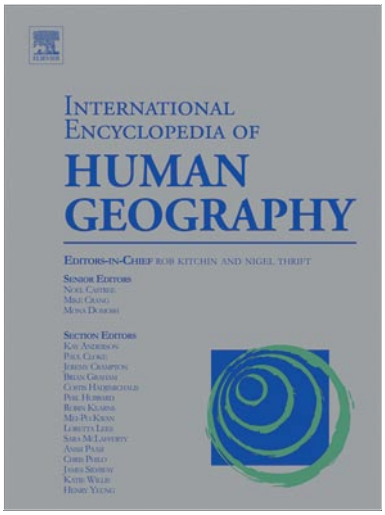

Delmelle E. 2009. Point Pattern Analysis. In Kitchin R, Thrift N (eds) International Encyclopedia of Human Geography, Volume 8, pp. 204–211. Oxford: Elsevier.
ISBN: 978-0-08-044911-1

Point Pattern Analysis

E. Delmelle, University of North Carolina at Charlotte, NC, USA

Glossary

Delaunay Tessellation A geometrical triangulation technique based on Thiessen polygons connecting nearby points, resulting in triangles. This method is extremely useful in determining the neighbors in the direct vicinity of an event.

Monte Carlo Simulation A process by which a set of events is simulated randomly several times. The ultimate goal is to derive a sampling distribution via simulation of the null hypothesis. Monte Carlo simulations create several realizations of a phenomenon as it potentially varies over space.

Moving Window A filtering process, in which the value at the center of the moving window is computed as a function of the surrounding cells, or events located within the window.

Poisson Process A nondeterministic process used to model the locations of random incidents in geographic space. Random incidents are independently occurring. The Poisson parameter λ stands for the intensity of the process.

Smoothing A process (or result of a process) where values are averaged based on surrounding values, following a moving window for instance. Smoothing also refers to procedures that amend an existing grid.

Basic Visual Exploration of Point Patterns

A spatial point pattern refers to data in the form of points, where a point denotes the location of an event. In geography, it is usually desirable to analyze whether these particular events, such as crimes, car accidents, fires, emergency calls and diseases for instance, exhibit a spatial pattern (e.g., hot spot), in the hope to better understand the underlying process that generated the events. Such events are considered discrete, because they occur at specific locations. A visual inspection of a map showing the locations of those events (e.g., scatter plot) may not always bring a correct interpretation of the true pattern, especially when events occur repeatedly at the same location. True clusters may go unnoticed. Events that occur repeatedly in time at the same location can be represented using a bubble plot, where the size of the bubble is a function of the event frequency.

General Descriptive Methods

A wide variety of descriptive statistical techniques exist to describe the geographical characteristics of a point pattern, such as its central tendency (the 'center' of the point pattern) or its dispersion (the degree of separation or clustering among the points). To measure central tendency, a first step consists of computing the mean center of a set of events. The mean center is the central or average location of a set of points, computed as the mean x- and y-coordinate values for all the events in the study region

$$\bar{x}_{\text{mc}} = \frac{\sum_{i=1}^{n} x_i}{n}, \bar{y}_{\text{mc}} = \frac{\sum_{i=1}^{n} y_i}{n} \quad [1]$$

where $\bar{x}_{\text{mc}}$ and $\bar{y}_{\text{mc}}$ are the coordinates of the spatial mean, and x_i and y_i are the coordinates of an event i, and n is the number of events. The mean center is significantly affected by the presence of outliers as well as the frequency of occurrence in the case an incident happens more than once at the same location. The introduction of weights in eqn [1] can reflect the importance of some events, which extends eqn [1] to the notion of weighted spatial mean. The weight is an interval or ratio value associated with a feature attribute. The weighted mean center is given by

$$\bar{x}_{\text{wmc}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \bar{y}_{\text{wmc}} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \quad [2]$$

where the subscript 'wmc' stands for weighted mean center. The addition of weight is particularly useful for areal data since some regions can be larger, or simply more important than others. The weighted mean center is an important indicator, especially in locating facilities, which will serve the entire population. The location of a new emergency service is an excellent example since we try to minimize the distance to the population.

Dispersion of Point Distribution

The spatial mean indicates the central location of a set of events, but does not reflect the dispersion of the point distribution. Indeed, two point patterns may have the same mean center, yet one may be very clustered around that mean center, whereas the other may have a highly dispersed, uniform pattern. The standard distance *SD*

indicates the dispersion of point events from the spatial mean.

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x_{mc})^2 + \sum_{i=1}^{n}(y_i - y_{mc})^2}{n}} \quad [3]$$

When point events are very dispersed, the *SD* value will be higher than when incidents are clustered around the spatial mean.

Orientation of Point Distribution

In which direction do events tend to cluster? To illustrate the spatial spread of a set of point locations, the use of a standard deviational ellipse is a common and effective visualization tool, because it captures the directional bias in a point distribution. For instance, events occurring along a road network, such as bike accidents, will exhibit a linear pattern. The deviational ellipse is characterized by three parameters: the angle of rotation θ, deviation along major axis a, and the deviation along the minor axis b. As a result, we obtain the directions of maximum and minimum spread. Figure 1 illustrates the mean center and directional ellipse for two distinct point patterns. The point pattern represents the origins of patients being treated at a city hospital in Cali, Colombia for the months of September and July 2004. A larger ellipse denotes that patients are clearly more dispersed in September; the north-northeast (NNE) direction of the ellipse is affected by the presence of patients in the northern and northwestern part of the city. In July, however, the ellipse is compacted (closer to a circle), and points in the east-northeast (ENE) direction. As a result, the mean center is closer to the hospital.

The absolute coordinate values of event i (x_i, y_i) are transformed to relative values (x_i', y_i') based on the location of the mean center. The center of the transformed coordinate becomes (0, 0).

$$\begin{aligned} x_i' &= x_i - x_{mc} \\ y_i' &= y_i - y_{mc} \end{aligned} \quad [4]$$

The angle of rotation θ is given by:

$$\tan\theta = \frac{\left(\sum_{i=1}^{n} x_i'^2 - \sum_{i=1}^{n} y_i'^2\right) + \sqrt{\left(\sum_{i=1}^{n} x_i'^2 - \sum_{i=1}^{n} y_i'^2\right)^2 + 4\left(\sum_{i=1}^{n} x_i' y_i'\right)^2}}{2\sum_{i=1}^{n} x_i' y_i'} \quad [5]$$

where a negative tangent implies a rotation of the angle counterclockwise, and a positive tangent implies a clockwise rotation, respectively. Finally, to reconstruct the ellipses, the deviations along the x- and y-axes must be known:

$$\begin{aligned} \delta_x &= \sqrt{\frac{\sum_{i=1}^{n}(x_i' \cos\theta - y_i' \sin\theta)^2}{n}}, \\ \delta_y &= \sqrt{\frac{\sum_{i=1}^{n}(x_i' \sin\theta + y_i' \cos\theta)^2}{n}} \end{aligned} \quad [6]$$

General Clustering Methods

Quadrat Analysis

The general descriptive techniques discussed in the previous section do not inform on the clustering level of

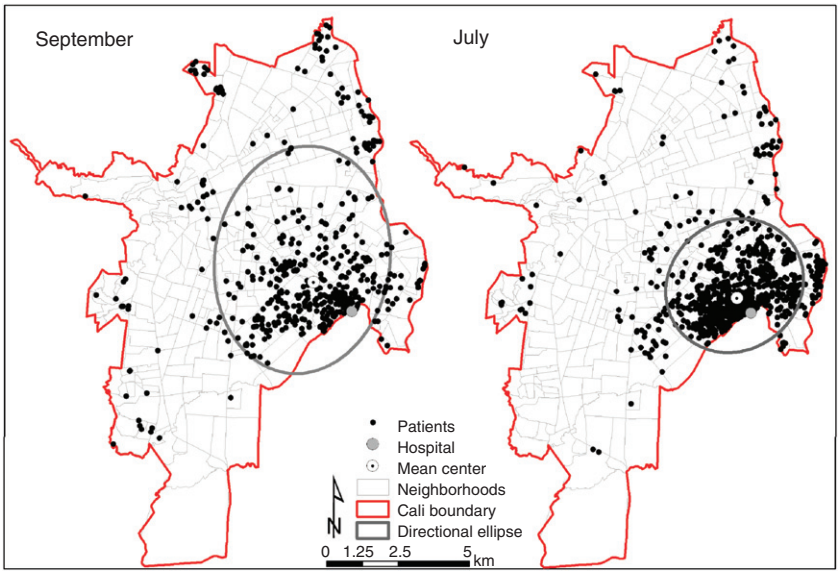

Figure 1 Patients being treated at a city hospital in Cali, Colombia, for the months of July and September 2004.

the data; rather they indicate the potential spread and orientation of the dataset. One spatial statistical method to measure potential clustering among events is a quadrat analysis. It consists of overlaying a uniform grid, which is made out of equal-sized quadrats, onto the map containing point incidents. The number of events within each cell (or quadrat) is determined. Essentially, the analysis compares the actual count of events within each quadrat with the expected frequency of occurrence if the process generating those events was random in nature (known as complete spatial randomness or CSR). Complete spatial random processes, also called Poisson processes, are characterized by the fact that each point is likely to occur at any location. In general, we are interested in the hypothesis that the current phenomenon does not follow a Poisson process – in other words, that points are more clustered or more dispersed than would be expected under a Poisson process.

Define $P(X=x)$ as the probability of finding x events within a specific quadrat. Assuming a Poisson process, this probability can be calculated as follows:

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad [7]$$

where λ is the intensity of the process. The true value of the term λ is usually unknown, but can be computed as the average number of events per quadrat. For instance, if a grid of 100 cells is imposed over a study area, and 67 events occur within this area, a good estimate of intensity is $\hat{\lambda} = .67$. The probability of finding no event within a cell is $P(X=0)=.5117$, one event $P(X=1)=.3428$, and two events $P(X=2)=.1149$.

Determining the presence or absence of clustering

The variance mean ratio (VMR) is a formula for measuring the degree of clustering within a study region based on quadrat analysis.

$$\text{VMR} = \frac{\frac{\sum_{i=1}^{m}(x_i - \lambda)^2}{m-1}}{\lambda} \quad [8]$$

where m is the number of cells of equal size, λ the mean number of points per cell, and x_i the number of cells containing i events. The numerator stands for the variance of the frequency of cells containing i events, and the VMR index is then used to standardize this variance relative to the mean cell frequency. A VMR value less than 1 characterizes patterns that exhibit a tendency toward uniformity. However, when the VMR value is greater than 1, cells are characterized by either a much greater or a much lower number of events than expected which indicates a highly clustered spatial pattern. A VMR value close to 1 is typical of a random, Poisson process.

Example

Figure 2 illustrates a simulated point process of 100 events. Half of those points are generated in a random fashion, while the remaining half is explicitly clustered within three blocks, two of size 20 by 20, and one of size 7 by 7. The size of the study region is 100 by 100 units. If we divide the area in four rows and four columns for instance, it creates 16 squared cells of 25 side units each. The corresponding VMR ratio is 2.54, characterizing a pattern of strong spatial clustering, since there is a substantially greater and lower number of events in each of these 16 cells than expected.

Advantages and limitations of the technique

Besides the relative easiness to implement the quadrat analysis and the quick results it provides, the technique has some drawbacks, mostly related to cell size, and the ability of the method to differentiate different spatial patterns of points.

The size of the quadrat is of paramount importance. A cell too small in size may cause a high variability in quadrat counts, which will result in several empty cells and consequently clustering may go unnoticed. When the cell size is too coarse on the other hand, the within-cell patterns will be missed. Figure 3 shows the variation of the VMR value as a function of the cell size. As could be expected, when the cell size is too small, the spatial pattern is unnoticed. It is usually desirable to fix the cell size to obtain an average of 1.6 to 2 events per cell. As opposed to the nearest neighbor approach (see below), quadrat analysis does not look at the interseparation distance between events. Additionally, the technique solely relies on frequency counts and the spatial arrangement of the events is not explicitly

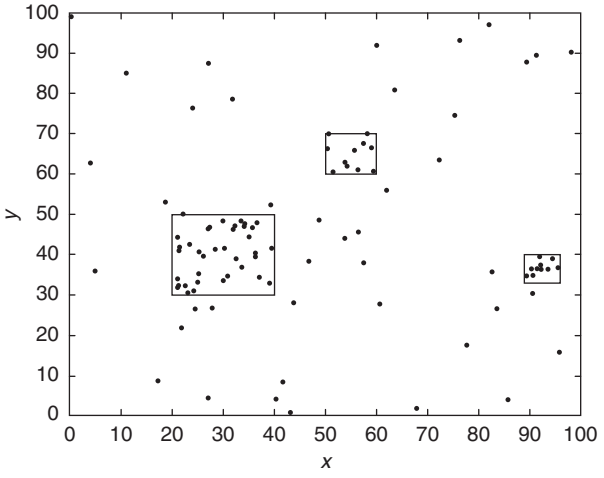

Figure 2 Simulated point pattern where half of the points are generated at random, and the other half is clustered within the three boxes.

considered. In Figure 4, two distinct patterns of incidents have been simulated. However, their VMR value is exactly the same. In the right Figure 4b, a clear clustering occurs in the northwestern part of the study area, while on the left Figure 4a, incidents seem to be spread out.

Nearest neighbor

The nearest-neighbor technique avoids the problem of quadrat size determination. Essentially, the nearest-neighbor statistic is used to test whether a set of incidents are closer together than would be expected by random distribution. The statistic computed R is the ratio between the observed average distance between all events and the expected value if the events were distributed in a random fashion:

$$R = \frac{R_o}{R_e} = \frac{\bar{d}}{\frac{1}{2\sqrt{\lambda}}} \quad [9]$$

where $\bar{d}$ is the mean of the distances from incidents to their nearest neighbor and λ is the total number of incidents in the area. The term R_o denotes the observed average distance between incidents and their nearest neighbors, while R_e is the expected distance between incidents under random circumstances. When the nearest-neighbor statistic R is equal to 0, all points are in one location. When $R_o = R_e$, we have a perfectly random pattern. For a square area, when $R = 2.14$, we have a perfectly uniform pattern where incidents are spread out evenly.

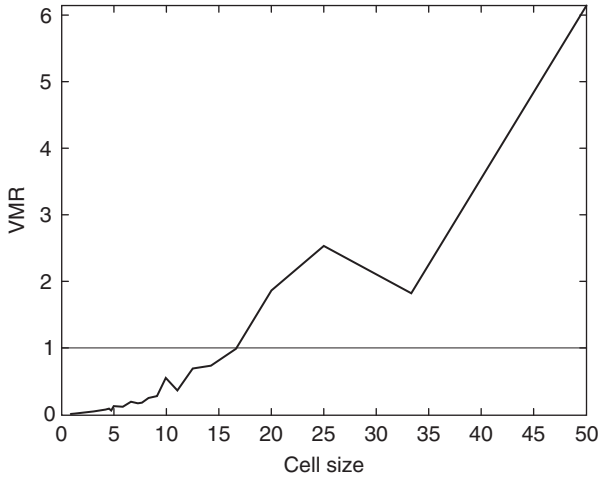

Figure 3 Change in VMR with varying cell size. Uniform patterns are observed when the quadrat size is small, while clustering is obvious at coarser scale (cell size greater than 17), and maximum for a cell size of 50 (study area divided into 4 quadrats).

Example

The point pattern from Figure 2 exhibits a R_o-value of 3.186, a R_e-value of 4.896, and an R-value of 0.651, which confirms that the pattern is clustered. Confirming the presence of clustering can also be accomplished using Monte Carlo methods, by simulating a random pattern many times, and obtaining the R_e-value based on those random patterns. For instance, following 1000 Monte Carlo simulations, an average R_e-value of 3.731 was obtained, and R was equal to 0.854, which confirms that the point pattern was clustered. The Monte Carlo approach is especially useful when the study area is irregularly shaped.

Advantages and limitations of the technique

The shape of the study area will decrease the nearest-neighbor statistic, especially if it is narrow, because events will necessarily be next to one another. Another problem commonly noted in the literature is that

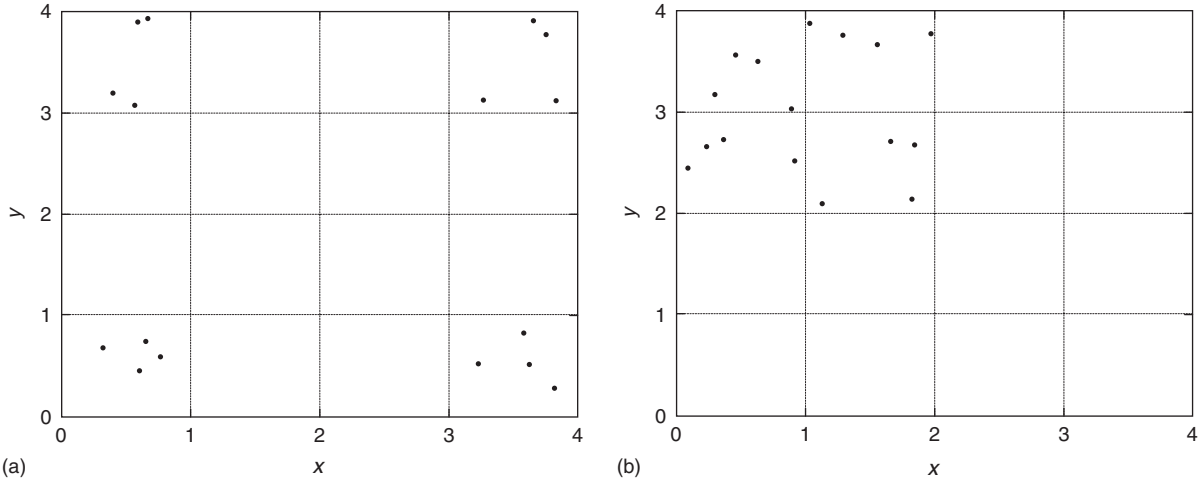

Figure 4 Two very distinct spatial patterns exhibiting a similar VMR value of 3.2 (cell size is of unit 1).

clustering may only be detected on a relatively small scale. In that case, it is possible to extend the statistic to higher orders (second or third nearest neighbor). It should be noted that it is possible to compute the cumulative distribution of separating distances, which in turn sheds light on the spatial scale of clustering. In Figure 5, a thick line denotes the cumulative distribution of the clustered dataset mapped in Figure 2. Not surprisingly, the curve increases sharply close to the origin since many events are located next to one another, which suggests clustering due to inter-event attraction within certain parts of the study area. The curve flattens out beyond a separating distance of 100 units, indicating an absence of spatial clustering at that scale. The dotted line is the cumulative distribution of a random point pattern. At short separating distances, this pattern exhibits a greater amount of pairs of points than in the case of a symmetric pattern. Finally, the last curve represents the cumulative distribution following a perfectly symmetric (regular) point pattern. Since the events are very spread out from one another, the potentiality for a spatial clustering pattern is always below a random point pattern.

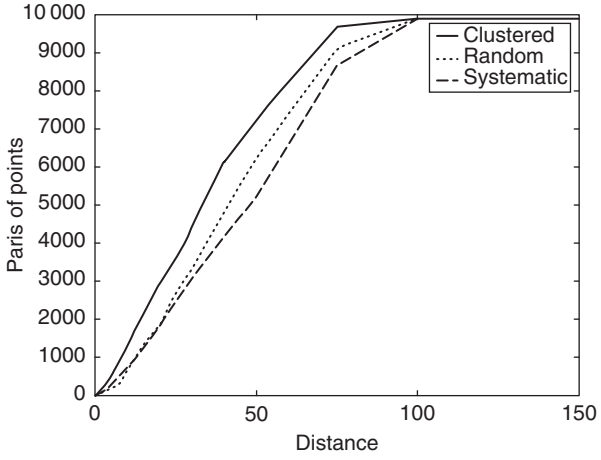

Figure 5 The cumulative distribution of pairs of points as a function of the separating distance (since $n = 100$, there are n^2 possibilities).

The nearest-neighbor technique solely relies on the distances between events, not whether there is a high concentration of events next to each other. In Figure 6 for instance, two distinct patterns of incidents have been simulated. Although their nearest-neighbor values are exactly the same, their spatial patterns are totally different. Patterns may vary when calculating the neighbor statistic at different orders of distance (second or third nearest neighbor for instance).

Most geographic information systems (GISs) offer the flexibility to compute this statistic. Since a complete search for the nearest neighbor of each point is carried out, the procedure can be time consuming when n is large. It is however possible to construct a Delaunay tessellation of the n events and search for nearest-neighbor distances within the tessellation.

K-Function

The K-Function provides an alternative to the nearest-neighbor statistic as a technique aimed at determining the amount of clustering at a wider range of scales. To compute the statistic, a circle of a specified radius (h) is placed over each point in the set (i). Points within this circle (j) are counted and the circle is then moved to the next point to begin the process again. When this has been done for all existing points n, the radius is then expanded and the process is repeated until a specified maximum radius is reached.

$$K(h) = \frac{A}{n^2} \sum_{i \neq j} \sum I_h(d_{ij}) \qquad [10]$$

where d_{ij} is the distance between two events i and j within the study region, A the size of the study area, and $I_h(d_{ij})$ an

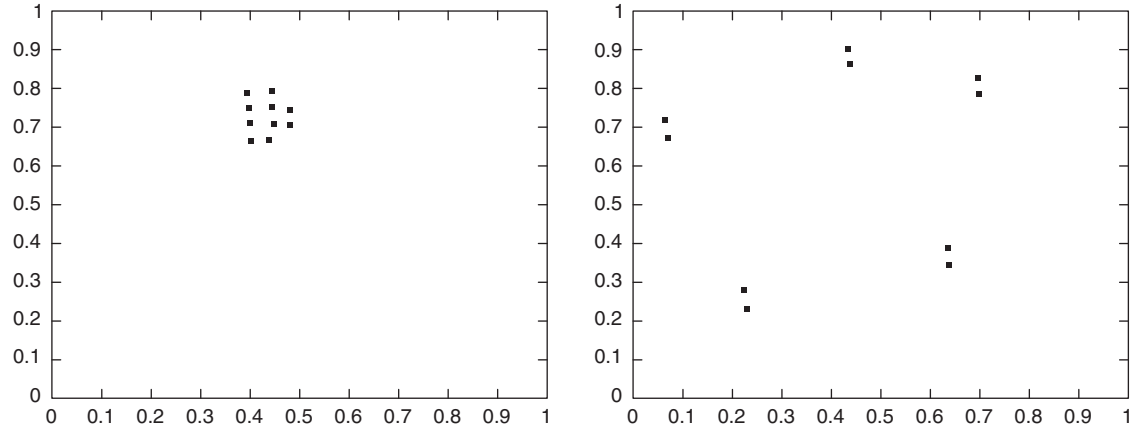

Figure 6 Two very distinct spatial patterns exhibiting a similar nearest-neighbor value.

indicator function defined as:

$$I_h(d_{ij}) = \begin{cases} 1 & \text{if} \quad d_{ij} \leq h, \\ 0 & o.w. \end{cases} \qquad [11]$$

In general, a high value of $K(h)$ denotes clustering. The value of $K(h)$ can be graphed against the distance (h) to show the scale(s) at which the point pattern exhibits randomness, clustering, or dispersion. These values can be tested using Monte Carlo methods in which K functions are computed for a large number of randomly generated point patterns. For each h value, the 5% highest and 5% lowest K-values for these random patterns form the upper and lower significance envelopes under the null hypothesis of CSR. Figure 7 illustrates this concept, where the dotted lines represent the 5% and 95% confidence intervals, and the thick line represents the observed K-function values (for different h-values) for the dataset of Figure 2. When the observed value for a given h is between the upper and lower envelopes, we conclude that the point pattern is random at that spatial scale (h). When the K-function is above the envelopes, we observe a clustered pattern, whereas the pattern is spread out when the K-function is below the envelopes. The K-function value increases as h increases since more observations are counted within the radius h.

Advantages and limitations of the technique

The K-function is superior to the nearest-neighbor technique in that it analyzes possible point patterns at different scales, while the nearest-neighbor approach uses distances only to the closest events, hence nearest neighbor only considers the smallest scales of patterns. The K-function is however relatively time consuming to compute, and it is recommended to use specific software, such as CrimeStat.

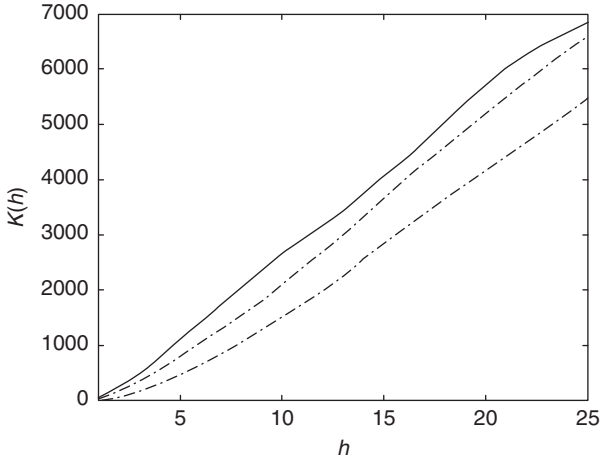

Figure 7 The K-function for the clustered dataset of Figure 2. Note the lower and upper envelopes, which were obtained after 100 Monte Carlo simulations.

Kernel Density Estimation

Neither the nearest neighbor, nor the K-function, nor the quadrat analysis, identify locations of clusters; rather, they determine an overall tendency toward clustering, randomness, or dispersion. Kernel density mapping provides a means for visually identifying clustered areas. An advantage of the density mapping is that most GIS softwares support the method.

To calculate this density, the entire region is divided into a grid and a search radius is drawn around each grid point g in a similar way to a moving window average. A suitable window – defined by the radius size τ – is moved over the fine grid of locations, and the intensity at each grid point is estimated from the event count per unit area within the window centered on that grid point. The total number of events i that fall within the search radius is divided by the size of the window resulting in a density value for each grid point. Kernel density mapping extends this method by assigning weights when performing the search so that points closer to the center of the window receive a higher weight than those further away. Mathematically, the kernel density at a grid point is denoted $\hat{\lambda}_\tau(g)$ and can be estimated as follows:

$$\hat{\lambda}_\tau(g) = \sum_{h_i \leq \tau} \frac{3}{\pi \tau^2}\left(1 - \frac{h_i^2}{\tau^2}\right) \qquad [12]$$

with h_i the separating distance between an event i and the grid point g. The bandwidth τ determines the amount of smoothing. The size of the bandwidth τ will affect the outcome of the map as a smaller search area will result in more distinct events to be highlighted. On the other hand, a larger radius can identify broad zones where a high number of incidents exist. A bandwidth that is too large will stretch the kernel and the surface will appear flat. The choice of the bandwidth may depend on the purpose of the study.

Example

Figures 8 and 9 represent the surface following a kernel density passed on the events from Figure 2, where bandwidths of 5 units and 10 units were used, respectively. The peaks denote regions where there is a strong concentration of events, which coincide with the three blocks highlighted in Figure 2. As expected, the intensity of the peaks is higher when the bandwidth is smaller, and a greater smoothing occurs when the bandwidth increases. Interestingly enough, many small peaks appear on Figure 8 when the bandwidth is $\tau = 5$, which suggests that this bandwidth was not appropriate, as those peaks correspond to a single event. The Kernel map appears more flat in **Figure 9**, because a greater bandwidth ($\tau = 10$) smoothes out the concentration of events.

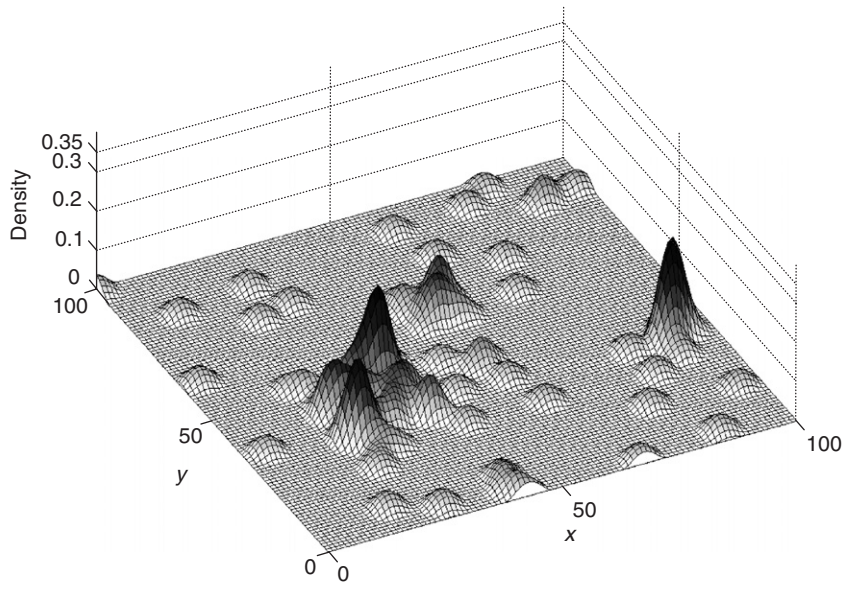

Figure 8 The kernel density surface for the clustered point dataset of **Figure 2**, calculated with a bandwidth $\tau = 5$.

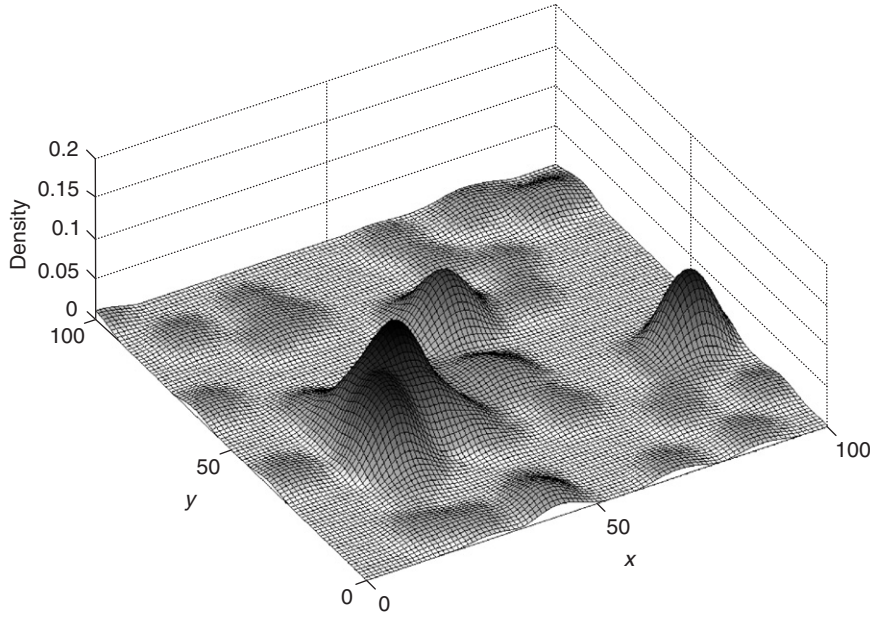

Figure 9 The kernel density surface for the clustered point dataset of **Figure 2**, calculated with a bandwidth $\tau = 10$.

Spatial Patterns on a Network

One drawback to using these aforementioned spatial clustering methods is that they assume that events can be located anywhere in the study area. However, many point events occur on a network, such as car crashes or pedestrian accidents. Such events are most often restricted to the existing road network; so ideally, the analysis should focus solely on the places where it is possible for accidents to occur. Recently, network-based point pattern analysis methods have been developed that provide a more accurate computation of accident clustering. However, such methods remain computationally intensive and to this date they are not supported by any GIS.

Conclusion

The techniques presented in this article are excellent tools to analyze spatial point patterns, and remain relatively easy to perform, either within a GIS or by

programming. However, one must remain careful, and consider the various limitations of each technique. For instance, the nearest-neighbor distance analysis is prone to errors associated with irregular boundaries, while quadrat analysis is prone to errors associated with edge effects and variability within quadrat cells. Attention must also be paid to a few key issues such as spatial scale, edge effects, and events occurring on a network. It is usually desirable to conduct a test for clustering at different scales to evaluate the magnitude of this clustering. These issues are being researched now; however, most of them are not available yet in commercial GIS.

See also: Edge Effects; Monte Carlo Simulation; Scale Analytical; Spatial Autocorrelation; Spatial Clustering, Detection and Analysis of; Spatial Data Mining, Cluster and Pattern Recognition; Spatial Data Mining, Geovisualization; Spatial Filtering/Kernel Density Estimation.

Further Reading

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. New York: John Wiley.

Chou, Y. H. (1996). *Exploring Spatial Analysis in GIS*. Santa Fe, NM: Onword Press.

Diggle, P. (1983). *Statistical Analysis of Spatial Point Patterns*. London: Academic Press.

Fotheringham, S., Brunsdon, C. and Charlton, M. (2004). *Quantitative Geography*. London: Sage Publications.

Levine, N., Kim, K. and Nitz, L. (1995). Spatial analysis of Honolulu motor vehicle crashes: I. Spatial patterns. *Accident Analysis and Prevention* 27, 663–674.

Okabe, A. and Yamada, I. (2001). The K-function method on a network and its computational implementation. *Geographical Analysis* 33, 271–290.

Okabe, A., Yomono, H. and Kitamura, M. (1995). Statistical analysis of the distribution of points on a network. *Geographical Analysis* 27(2), 152–175.

Rogerson, P. A. (2001). *Statistical Methods for Geography*. London: Sage Publications.

Wong, D. and Lee, J. (2005). *Statistical Analysis of Geographic Information*. New York: Wiley.

Yamada, I. and Rogerson, P. A. (2003). An empirical comparison of edge effect correction methods applied to K-function analysis. *Geographical Analysis* 35, 97–109.

Yamada, I. and Thill, J.-C. (2004). Comparison of planar and network K-functions in traffic accident analysis. *Journal of Transport Geography* 12, 149–158.

Relevant Websites

http://www.icpsr.umich.edu
Inter-University Consortium for Political and Social Research (ICPSR): CrimeStat.