[AM-03-007] Point Pattern Analysis

Abstract

Point pattern analysis (PPA) focuses on the analysis, modeling, visualization, and interpretation of point data. With the increasing availability of big geo-data, such as mobile phone records and social media check-ins, more and more individual-level point data are generated daily. PPA provides an effective approach to analyzing the distribution of such data. This entry provides an overview of commonly used methods in PPA, as well as demonstrates the utility of these methods for scientific investigation based on a classic case study: the 1854 cholera outbreaks in London.

Keywords: basic analytical methods, density based methods, distance based methods, point data, spatial analysis

Author & citation

Yuan, Y., Qiang, Y., Bin Asad, K., and Chow, T. E. (2020). Point Pattern Analysis. The Geographic Information Science & Technology Body of Knowledge (1st Quarter 2020 Edition), John P. Wilson (ed.). DOI: 10.22224/gistbok/2020.1.13..

An earlier version can also be found at:

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., and Wentz, E. (2006). Point pattern analysis. The Geographic Information Science & Technology Body of Knowledge. Washington, DC: Association of American Geographers.

Explanation

1. Introduction
2. Methods for Point Pattern Analysis
3. Case Study: Cholera in London, 1854
4. Closing Remarks

1. Introduction

Point pattern analysis (PPA) studies the spatial distribution of points (Boots & Getis, 1988). Previous studies have developed various methods and measurements, such as density-based methods and distance-based methods, to analyze, model, visualize, and interpret the properties of point patterns. Generally, these properties can be divided into two categories: first-order properties and second-order properties. The former focuses on the characteristics of individual locations and their variations across space, whereas the latter focuses on properties that concern not only individual points, but also the interactions between points and their influences on one another. One example of second-order properties is the degree of dispersion (e.g., clustered, dispersed, or random) of a point pattern (Oyana & Margai, 2016). In general, density-based methods, such as kernel density, mostly address first-order properties of point patterns. Distance-based methods, on the other hand, consider the distance between point pairs and therefore measure second-order

properties. This entry reviews both types of methods in PPA and illustrates these methods based on a classic case study of the 1854 cholera outbreaks in London (Snow, 1855).

2. Methods for Point Pattern Analysis

2.1 Descriptive Statistics

In PPA, descriptive statistics provide a summary of the basic characteristics of a point pattern, such as its central tendency and dispersion. Central tendency focuses on extracting the central or typical location from a point pattern, providing an estimate of the location around which all the points are spread (O'Sullivan & Unwin, 2010). Two commonly used measures of central tendency are mean center and median center (Gimond, 2019). Mean center averages the (x, y) coordinates of all points in the study area (Equation 1):

$$(\mu_x, \mu_y) = (\frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n}) \quad (1)$$

where (μx, μy) are the coordinates of the mean center, (xi, yi) represent the coordinates of a given point i, and n is the total number of points.

In spatial statistics, median center is the location that minimizes the sum of distances traveled to all points in the study area, and it is calculated using an iterative procedure introduced by Kulin and Kuenne (1962). The algorithm starts with an assigned point as the initial location (e.g., the median center), and then the new coordinates of the median center (x', y') are updated as follows (Rogerson, 2019):

$$x' = \frac{\sum_{i=1}^{n} \frac{w_i x_i}{d_i}}{\sum_{i=1}^{n} \frac{w_i}{d_i}}, y' = \frac{\sum_{i=1}^{n} \frac{w_i y_i}{d_i}}{\sum_{i=1}^{n} \frac{w_i}{d_i}} \quad (2)$$

where di is the distance between point (xi, yi) and the median center from the previous round, and wi is the weight assigned to point (xi, yi). When calculating an unweighted mean center, wi is a constant for all points in the study area. This process is repeated until the newly computed median center is not significantly different from the prior one (i.e., the distance between the newly computed median center and the previous median center is smaller than a predefined threshold). The mean center and the median center of a point pattern often do not coincide (Figure 1), and the median center is often considered a more robust indicator and less impacted by outlier points than the mean center is.

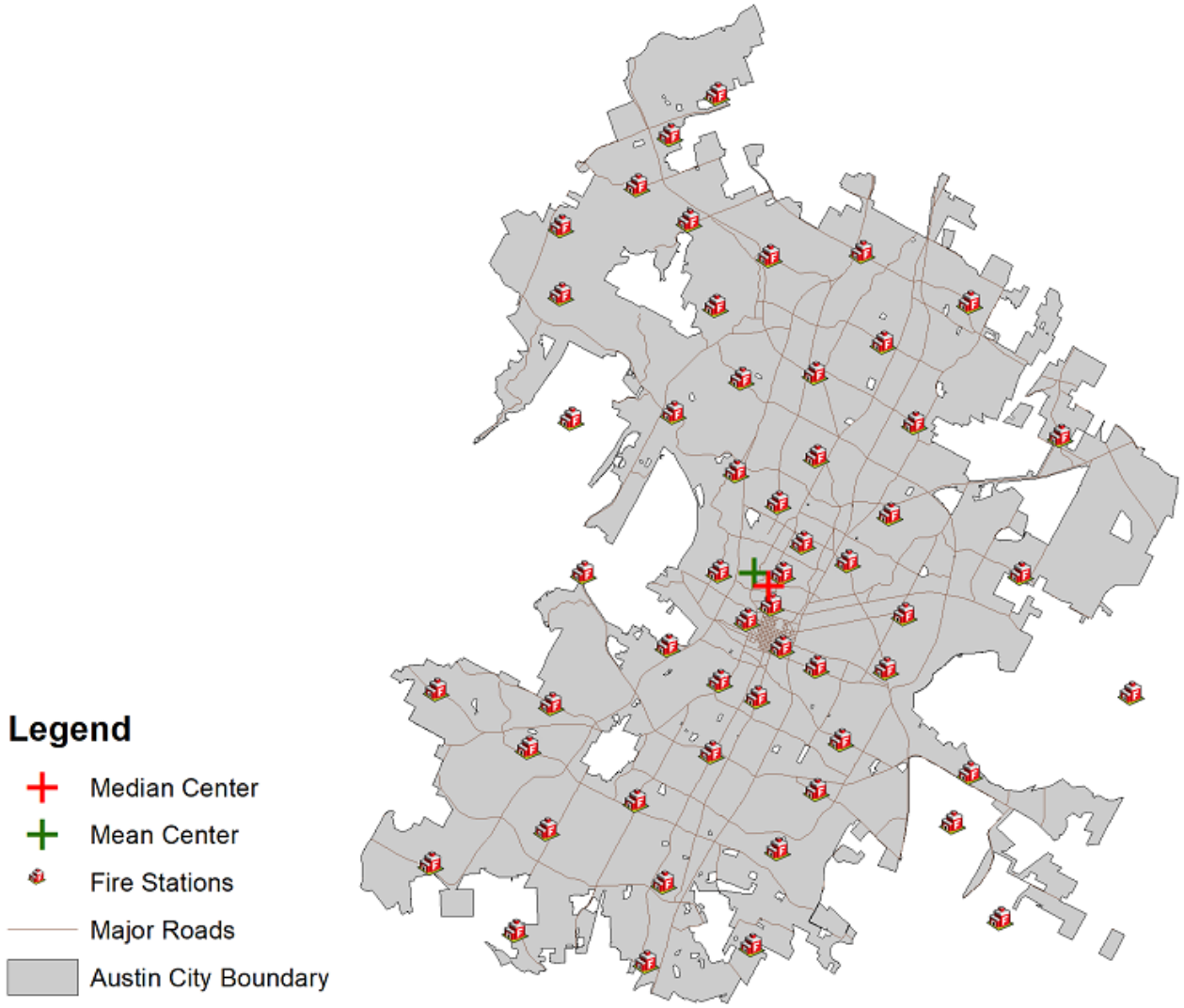


Figure 1: The mean center and media center of fire stations in Austin, Texas. Note that some fire stations are in the Austin extraterritorial jurisdiction (ETJ) area, and therefore are located outside of the city boundary. Data source: data.AustinTexas.gov. Map source: authors.

In addition to measuring central tendency, researchers have also developed several indicators to measure how spread out a point pattern is (e.g., the dispersion). These indicators include standard distance and standard deviational ellipse.

- **Standard distance**: Standard distances are defined similarly to standard deviations. This indicator measures how dispersed a group of points is around its mean center (Gimond, 2019). Equation 3 shows how standard distances are calculated.

$$d = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu_x)^2 + \sum_{i=1}^{n}(y_i - \mu_y)^2}{n}} \quad (3)$$

where (μx, μy) are the coordinates of the mean center, (xi, yi) represent the coordinates of

a given point i, and n is the total number of points.

- **Standard deviational ellipse**: Although standard distance can show the degree of dispersion of a point pattern, it only calculates an isotropic metric and does not show any directional effect. To this end, researchers use standard deviational ellipses to calculate separate standard distances for two perpendicular axes. The center of the ellipse is the mean center, the major elliptical axis follows the direction with the greatest dispersion, and the length of each orthogonal axis is determined by the corresponding standard distance along that direction (Figure 2). The standard deviational ellipse is very useful in representing point patterns that follow a directional orientation (ESRI, 2018). The standard deviational ellipse can be calculated by using the point locations or by assigning weights (w) to different points based on their attributes (Gatrell et al., 1996). This is called a weighted standard deviational ellipse. The rotated semi-major (σx) and semi-minor (σy) axes of a weighted directional distribution can be calculated as follows (Wang, Shi, & Miao, 2015):

(4)

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{1}^{n}(\tilde{y}_i sin\theta + \tilde{x}_i cos\theta)^2} \quad (5)$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum_{1}^{n}(\tilde{y}_i cos\theta + \tilde{x}_i sin\theta)^2} \quad (6)$$

where w is the weight matrix, (xi, yi) are the coordinates of point i, (μx, μy) represent the (weighted) mean center, and the rotation angle θ can be calculated by:

$$tan\theta = \frac{(\sum_{i=1}^{n}\tilde{x}_i^{\,2} - \sum_{i=1}^{n}\tilde{y}_i^{\,2}) + \sqrt{(\sum_{i=1}^{n}\tilde{x}_i^{\,2} - \sum_{i=1}^{n}\tilde{y}_i^{\,2}) + 4(\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i)^2}}{2\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i} \quad (7)$$

The interpretation of the standard deviational ellipse is similar to the interpretation of standard deviations; the major and minor axes show the dispersion of points along these two directions. For normally distributed data, the values within one standard deviation of the mean are approximately 68% of the dataset. Within two and three standard deviations of the mean, the percentage of data rises to 95% and 99.7%. This is called the 68–95–99.7 rule in statistics. However, this rule does not necessarily hold for spatial data. In a two-dimensional space, the percentages change to 63%, 98%, and 99.9% for one, two, and

three standard deviations from the mean, respectively (ESRI, 2018).

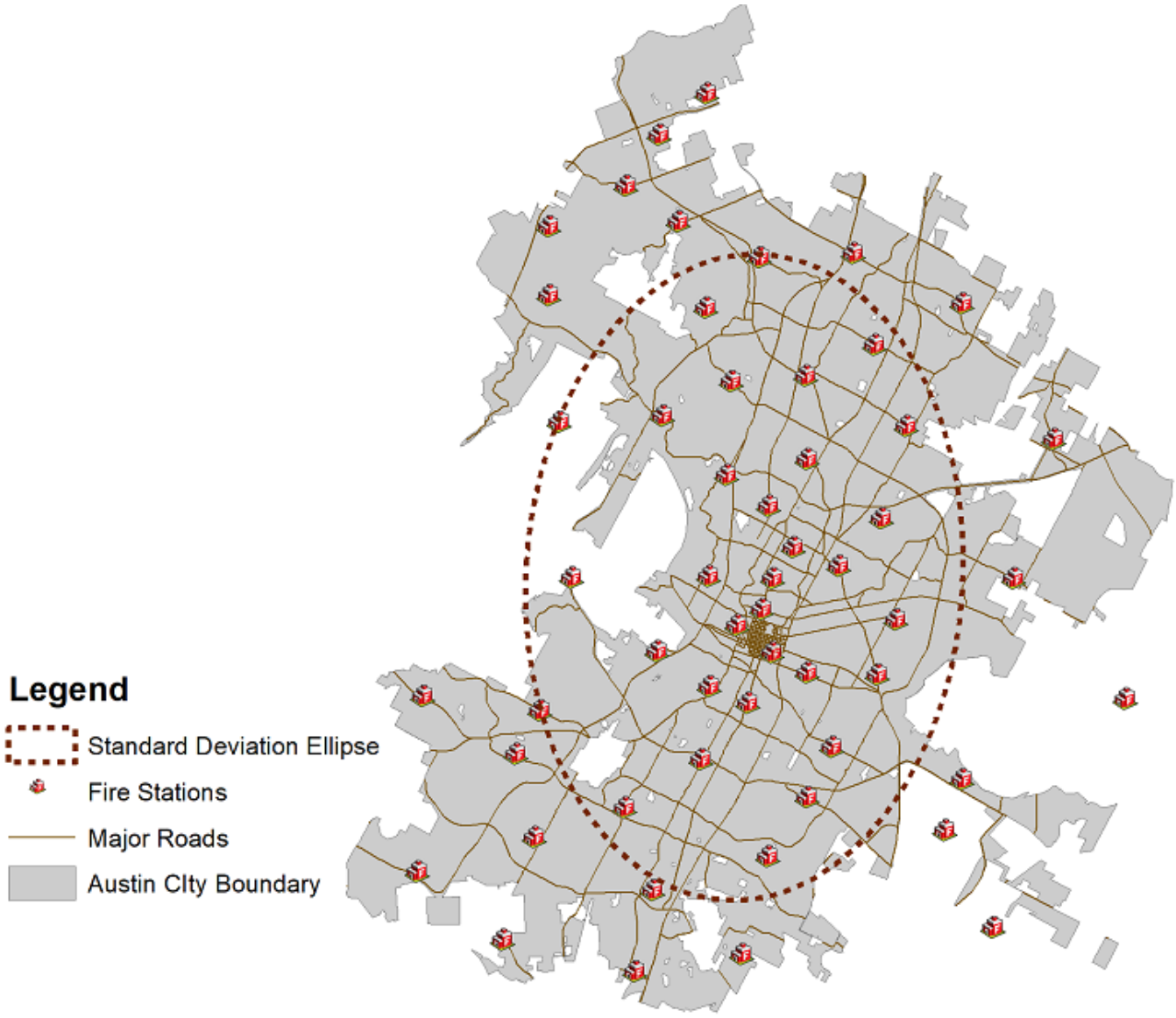


Figure 2: Standard deviation ellipse created based on Austin fire station locations. The points are more dispersed along the north-south direction. Map source: authors.

2.2 Distance-based Measures

As mentioned in the introduction, distance-based measures analyze the spatial distribution of points using distances between point pairs, and they are often considered a direct indicator of the second-order property. Although most distance-based measures are based on the Euclidean distance, various distance-based measures have been developed to analyze spatial patterns using non-Euclidean distances (Lamb, Downs, & Lee, 2016; Okabe & Yamada, 2001). This subsection describes several commonly used distance-based measures based on the Euclidean distance.

2.2.1 Nearest-Neighbor Distance

Nearest-Neighbor Distance (NND) is the distance between a point and its closest neighboring point. NND is also known as the first-order nearest neighbor. In addition, distance can be calculated for the kth nearest neighbor, which is called the kth-order NN or KNN. The mean of NND between all point pairs is used as a global indicator to measure the overall pattern of a point set (Clark & Evans, 1954). The mean NND of a given point collection can be compared with the expected NND from points following complete spatial randomness (CSR) to test the significance of the pattern (e.g., how clustered or dispersed the point pattern is). CSR describes a point pattern where all points in the study area occur randomly. It is often generated by Monte Carlo simulation in practice.

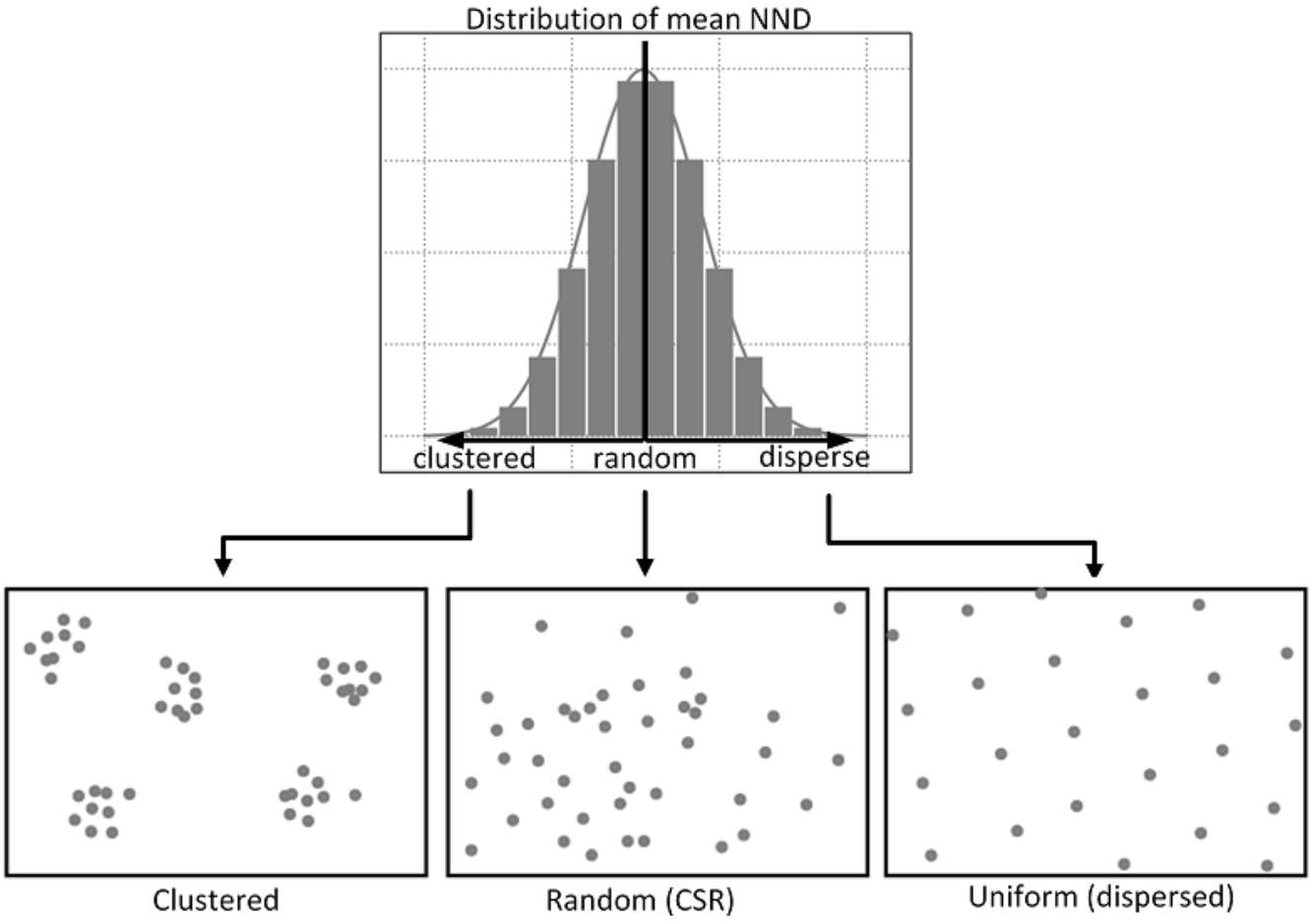


Figure 3: Relations between different point patterns and mean nearest neighbor distance (NND). Image source: authors.

The mean NND of points in CSR form a Poisson distribution as shown in Figure 3 (Smith, Goodchild, & Longley, 2007). The standard deviation (z-score) of the NND from the expected value indicates the probability that the point pattern is not from a random

process. The mean NND $(\overline{D})$ of a point set can be calculated by Equation 8:

$$\overline{D} = \frac{\sum_{i=1}^{n} d_i}{n} \quad (8)$$

where di is the NND of point i, and n is the number of points. The expected mean NND in CSR can be calculated as:

$$\overline{D}_E = \frac{0.5}{\sqrt{\frac{n}{A}}} \quad (9)$$

where A is the area of the minimum bounding box of the point set, and the z-score of the mean NND can be calculated as

$$z = \frac{\overline{D} - \overline{D}_E}{SE} \quad (10)$$

where

$$SE = \frac{0.261356}{\sqrt{n^2/A}} \quad (11)$$

The z-score indicates the significance of the point pattern, either clustered or dispersed.

2.2.2 Distance Functions

- **G Function**: Although the mean NND can measure the degree of clustering of a point set using a single-value metric, it provides limited information about the complexity of point patterns at different spatial scales. Several distance functions have been developed to describe more detailed variations of a point pattern. The G function is the simplest one, which calculates the cumulative frequency distribution of the NND of a point pattern. The G function can be written as:

$$G(d) = \frac{sum(D_{ij} < d)}{n} \quad (12)$$

where sum(Dij < d) stands for the number of point pairs i and j with a distance smaller than d, and n represents the total number of points (Figure 4).

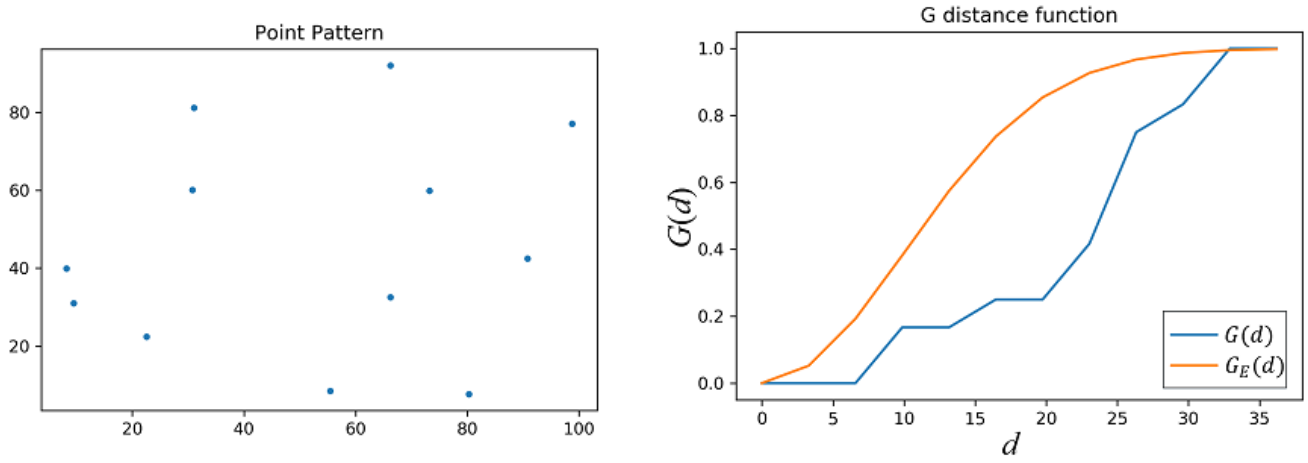


Figure 4: A point set (a) and its G function (b), where G(d) is the G function of the point set. GE(d) is the G function of the same number of points under the CSR assumption. Image source: authors.

Because the G function is cumulative, it increases as the distance d increases. The shape of the G function provides information regarding how a point pattern clusters. If points are clustered, the G function increases rapidly at short distances. If points tend to be dispersed, the G function increases slowly up to the distance at which most points are spaced, and then it starts to increase rapidly.

- **F Function**: The F function first generates a few random points (denoted as P) in the study area, and then it determines the minimum distance from each random point in P to any original points (denoted as O) in the study area. The F function is written as:

$$F(d) = \frac{sum[d_{min}(p_i, s) < d]}{n} \quad (13)$$

where F(d) indicates the value of the F function at distance d, and sum[dmin(pi,s) < d] is the number of points in P with a minimum distance to any point in O smaller than d. The advantage of the F function is that one can increase the number of the randomly generated points to obtain a smoother curve.

- **K Function**: The G and F functions only consider the nearest neighbor for each point and ignore the distances to other points, so they cannot be used to analyze point patterns at multiple scales (distances). In addition, they do not reflect the local variances of a point pattern. For instance, a point set may consist of evenly spaced

clusters, meaning the points are clustered in local areas, but the clusters are evenly distributed spatially. The Ripley's K function is a powerful approach to identify the multi-scale patterns of points. The three aspects/steps/factors of Ripley's K function are: (i) Construct a circle with a radius d around each point i; (ii) Count the total number (n) of points that fall inside any of the circles (excluding the points at the circle centers); and (iii) Increment d by a small fixed amount and repeat the first two steps.

The K function is written as:

$$K(d) = \frac{R}{n^2} \sum \sum_{i \neq j} \frac{I_d(d_{ij})}{w_{ij}} \quad (14)$$

where R is the size of the study area, and wij is an indicator for edge correction. wij equals 1 if the circle centered at point i passes through point j (i.e., with a radius of dij) and completely falls inside the study area. If part of this center falls outside the study area, wij is the proportion of the circumference of the circle that falls in the study area. Id(dij) equals 1 if the distance between point i and j (dij) is smaller than d, and otherwise equals 0.

The observed K(d) function minus the expected values under CSR leads to the L function, which measures the deviation of a point pattern from CSR. While other distance functions (G, F, and K) are monotonically increasing, the L function may either decrease or increase at different distances.

The confidence intervals around the expected values can be computed by Monte Carlo simulation (Smith et al., 2007). The significance of the resulting point patterns at varying distances can be analyzed by comparing the K function of the point sample with the expected function under CSR. For instance, the K function in Figure 5 shows a significant clustered pattern at short distances, but significantly dispersed at longer distances. More information about the statistical interpretation of the functions can be found in O'Sullivan and Unwin (2010) and Smith et al. (2007).

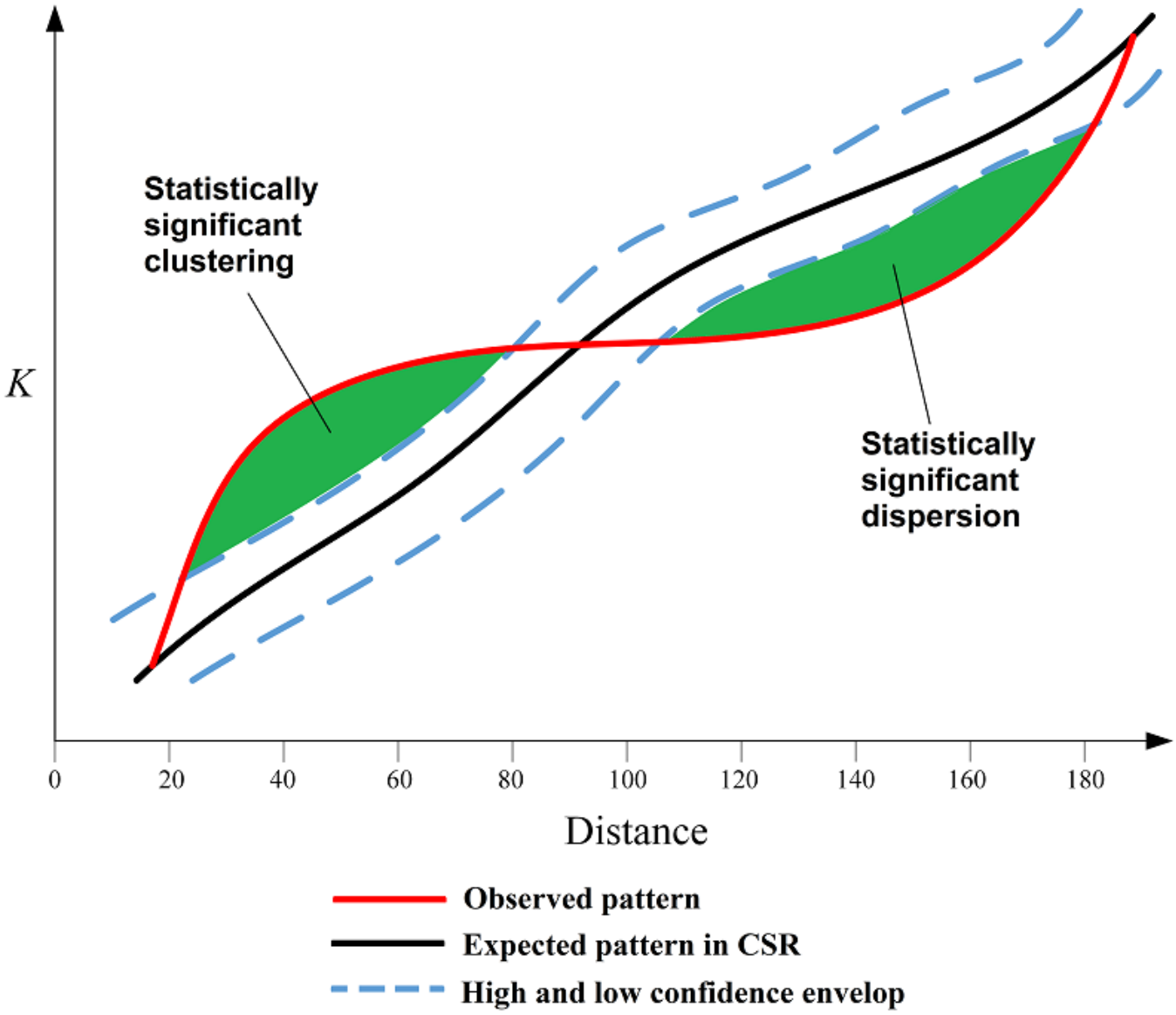


Figure 5: Statistical significance of Ripley's K function. Image source: authors.

2.3 Density-based Measures

In addition to distance-based measures, researchers apply various density-based measures to investigate the variations of point densities across space. Density measures can be divided into two categories: global density and local density. Global density refers to the ratio between the observed number of points relative to the size of the study area. It can be simply calculated using the following equation:

$$\lambda = \frac{n}{a} \tag{15}$$

where n is the number of points, and a is the size of the study area.

Global density provides a single metric for the intensity of point density across the whole

study area, but it is not capable of capturing local variations. Local density, however, shows varying point densities at different locations in the study area. The two most commonly used density-based measures are quadrat density and kernel density.

2.3.1 Quadrat Density

For quadrat density analysis, the study area is divided into smaller sub-regions (i.e., quadrats), and then the point density is computed for each sub-region (Equation 15). For example, in Figure 6, the study area has been divided into 4*5 uniformly shaped quadrats of 81 km2, and the top left quadrat has a density of 1/81. In addition, quadrats can be in different shapes, such as hexagons, squares, triangles, and Thiessen polygons (Gimond, 2019). The result of quadrat density analysis is highly sensitive to the selection of quadrat shape and size and the corresponding number. Small quadrats can lead to many quadrats with very few or no points in them, whereas excessively large quadrats cannot capture subtle changes at a fine scale (Anderson & Marcus, 1993).

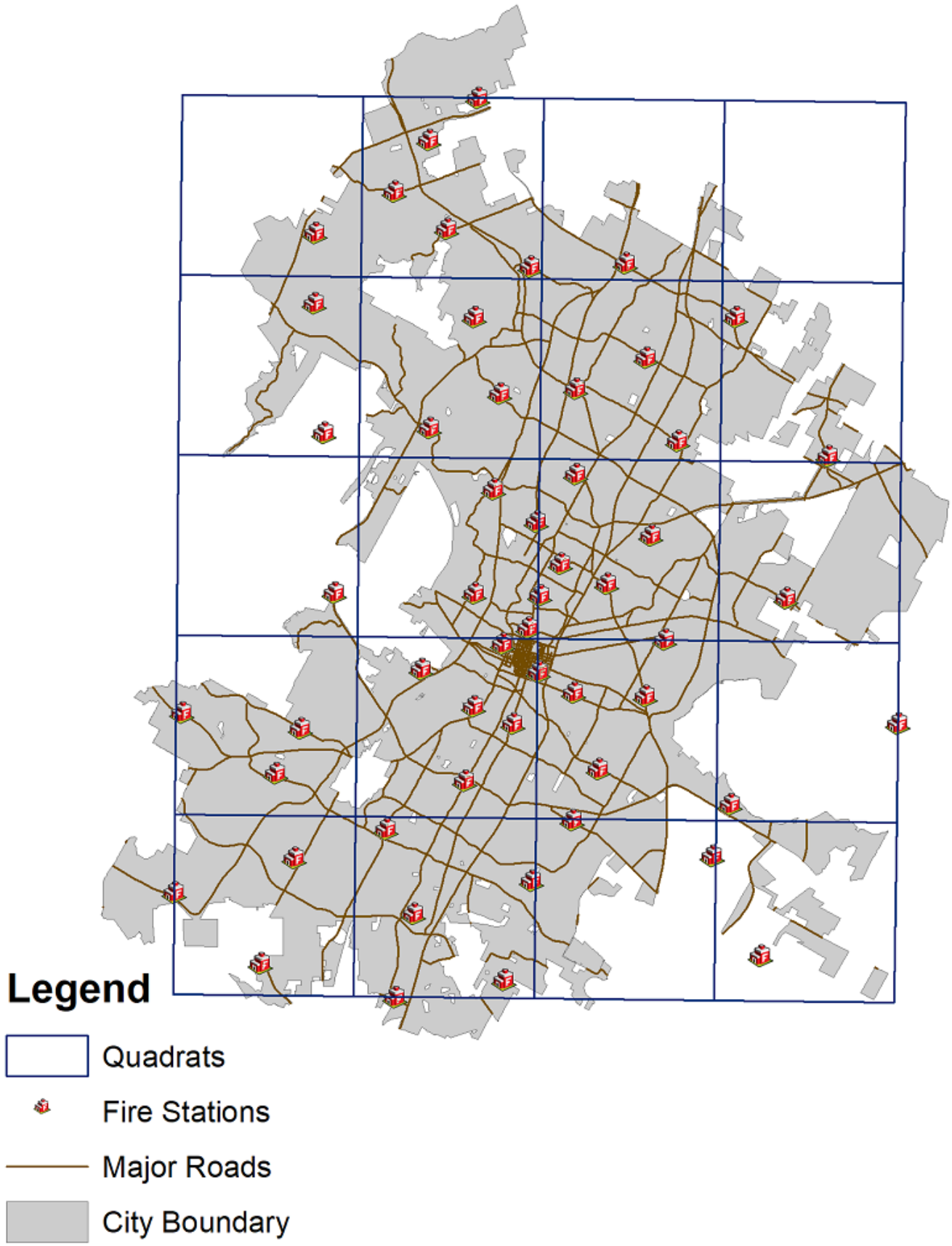


Figure 6. Quadrat density analysis of Austin fire stations. Map source: authors.

2.3.2 Kernel Density

Unlike quadrat density analysis, which assumes that the density of events is uniform within each quadrat, Kernel Density Estimation (KDE) is based on the assumption that every location has a density, and the estimate of densities not only relies on the occurrence of events, but also a predefined mathematical equation (i.e., the kernel) (O'Sullivan & Unwin, 2010). More specifically, it estimates the local density of points in a non-parametric and continuous way by counting the number of events in a region (i.e., the search window) that is centered at the location where the density is calculated. During the calculation, only points within the search window are counted, and nearby points are often weighted more heavily than distant points (O'Sullivan & Unwin, 2010). There are different kernel functions to assign weights to the points. Commonly used kernel functions include the linear kernel, polynomial kernel, uniform kernel, Gaussian kernel, exponential kernel, etc. Because KDE creates a continuous surface of point densities, it is particularly useful for transforming discrete observations to a continuous variable (Figure 7).

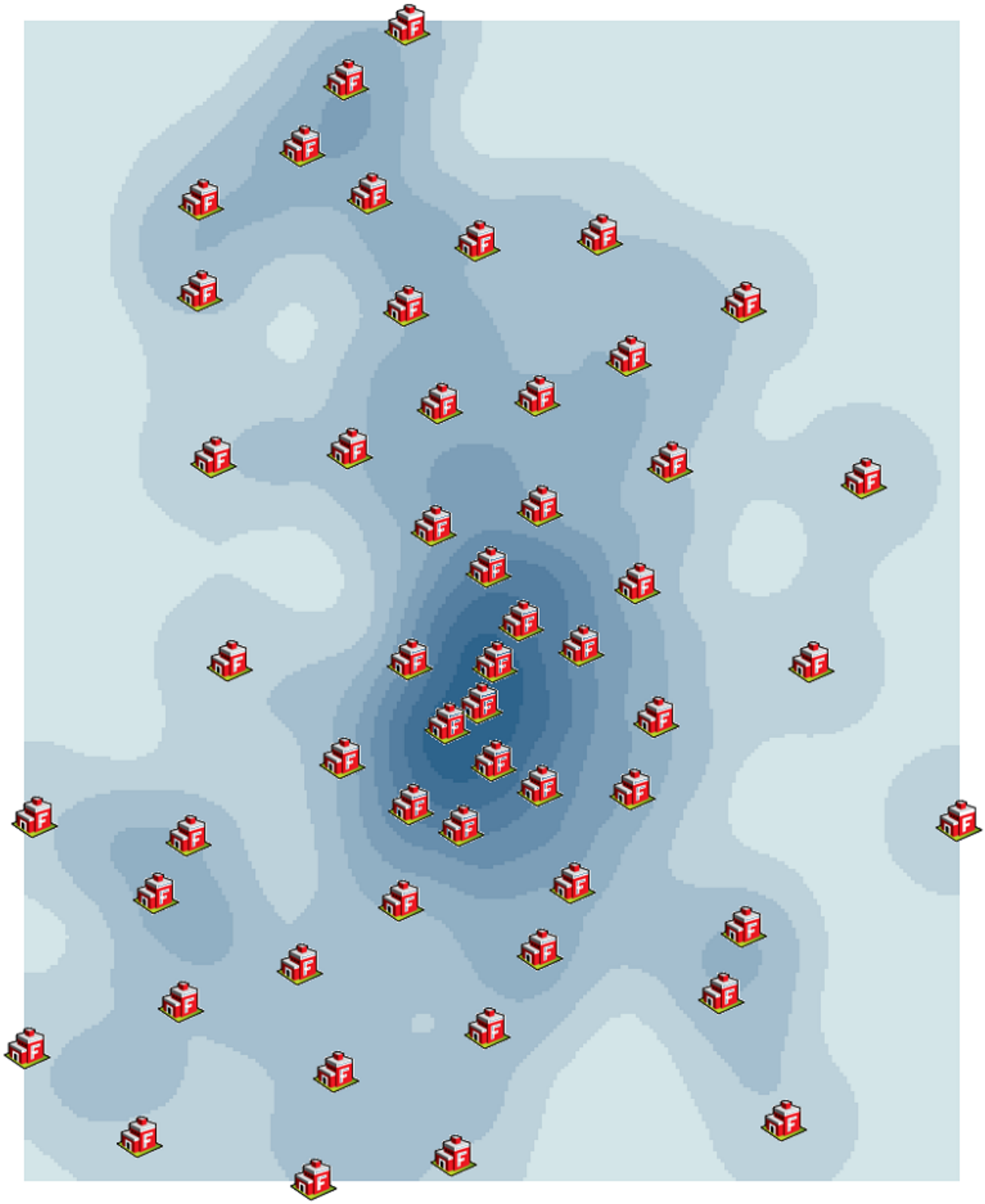


Figure 7: Kernel density estimation of Austin fire stations using a linear kernel function. Image source: authors.

It is important to note that the size and extent of the study area can affect the calculation

of both distance-based and density-based measures, although not in the same manner or to the same degree. The study area can be defined as the minimum bounding box of a point set (i.e., similar to the nearest neighborhood statistics in Equation 11). It can also be defined in many other ways, such as using administrative boundaries, land use patterns, or physical geographic entities (e.g., a forest). Many of the indicators in PPA are sensitive to changes in the study area. This is often referred to as the "edge effect." For example, for NND, unobservable feature points outside of the study area can cause substantial biases in calculating the distances. When calculating global or local densities, the size of the study area will change the entire calculation. Therefore, boundary conditions of the study area should be chosen carefully in PPA. A common way to mitigate the edge effect is to assign lower weights to points close to the boundary of the study area.

2.4 Tools and Packages for PPA

There are many software packages and tools developed to analyze and model point patterns. Table 1 lists a few commonly used ones:

Table 1. Tools and Packages for Point Pattern Analysis (PPA)

Package	Programming Language	Description
ArcMap	None required	Includes a comprehensive toolkit for PPA, such as the calculation of descriptive statistics, distance-based measures, density-based measures, etc.
ArcGIS Pro	None required	Includes a comprehensive toolkit for PPA, such as the calculation of descriptive statistics, distance-based measures, density-based measures, etc.
ArcPy	Python	A Python package that allows users to conduct the spatial analysis function of ArcGIS in Python.
CrimeStat	None required	A package specialized in point-based crime data analysis.
PySAL	Python	A geospatial data science tool in Python that includes a sub-package "pointpats" for PPA.
QGIS	None required	An open-source geographic information system that includes customized plugins for PPA.
spatstat	R	An R package specialized in spatial point pattern analysis in both 2-dimensional and 3-dimensional spaces.

3. Case Study: Cholera in London, 1854

To demonstrate the effectiveness of PPA, this section uses the deadly cholera outbreaks in 1854 London as an example (Snow, 1855). At the time, many people believed this infectious disease was spread by inhaling mysterious "miasmas" (i.e., bad air). By mapping the cholera deaths near the Soho neighborhood, however, Dr. John Snow hypothesized that it could be caused by oral consumption of contaminated food or water sources. Dr. Snow purportedly stopped the epidemic by removing the pump handle at Broad Street. One could also test Dr. Snow's suspicion of the Broad Street pump by deriving the mean center of all cholera death cases in the Soho neighborhood. It was clear that the mean center and the median center of all cholera cases in Soho were centralized around the water pump at Broad Street (Figure 8). Using this dataset, the median center was not far from the mean

center as well. Moreover, the directional distribution (i.e., the 1st and 2nd standard deviations) fits the pattern of these cases very well. The cholera data used in this case study has a count attribute for each address, so it was used as a weight to calculate the point pattern analysis. Hence, Equation 1 for the mean center (μx, μy) becomes

$$(\mu_x, \mu_y) = (\frac{\sum_{i=1}^{n} w_i x_i}{n}, \frac{\sum_{i=1}^{n} w_i y_i}{n}) \quad (16)$$

where wi is the death count at location i as the weight, and n is the total number of points. The weighted mean center can be derived as:

μx = (1,114,518 ´ 3) + (1,114,522 ´ 2) + … + (1,114,551 ´ 1)/489 = 1,114,624

μy = (5,744,204 ´ 3)+ (5,744,198 ´ 2) + … + (5,744,032 ´ 1)/489 = 5,744,215

The median center can be derived iteratively as defined in Section 2.1.

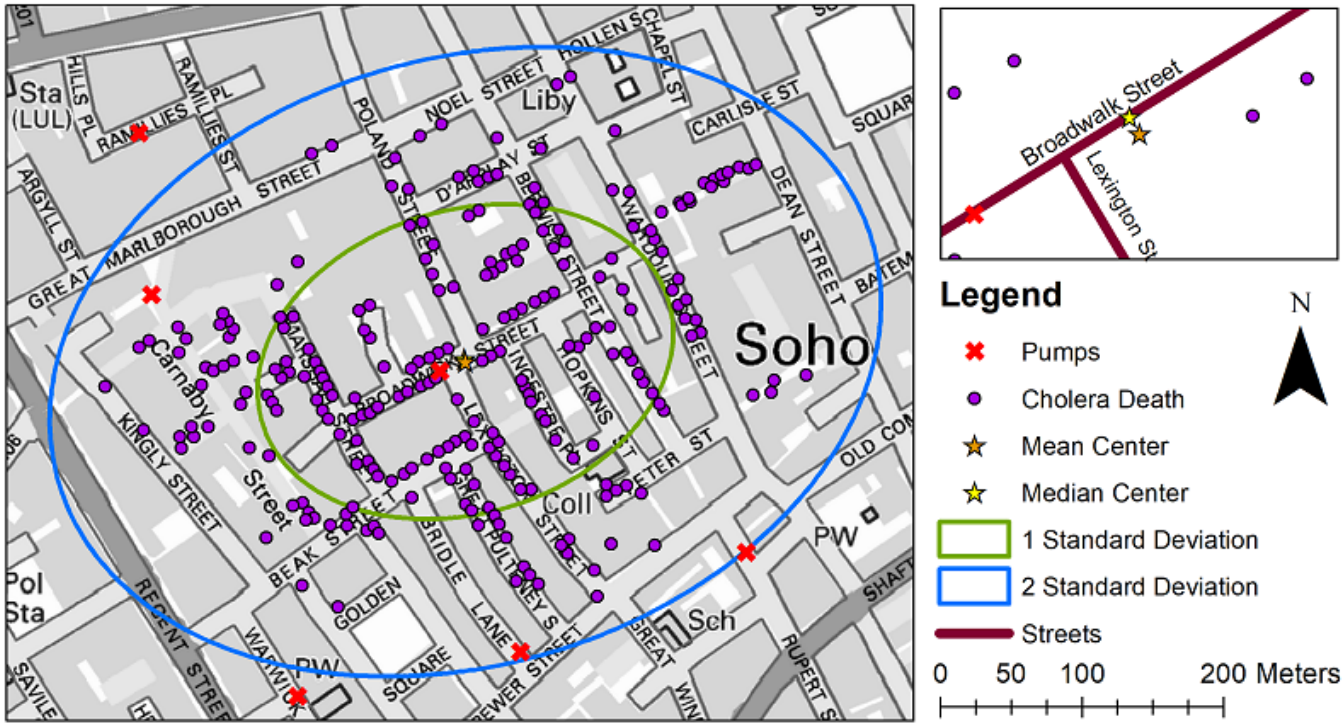


Figure 8: Map of cholera cases near Soho, London in 1854 (Snow, 1855). Data source: Robin Wilson's Blog. Map source: authors.

One may also analyze the point pattern of the London cholera outbreak based on spatial statistics. By comparing the expected and observed distances between the nearest neighbor among all cholera cases, the ratio of the average nearest neighbor is 0.76, indicating a significant clustering pattern at the 0.01 level (Figure 9).

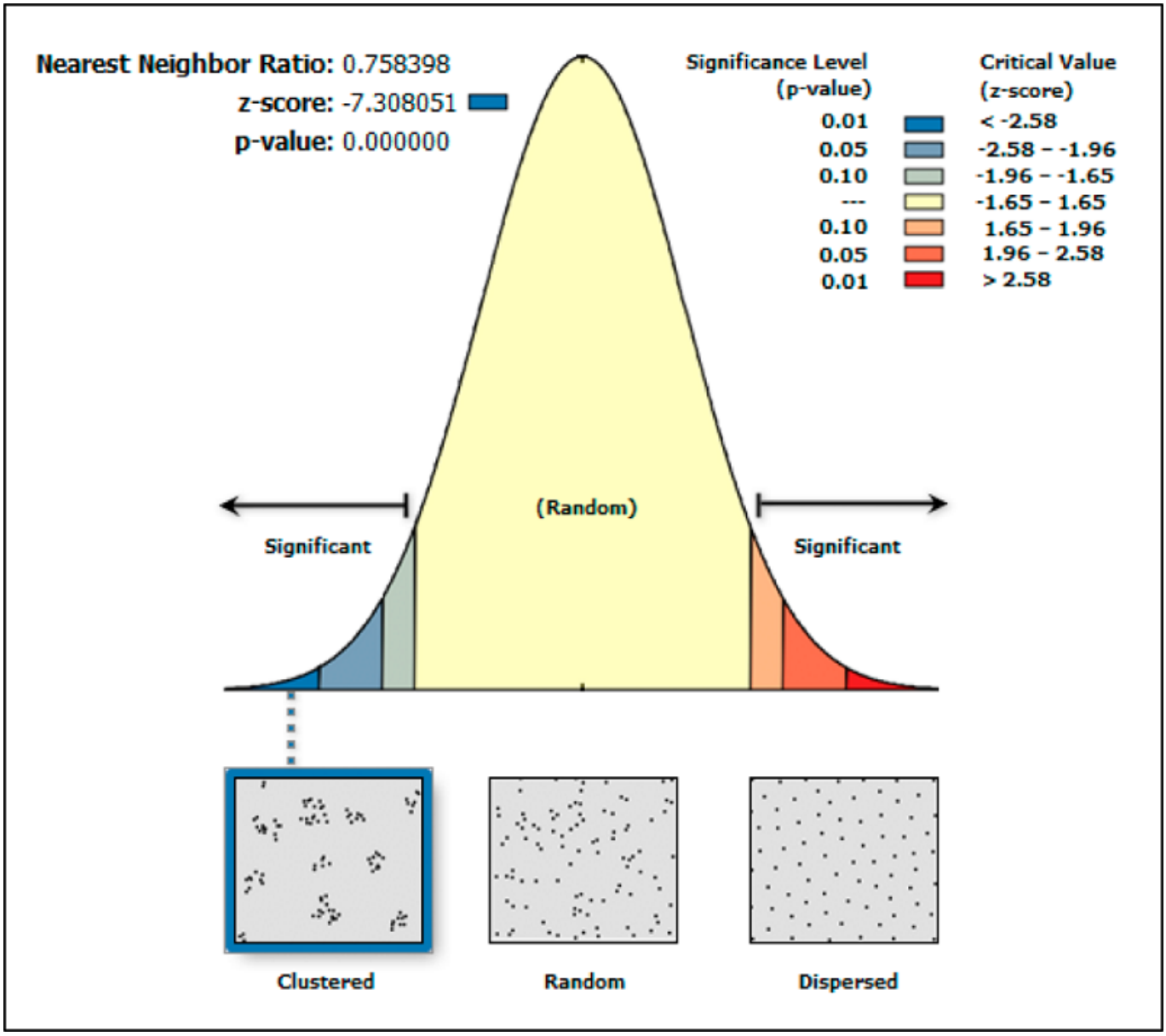


Figure 9: The results of the average nearest neighbor statistics. Image source: authors.

The London cholera outbreak's clustered pattern also makes it possible to explore the nearest neighbor of any of the outbreak's geographic features to examine their spatial relationship. For example, Thiessen polygons delineate the "area of influence" for each feature in set A (e.g., water pumps), so that any feature(s) of set B (e.g., cholera cases) that fall inside that particular Thiessen polygon would be closer than any other features of set A. As shown in Figure 10, most cholera cases were inside the Thiessen polygon of the Broad Street pump, which makes it a reasonable suspect based on their spatial relationship.

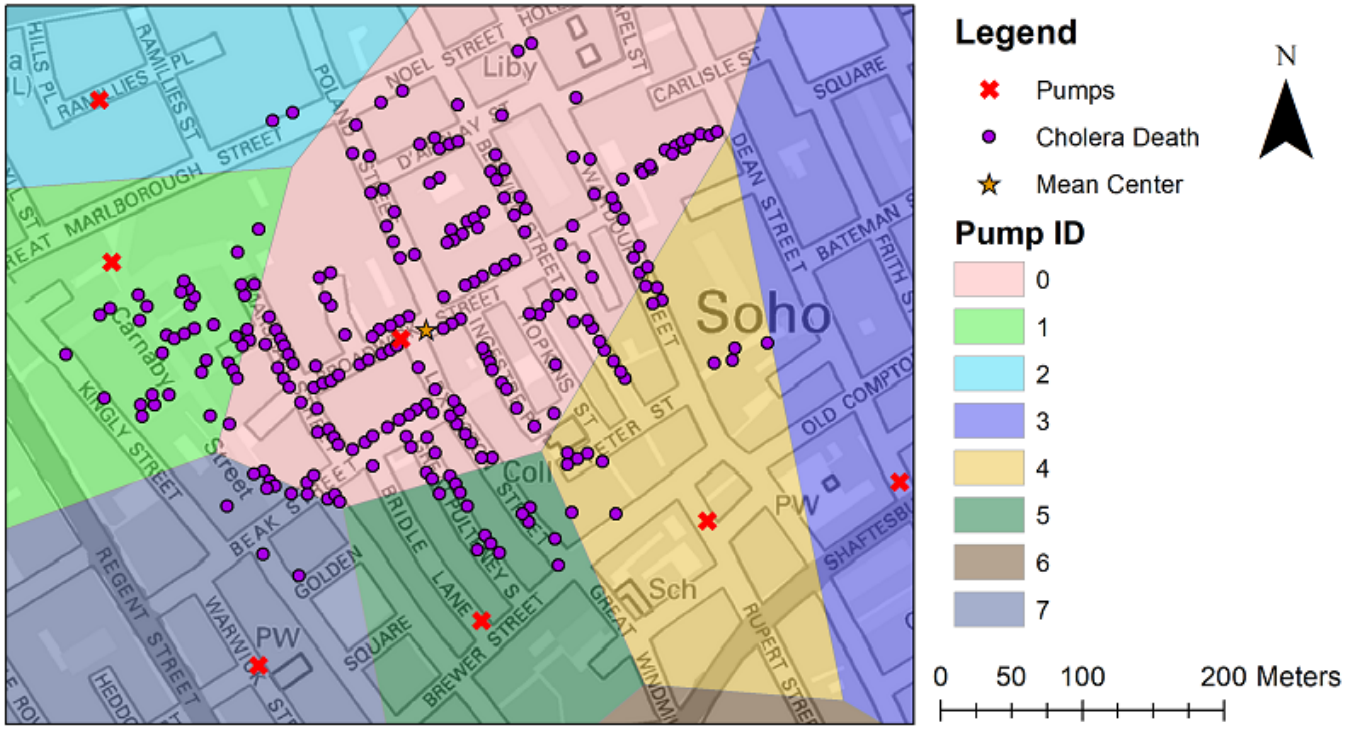


Figure 10: The Thiessen polygons of water pumps in 1854 London. Map source: authors.

As mentioned in Section 2.3, the density surface can provide another way to visualize the concentration of cholera cases. By measuring the number of cases per unit area over a region, the kernel density surface yet again pinpoints the highest concentration of cholera cases at the water pump at Broad Street (Figure 11). Here, the study area is chosen as the bounding box of the cholera cases. Note that the mean center of cholera cases also aligns well with the highest density, and it also falls inside the Thiessen polygon of the Broad Street water pump as well. Therefore, it is possible to adopt any of the aforementioned point pattern analyses to examine the geographic pattern of cholera cases and its spatial relationship with other geographic features, which allows us to examine probable theories and associated hypotheses to gain a better understanding of the underlying phenomenon.

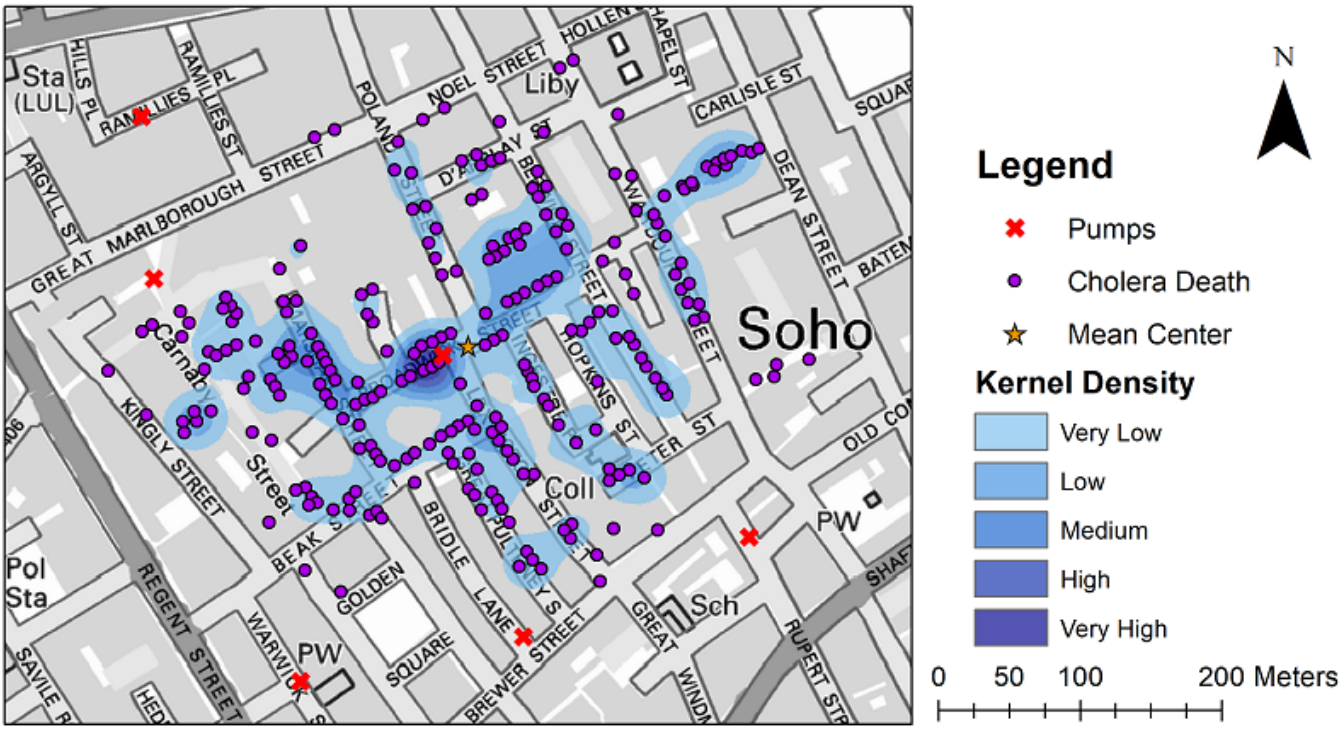


Figure 11: The kernel density of cholera cases in 1854 London. Map source: authors.

4. Closing Remarks

Point pattern analysis provides an effective way to visualize and interpret the distribution of point patterns across space. It is particularly useful for conducting exploratory analysis at an early stage of a research project. This entry reviewed commonly used methods for PPA, including descriptive statistics, distance-based measures, and density-based measures. These measurements provide effective tools for understanding the global and local patterns of point data.

References

Anderson, S., & Marcus, L. F. (1993). Effect of Quadrat Size on Measurements of Species Density. Journal of Biogeography, 20(4), 421-428.

Boots, B. N., & Getis, A. (1988). Point Pattern Analysis. Newbury Park, California: Sage Publications.

Clark, P. J., & Evans, F. C. (1954). Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations. Ecology, 35(4), 445-453.

de Smith, M. J., Goodchild, M. F., & Longley, P. A. (2007). Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools (2nd Edition). Troubador Publishing.

ESRI. (2018). How Directional Distribution (Standard Deviational Ellipse) Works.

Gatrell, A. C., Bailey, T. C., Diggle, P. J., & Rowlingson, B. S. (1996). Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology. Transactions of the Institute of British Geographers, 256-274.

Gimond, M. (2019, 08/09). Chapter 11 Point Pattern Analysis.

Kulin, H. W., & Kuenne, R. E. (1962). An Efficient Algorithm for the Numerical Solution of the Generalized Weber Problem in Spatial Economics. Journal of Regional Science, 4(2), 21-33.

Lamb, D. S., Downs, J. A., & Lee, C. (2016). The Network K-Function in Context: Examining the Effects of Network Structure on the Network K-Function. Transactions in GIS, 20(3), 448-460.

O'Sullivan, D. and Unwin, D. (2010) Geographic Information Analysis, 2nd Edition. John Wiley & Sons, Inc.

Okabe, A., & Yamada, I. (2001). The K-Function Method on a Network and Its Computational Implementation. Geographical Analysis, 33, 271-290.

Oyana, T. J., & Margai, F. M. (2016). Spatial Analysis: Statistics, Visualization, and Computational Methods. Boca Raton, Fla.: Taylor & Francis.

Rogerson, P. (2019). Statistical Methods for Geography (5th ed.). SAGE Publishers.

Snow, J. (1855). On the Mode of Communication of Cholera (2nd ed.). London, U.K.: J. Churchill.

Wang, B., Shi, W. Z., & Miao, Z. L. (2015). Confidence Analysis of Standard Deviational Ellipse and Its Extension into Higher Dimensional Euclidean Space. PLOS One, 10(3): e0118537.