattern is the result of a random process. This means that the probability of finding clusters in our data is minimal, and there is no spatial autocorrelation. To better understand the spatial random process and spatial random pattern, consider the following example. Imagine that you have 100 square paper cards. Each card has a single color: red, green or blue. If we throw (rearrange) all the cards onto the floor, we have our first spatial arrangement (spatial pattern). Is this spatial pattern of colors random, or is it clustered? It is most likely that cards will be scattered, and colors will be mixed. The probability that most of the red cards will be clustered together after our first throw is almost zero. If we repeat the throw 1,000 times, we might identify some clusters of red cards in specific regions occurring, say, in 10 out of the 1,000 throws. This shows that the probability of obtaining a cluster of reds is 1%. In other words, the random process (throwing cards on the floor) generated a random spatial arrangement of objects (spatial pattern) for 99.0% of the trials. Thus, the spatial pattern of the objects is random. The statistical procedure is not as simple as that described here, as it includes additional calculations and assumptions.

For example, to reject or not reject the null hypothesis of complete spatial randomness, we use two metrics: a z-score and *p*-value (see Section 2.5.5). The z-score is the critical value used to calculate the associated *p*-value under a standard normal distribution. It reflects the number of standard deviations at which an attribute value lies from its mean. The *p*-value is a probability that defines the confidence level. When the *p*-value is very small, the probability that the observed pattern was created randomly is minimal (Mitchell 2005). It indicates the probability that the observed pattern is the result of a random process.

A small *p*-value in spatial statistical language for the complete spatial randomness hypothesis is translated as follows: It is very unlikely that the observed pattern is the result of a random process, and we can reject the null hypothesis. There is strong evidence that there is an underlying spatial process at play. This is precisely what a geographer is interested in: locating and explaining spatial processes affecting the values of any spatial arrangement or spatial phenomenon.

If the data exhibit spatial randomness, there is no pattern or underlying process. Further geographical analysis is thus unnecessary. If the data do not exhibit spatial randomness, then they do not have the same probability of occurrence in space (first-order effect presence) and/or the location of an event depends on the location of other events (second-order effect presence). In this case, spatial analysis is very important, as space has an effect on event occurrences and their attribute values.

To conclude, when first-order effects exist, the location is a major determinant of event occurrence. When second-order effects exist, the interactions among events are largely influenced by their distance. The absence of a second-order stationary process leads to either uniformity or clustering.

3.3 Point Pattern Analysis Methods

There are two main (interrelated) methods of analyzing point patterns, namely the distance-based methods and the density-based methods.

- **Distance-based** methods employ the distances among events and describe second-order effects. Such methods include the nearest neighbor method

(Clark & Evans 1954; see Section 3.3.1), the *G* and *F* distance functions, the Ripley's *K* distance function (Ripley 1976) and its transformation, the *L* function (see Section 3.3.2).

- **Density-based** methods use the intensity of events occurrence across space. For this reason, they describe first-order effects better. Quadrat count methods and kernel estimation methods are common density-based methods. In quadrat count methods, space is divided into a regular grid (such as a grid of squares or hexagons) of a unitary area. Each unitary region includes a different number of points due to a spatial process. The distribution analysis and its correspondence to a spatial pattern are based on probabilistic and statistical methods. Another, more widely used method is the kernel density estimation (KDE; see Section 3.3.3). This method is better than the quadrat method because it provides a local estimation of the point pattern density at any location of the study area, not only for the locations where events occur (O'Sullivan & Unwin 2003 p. 85).

Another point pattern analysis approach (O'Sullivan & Unwin 2003 p. 127) involves a combination of density and distance, thereby creating proximity polygons. Proximity polygons, such as Delaunay triangulation, and related constructions, such as the Gabriel graph and the minimum spanning tree of a point pattern, display interesting measurable properties. For example, analyzing the distribution of the area among polygons that are close shows how evenly spaced the events in question are. Furthermore, the number of an event's neighbors in Delaunay triangulation and the lengths of the edges are indicators of how the events are distributed in space.

Finally, another way of analyzing events is by using hot spot analysis. This is mainly used to identify if clusters of values of a specific variable are formed in space. It is most commonly used with polygon features and is explained in detail in Section 4.4.2. When considering point features with no associated variables, it is interesting to study the spatial objects' arrangement in terms of intensity levels. This is achieved using optimized hot spot analysis (see Section 4.4.4 for a more in-depth analysis). Strictly speaking, this type of analysis does not assess the type of point pattern or the spatial process that generated it, but it does analyze the location of the events in order to identify any spatial autocorrelation.

3.3.1 Nearest Neighbor Analysis

Definition

Nearest neighbor analysis (also called average nearest neighbor) is a statistical test used to assess the spatial process from which a point pattern has been generated. It is calculated based on the formula (3.13) (Clark & Evans 1954, O'Sullivan & Unwin 2010 p. 142):

$$R = \frac{obsereved\ mean\ distance}{expected\ mean\ distance} = \frac{\bar{d}_{min}}{E(d)} = 2\bar{d}_{min}\sqrt{n/a} \tag{3.13}$$

where

$$\bar{d}_{min} = \frac{\sum_{i=1}^{n} d_{min}(s_i)}{n} \tag{3.14}$$

$$E(d) = 1/\left(2\sqrt{n/a}\right) \tag{3.15}$$

$\bar{d}_{min}$ is the average nearest neighbors distance of the observed spatial pattern
$d_{min}(s_i)$ the distance of event s_i to its nearest neighbor
n is the total number of events
$E(d)$ is the expected value for the mean nearest neighbor distance under complete spatial randomness
a is the area of the study region, or if not defined the minimum enclosing rectangle around all events

Why Use

To decide if the point pattern is random, dispersed or clustered.

Interpretation

A nearest neighbor ratio R less than 1 indicates a process toward clustering. A value larger than 1 indicates that the pattern is dispersed due to a competitive process. A value close to 1 reflects a random pattern.

Discussion and Practical Guidelines

In practice, this method identifies whether the random, dispersed or clustered pattern better describes a given set of events (also called observed spatial pattern). The statistic tests the null hypothesis that the observed pattern is random and is generated by complete spatial randomness. The method compares the observed spatial distribution to a random theoretical one (i.e., Poisson distribution; Oyana & Margai 2015 p. 153). The test outputs the observed mean distance, the expected mean distance (through a homogeneous Poisson point process), the nearest neighbor ratio R, the p-value and the z-score. The expected mean distance is the mean distance the same number of events would most probably have if they were randomly scattered in the same study area. The p-value is the probability that the observed point pattern is the result of complete spatial randomness. The smaller the p-value, the less likely it is that the observed pattern has been generated by complete spatial randomness. If the p-value is larger than the significance level, we cannot reject the null hypothesis (as there is insufficient evidence), but we cannot accept it either (for more on how to interpret *p*-values, see Section 2.5.5). In this case, we have to use other methods (such as those presented later) to determine the observed pattern's type.

A disadvantage of this method is that it summarizes the entire pattern using a single value. It is therefore used mainly as an indication of whether a clustered or a dispersed pattern exists and less often to locate where this process takes place. In addition, this method is highly influenced by the study area's size. A large size might indicate clustering for events for which a smaller size would probably reveal dispersion. Potential case studies include the tracing of clustering in a disease outbreak or identifying if customers of a product are dispersed throughout a region.

3.3.2 Ripley's *K* Function and the *L* Function Transformation

Definition

Ripley's *K* function is a spatial analysis method of analyzing point patterns based on a distance function (Ripley 1976). The outcome of the function is the expected number of events inside a radius of *d* (Oyana & Margai 2015 p. 163). It is calculated as a series of incremental distances *d* centered on each of the events in turn (3.16).

$$K(d) = \frac{a}{n^2}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1, j\neq i}^{n}\frac{I_d\left(d_{ij}\right)}{w_{ij}} \tag{3.16}$$

where

- d is an incremental distance value defined by the user (range of distances that the function is calculated).
- d_{ij} is the distance between a target i event and a neighboring j event.
- n is the total number of events.
- a is the area of the region containing all (n) features.
- I_d is an indicator factor. I_d is 1 if the distance between the i and j events is less than d, otherwise it is 0.
- $w_{i,j}$ are weights most often calculated as the proportion of the circumference of a circle with radius d round the target event (Oyana & Margai 2015 p. 163). They are used to correct for edge effects.

To better understand how this function works, imagine that, for each i event (target), we place a circle of d radius and then count how many events lie inside this circle. The total count is the value of the indicator factor I_d. This procedure is used for all events and for a range of distances d (e.g., every 10 m starting from 50 m to 100 m).

In practice, the original Ripley's *K* function produces large values when the distance increases. Many mathematical transformations of the Ripley's *K* function exist to account for such a problem. A widely used transformation is the *L* function (Bailey & Gatrell 1995; O'Sullivan & Unwin 2010 p. 142). This allows for fast computation and also makes the *L* function linear under a Poisson distribution, which enables easier interpretation, as in (3.17) (Oyana & Margai 2015 p. 163):

$$L(d) = \sqrt{\frac{K(d)}{\pi}} - d \tag{3.17}$$

Why Use

K and *L* functions are used, (a) to decide if the observed point pattern is random, dispersed or clustered at a specific distance or a range of distances and (b) to identify the distance at which the clustering or dispersion is more pronounced.

Interpretation

With the *L* function (Eq. 3.17), the expected value in the case of complete spatial randomness is equal to the input distance *d*. This means that for a random pattern, the *L* function equals to zero; for a clustered pattern, it gets a positive value; and for a dispersed pattern, it gets a negative value. To avoid negative and close-to-zero values, we can omit *d* from Eq. (3.17) and plot *L* against *d* to get a graph like the one in Figure 3.14 (ArcGIS applies this approach with a slightly different denominator for *K*). With this *L* transformation, the expected value is again equal to distance *d*, which can now be used as a reference line at a 45° angle. When the observed value of the *L* function is larger than the expected value for a particular distance or range of distances (see the area above the expected line in Figure 3.14), then the distribution is more clustered than a random distribution.

When the observed value is larger than the upper confidence envelope value, then the spatial clustering is statistically significant (see Figure 3.14; Mitchell 2005). On the other hand, when the observed value of the *L* function is smaller than the expected (the area below the expected line), then the distribution is more dispersed than a distribution that would be the result of complete spatial randomness. Moreover, when the observed value is smaller than the lower confidence interval envelope, the spatial dispersion is statistically significant for this distance (see Figure 3.14).

The confidence envelope is the area in which the expected values would lie in a random pattern for a specific confidence level (see the Discussion and Practical Guidelines section). To simulate this process, the points are randomly distributed many times (e.g., 99 or 999 times) using the Monte Carlo approach. Each one of these times is called a permutation, and the expected value is

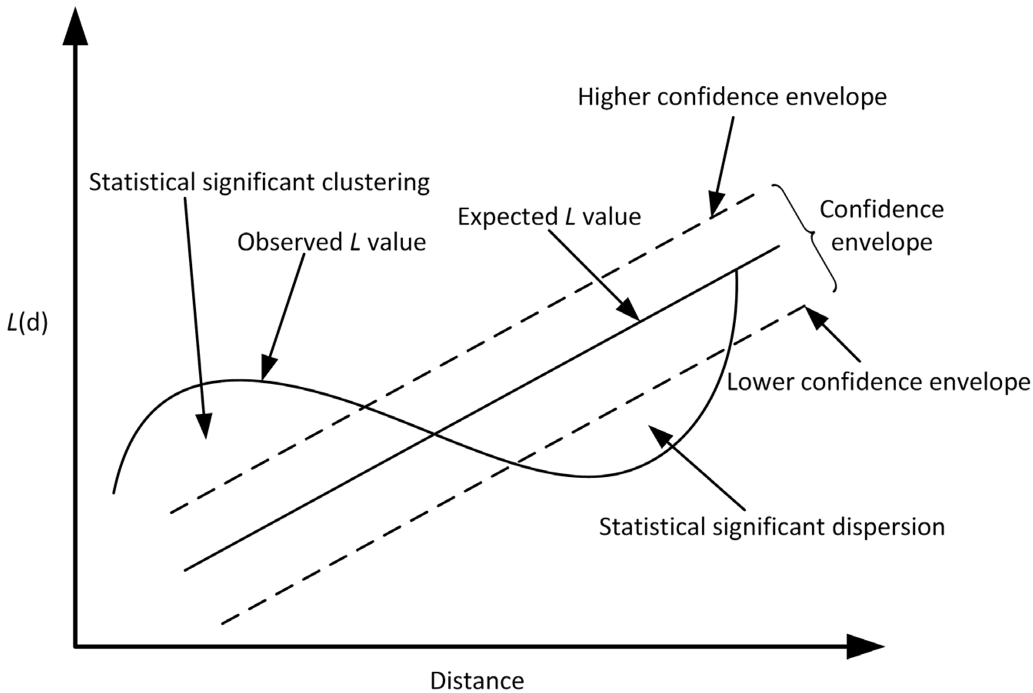

Figure 3.14 *L* function transformation of Ripley's *K* function over a range of distances.

calculated for each permutation, creating a sampling distribution. For 999 permutations, the *p*-value is typically set to 0.001; for 99 permutations, it is set to 0.01. These reflect 99% and 99.9% confidence intervals, respectively. When a value lies outside the envelope area, the null hypothesis for complete spatial randomness is rejected.

Discussion and Practical Guidelines

The process used to calculate the *L* function occurs in the following steps (Oyana & Margai 2015 p. 164):

1. Calculate the observed value *K* (using the *K* function) of the observed set of events. The expected value is obtained by placing a buffer with radius *d* over a target event and then counting how many events lie inside this zone. Follow the same process for each single event in the dataset. Then calculate the average count of events within a range of distances.
2. Transform the *K* function estimates to an *L* function to make it linear.
3. With the *L* transformation, the expected *K* value is equal to distance *d*.
4. Determine the confidence envelope by estimating the minimum and maximum *L* values using permutations and the related significance levels under the null hypothesis of complete spatial randomness.

5. Plot L on a graph to reveal if clustering or dispersion is evident at various distances by comparing the expected to the observed values and according to whether the observed values lie inside or outside the confidence envelope.

The K and L functions are sensitive to the definition of the bounding rectangle in which the features are contained. This raises two issues. First, the study region area has a direct effect on the density parameter; second, it gives rise to the edge effect problem (see Section 1.3), in which the events that lie in the edge of the study area tend to have fewer neighbors than those lying in central locations. In the first case, identical arrangements of events are likely to yield different results for various study area sizes. For example, if the size doubles but the points are concentrated in the center, this might result in a clustered pattern. For the exact same event arrangement, however, if the study area is defined by the minimum enclosing rectangle, the pattern might be dispersed or random.

On the other hand, a tight study area can cause the edge effect problem. Events close to the region boundaries tend to have larger nearest neighbor distances, although they might have neighbors just outside the boundaries that lie in closer proximity than those lying inside (O'Sullivan & Unwin 2003 p. 95). Various methods have been developed to account for edge effects (known as edge correction techniques), including guard zones, simulating outer boundaries, shrinking the study area and Ripley's edge correction formula.

The Monte Carlo simulation approach is a more reliable method of accounting for edge effects and the study region area effect. A Monte Carlo procedure generates and allocates n events randomly over the study region hundreds or thousands of times (permutations), creating a sampling distribution. The observed pattern is then compared with the patterns generated under complete spatial randomness through Monte Carlo simulation. The results of complete spatial randomness are also used to construct an envelope inside of which the L value of a random spatial pattern is expected to lie. If there is a statistically significant difference between the observed and the simulated patterns (those lying outside the envelope), then we may reject the null hypothesis that the observed pattern is the result of complete spatial randomness. Since each permutation is subject to the same edge effects, the obtained sampling distribution of the Monte Carlo procedure accounts for both edge and study region area effects simultaneously without the need to apply other corrections (O'Sullivan & Unwin 2010 p. 150).

Spatial patterns change when studied at multiple distances and spatial scales, reflecting the existence of particular spatial processes at specific distance ranges (Mitchell 2005). In other words, a spatial pattern might be random for some range of distances and clustered for others. Ripley's K function illustrates how the spatial clustering or dispersion of features' centroids changes when the neighborhood size changes by summarizing the clustering or dispersion over a range of different distances something helpful for assessing the proper scale of analysis for the problem at hand. This is similar to how the scale of analysis is defined in spatial autocorrelation studies (see Chapter 4). When we analyze a point distribution in different distances, we may trace the distance at which the clustering starts and ends and the distance at which the dispersion begins and ends. Thus, based on the scope of the research, for a single-point dataset, we may address different questions relating to the scale of analysis and the patterns of clustering at various distances (i.e., study how the clustering of crime changes at different distances). In addition, using this type of function does not provide a single description for the pattern in question (as we obtain when we use the average nearest neighbor analysis) but provides deeper insight, as we describe it according to the distance desired.

3.3.3 Kernel Density Function

Definition

Kernel density estimation is a nonparametric method that uses kernel functions to create smooth maps of density values, in which the density at each location indicates the concentration of points within the neighboring area (high concentrations as peaks, low concentrations as valleys; see Figure 3.15). The kernel density at a specific point is estimated using a kernel function with two or more dimensions. The term "kernel" in spatial statistics is used mainly to refer to a window function centered over an area that systematically moves to each location while calculating the respected formula. Kernel functions weigh events according to distance so that those that are closer are weighted more than events further away (O'Sullivan & Unwin 2003 p. 87).

A typical kernel function (Gaussian) looks like a bell, is centered over each event and is calculated, within bandwidth (h), based on Eq. (3.18) (Wang 2014 p. 49, Oyana & Margai 2015 p. 167; see Figure 3.15):

$$\hat{f}(x) = \frac{1}{nh^r} \sum_{i=1}^{n} k\left(\frac{d}{h}\right) \tag{3.18}$$

where

h is the bandwidth (the radius of the search area)
n is the number of events inside the search radius
r is the data dimensionality
$k()$ is the kernel function
d is the distance between the central event s_i and the s event inside the bandwidth

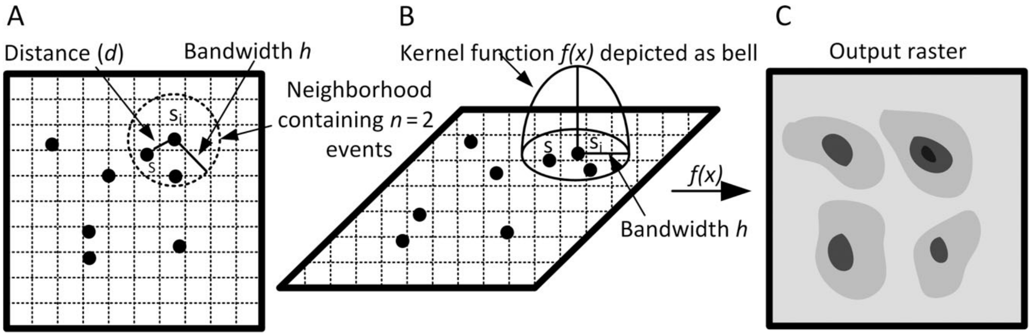

Figure 3.15 (A) Kernel density function is calculated for every single event in turn. (B) We can imagine a kernel as a bell with the same total weight of one unit. The total weight of all bells equals the total number of points (Longley et al. 2011 p. 372). Failing to retain the total population would suggest that there are more or fewer events in the study area than there really are (O'Sullivan & Unwin 2003 p. 87). The shape of the kernel depends on the bandwidth parameter *h*. A large bandwidth results in a broader and lower height kernel, while a smaller bandwidth leads to a taller kernel with a smaller base. Large bandwidths are more suitable for revealing regional patterns, while smaller bandwidths are more suitable for local analysis (Fotheringham et al. 2000 p. 46). When all events are replaced by their kernels and all kernels are added, then a raster density surface is created, as seen in C. (C) Darker areas indicate higher intensity while lighter areas indicate lower intensity. The output raster allows for a better understanding of how events are arranged in space, as it renders every single cell, creating a smooth surface that also reflects the probability of event occurrence.

The kernel function calculates the probability density of an event at some distance from a reference point (Oyana & Margai 2015 p. 168).

Various kernel types can be plugged into the preceding formula, including normal, uniform, quadratic and Gaussian types. ArcGIS uses the quadratic kernel function described in Silverman (Silverman 1986, Wang 2014 p. 50). The functional form is as shown in (3.19):

$$\hat{f}(x) = \frac{1}{nh^2\pi}\sum_{i=1}^{n}\left(1-\frac{d^2}{h^2}\right)^2 \tag{3.19}$$

Why Use

Kernel density estimation is used to create smooth surfaces that depict, first, the density of events and, second, an estimation of areas of higher or lower event occurrence intensity. Through the raster map visualization of a kernel density estimation, one can quickly locate hot spot and cold spot areas and places of high or lower density (see Figure 3.16). It can also be applied for cluster analysis of weighted or unweighted events in epidemiology, criminology, demography, ethnology and urban analysis.

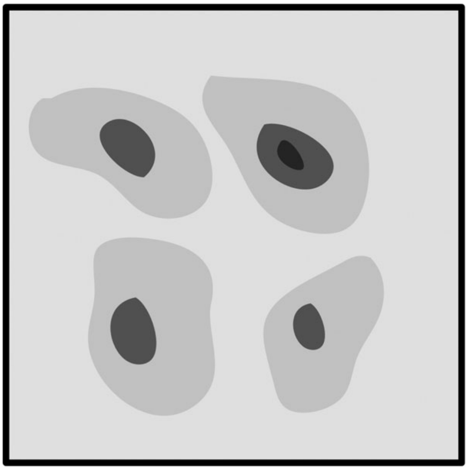

Figure 3.16 Kernel density estimation. A raster output depicts the intensity of events in the study area.

Interpretation

Kernel density estimates the density of the point features, rather than their values (e.g., temperature or height; Silverman 1986, Bailey & Gatrell 1995). The density estimates can be displayed by either surface maps or contour maps that show the intensity at all locations (see Figure 3.15C). Peaks reveal higher concentrations and valleys lower densities.

Discussion and Practical Guidelines

The kernel function provides an interpolation technique for assessing the impact of each event on its neighborhood (as defined by the bandwidth; Oyana & Margai 2015 p. 168). In practice, the kernel function is applied to a limited area around each event defined by the bandwidth *h* to "spread" its effect across space (see Figure 3.15A). Outside the specified bandwidth, the value of the function is zero. For example, the quadratic kernel function falls off gradually to zero with distance until the radius is reached. Each individual event's impact is depicted as a 3-D surface (e.g., like a bell; see Figure 3.15B). The final density is calculated by adding, for each raster cell, the intersections of the individuals' surfaces.

The main advantage of kernel density estimation is that the resulting density function is continuous at all points along the scale. In addition, kernel density estimation weighs nearby events more heavily than those lying further away using a weighting function, thus applying an attenuating effect in space. This is a major difference with a typical point density calculation, in which density is calculated for each point based on the number of surrounding objects divided by the surrounding area, with no differentiation based on proximity or any other weighting scheme. Kernel density spreads the known quantity to the raster cell inside the specified radius.

A weak point of this technique is the subjectivity in the definition of the bandwidth factor h. The selection of an appropriate value for h is usually made by trial and error. Large bandwidth h values create a less-detailed raster. On the other hand, small values create a raster with a more realistic density (see note on Figure 3.15B; Mitchell 2005). The selection of the appropriate bandwidth depends on the problem at hand and the desired analysis scale. As mentioned in the previous section, both Ripley's K and its transformed L function provide a graph depicting the distances at which clustering or dispersion is more pronounced. One can first apply the L function and then select the distance that better fits the scope of the analysis which also reflects the spatial process at play. Optimized hot spot analysis may also be used to identify the distance at which spatial autocorrelation is more evident and then be used as the bandwidth h (see Section 4.4.4). Another approach is to use as the initial bandwidth value the one that results from formula (3.20) (ESRI 2014):

$$h = 0.9 \min\left(SD, \sqrt{\frac{1}{\ln(2)}} D_m\right) n^{-0.2} \tag{3.20}$$

where

h is the bandwidth (the radius of the search area),
n is the number of events. If a count field is used then, n is the sum of the counts
D_m is the median distance
SD is the standard distance
min means that the minimum value between SD and $\sqrt{\frac{1}{\ln(2)}} D_m$ will be used in (3.20)

Locational outliers should be removed when selecting the appropriate bandwidth value, as they might lead to a very smooth raster layer. Finally, a weight may also be given to each event (e.g., number of accidents at intersections or the number of floors in a building). The weight is the quantity that will spread across the study region through the kernel density estimation and the created surface. This type of analysis might be useful for cross-comparisons with other aerial data – for example, to study respiratory problems of people (weighted or unweighted events) in relation to PM2.5 (particulate matter) by using an air pollutant raster map to identify potential spatial correlations and overlapping.

3.4 Chapter Concluding Remarks

- Point pattern analysis identifies if a point pattern is random, clustered or dispersed. For analyzing the spatial distribution of the attribute values of spatial objects, we apply spatial autocorrelation methods presented in Chapter 4.

- Standard distance is a measure of dispersion (spread). It is a distance expressing the compactness of a set of spatial objects.
- The standard deviational ellipse is more descriptive and more widely used than standard distance.
- Spatial outliers reveal anomalies in a spatial dataset that should be further examined.
- By describing the spatial distribution of events, we attempt to identify the spatial process that leads to this formation.
- Complete spatial randomness process (also called the independent random process, or IRP), is a process wherein spatial objects (or their attribute values) are scattered over the geographical space randomly.
- The first-order spatial variation effect occurs when the values or the location of spatial objects vary from place to place due to local effect of space.
- The second-order spatial variation effect occurs when there is an interaction among nearby locations. The location, or the value of an observation, is highly influenced by a nearby location or the value of a neighboring observation (the independence assumption of IRP no longer holds).
- Stationarity is the state where (a) every location in space has an equal probability of event placement through a spatial process and (b) there is no interaction among events (independence).
- An anisotropic process occurs when intensity varies according to specific directions.
- *L* and *K* functions summarize clustering or dispersion over a range of different distances. This assists in assessing the proper scale of analysis.
- The main advantage of kernel density estimation is that the resulting density function is continuous at all points along the scale.
- Kernel density estimation weights nearby events more heavily than those lying further away using a weighting function, thus applying an attenuating effect in space.
- Kernel density estimates the density of the points rather than their values.
- The kernel function calculates the probability density of an event at some distance based on an observed reference point event.
- A weak point of kernel density estimation is the subjectivity in the definition of the bandwidth factor *h*.