Point Pattern Distance Based Methods

Sergio Rey

Geographic Information Analysis
School of Geographical Sciences and Urban Planning
Arizona State University

Outline

Outline

Nearest Neighbors

Concept

- Spatial neighbor(s) for each point
- Distance between a point and its nearest neighbor(s)

Applications

- Statistics: clustering of points
- Statistics: specification of spatial weights
- Networks: nearest neighbor graphs
- Computer Science: nearest neighbor problems

Nearest Neighbor Definition

Nearest neighbor problem

Given a set S of n points in a plane, the nearest neighbor problem for a point u is to find the nearest neighbor $N(u)$ such that $d_{u,N(u)} \leq d_{u,j} \; \forall \; j \in S \setminus u$.

All nearest neighbors problem

Given a set S of n points in a plane, the all nearest neighbors problem is to find a nearest neighbor $N(u)$ of every $u \in S$.

Outline

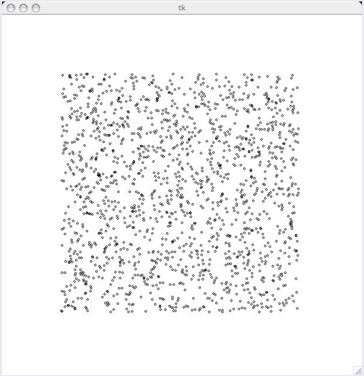
Nearest Neighbor Concepts

Types of Nearest Neighbors

- Nearest neighbors
- Mutual nearest neighbors
- k-nearest neighbors

Nearest Neighbors

Points



Nearest Neighbor Graph

Mutual Nearest Neighbors

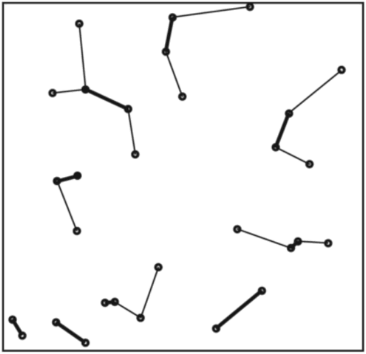


Figure 3.2 Network of first nearest neighbours for an artificial example of point pattern (NN). Mutual nearest neighbour pairs have bold lines, and form a subgraph (MNN) of the example.

Source: Dale and Fortin, R. (2014) "Spatial Analysis"

k-Nearest Neighbors

Figure 3: (**a**) $k = 1$ neighbours; (**b**) $k = 2$ neighbours; (**c**) $k = 4$ neighbours

Source: Bivand, R. (2015) “Creating neighbors.”

Outline

Mean Nearest Neighbor Statistic

$d_{min}(s_i)$

$$d_{min}(s_i) = min(d_{i,1}, d_{i,2}, \ldots, d_{i,n}) \quad (1)$$

$d_{min}(s_i)$ is the distance between i and its nearest neighbor event.

Test Statistic

$$\bar{d}_{min} = \frac{1}{n} \sum_{i=1}^{n} d_{min}(s_i) \quad (2)$$

Originally suggested by Clark and Evans (1954)

Mean Nearest Neighbor Statistic

$d_{min}(s_i)$

$$d_{min}(s_i) = min(d_{i,1}, d_{i,2}, \ldots, d_{i,n}) \tag{1}$$

$d_{min}(s_i)$ is the distance between i and its nearest neighbor event.

Test Statistic

$$\bar{d}_{min} = \frac{1}{n} \sum_{i=1}^{n} d_{min}(s_i) \tag{2}$$

Originally suggested by Clark and Evans (1954)

Mean Nearest Neighbor Statistic Distribution

$\bar{d}_{min} \sim N(\mu, \sigma^2)$

$$\mu = E[\bar{d}_{min}] = 0.5(n^{-1}|A|)^{1/2} + (0.051 + 0.042n^{-1/2})n^{-1}P \qquad (3)$$

$$\sigma^2 = V[\bar{d}_{min}] = 0.070n^{-1/2}|A| + 0.037(n^{-5}|A|)^{1/2}P \qquad (4)$$

where $|A|$ and P are the area and perimeter of the study area, respectively.

Issues

- Approximation, not an exact result.
- Dependence of nearest neighbor distances is ignored.
- Distribution of $d_{min}(s_i)$ ignored (only first moment).

Mean Nearest Neighbor Statistic Distribution

$\bar{d}_{min} \sim N(\mu, \sigma^2)$

$$\mu = E[\bar{d}_{min}] = 0.5(n^{-1}|A|)^{1/2} + (0.051 + 0.042n^{-1/2})n^{-1}P \quad (3)$$

$$\sigma^2 = V[\bar{d}_{min}] = 0.070n^{-1/2}|A| + 0.037(n^{-5}|A|)^{1/2}P \quad (4)$$

where $|A|$ and P are the area and perimeter of the study area, respectively.

Issues

- Approximation, not an exact result.
- Dependence of nearest neighbor distances is ignored.
- Distribution of $d_{min}(s_i)$ ignored (only first moment).

Mean Nearest Neighbor Statistic Distribution

$\bar{d}_{min} \sim N(\mu, \sigma^2)$

$$\mu = E[\bar{d}_{min}] = 0.5(n^{-1}|A|)^{1/2} + (0.051 + 0.042n^{-1/2})n^{-1}P \quad (3)$$

$$\sigma^2 = V[\bar{d}_{min}] = 0.070n^{-1/2}|A| + 0.037(n^{-5}|A|)^{1/2}P \quad (4)$$

where $|A|$ and P are the area and perimeter of the study area, respectively.

Issues

- Approximation, not an exact result.
- Dependence of nearest neighbor distances is ignored.
- Distribution of $d_{min}(s_i)$ ignored (only first moment).

Mean Nearest Neighbor Statistic Distribution

$\bar{d}_{min} \sim N(\mu, \sigma^2)$

$$\mu = E[\bar{d}_{min}] = 0.5(n^{-1}|A|)^{1/2} + (0.051 + 0.042n^{-1/2})n^{-1}P \qquad (3)$$

$$\sigma^2 = V[\bar{d}_{min}] = 0.070n^{-1/2}|A| + 0.037(n^{-5}|A|)^{1/2}P \qquad (4)$$

where $|A|$ and P are the area and perimeter of the study area, respectively.

Issues

- Approximation, not an exact result.
- Dependence of nearest neighbor distances is ignored.
- Distribution of $d_{min}(s_i)$ ignored (only first moment).

Mean Nearest Neighbor Statistic Distribution

$\bar{d}_{min} \sim N(\mu, \sigma^2)$

$$\mu = E[\bar{d}_{min}] = 0.5(n^{-1}|A|)^{1/2} + (0.051 + 0.042n^{-1/2})n^{-1}P \quad (3)$$

$$\sigma^2 = V[\bar{d}_{min}] = 0.070n^{-1/2}|A| + 0.037(n^{-5}|A|)^{1/2}P \quad (4)$$

where $|A|$ and P are the area and perimeter of the study area, respectively.

Issues

- Approximation, not an exact result.
- Dependence of nearest neighbor distances is ignored.
- Distribution of $d_{min}(s_i)$ ignored (only first moment).

Outline

Edge Effects

Problems

- For points close to the boundary intensity is underestimated.
- Neighboring points are outside the study region.

Solutions

- Buffer the points
- Edge corrections
- Monte Carlo Simulations

Edge Effects

Problems

- For points close to the boundary intensity is underestimated.
- Neighboring points are outside the study region.

Solutions

- Buffer the points
- Edge corrections
- Monte Carlo Simulations

Edge Effects

Problems

- For points close to the boundary intensity is underestimated.
- Neighboring points are outside the study region.

Solutions

- Buffer the points
- Edge corrections
- Monte Carlo Simulations

Edge Effects

Problems

- For points close to the boundary intensity is underestimated.
- Neighboring points are outside the study region.

Solutions

- Buffer the points
- Edge corrections
- Monte Carlo Simulations

Edge Effects

Problems

- For points close to the boundary intensity is underestimated.
- Neighboring points are outside the study region.

Solutions

- Buffer the points
- Edge corrections
- Monte Carlo Simulations

Edge Effects

Problems

- For points close to the boundary intensity is underestimated.
- Neighboring points are outside the study region.

Solutions

- Buffer the points
- Edge corrections
- Monte Carlo Simulations

Edge Effects

Problems

- For points close to the boundary intensity is underestimated.
- Neighboring points are outside the study region.

Solutions

- Buffer the points
- Edge corrections
- Monte Carlo Simulations

Outline

Nearest Neighbor G Function

$G(d)$

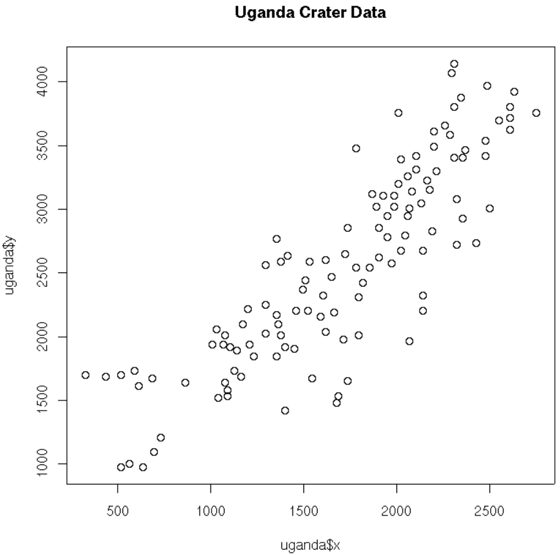
$$G(d) = \sum_{i=1}^{n} \Phi_i^d / n \tag{5}$$

where

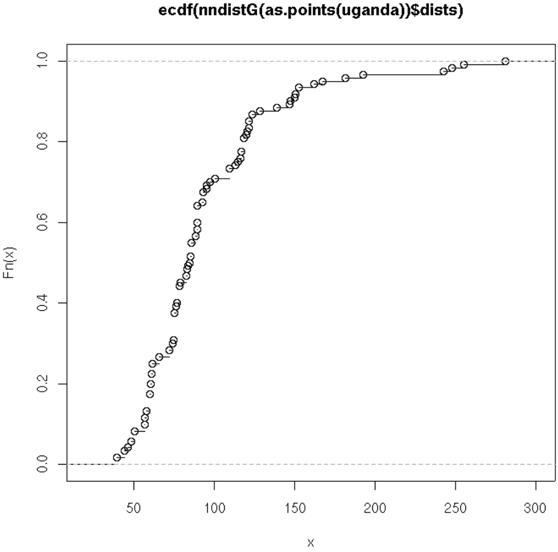
$$\Phi_i^d = \begin{cases} 1 & if\ d_{min}(s_i) < d \\ 0 & otherwise \end{cases} \tag{6}$$

$G(d)$ is the proportion of nearest neighbor distances that are less than d.

Uganda Crater Data



Nearest Neighbor G Function



G Function Intepretation

Shape

- G increasing rapidly at small distances points to *clustering*.
- G increases slowly points to *uniformity*.
- Both are deviations from CSR.

Compare G to that from a CSR Process

- Theoretical G
- Homogeneous Poisson process
- Density equal to density of actual pattern
- Empirical distribution against theoretical distribution
  - Should be a 45 degree line if process is CSR
  - Above the line = clustering
  - Below the line = dispersion

G Function Intepretation

Shape

- G increasing rapidly at small distances points to *clustering*.
- G increases slowly points to *uniformity*.
- Both are deviations from CSR.

Compare G to that from a CSR Process

- Theoretical G
- Homogeneous Poisson process
- Density equal to density of actual pattern
- Empirical distribution against theoretical distribution
  - Should be a 45 degree line if process is CSR
  - Above the line = clustering
  - Below the line = dispersion

G Function Intepretation

Shape

- G increasing rapidly at small distances points to *clustering*.
- G increases slowly points to *uniformity*.
- Both are deviations from CSR.

Compare G to that from a CSR Process

- Theoretical G
- Homogeneous Poisson process
- Density equal to density of actual pattern
- Empirical distribution against theoretical distribution
  - Should be a 45 degree line if process is CSR
  - Above the line = clustering
  - Below the line = dispersion

G Function Intepretation

Shape

- G increasing rapidly at small distances points to *clustering*.
- G increases slowly points to *uniformity*.
- Both are deviations from CSR.

Compare G to that from a CSR Process

- Theoretical G
- Homogeneous Poisson process
- Density equal to density of actual pattern
- Empirical distribution against theoretical distribution
  - Should be a 45 degree line if process is CSR
  - Above the line = clustering
  - Below the line = dispersion

G Function Intepretation

Shape

- G increasing rapidly at small distances points to *clustering*.
- G increases slowly points to *uniformity*.
- Both are deviations from CSR.

Compare G to that from a CSR Process

- Theoretical G
- Homogeneous Poisson process
- Density equal to density of actual pattern
- Empirical distribution against theoretical distribution
  - Should be a 45 degree line if process is CSR
  - Above the line = clustering
  - Below the line = dispersion

G Function Intepretation

Shape

- G increasing rapidly at small distances points to *clustering*.
- G increases slowly points to *uniformity*.
- Both are deviations from CSR.

Compare G to that from a CSR Process

- Theoretical G
- Homogeneous Poisson process
- Density equal to density of actual pattern
- Empirical distribution against theoretical distribution
  - Should be a 45 degree line if process is CSR
  - Above the line = clustering
  - Below the line = dispersion

G Function Intepretation

Shape

- G increasing rapidly at small distances points to *clustering*.
- G increases slowly points to *uniformity*.
- Both are deviations from CSR.

Compare G to that from a CSR Process

- Theoretical G
- Homogeneous Poisson process
- Density equal to density of actual pattern
- Empirical distribution against theoretical distribution
  - Should be a 45 degree line if process is CSR
  - Above the line = clustering
  - Below the line = dispersion

G Function Intepretation

Shape

- G increasing rapidly at small distances points to *clustering*.
- G increases slowly points to *uniformity*.
- Both are deviations from CSR.

Compare G to that from a CSR Process

- Theoretical G
- Homogeneous Poisson process
- Density equal to density of actual pattern
- Empirical distribution against theoretical distribution
  - Should be a 45 degree line if process is CSR
  - Above the line = clustering
  - Below the line = dispersion

G Function Intepretation

Shape

- G increasing rapidly at small distances points to *clustering*.
- G increases slowly points to *uniformity*.
- Both are deviations from CSR.

Compare G to that from a CSR Process

- Theoretical G
- Homogeneous Poisson process
- Density equal to density of actual pattern
- Empirical distribution against theoretical distribution
  - Should be a 45 degree line if process is CSR
  - Above the line = clustering
  - Below the line = dispersion

G Function Intepretation

Shape

- G increasing rapidly at small distances points to *clustering*.
- G increases slowly points to *uniformity*.
- Both are deviations from CSR.

Compare G to that from a CSR Process

- Theoretical G
- Homogeneous Poisson process
- Density equal to density of actual pattern
- Empirical distribution against theoretical distribution
  - Should be a 45 degree line if process is CSR
  - Above the line = clustering
  - Below the line = dispersion

G Function Intepretation

Shape

- G increasing rapidly at small distances points to *clustering*.
- G increases slowly points to *uniformity*.
- Both are deviations from CSR.

Compare G to that from a CSR Process

- Theoretical G
- Homogeneous Poisson process
- Density equal to density of actual pattern
- Empirical distribution against theoretical distribution
  - Should be a 45 degree line if process is CSR
  - Above the line = clustering
  - Below the line = dispersion

G Function Intepretation

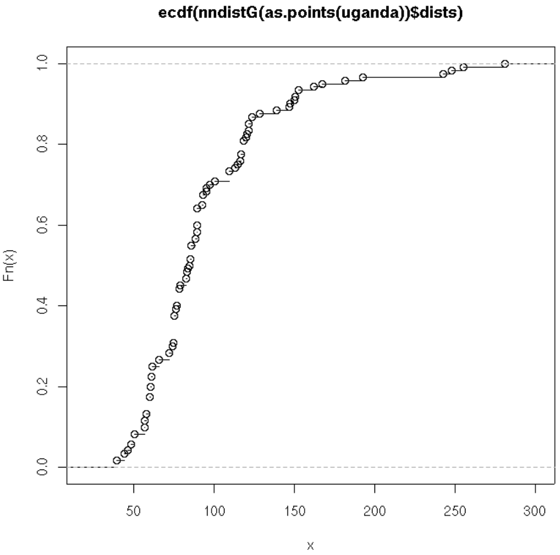
Shape

- G increasing rapidly at small distances points to *clustering*.
- G increases slowly points to *uniformity*.
- Both are deviations from CSR.

Compare G to that from a CSR Process

- Theoretical G
- Homogeneous Poisson process
- Density equal to density of actual pattern
- Empirical distribution against theoretical distribution
  - Should be a 45 degree line if process is CSR
  - Above the line = clustering
  - Below the line = dispersion

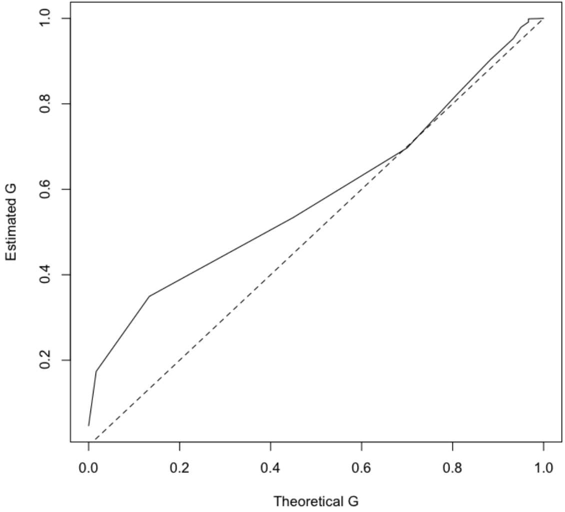
Nearest Neighbor G Function



Estimated vs. Theoretical G Function: Code

```
> library(splancs)
> data(uganda)
> plot(Ghat(as.points(uganda), seq(20, 500, 20)),
+ Fzero(pdense(as.points(uganda),  uganda$poly),
+ seq(20, 500, 20)), type="l",
+ xlab="Theoretical G",
+ ylab="Estimated G")
> lines(c(0,1),c(0,1),lty=2)
```

Estimated vs. Theoretical G Function



Outline

Nearest Neighbor F Function

- G function is sensitive to n
  - Can be rough
  - Takes on stepped appearance for small n
- Alternative approach is to generate N random points in the domain
  - Analyze the distribution of nearest event neighbor distances
  - Closest event to each point.
- Can be used for small n data sets

Nearest Neighbor F Function

- G function is sensitive to n
  - Can be rough
  - Takes on stepped appearance for small n
- Alternative approach is to generate N random points in the domain
  - Analyze the distribution of nearest event neighbor distances
  - Closest event to each point.
- Can be used for small n data sets

Nearest Neighbor F Function

- G function is sensitive to n
  - Can be rough
  - Takes on stepped appearance for small n
- Alternative approach is to generate N random points in the domain
  - Analyze the distribution of nearest event neighbor distances
  - Closest event to each point.
- Can be used for small n data sets

Nearest Neighbor F Function

- G function is sensitive to n
  - Can be rough
  - Takes on stepped appearance for small n
- Alternative approach is to generate N random points in the domain
  - Analyze the distribution of nearest event neighbor distances
  - Closest event to each point.
- Can be used for small n data sets

Nearest Neighbor F Function

- G function is sensitive to n
  - Can be rough
  - Takes on stepped appearance for small n
- Alternative approach is to generate N random points in the domain
  - Analyze the distribution of nearest event neighbor distances
  - Closest event to each point.
- Can be used for small n data sets

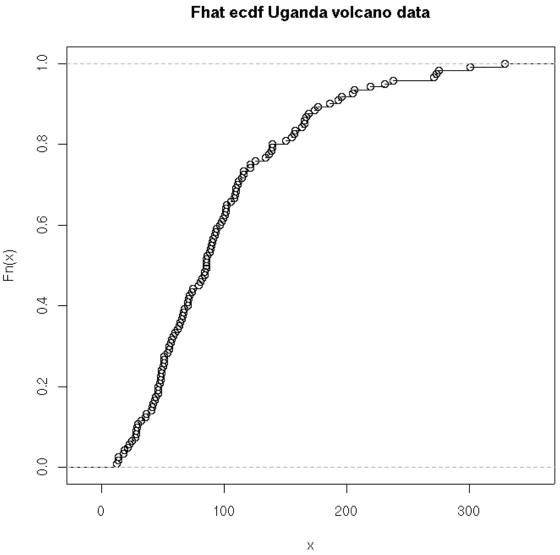
Nearest Neighbor F Function

- G function is sensitive to n
  - Can be rough
  - Takes on stepped appearance for small n
- Alternative approach is to generate N random points in the domain
  - Analyze the distribution of nearest event neighbor distances
  - Closest event to each point.
- Can be used for small n data sets

Nearest Neighbor F Function

- G function is sensitive to n
  - Can be rough
  - Takes on stepped appearance for small n
- Alternative approach is to generate N random points in the domain
  - Analyze the distribution of nearest event neighbor distances
  - Closest event to each point.
- Can be used for small n data sets

Nearest Neighbor F Function



Outline

Nearest Neighbor J Function

$$J(d) = (1 - G(d))/(1 - F(d)) \tag{7}$$

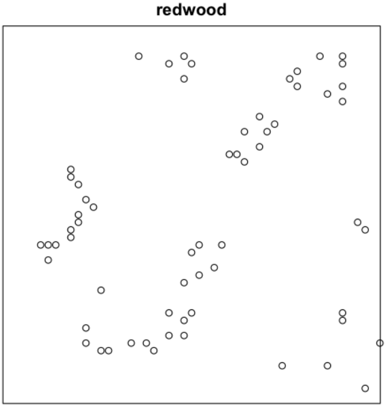
- $J(d) < 1$ points to spatial clustering
- $J(d) > 1$ points to spatial regularity

Nearest Neighbor J Function

$$J(d) = (1 - G(d))/(1 - F(d)) \tag{7}$$

- $J(d) < 1$ points to spatial clustering
- $J(d) > 1$ points to spatial regularity

Redwood

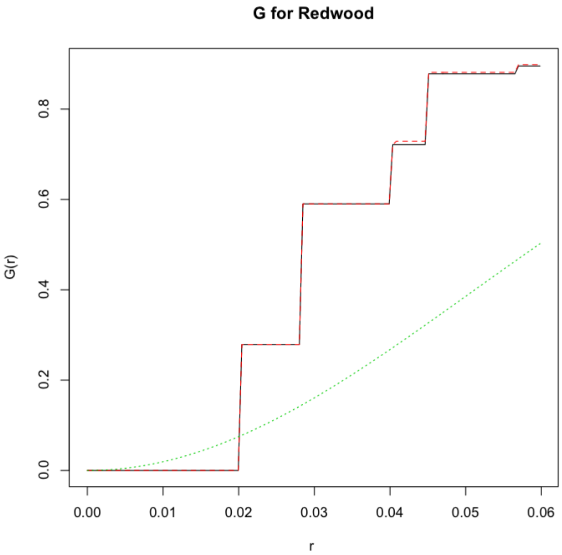


R code

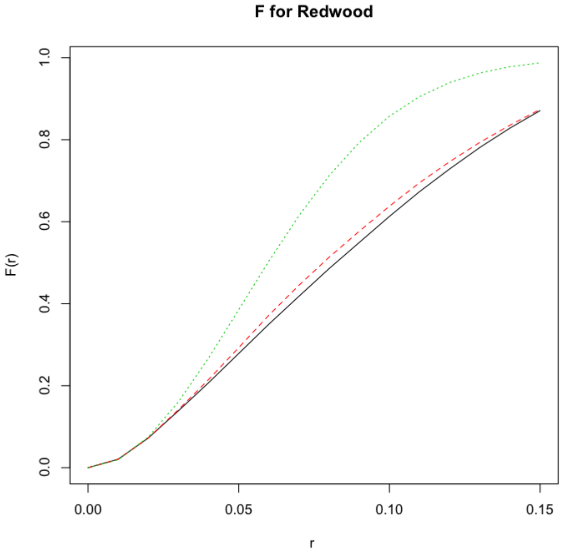
```
> gr=Gest(redwood)
> plot(gr,main="G for Redwood")
     lty col
km     1   1
rs     2   2
theo   3   3
```

- *km*: spatial Kaplan-Meier estimator of $G(r)$
- *rs*: the reduced sample edge correction estimator of $G(r)$
- *theo*: the theoretical value of $G(r)$ for a CSR process

Nearest Neighbor G Function



Nearest Neighbor F Function

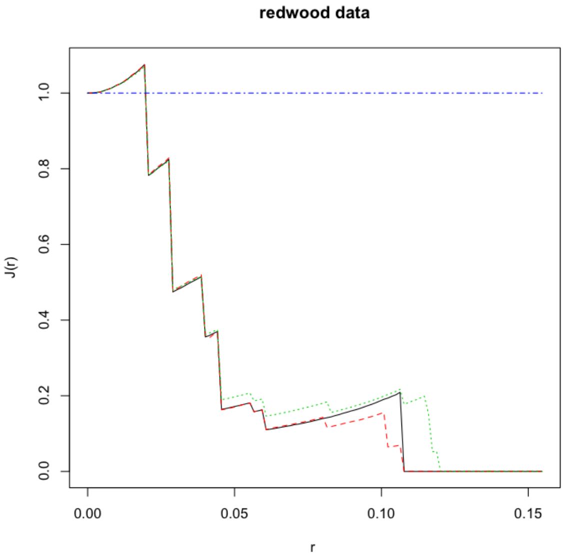


R code: J Function

```
>     J <- Jest(redwood, 0.01)
>     plot(J, main="redwood data")
      lty col
km      1   1
rs      2   2
un      3   3
theo    4   4
>     # values are below J= 1, indicating clustered pattern
```

- *km*: spatial Kaplan-Meier estimator of $G(r)$
- *rs*: the reduced sample edge correction estimator of $G(r)$
- *un*: the uncorrected estimate of $J(r)$ computed from the uncorrected estimates of F and G
- *theo*: the theoretical value of $J(r)$ for a CSR process

Nearest Neighbor J Function



F function simulation