University of California

Los Angeles

Methods for Spatial Analysis on a Network

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Victor Ying

2013

Abstract of the Thesis

Methods for Spatial Analysis on a Network

by

Victor Ying

Master of Science in Statistics

University of California, Los Angeles, 2013

Professor Mark Stephen Handcock, Chair

Network point patterns are usually analyzed by methods that assume a continuous plane and Euclidean distance. These methods fail to account for the constraint that network spatial phenomena must lie on a network. This paper proposes three statistical methods, called the network (inter-event distance) H-function method, network (nearest-neighbor distance) G-function method, and network (point-to-nearest-event/empty-space distance) F-function method. We do so by extending the existing H-, G-, and F-functions defined on a continuous plane with Euclidean distance, formulating these methods on a linear network with the shortest-path distance.

The thesis of Victor Ying is approved.

Qing Zhou

Nicolas Christou

Ying Nian Wu

Mark Stephen Handcock, Committee Chair

University of California, Los Angeles
2013

To my mother, father, grandparents, Krystal, Alan, Diana, Sara, Xu Huei, Molly, Jumbo, Oscar, Judy, David, Leni, Eric, David Jr., Derek, Fanny, and Johnny.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I wish to thank Drs. Mark Steven Handcock, Ying Nian Wu, Nicolas Christou, Qing Qhou, Mark Hansen, Hongquan Xu, Akram Mousa Almohalwas, Rick Paik Schoenberg, Sam Behseta, James O. Friel, George Jennings, Frédéric Brulois, Angel R. Pineda, Alfonso F. Agnew, Laurie Casaus, Perry Nicassio, Maciej Zworski, Ms. Glenda Jones, and Mr. Gary Dell'Osso.

CHAPTER 1

Motivation

Real world phenomena often occur either on or alongside a network. We refer to them as *network spatial phenomena.* Examples include the distribution of street crimes in Chicago (Figure 1.1) and spider webs on a brick wall (Figure 1.2). These phenomena are usually analyzed with methods designed for a continuous plane and Euclidean distance [15] – techniques called *planar spatial methods.* Analyses made using planar spatial methods are called *planar spatial analyses.* The justification for using planar spatial methods for the analysis of network spatial phenomena are that: (1) computing the Euclidian distance on a plane can often be easier than computing the network distance, and that (2) the network distance can sometimes be approximated by Euclidian distance [15]. While the first reason remains somewhat true, the validity of the second is questionable. For example, it is clear that the points in Figure 1.3a are not randomly distributed when the points are distributed on a plane. When the points are distributed on the network (as in Figure 1.3b), however, this is no longer true. The points are, in fact, randomly generated on the network. Even the first reason for using planar spatial methods is allayed by the fact that Geographical Information Systems (GIS) and advances in software that have made computation of the shortest-path distance, the prime example of a distance metric on a network, easy.

This illustrates the danger of analyzing network spatial phenomena using planar spatial methods. What should be used instead are methods that assume a network space using the shortest-path distance – the so-called *network spatial* *methods.* In this thesis, I develop analog network spatial methods based on existing planar spatial ones.

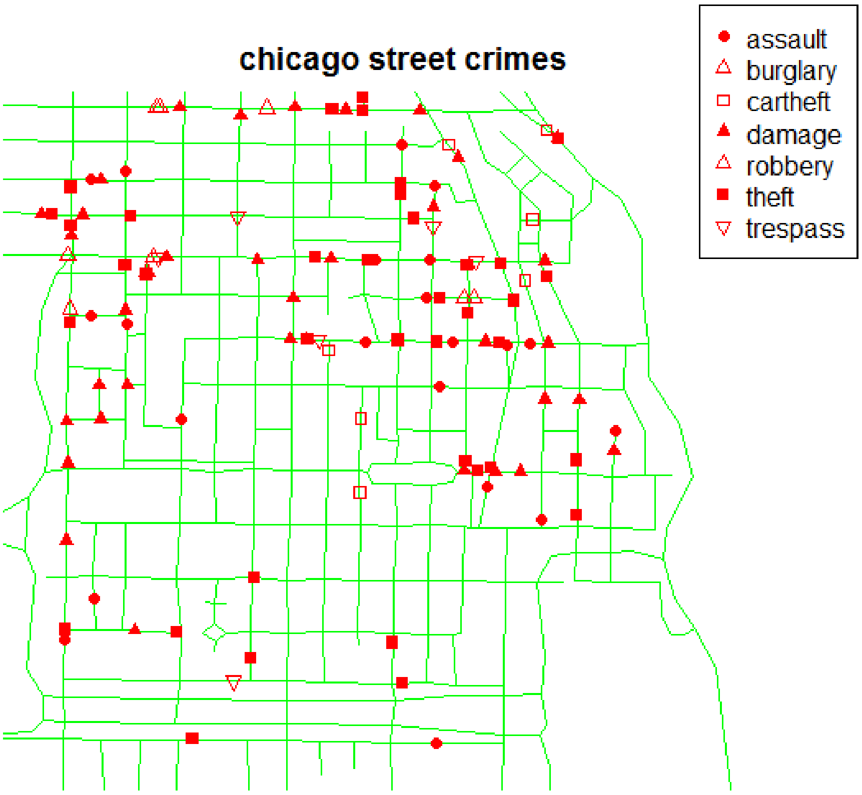


Figure 1.1: A record of street crimes in an area of Chicago. The points represent where crimes have occured.

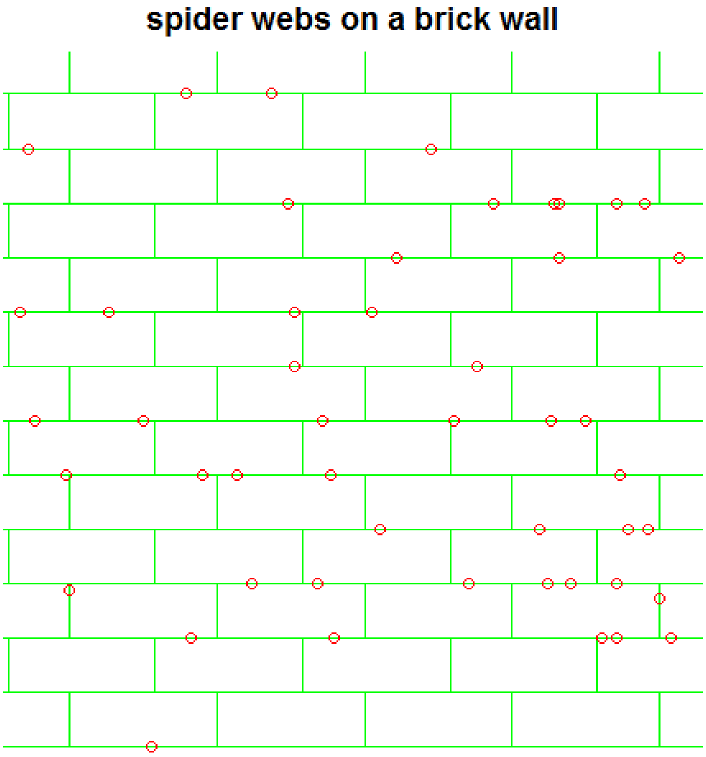
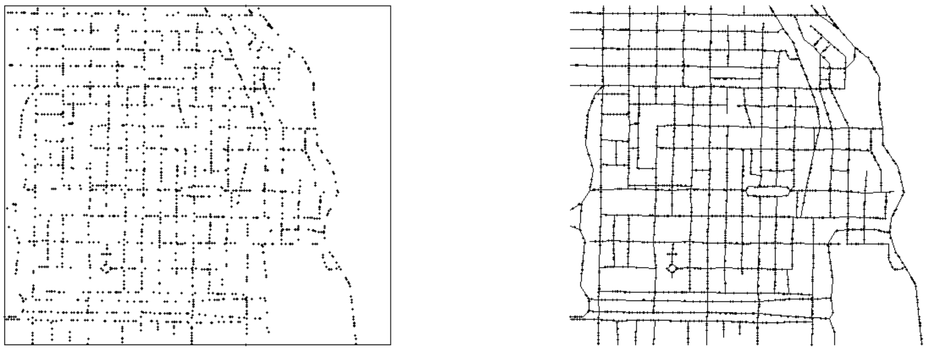


Figure 1.2: The plotted locations of 48 webs of the spider *Oecobius annulipes* on the mortar lines of a brick wall.



(a) (b)

Figure 1.3: (a) Non-randomly distributed points on a plane; (b) randomly distributed points on a network.

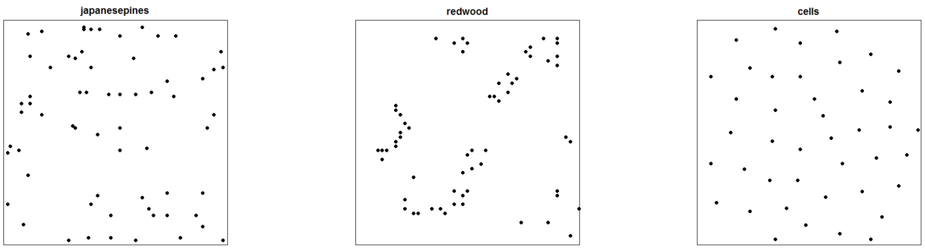
CHAPTER 2

Review of Spatial Statistics in $\mathbb{R}^2$

A spatial point pattern is a data-set that contains a set of points/objects/events distributed within a region of space. It might be thought of as consisting of a set of locations in a defined region at which events of interest have been recorded [1]. Such data occur in many different contexts. The points could, for example, represent trees, cells, crimes, forest fires, people, gorilla nests, etc. They can exist in a region of a one-dimensional line, two-dimensional plane, on the Earth's surface, or a three-dimensional volume; they can be points in space-time as well (for example, earthquake epicenter locations and time). We refer to the set of points within the region of interest as a 'spatial point pattern,' while the term 'spatial point process' refers to the stochastic mechanism that generates a random set of points in space.

Spatial point process methods are used to analyze point pattern datasets to answer questions such as whether the point pattern is exhibiting 'independence'/ 'complete randomness,' 'aggregation'/'clustering,' or 'regularity'/'repulsion' [1]. Independence or complete randomness refers to data with no clear structure, where the points are equally likely to occur anywhere within the study region; the term 'complete spatial randomness' (CSR) is used to describe this type of point pattern. The terms clustered and aggregated refer to points having a tendency to be located close to each other; regularity refers to patterns where the points have a tendency to be located away from each other. These three types of patterns are illustrated in Figure 2.1. Figure 2.1a, from Numata [14] via Diggle [10] and Baddeley &

Turner [4], gives the locations of 65 Japanese black pine saplings in a square sampling region in a natural forest. The absence of a clear structure here suggests it might be considered a completely random pattern. The second (Figure 2.1b), extracted by Ripley [17] from Strauss [18] and presented in Diggle [10], presents the locations of 62 seedlings and saplings of California redwood trees in a square sampling region. They appear to be 'clustered' or 'aggregated'. The third and final figure (Figure 2.1c), from Ripley via Diggle, records the locations of the centers of 42 biological cells in a histological section observed under optical microscopy; the cells here appear to exhibit a regular pattern.



(a) Independent: locations of Japanese black pine saplings [14, 10, 4].

(b) Aggregated: locations of 62 seedlings and saplings of California redwood trees [18, 17, 4].

(c) Regular: locations of the centers of 42 biological cells [17, 4].

Figure 2.1: Examples of independent/completely random, clustered/aggregated, and regular point patterns, respectively.

2.1 Complete Spatial Randomness (CSR)

CSR of a spatial point pattern implies two things:

1. The number of events in any two-dimensional region A with area $|A|$ follows a Poisson distribution with mean $\lambda|A|$, where the constant λ is the *intensity*

(that is, mean number of events per unit area)

2. The n events (x_i, where $i = 1, \ldots, n$) in the region A are an independent sample from the uniform distribution on A.

Property 1 of CSR implies that the intensity is constant and does not vary over the 2D area. Property 2 implies that the events do not interact– that is, the existence of an event at x neither encourages nor inhibits the occurrence of any other event around x.

Preliminary analysis of a point pattern dataset often begins with a test of CSR. Although CSR is of limited scientific interest in itself, there are several reasons for beginning with a test of CSR:

1. Rejection of CSR is a minimal prerequisite to any serious attempt to model an observed pattern.
2. The tests might provide insight into the point pattern and suggest plausible alternative models.
3. CSR acts as a dividing hypothesis between regular and aggregated patterns [1, 10, 5].

2.2 Edge Effects

Edge effects often arise in the analysis of spatial point patterns. Edge effects occur when the region of observation, A, is part of a larger region in which we do not have data. The issue is that unobserved events outside of A might interact with observed events within A. Since unobserved events are not observed, it is difficult to assess their effects upon points within our window. That is, while a point process, $\mathbf{X}$, might extend throughout 2-D space, it is only observed within A, and, by confining observations to A, the observed distance $d(u, \mathbf{x}) = d(u, \mathbf{X} \cap A)$

from the point u to the nearest point inside A might end up being greater than the true distance $d(u, \mathbf{X})$ from u to the nearest point of the complete point process. This is illustrated in Figure 2.2. For some kinds of exploratory analysis edge effects can be ignored, but for others corrections will be required.

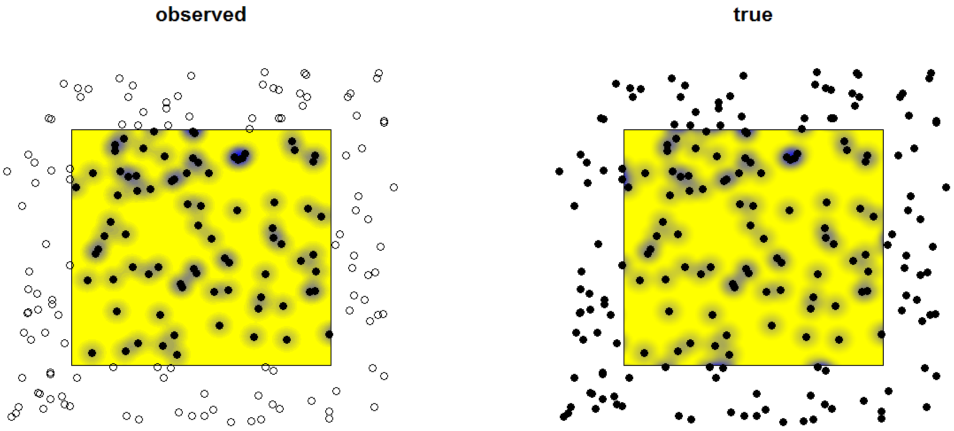


Figure 2.2: Edge effects illustrated [4, p. 117].

2.3 Monte Carlo Tests and Simulation Envelopes

2.3.1 Pointwise Monte Carlo Tests

Monte Carlo tests are often useful when the use of other methods is not feasible. It consists of ranking the value, v_1, of a test statistic V against a corresponding set of values $v_i : i = 2, \ldots, s$, generated by independent random sampling from the distribution of V under a simple null hypothesis $\mathcal{H}$. That is, to assess the significance of the observed v_1, we carry out $s - 1$ simulations and calculate the corresponding quantities $v_2, \ldots, v_s$. The significance level is evaluated from the rank of v_1 amongst the order-statistics $v_{(1)} < \cdots < v_{(s)}$ [7, 13]. Let $v_{(j)}$ denote the jth largest amongst the $v_i : 1, \ldots, s$. Then, under $\mathcal{H}$, each of the s possible rankings of v_1 are equally likely and we have

$$P\{v_1 = v_{(j)}\} = s^{-1} : j = 1, \ldots, s$$

Hence, for an upper-tailed test, rejection of $\mathcal{H}$ on the basis that v_1 ranks the kth largest or higher gives a one-sided test of size k/s. For a lower-tailed test, rejection on the basis that v_1 ranks kth smallest or lower also gives a test of size k/s. For a two-tailed/two-sided test, the p-value is 2 times the smaller of the one-tailed p-values above [11]. These results all assume that the values of v_i are all different and that there are no ties so the ranking is unambiguous. If V is discrete so that ties are possible, then either choose the least extreme rank for v_1 as a conservative rule or use a randomized p-value to break the tie.

To help interpret the significance (or lack thereof) of the observed test statistics $\hat{V}$ against the value expected under CSR, *simulation envelopes* are often used. To construct pointwise simulation envelopes, we proceed as follows. Generate $s - 1$ independent simulations of CSR inside the study region A. Compute the estimated V functions for each of these realizations, that is $\hat{V}^{(j)}(r)$ for $j = 2, \ldots, s-1$. Define the pointwise upper and lower envelopes of these simulated curves to be

$$\begin{aligned} L(r) &= \min_j \hat{V}^{(j)}(r) \\ U(r) &= \max_j \hat{V}^{(j)}(r). \end{aligned}$$

For any fixed value of r, if the data come from a uniform Poisson process then the probability that $\hat{V}(r)$ lies outside the envelope $[L(r), U(r)]$ is $2/s$; that is, the test that rejects the null hypothesis of CSR when $\hat{V}(r)$ lies outside $[L(r), U(r)]$, has significance level $\alpha = 2/s$. Thus, with 39 simulations, we have a test of size $\alpha = 2/40 = 0.05$. If one prefers to use the pointwise order statistic (that is, the pointwise kth largest and kth smallest values), then the resulting test is of size $\alpha = 2k/s$ [3, pp. 132-133].

2.3.2 Simultaneous Monte Carlo Tests

Note that the pointwise test above requires that r be fixed in advance. If we had based our decision on whether the empirical function (say V in the notation of the previous subsection) *ever* wandered outside the envelope, then we would have chosen the value of r in a data-dependent way, and the true significance level would be higher (that is, less significant). To avoid this problem, we can construct *simultaneous critical bands* with the property that, under $\mathcal{H}$, the probability of $\hat{V}$ ever wandering outside the critical bands is exactly 5% (or any other desired α).

Simultaneous critical bands are created by first computing, for each of the $s-1$ estimated $\hat{V}(r)$, its maximum deviation from the theoretical (under CSR) V function:

$$D^{(j)} = \max_r |\hat{V}^{(j)}(r) - V_{pois}(r)|, \tag{2.1}$$

where $V_{pois}(r)$ is the theoretical V function under the null hypothesis of CSR and j runs from 2 to s. For each of the $M = s-1$ simulated datasets, compute a value of $D^{(j)}$. Among these, determine the largest $D^{(j)}$ and refer to this maximum value as $D_{\max}$. The upper and lower limits are

$$L(r) = V_{pois}(r) - D_{\max} \tag{2.2}$$

$$U(r) = V_{pois}(r) + D_{\max}. \tag{2.3}$$

The observed empirical $\hat{V}^{(1)}(r)$ function then exceeds these limits only if the $D^{(1)}$ for the data is greater than $D_{\max}$. Under the null hypothesis $\mathcal{H}$, this occurs with probability $1/(M+1)$. For a test of size $\alpha = 5\%$, use $M = s-1 = 19$ simulations.

2.4 Preliminary Testing, Exploratory Data Analysis (EDA)

For reasons discussed in Section 2.1, it is often useful to begin an analysis of a point pattern dataset with a test of CSR. The remainder of this chapter will describe a number of different test that have been proposed to do so. The main classical techniques, which will be described in this chapter, are the *distance methods*, which are based on measuring distances between points. Three that we may consider are

- **pairwise distances** $t_{ij} = \|x_i - x_j\|$ between all distinct pairs of points x_i and x_j $(i \neq j)$ in the pattern. They are also referred to as *inter-event distance*. We shall refer to them as pairwise distances here;
- **nearest neighbor distances** $t_i = \min_{j \neq i} t_{ij}$, the distance from a point x_i to its nearest neighbor;
- **empty space distances** $d(u, \mathbf{x}) = \min_i\{\|u - x_i\| : x_i \in \mathbf{x}\}$, the distance from a fixed reference location $u \in \mathbb{R}^2$ in the observation window to the nearest data point x_i in the point pattern $\mathbf{X}$. (They are also known as point-to-nearest-event distances, but we shall continue to use the term empty space distances.)

2.4.1 Pairwise Distances

The first summary description we consider is the empirical distribution of *pairwise distances*. For a pattern of n events in a region A, there are $\frac{1}{2}n(n-1)$ pairwise distances from event x_i to event x_j. Assuming the point process $\mathbf{X}$ is stationary, we define the cumulative distribution function (CDF) of pairwise distances T to be [9, p. 1]:

$$H(r) = \mathbb{P}\{t_{ij} \leq r\} \tag{2.4}$$

The empirical distribution function (EDF) of the observed inter-event distances is

$$\begin{aligned} \hat{H}_1(r) &= \left\{\frac{1}{2}n(n-1)\right\}^{-1} \#(t_{ij} \leq r) &(2.5)\\ &= \frac{2}{n(n-1)} \sum_{i \neq j} \mathbf{1}\{\|x_i - x_j\| \leq r\}, &(2.6) \end{aligned}$$

where $\#(t_{ij} \leq r)$ means 'the number of t_{ij} less than or equal to r', x_i and x_j are events in the observed point pattern, and $\mathbf{1}\{\cdot\}$ is the indicator function. It is negatively biased, since we can never observe a pairwise distance greater than the diameter of the window [3, p 125].

To interpret this estimate, we compare it to what would be expected under CSR. In general, the theoretical distribution function of T between two events independently and uniformly distributed in A will depend on the size and shape of A but is known for the most common cases when A is a square or circle [10]. When A is a square of unit side, the theoretical distribution function of T is

$$H_{pois}(r) = \begin{cases} \pi r^2 - 8r^3/3 + r^4/2, & \text{if } 0 \leq r \leq< 1, \\ 1/3 - 2r^2 - r^4/2 + 4(r^2-1)^{\frac{1}{2}}(2r^2+1)/3 & \\ \quad +2r^2 \sin^{-1}(2r^{-2}-1), & \text{if } 1 < r \leq \sqrt{2}. \end{cases} \tag{2.7}$$

When A is a circle of unit radius, it is

$$H_{pois}(r) = 1+\pi^{-1}\{2(r^2-1)cos^{-1}(r/2)-r(1+r^2/2)\sqrt{(1-r^2/4)}\}, \text{for all } 0 \leq r \leq 2. \tag{2.8}$$

If A is neither a square nor circle, then we can proceed by using Monte Carlo tests with simulation envelopes (see below).

Typically we compare $\hat{H}_1(r)$ with $H_{pois}(r)$. If we create a plot of $\hat{H}_1(r)$ as ordinate against $H_{pois}(r)$ as abscissa, then the plot should be roughly linear for data compatible with CSR. Values $\hat{H}_1(r) > H_{pois}(r)$ for small values of $H_{pois}(r)$ means an excess of small inter-event distances, which suggests a clustered pattern.

Monte Carlo tests with simulation envelopes can also be used to assess the significance or lack thereof of departures from linearity. Two of the many approaches are as follows:

1. Choose r_0 and define

$$v_i = \hat{H}_i(r_0).$$

Then proceed as in Section 2.3.

2. Define

$$v_i = \int \{\hat{H}_i(r) - H_{pois}(r)\}^2 \tag{2.9}$$

to be a measure of divergence between $\hat{H}_i(r)$ and $H_{pois}(r)$ over the entire range of r and again proceed as in Section 2.3.

If the region A is one where the theoretical distribution function $H_{pois}(r)$ is unknown, a test can still be performed by replacing $H_{pois}(r)$ in equation 2.9 with

$$\overline{H}_i(r) = (s-1)^{-1} \sum_{j \neq i} \hat{H}_j(r). \tag{2.10}$$

2.4.2 Nearest-Neighbor Distances

Assuming the point process $\mathbf{X}$ is stationary (invariant under translations), we define the CDF of nearest neighbor distances to be:

$$G(r) = \mathbb{P}\{d(u, \mathbf{X}\ \{u\}) \leq r | u \in \mathbf{X}\}, \tag{2.11}$$

where u is an arbitrary event and $d(u, \mathbf{X}\ \{u\})$ is the shortest distance from u to its nearest neighbor in the point pattern $\mathbf{X}$. The EDF of observed nearest-neighbor distances

$$G_1^*(r) = n^{-1} \sum_i \mathbf{1}\{t_i \leq r\} \tag{2.12}$$

is a positively biased estimator of $G(r)$, since confining observations to a window A means the observed nearest-neighbor distances are generally greater than the actual nearest neighbor distances of points in the entire point process $\mathbf{X}$ [3, pp. 122-123]. A correction for this 'edge effect bias' is required. It will typically be a weighted versions of the EDF

$$\hat{G}_1(r) = \sum_i e(u_j, r)\mathbf{1}\{t_i \leq r\}, \tag{2.13}$$

where $e(u, r)$ is an edge correction so that $\hat{G}$ is unbiased.

For a homogeneous Poisson process of intensity λ the theoretical CDF of nearest neighbor distances is known to be [3, p 123]

$$G_{pois}(r) = 1 - \exp(-\lambda \pi r^2). \tag{2.14}$$

Again, we can compare $\hat{G}_1(r)$ with $G_{pois}(r)$ by plotting one against the other and comparing with simulation envelopes as in Section 2.4.2. Values of $\hat{G}_1(r) > G_{pois}(r)$ imply shorter observed nearest-neighbor distances than for a Poisson process, suggesting a clustered pattern. Values of $\hat{G}_1(r) < G_{pois}(r)$ suggest a regular pattern.

The observed EDF $\hat{G}_1(r)$ can also be compared with simulation envelopes from simulated EDFs $\hat{G}_i(r) : i = 2, \ldots, s$ exactly as in Section 2.4.1.

2.4.3 Empty-Space Distances and the *F* Function

Assuming $\mathbf{X}$ is stationary, we define the CDF of empty space distances to be

$$F(r) = \mathbb{P}\{d(u, \mathbf{X}) \leq r\}, \tag{2.15}$$

where u is an arbitrary reference location.

The EDF of observed empty space distances on a grid of m sample points, $u_j : j = 1, \ldots, m$, in A is defined to be

$$F^*(r) = m^{-1} \sum_j \mathbf{1}\{d(u, \mathbf{X}) \leq r\}. \tag{2.16}$$

For reasons explained previously (in Section 2.2), it is a positively biased estimator of $F(r)$. Edge corrections are usually weighted versions of the EDF

$$\hat{F}(r) = \sum_j e(u_j, r) \mathbf{1}\{d(u_j, \mathbf{x}) \leq r\}, \tag{2.17}$$

where $e(u, r)$ is an edge correction so that $\hat{F}(r)$ is unbiased.

Again, we can compare the estimated observed $\hat{F}(r)$ with $F_{pois}(r)$. The theoretical CDF of empty space distances for a homogeneous Poisson process is provided by Diggle

$$F_{pois}(r) = 1 - \exp(-\lambda \pi r^2). \tag{2.18}$$

(This is identical to the nearest-neighbor function for the Poisson process. Hence F is equivalent to G. Interpretation is reversed, however.) Values of $\hat{F}(r) > F_{pois}(r)$ imply that empty space distances observed in the point pattern are shorter than for a Poisson process, suggesting a regular pattern.; $\hat{F}(r) < F_{pois}(r)$ suggest a clustered pattern.

The observed EDF $\hat{F}_1(r)$ can be analyzed by plotting against $F_{pois}(r)$ or comparing with simulation envelopes from simulated EDFs $\hat{F}_i(r) : i = 2, \ldots, s$ as in Section 2.4.1.

2.4.4 Pairwise Distances Revisited and the K function

If the distribution of a point process X is invariant under translation, we say that it is *stationary.* Diggle reports that Ripley defined the K-function for a stationary point process so that $\lambda K(r)$ is the expected number of other points of the process. Ripley's K-function is defined as

$$K(r) = \lambda^{-1} E[N_0(r)], \tag{2.19}$$

where $N_0(r)$ is the number of other events within a distance r of a typical or arbitrary point/event of the process. For a homogeneous Poisson process we have

$$K_{pois}(r) = \pi r^2. \tag{2.20}$$

Most estimators of K are weighted and renormalized EDFs of the pairwise distances

$$\hat{K}_1(r) = \frac{1}{\hat{\lambda}\text{area}(W)} \sum_i \sum_{j \neq i} \mathbf{1}\{\|x_i - x_j\| \leq r\} e(x_i, x_j; r), \tag{2.21}$$

where $e(x_i, x_j; r)$ is an edge correction. Without an edge correction, Equation (2.21) is equal to Equation (2.5) up to a multiplicative constant factor. Thus the K function is equivalent to the H function. We can compare the estimate $\hat{K}_1(r)$ with K_{pois} and/or with simulation envelopes as in Section 2.4.1. $\hat{K}_1(r) > \pi r^2$ suggests clustering; $\hat{K}_1(r) < \pi r^2$ suggests a regular pattern.

2.4.5 Pair Correlation Function g

Another summary function is the *pair correlation function* (pcf)

$$g(r) = \frac{K'(r)}{2\pi r}, \tag{2.22}$$

where $K'(r)$ is the derivative of K. It is roughly interpreted as the probability of observing a pair of points separated by a distance r, divided by the corresponding probability in a Poisson process. While in some ways easier to interpret, it is more difficult to estimate. We proceed by inspecting a plot of the estimated pcf. The value $g(r) = 1$ suggests complete randomness, since $g_{pois} \equiv 1$, while values $g(r) > 1$ imply clustering (at distance r) and values $g(r) < 1$ imply a regular pattern. [3].

2.4.6 Quadrat Counts

Quadrat counting is an alternative to the distance-based approaches discussed so far. It involves dividing the observation window A into m rectangular subregions (or quadrats) of equal size and and counting the number of points in each rectangle to test homogeneity (that is, CSR). Specifically, we assume A is the unit square and partition it into a regular $k \times k$ grid of quadrats so that $m = k^2$. Let $n_i : i = 1, \ldots, m$ be the number of points counted in each quadrat and $\bar{n} = n/m$ be the sample mean of the n_i. Departures from CSR are assessed based on Pearson's

χ-squared statistic

$$\chi^2 = \sum_{i=1}^{m} (n_i - \bar{n})^2 / \bar{n}, \tag{2.23}$$

which has a χ^2_{m-1} null distribution provided $\bar{n}$ is not too small [10].

CHAPTER 3

Review of Linear Point Processes

In this chapter, we will first present the basic definitions and terminology for a linear network and related geometrical objects. We will then provide a summary of homogeneous and inhomogeneous Poisson processes restricted to a linear network.

3.1 Basic Definitions for a Network

Definition 1. A set $l \subset \mathbb{R}^2$ is called a *line segment* in the plane with endpoints u and v iff l can be expressed as

$$l = l_{u,v} = \{tu + (1-t)v : 0 \leq t \leq 1\},$$

for some points $u, v \in \mathbb{R}^2, u \neq v$. $l_{u,v}$ is referred to as the line segment between u and v, and the length of the line segment is defined as

$$|l| = |l_{u,v}| = \|u - v\|,$$

where $\|.\|$ is the Euclidean distance in $\mathbb{R}^2$.

Definition 2. A *linear network*, L, in $\mathbb{R}^2$ is the union

$$L = \bigcup_{i=1}^{n} l_i,$$

of a finite collection of line segments $l_1, \ldots, l_n$ in the plane, where $l_i = l_{u_i,v_i}$ are segments of finite length in $\mathbb{R}^2$ such that $|l_i \cap l_j| = 0$ for $i \neq j$, and the segments are maximal in the sense that $l_i \cup l_j$ is not a line segment for any $i \neq j$.

Definition 3. Let $L = \cup_{i=1}^{n} l_i$ be a linear network. The total *length* of all line segments in L is defined as

$$|L| = \sum_{i=1}^{n} |l_i|.$$

Definition 4. Let L be a linear network. A point $v \in L$ is called an *intersection vertex* if $v = l_i \cap l_j$ for some $i, j, i \neq j$. It is called a *free vertex* if $l_i = l_{x,v}$ (or $l_{v,x}$) for some i, x, and $l_i \cap l_j \neq v$ for any $j \neq i$. The union of intersection vertices and free vertices is called the set of *vertices* of L.

Definition 5. Let L be a linear network and $u, v, \in L$. The *path* between two points, u and v, in L is a sequence

$$x_0, x_1, \ldots, x_m$$

such that

$$x_0 = u,$$
$$x_k = v,$$
$$[x_i, x_{i+1}] \subset L \text{ for each } i = 0, \ldots, m-1.$$

and is denoted by $P(u, x_1, \ldots, x_{k-1}, v)$.

Definition 6. If $P(u, x_1, \ldots, x_{k-1}, v)$ is a path in a linear network L, then the *length of the path* is defined as

$$\|u - x_1\| + \|x1 - x_2\| + \ldots + \|x_{k-1} - v\|.$$

Definition 7. The *shortest path distance* $d_L(u, v)$ between two points, u and v in L is the length of the shortest possible path from u to v. If no path from u to v exists then we have $d_L(u, v) = \infty$.

Definition 8. The *disc*, $b_L(u, r)$, of radius $r > 0$ and center point u in L is defined as the set of all points v in the network whose shortest path distance from u is less than or equal to r:

$$b_L(u, r) = \{v \in L : d_L(u, v) \leq r.\}$$

Definition 9. The *relative boundary*, $d_L(u,v)$, of $b_L(u,r)$ is the set of points lying exactly r units away from u:

$$\partial b_L(u,r) = \{v \in L : d_L(u,v) = r\}.$$

Definition 10. The number of points in the relative boundary $\partial b_L(u,r)$ is known as the *circumference* and is denoted $m(u,r)$. It is the number of points of L lying exactly r units away by the shortest path from u:

$$\begin{aligned} m(u,r) &= \#\partial b_L(u,r) && (3.1)\\ &= \text{ number of points } v \text{ in } L \text{ such that } d_L(u,v) = r && (3.2)\\ &= N_L\{v : d_L(u,v) = r.\} && (3.3) \end{aligned}$$

It is finite for all $r < \infty$ and, by convention, we set $m(u,\infty) = \infty$.

Definition 11. Given a linear network L, the quantity

$$R \equiv d_{max} = \min_{u \in L} \max_{v \in L} d_L(u,v)$$

is analogous to and can be interpreted as the *circumradius* of the network– that is, the radius of the smallest disc containing the entire network.

3.2 Point Processes in a Linear Network

A point process on a linear network L is a stochastic mechanism that generates a finite set of point on L where both the number of points and their locations are random. Let $\mathbf{x} = \{x_1, \ldots, x_n\}, x_i \in L, n \geq 0$ denote a realization of the point process $\mathbf{X}$ on the linear network L.

Definition 12. If $\mathbf{X}$ is a point pattern in a linear network L and $X \subseteq L$, then $N_{\mathbf{X}}(S)$ is defined as the number of points in $\mathbf{X}$ and S. That is,

$$N_{\mathbf{X}}(S) = \text{number of points in } \mathbf{X} \cap S. \qquad (3.4)$$

A feature of a point process that gives a basic idea of the distribution of a point pattern is the *intensity*, which measures the mean rate of occurrence of points per unit length. The intensity may vary at different locations.

Definition 13. A point process $\mathbf{X}$ on a linear network L is said to have *[first moment] intensity function* $\lambda(u), u \in L$ if, for any line segment $S \subseteq L$,

$$\mathbb{E}[N_{\mathbf{X}}(S)] = \int_S \lambda(u) d_1 u, \tag{3.5}$$

where $\int_S \lambda(u) d_1 u$ is the one-dimensional integration over the line segment.

$\lambda(u)$ can be interpreted intuitively as the expected number of random points per unit length in a small neighborhood or interval about u:

$$\lambda(u) = \lim_{|\delta u| \to 0} \frac{\mathbb{E}[N_{\mathbf{X}}(\delta u)]}{|\delta u|},$$

where δu is a small interval around u. When the expected number of points per unit length is constant regardless of location (that is, when it is equal to some constant λ), then we say that it is first-order stationary (homogeneous) with intensity function $\lambda(u) = \lambda$, where λ can be interpreted as the expected number of points per unit length. Formally we have the following:

Definition 14. Let L be a linear network and $\mathbf{X}$ be a point process in L. Then we say that $\mathbf{x}$ is *first-order stationary* if, for any $S \subseteq L$,

$$\mathbb{E}[N_{\mathbf{X}}(S)] = \lambda \cdot |S|,$$

where λ is the intensity of $\mathbf{X}$.

CHAPTER 4

Exploratory Data Analysis on a Linear Network

Chapter 4 describes how the methods in Section 2.4 can be applied to linear point processes. The approach used in this chapter are the Monte Carlo methods. First, we define new network spatial methods based on existing planar spatial methods discussed in Section 2.4 by substituting the Euclidian distances used there with the shortest-path distances on a network. We take the linear network under consideration and repeatedly simulate actualizations of a completely random point process on that network. Calculating the summary description for each actualization, we determine the upper and lower envelopes for the summary description. We can then compare our observed summary description with the upper and lower envelopes expected under CSR.

We will need the following definitions:

- the **network pairwise distances**, $d_{ij} = \|x_i - x_j\|_l = d_L(x_i, x_j)$, between all distinct pairs of points x_i and x_j $(i \neq j)$ in the pattern is the length of the shortest path between point x_i and x_j (that is, the "as the ambulance flies" distance). Okabe & Yamada (2001) refer to this length as the *network distance*, or simply *distance* when a network is understood. We shall follow that convention here.

- **network nearest neighbor distances**, $d_i = \min_{j \neq i} d_{ij}$, is the distance from a point x_i to its nearest neighbor in the network L;

- **network empty space distances**, $d_L(u, \mathbf{x}) = \min_i\{\|u - x_i\|_l : x_i \in \mathbf{x}\}$, is the distance from a fixed reference location $u \in L$ in the observation window to the nearest data point, x_i, in the point pattern, $\mathbf{X}$, in the network, L.

4.1 Pairwise Distances on a Network

The first summary description we consider is the empirical distribution of pairwise distances on a linear network. For a pattern of n events in a region A, there are $\frac{1}{2}n(n-1)$ pairwise distances from event x_i to event x_j. We define the CDF of pairwise distances D to be:

$$H_l(r) = \mathbb{P}\{d_{ij} \leq r\}. \tag{4.1}$$

Note that this is just our earlier definition of theoretical pairwise distance described in Section 2.4.1 with the Euclidian distance replaced by the shortest-path distance on a network.

Similarly, we define the EDF of the observed inter-event distances to be:

$$\begin{aligned} \hat{H}_l(r) &= \left\{\frac{1}{2}n(n-1)\right\}^{-1} \#(d_{ij} \leq r) && (4.2) \\ &= \frac{2}{n(n-1)} \sum_{i \neq j} \mathbf{1}\{\|x_i - x_j\|_l \leq r\}, && (4.3) \end{aligned}$$

where $\#(d_{ij} \leq r)$ means 'the number of d_{ij} less than or equal to r', x_i and x_j are events in the observed network point pattern, $\mathbf{1}\{\cdot\}$ is the indicator function, and $\|x_i - x_j\|_l$ is the shortest-path distance between x_i and x_j. It is negatively biased, since we can never observe a pairwise distance greater than the diameter of the window [3, p 125].

4.2 Nearest-Neighbor Distances on a Network

We define the CDF of nearest neighbor distances on a network to be:

$$G(r) = \mathbb{P}\{d_L(u, \mathbf{X}\ \{u\}) \leq r | u \in \mathbf{X}\}, \tag{4.4}$$

where u is an arbitrary event and $d_L(u, \mathbf{X} \ \{u\})$ is the shortest distance from u to its nearest neighbor in the point pattern $\mathbf{X}$ on a linear network L. The EDF of observed nearest-neighbor distances is

$$\hat{G}_l(r) = n^{-1} \sum_i \mathbf{1}\{d_i \leq r\}, \tag{4.5}$$

a positively biased estimator of $G(r)$ for the same reasons as before.

4.3 Empty-Space Distances on a Network

The CDF of empty space distances on a network is

$$F_l(r) = \mathbb{P}\{d_L(u, \mathbf{X}) \leq r\}, \tag{4.6}$$

where u is an arbitrary reference location.

The EDF of observed empty space distances on a grid of m sample points, $u_j : j = 1, \ldots, m$, in A is defined to be

$$\hat{F}_l(r) = m^{-1} \sum_j \mathbf{1}\{d_L(u, \mathbf{X}) \leq r\}, \tag{4.7}$$

where Diggle [10] recommends using the same number of grid points, m, as there are events in the observed pattern. $\hat{F}_l(r)$ is positively biased.

4.4 The Network K Function

There is already an analog of the planar K function for use on linear networks. It was defined by Okabe & Yamada in 2001 as [16, 1]:

$$\hat{K}_{OY}(r) = \frac{1}{\hat{\lambda}^2 |L|} \sum_i \sum_{i \neq j} \mathbf{1}\{d_L(x_i, x_j) \leq r\}. \tag{4.8}$$

It is equal to the $\hat{H}_l$ function (Equation 4.2) up to a multiplicative constant. We will show in the next chapter that the two are essentially the same. An edge corrected version was later developed by Ang [2].

4.5 The Pair Correlation Function g on a Network

An analog of the pcf for linear networks is defined by Ang et al. [2] as:

$$\rho_l(r) = \frac{1}{\sum_i 1/\hat{\lambda}(x_i)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{k(d_l(x_i, x_j) - r)}{\hat{\lambda}(x_i)\hat{\lambda}(x_j) m(x_i, d_l(x_i, x_j))} \tag{4.9}$$

where k is a kernel on $\mathbb{R}$ and $m(\cdot, \cdot)$ is the circumference as defined in (3.1).

4.6 Quadrat Counting on a Network

The quadrat counting approach used to analyze spatial point patterns in $\mathbb{R}^2$ does not seem to lend itself for use on networks. One approach might be to partition the network into m equal segments each of length $|L|/m$, where $|L|$ is the total length of the network. If we treat each segment as a "quadrat," then we should be able to proceed with the analysis as in Section 2.4.6.

CHAPTER 5

Data & Analysis

Our goal is to use our newly proposed summary functions to determine whether our data sets can be modeled as a homogeneous Poisson process. If the data were not compatible with CSR, then the question would be whether they are aggregated or exhibit regularity. We also seek to compare the results of planar summary functions with the results of their network analogs – by first treating each point pattern as a two-dimensional rectangle, analyzing it with a planar summary function, and then analyzing it as a point pattern on a network with a network summary function. We will determine if there is any difference and, if so, try to explain why.

5.1 Data

5.1.1 Chicago Street Crime Data

Figure 1.1 shows a record of street crimes in an area of Chicago in a two week period from April 25, 2002 to May 8, 2002, available in the `spatstat` package for the R programming language. It was reported by the Chicago Weekly News, manually digitized by Adrian Baddeley, and can be accessed by entering:

```
> data(chicago)
```

in R. The data give the location of each crime, all of which occurred either on or near the network of streets. It also gives the type of crime.

5.1.2 Spider Nest Data

The first data set (presented in Figure 1.2) gives the positions of 48 webs of the spider *Oecobius navus* on the network of mortar lines of a brick wall. The data came from the positions of webs on two different sections of wall, each sampled six times over a period of six weeks for a total of twelve point patterns. This thesis will focus on only the first point pattern – point pattern QI 2705. We wish to elicit information about the habitat preferences, resource requirements, and co-specific interaction of the species.

5.2 Analysis

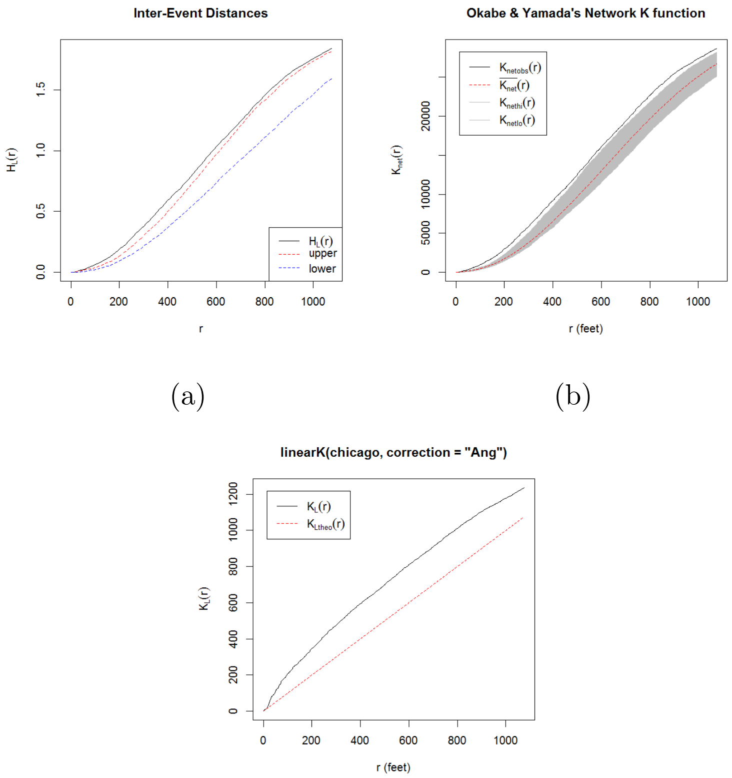
5.2.1 Pairwise Distances and the Network K Function

5.2.1.1 Analysis of Chicago Street Crimes

As discussed in Section 4.4, the pairwise distance H function is equal to Okabe & Yamada's Network K function up to a multiplicative constant. This is illustrated in Figure 5.1, where the two plots appear identical. Figure 5.2a shows an EDF plot of pairwise distances along with the upper and lower boundaries expected under 99 simulations of CSR. We see that the empirical network H function lies well above the upper boundary throughout its range; we conclude that these data are incompatible with a completely random spatial distribution of crimes. The same conclusion is reached with a visual inspection of Ang's geometrically-corrected network K-function.

5.2.1.2 Analysis of Spider Nests

For the spider nest data, we again plot the empirical H-function along with the upper and lower envelopes expected under 99 simulations of CSR. The EDF plot,

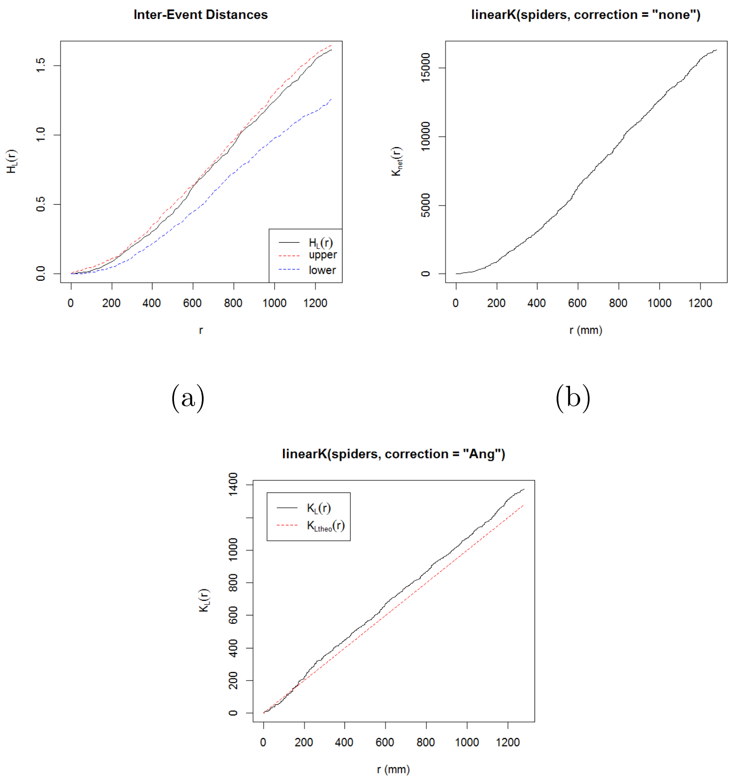


(a)

(b)

(c)

Figure 5.1: (a) Empirical network H function for the chicago dataset with upper and lower envelopes expected under CSR. (b) Okabe & Yamada's network K function for the chicago dataset. (c) Ang's geometrically corrected network K function for the chicago dataset.



(a) (b)

(c)

Figure 5.2: (a) Empirical network H function for the spider nest dataset with upper and lower envelopes expected under CSR. (b) Okabe & Yamada's network K function for the spider dataset. (c) Ang's geometrically corrected network K function for the spider dataset.

Figure 5.2, shows that the summary function never wanders beyond the upper and lower envelopes. This indicates that the data are compatible with CSR.

5.2.2 Nearest-Neighbor Distances

5.2.2.1 Analysis of Chicago Street Crimes

Figure 5.3 shows the EDF plot of network nearest neighbor distances for the Chicago street crimes dataset with corresponding lower and upper envelopes from

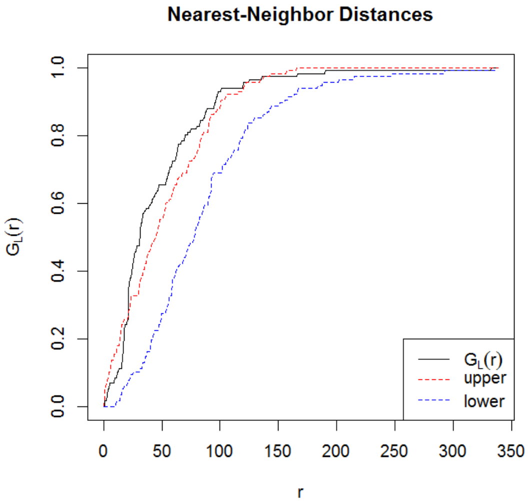


Figure 5.3: EDF plot of nearest neighbor distances for Chicago street crime data. Solid curve represents data. Dashed curves represent upper and lower envelopes from 99 simulations of CSR.

99 simulations of CSR. It shows an excess of small nearest neighbor distances, which is characteristic of aggregated patterns. Therefore we feel that there is ample evidence for rejecting the hypothesis of CSR in favor of an aggregated alternative.

5.2.2.2 Analysis of Spider Nests

For the spider nests, the EDF plot in Figure 5.4 shows that the data lies within the upper and lower envelopes throughout its range. It suggests acceptance of CSR.

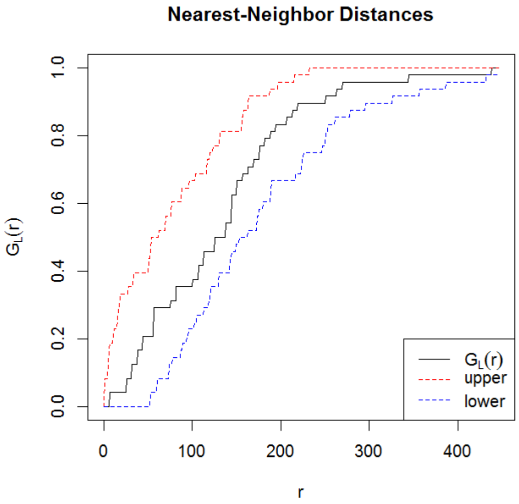


Figure 5.4: EDF plot of nearest neighbor distances for spider nests data. Solid curve represents data. Dashed curves represent upper and lower envelopes from 99 simulations of CSR.

5.2.3 Empty-Space Distances

5.2.3.1 Analysis of Chicago Street Crimes

Figure 5.5 shows the EDF plot for an empty-space analysis of Chicago street crime data, using a grid of $m = 116$ points, where m was determined by following Diggle's rule of thumb of choosing as many grid points as there are events in the observed pattern. $\hat{F}_l(x)$ lies below the lower simulation envelope for most of its range. This is typical of an aggregated pattern and contrasts with the behavior of $\hat{G}_l(x)$ for these data shown in Figure 5.3.

5.2.3.2 Analysis of Spider Nests

Figure 5.6 show the corresponding EDF plot for the spider nests data, using a grid of $m = 48$ points. Again, we use Diggle's rule of thumb for determining the number of grid points to use. We see that $\hat{F}_l(x)$ lies between the simulation

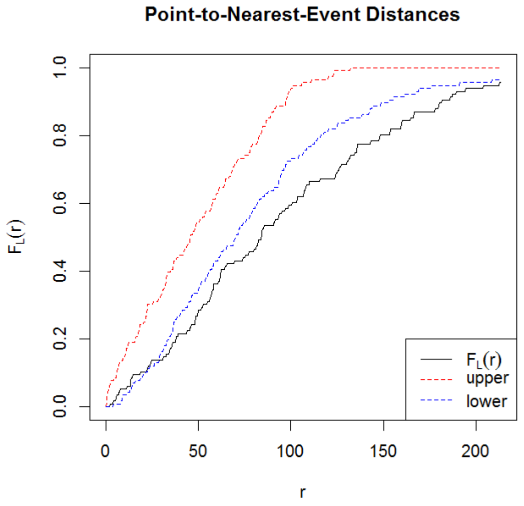


Figure 5.5: EDF plot of empty space distances for Chicago street crime data. Solid curve represents data. Dashed curves represent upper and lower envelopes from 99 simulations of CSR.

envelopes throughout its range. As in our previous analyses of the data, CSR is accepted.

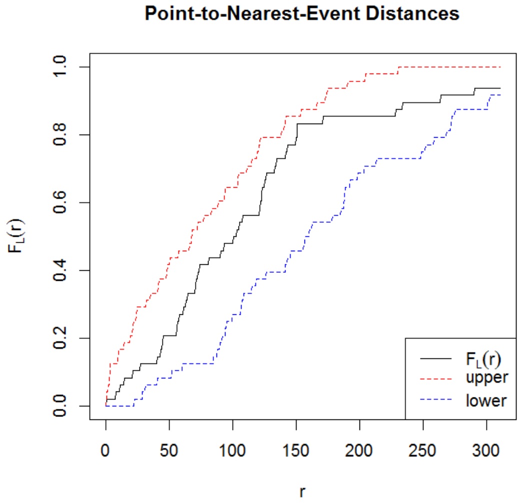


Figure 5.6: EDF plot of empty-space distances for spider nests data. Solid curve represents data. Dashed curves represent upper and lower envelopes from 99 simulations of CSR.

CHAPTER 6

Closing Remarks

6.1 Discussion

The techniques used in this paper can be improved upon. Just as Ang [1, 2] was able to improve Okabe & Yamada's Network K-function [16] by adding a correction for edge effects, so could we modify the summary functions used here. Developing an analog of quadrat counting on spatial point patterns in $\mathbb{R}^2$ for use on network point patterns might prove useful as well. So might the development of exact Mont Carlo tests [10] of CSR on a network.

The values of the network summary-function are limited by being difficult to interpret. Even for a completely random point process on the network, the expected value of a network summary function might depend on the geometry of the network [1]. There is no evidence that network summary functions obtained from different networks are directly comparable. May we, for example, compare the spatial patterns of street crime in two different cities using the respective network summary functions?

The methods described in this paper also assume that the use of the shortest-path distance is our best choice of a distance metric. This might not always be the case. For example, on a directed network such as a street network containing one-way streets, the shortest path between two points might not even be traversable. In this case, we would not want to use the shortest-path distance, but rather the directed shortest-path distance. The methods also assume that the shortest path is unique. If it is not – for example with Y-shaped paths or distances at junctions – then it seems safe to choose any/either of the equidistant routes as the shortest path.

Finally, we reiterate the differences between using planar spatial methods and network spatial methods for for analyzing network spatial phenomena. In general, while models for networked point processes are similar to those for regular spatial processes in the sense that network spatial methods are extensions of planar spatial methods, using planar spatial methods on network data can lead to false conclusions. For example, it was shown by Yamada and Thill (2004) that applying the planar K function method to network data overestimates clustering tendency. If we had used the planar K function instead of the network K function to analyze the data set in Figure 1.3, we would have concluded that the data are aggregated when they are in fact random. Network summary functions should be used for network data [15].

6.2 Conclusion

We have adjusted for the fact that the events of a network point patterns are constrained to lie on a network. We did so by substituting the shortest-path distance for the Euclidean distance in the definition of some common summary functions – namely the inter-event distance function H, the nearest-neighbor distance function G, and the point-to-nearest-event/empty-space distance function F. We used graphical procedures to test for CSR by comparing observed summary functions to upper and lower envelopes expected under simulations of CSR.

References

[1] Ang, Q.W. Statistical Methodologies for Events in a Linear Network. *University of Western Australia.* 2010.

[2] Ang, Q.W., Baddeley, A. and Nair, G. Geometrically Corrected Second Order Analysis of Events on a Linear Network, with Applications to Ecology and Criminology. *Scandinavian Journal of Statistics*, 39(4): 591-617, 2011.

[3] Baddeley, A. Analysing spatial point patterns in 'R'. *CSIRO.* March 2009. URL: `http://www.csiro.au/files/files/pn0y.pdf`.

[4] Baddeley, A. and Turner, R. Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6): 1–42, 2005.

[5] Baddeley, A., Turner, R., Moller, J. and Hazelton, M. Residual analysis for spatial point processes. *Journal of the Royal Statistical Society*, B(6):617–666, 2005.

[6] Baddeley, A.J. and Gill, R.D. Kaplan-Mier estimators of interpoint distance distributions for spatial point processes. *Annals of Statistics*, 25:263-292, 1997.

[7] Besag, J. and Diggle, P. Simple Monte Carlo Tests for Spatial Pattern. *Journal of the Royal Statistical Society.* 26(3):327–333, 1977.

[8] Chiu, S.N. and Stoyan, D. Estimators of distance distributions for spatial patterns. *Statistica Neerlandica*, 52:239-246, 1998.

[9] Collins, L.B., Pluzhnikov, A. and Stein, M.L. Improvement of Inter-event Distance Tests of Randomness in Spatial Point Processes. Technical Report No. 384, Department of Statistics, University of Chicago, 1994.

[10] Diggle, P.G. *Statistical Analysis of Spatial Point Patterns.* Second Edition. New York: Hodder Arnold, 2003.

[11] Geyer, C. *Stat 5601 Examples.* 2007.
URL: `http://www.stat.umn.edu/geyer/old/5601/examp/perm.html`.

[12] Habiger, J.D. and Pea, E.A. Randomised P-values and nonparametric procedures in multiple testing. *Journal of Nonparametric Statistics.* 23(3):583-604, 2011.

[13] Hope, A.C.A. A Simplified Monte Carlo Significance Test Procedure. *Journal of the Royal Statistical Society.* 30(3):582-598, 1968.

[14] Numata, M. Forest vegetation in the vicinity of Choshi. Coastal flora and vegetation at Choshi, Chiba Prefecture. *Bulletin of Choshi Marine Laboratory, Chiba University*, 3:28–48, 1961.

[15] Okabe, A. and Satoh, T. “Spatial Analysis on a Network”. In *The Sage Book of Spatial Analysis,* edited by A.S. Fotheringham and P.A. Rogerson, pp. 443-464. Sage, Los Angeles, 2009.

[16] Okabe, A. and Yamada, I. The K-Function Method on a Network and Its Computational Implementation. *Geographical Analysis.* 33(3):271–290, 2001.

[17] Ripley, B.D. Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society*, B(39):172-212, 1977.

[18] Strauss, D.J. A model for clustering. *Biometrika*, 63:467–475, 1975.