Spatial Analysis and Modeling

Analysis of Point Patterns

Instructor: Rossano Schifanella

@UNITO

introduction

- **Objectives:**
  - To determine if there is a tendency for points to exhibit a systematic pattern (i.e. some form of regularity or clustering)
  - If there is a systematic pattern, then to examine at what spatial scale this pattern occurs and whether particular clusters are associated with proximity to particular sources of some factors.
  - To estimate how the intensity of points varies across the study region
  - To seek models to account for observed point patterns

introduction

- **Analysis Approach:**
  - Events may have attributes which can be used to distinguish types – but it is the location pattern that is analyzed
  - Patterns in event locations are the focus
  - Stochastic aspect is where events are likely to occur
  - Does a pattern exhibit clustering or regularity?
  - Over what spatial scales do patterns exist?

introduction

- Diseases
- Crime types
- Earthquake epicenters
- Plant distributions
- ...

definition

Characteristics:

- set of n point locations with recorded “events”, e.g., locations of trees, disease or crime incidents $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_i, \ldots, \mathbf{s}_n\}$
- point locations correspond to all possible events or to subsets of them
- attribute values also possible at same locations, e.g., tree diameter, magnitude of earthquakes (*marked point pattern*) $W = \{w_1, \ldots, w_i, \ldots, w_n\}$

Analysis objectives:

- detect spatial clustering or repulsion, as opposed to complete randomness, of event locations (in space and time)
- if clustering detected, investigate possible relations with nearby “sources”

Basic Assumptions:

- **Data present a complete set of events in the study region R, which is called mapped point pattern**
  - i.e. all relevant events occurred in R have been recorded
- **Remark**
  - Some point pattern analysis are directed towards extracting limited information about a point process, by recording events in a sample of different areas of the whole region, which is called sampled point pattern.
  - e.g. field studies in forestry, ecology or biology, where complete enumeration is not feasible.

Basic Assumptions:

- The study region R might be of any arbitrary shape. Some of the methods can be applied only to regions, which are a square or a rectangle.
- In order to eliminate edge effects, a suitable guard area between perimeter of the original study region and sub-region within which analysis is performed is left.
- In all cases, the final area selected for study is assumed to be in some sense representative of any larger region from which it has been selected.

- **Further issues:**
  - analysis of point patterns over large areas should take into account distance distortions due to map projections
  - boundaries of study area should not be arbitrary
  - analysis of sampled point patterns can be misleading
  - one-to-one correspondence between objects in study area and events in pattern

Centrography

- **Mean center of a points pattern:**
  - point with coordinates $\bar{s} = (\bar{x}, \bar{y})$

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad \bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

  - center of point pattern, or point with average x and y-coordinates

Centrography

- **Median center of a points pattern:**
  - center for minimum distance

$$s \in \{s_1, \ldots s_c\} s.t. min \sum_{i=1}^{n} |s_i - s_c|$$

- no closed form
- p-median problem (a typical problem in spatial optimization)
  - the problem of locating p “facilities” relative to a set of “customers” such that the sum of the shortest demand weighted distance between “customers” and “facilities” is minimised

dispersion of points distributions

Standard distance of a point pattern:

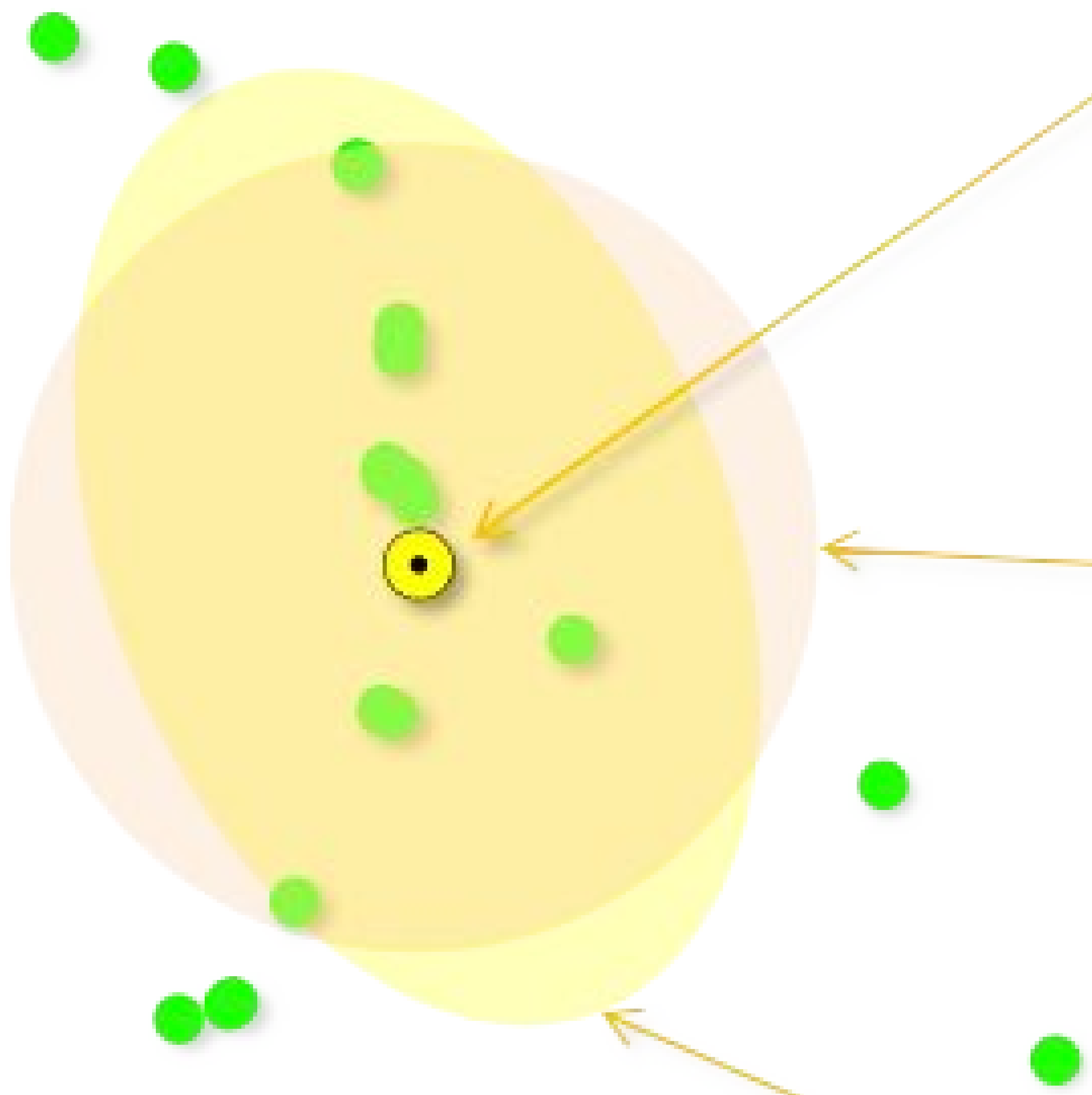
- average squared deviations of x and y coordinates from their respective mean:

$$d_{std} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}$$

- related to standard deviation of coordinates, a summary circle (centered at $\bar{s}$ with radius d_{std}) of a point pattern

Standard deviational ellipse:

- Taking directional effects into account for *anisotropy* cases
- Please refer to Levine and Associates, 2004 for calculations



Mean center

computed average X and Y coordinate values.

$$\bar{s} = \left(\frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n} \right)$$

Standard distance

measure of the variance between the average distance of the features to the mean center.

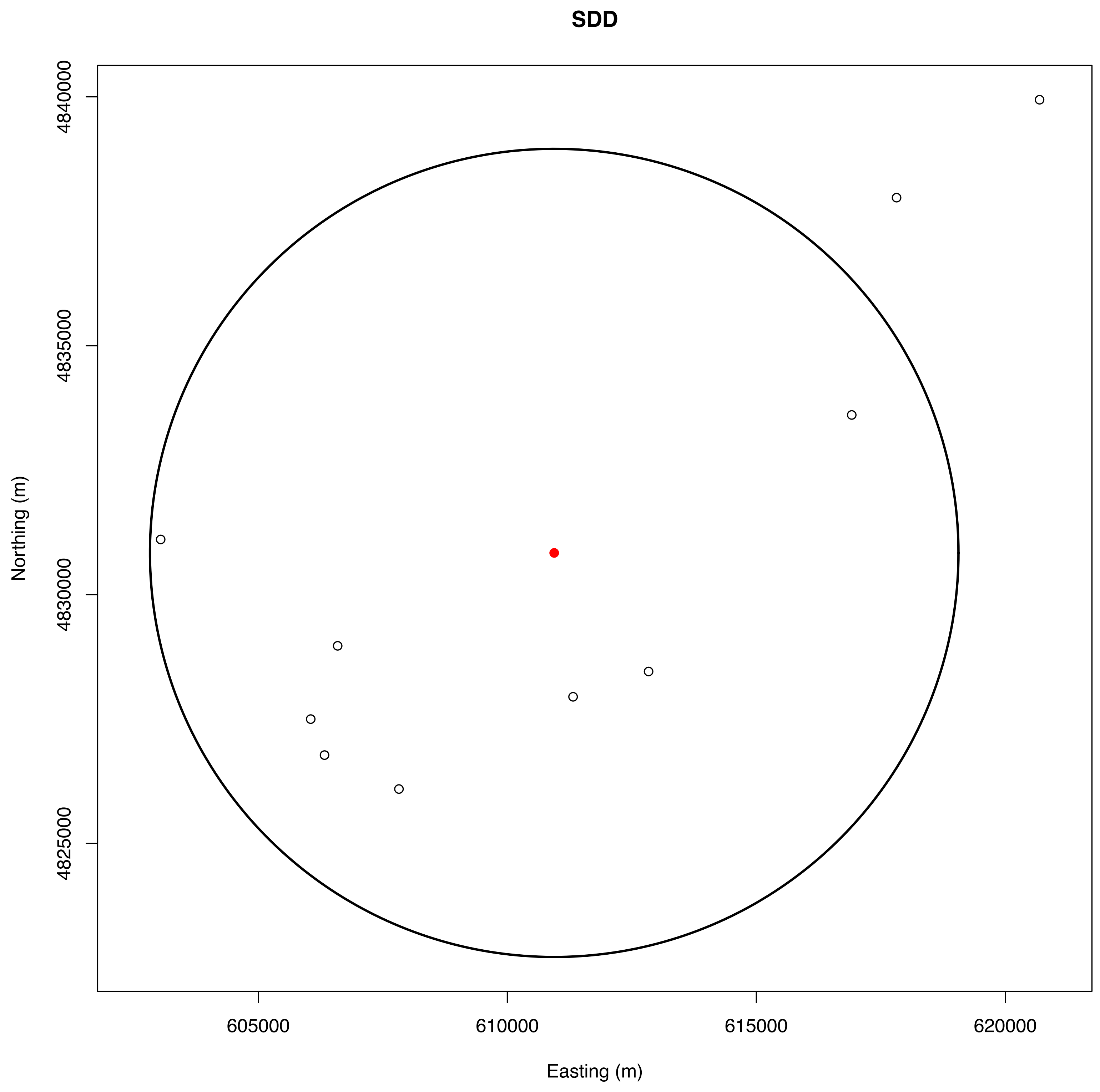
$$d = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu_x)^2 + (y_i - \mu_y)^2}{n}}$$

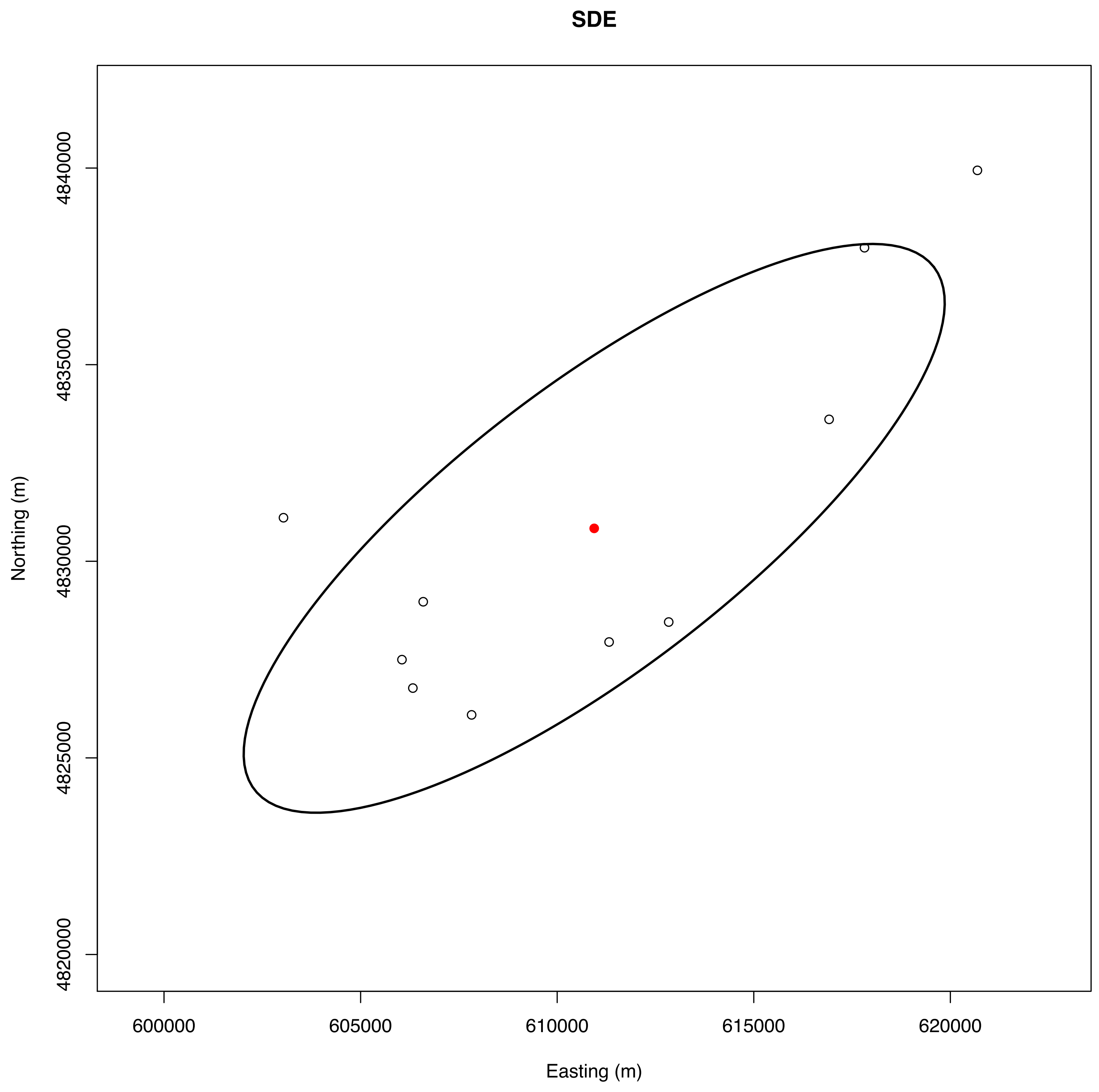
Standard deviational ellipse

separate standard distances for each axis.

$$\begin{cases} d_x = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \mu_x)^2}{n}} \\ d_y = \sqrt{\dfrac{\sum_{i=1}^{n} (y_i - \mu_y)^2}{n}} \end{cases}$$

Examples:





Remarks:

- indicates overall shape and center of point pattern
- *do not suffice to fully specify a spatial point pattern*

methods

1st order (i.e., intensity): absolute location of events on map:

- Quadrat methods
- Density Estimation (KDE)

2nd order (i.e., interactions): interaction of events:

- Nearest neighbor distance
- Distance functions G, K, F, L

quadrats methods

Consider a point pattern with n events within a study region A of area $|A|$

Global intensity:

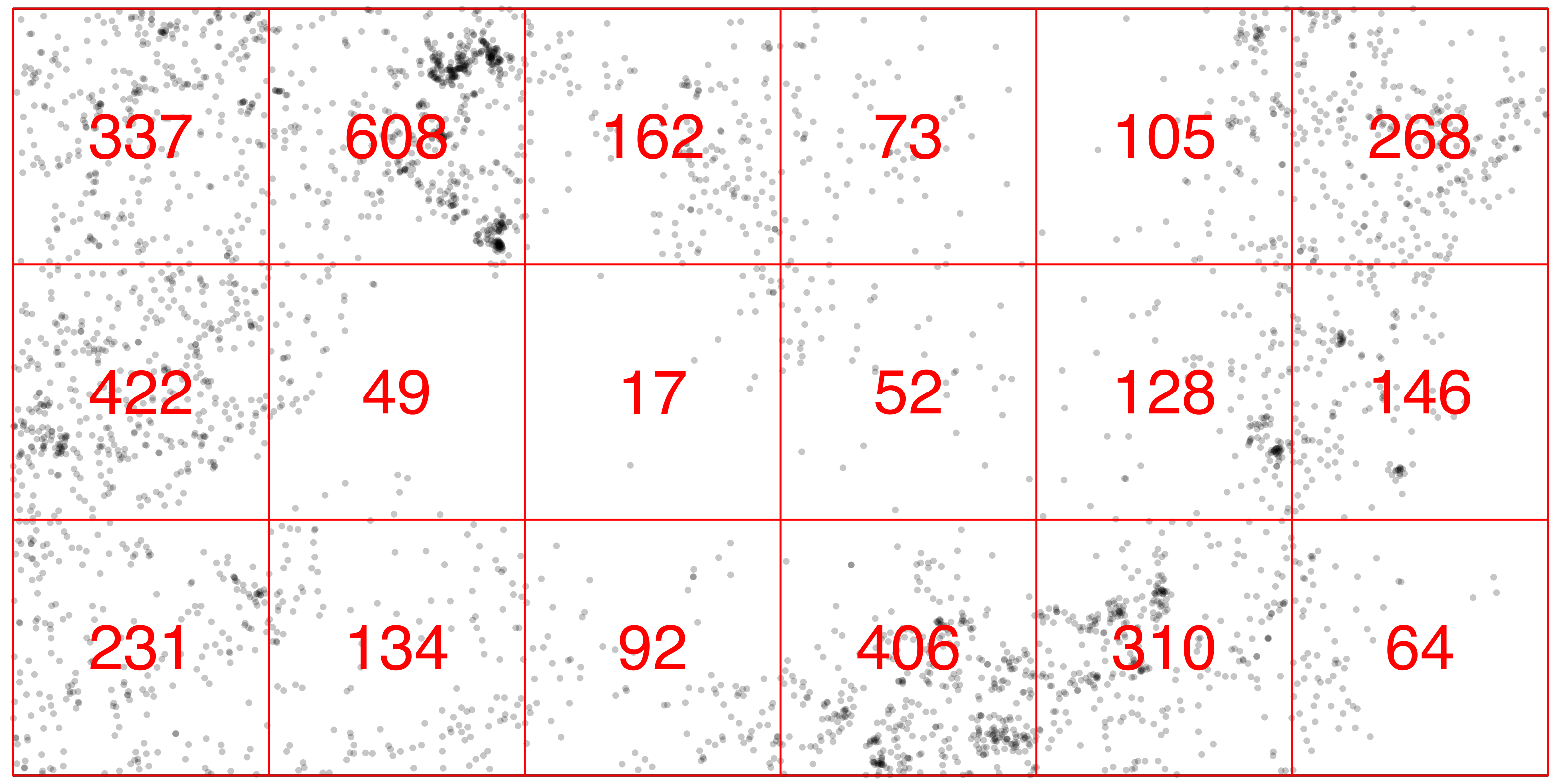
$$\hat{\lambda} = \frac{n}{|A|} = \frac{\#\text{of events within} A}{|A|}$$

Local intensity via quadrats

1. partition A into L sub-regions $A_l, l = 1, \ldots, L$ of equal area $|A_l|$ (also called quadrats)
2. count number of events $n(A_l)$ in each sub-region A_l
3. convert sample counts into estimated intensity rates as:

$$\hat{\lambda}(A_l) = \frac{n(A_l)}{|A_l|}$$

quadrats methods



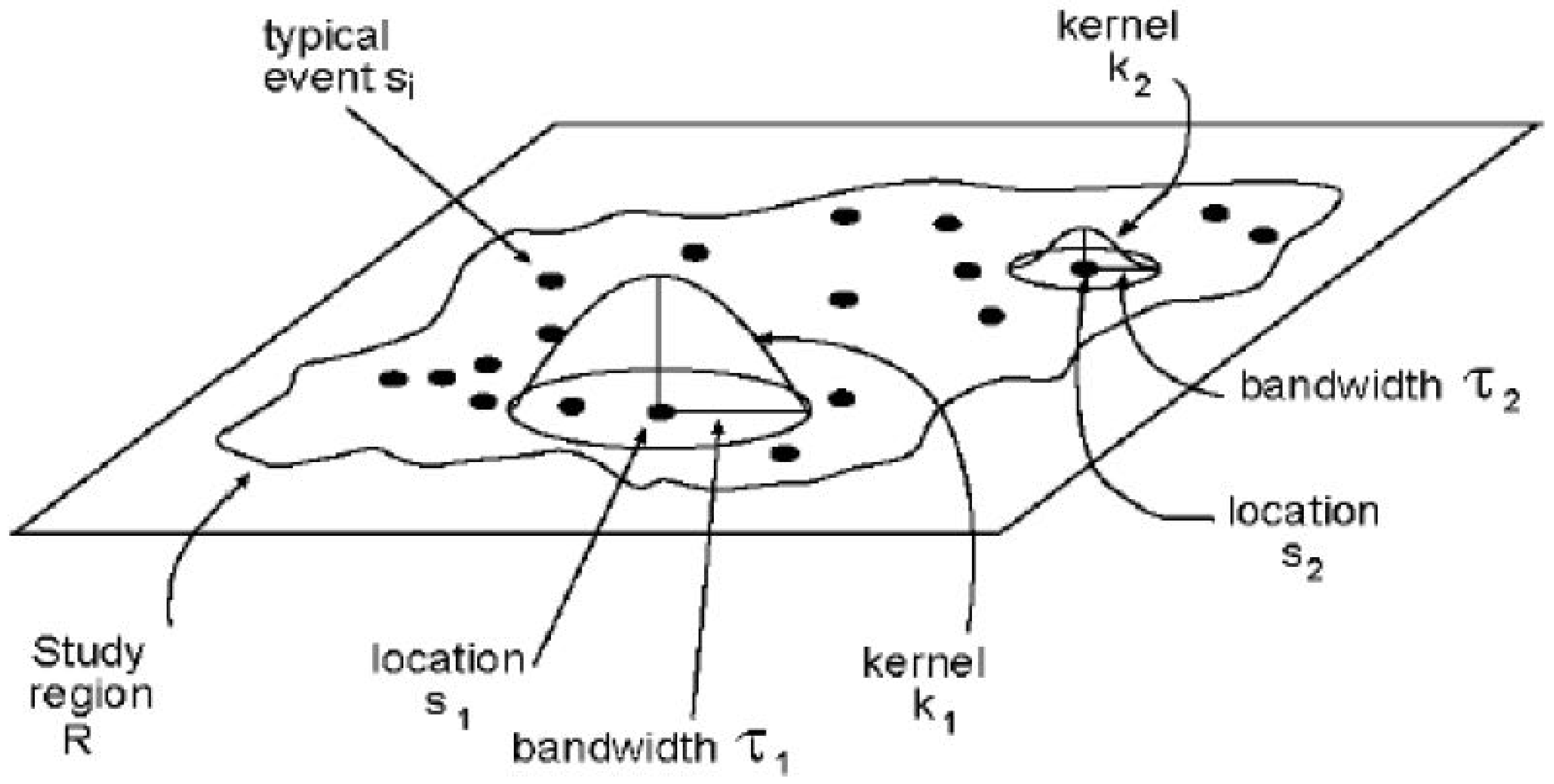
- estimated rates $\hat{\lambda}(A_l)$ over set of quadrats
- reveal large-scale patterns in intensity variation over A
- larger quadrats yield smoother intensity maps; smaller quadrats yield 'spiky' intensity maps
- size, origin, and shape of quadrats is critical (recall: *MAUP*)
- *only first-order effects are captured*

kernel density estimation (KDE)

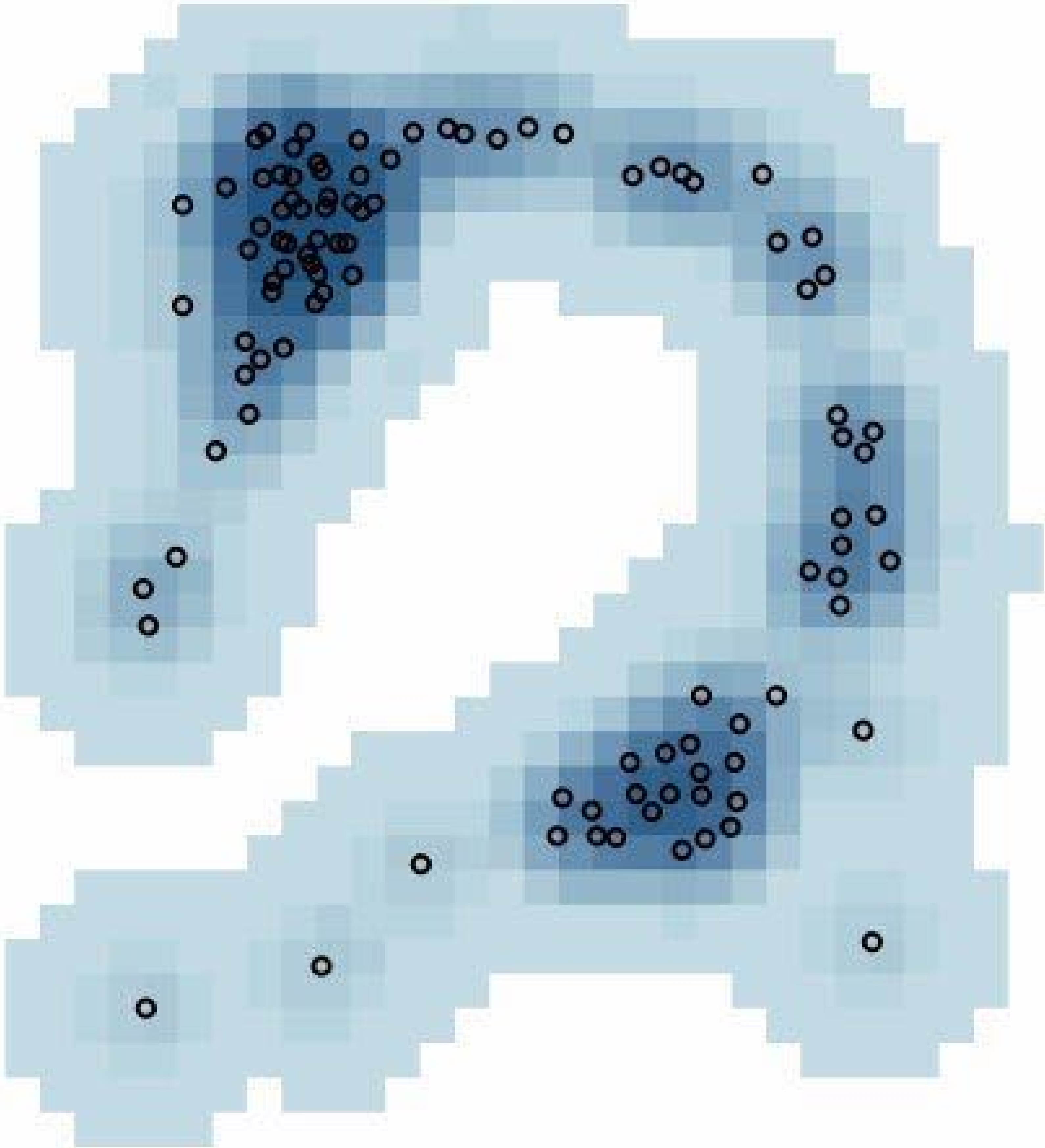
1. define a kernel $K(\boldsymbol{s}; r)$ of radius (or bandwidth) r centered at any arbitrary location $\boldsymbol{s}$
2. estimate local intensity at $\boldsymbol{s}$ as:

$$\hat{\lambda}(\boldsymbol{s}) = \frac{1}{n} \sum_{i=1}^{n} K(\boldsymbol{s}_i - \boldsymbol{s}; r)$$

3. repeat estimation for all points s in the study region to create a density map



kernel density estimation (KDE)



comments

- **Choice of kernel function is not critical (Diggle, 1985) Choice of bandwidth, or degree of smoothing critical:**
  - Small bandwidth → spiky results
  - Large bandwidth → loss of detail
- **Multi-scale analyses can use these bandwidth characteristics to investigate both broad trends and localized variation**
- **How to choose bandwidth: choose the degree of smoothing subjectively, by eye, or by formula (Diggle)**
- **could define local bandwidth based on function of presence of events in neighborhood of s (i.e., adaptive kernel estimation)**

distance-based descriptors point patterns

- Distances: accessing second order effects
  - Event-to-event distance: distance d_{ij} between event at arbitrary location s_i and another event at another arbitrary location s_j:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

  - Point-to-event distance: distance $\tilde{d}_{pj}$ between a randomly chosen point at location $\tilde{s}_p$ and an event at location s_j:

$$\tilde{d}_{pj} = \sqrt{(\tilde{x}_p - x_j)^2 + (\tilde{y}_i - y_j)^2}$$

  - Event-to-nearest-neighbour distance: distance $d_{min}(s_i)$ between an event at location s_i and its *nearest neighbor* event:

$$d_{min}(s_i) = min\{d_{ij}, j \neq i, j = 1, \ldots, n\}$$

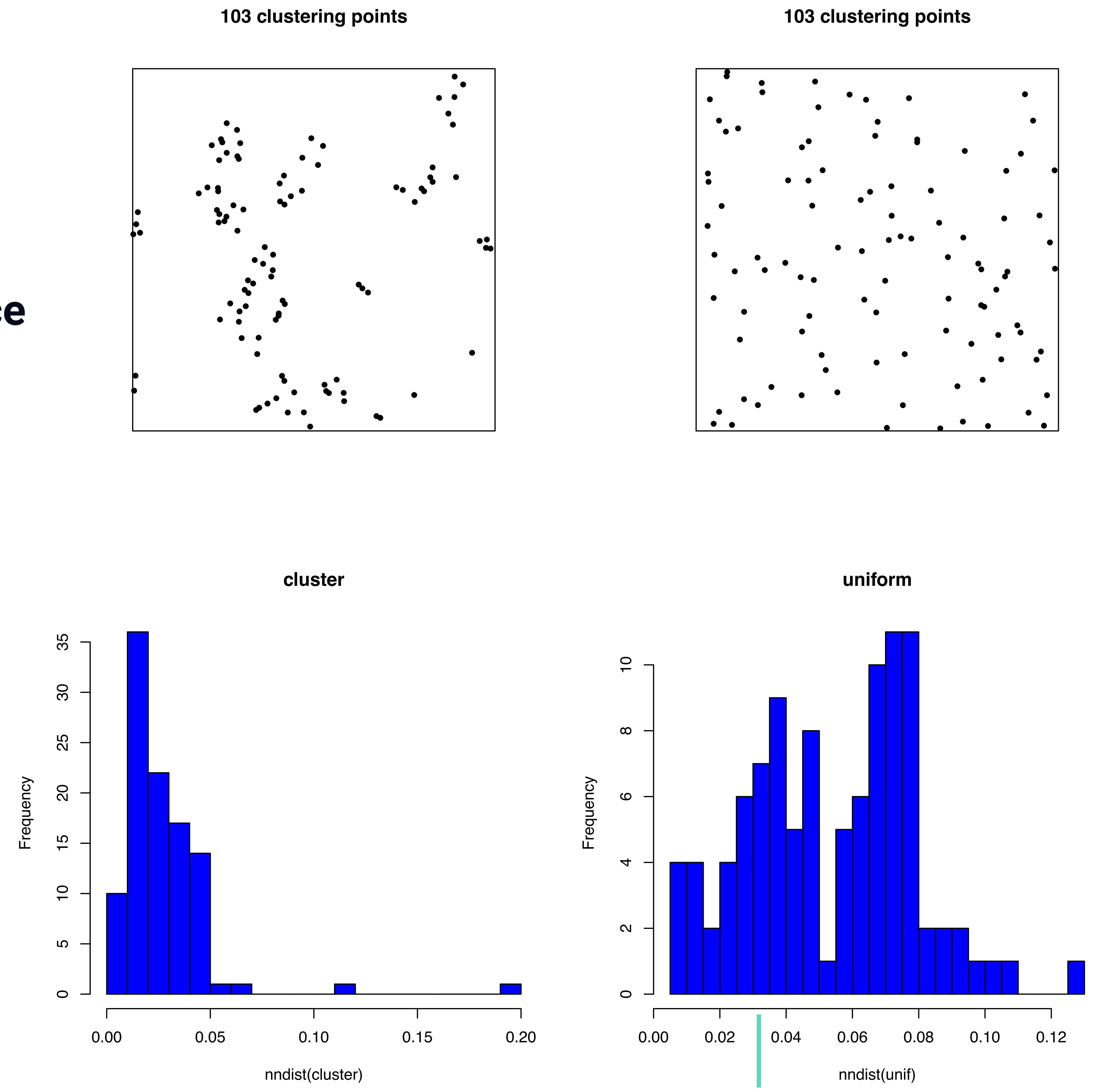
  - Point-to-nearest-neighbour distance (i.e., *empty space distance*): distance $d_{min}(\tilde{s}_p)$ between a randomly chosen point at location $\tilde{s}_p$ and its *nearest neighbor* event:

$$\tilde{d}_{min}(\tilde{s}_p) = min\{\tilde{d}_{pj}, j = 1, \ldots, n\}$$

nearest neighbour distances

- **Mean nearest neighbour distance**
  - Average of all $d_{min}(s_i)$ values

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



G function

- Definition: nearest neighbour distance function, i.e., proportion of event-to-nearest-neighbor distances $d_{min}(s_i)$ no greater than given distance cutoff d, estimated as:

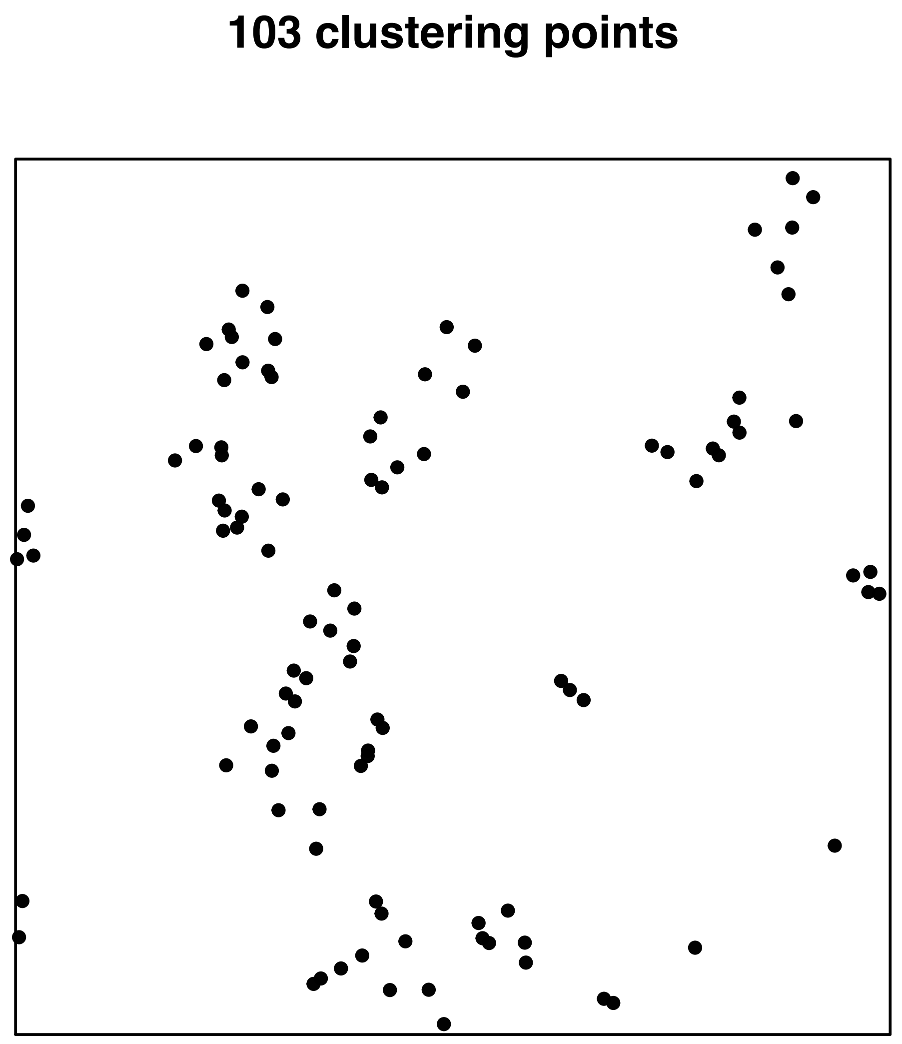
$$\hat{G}(d) = \frac{\#\{d_{min}(s_i) < d, i = 1, \dots, n\}}{n}$$

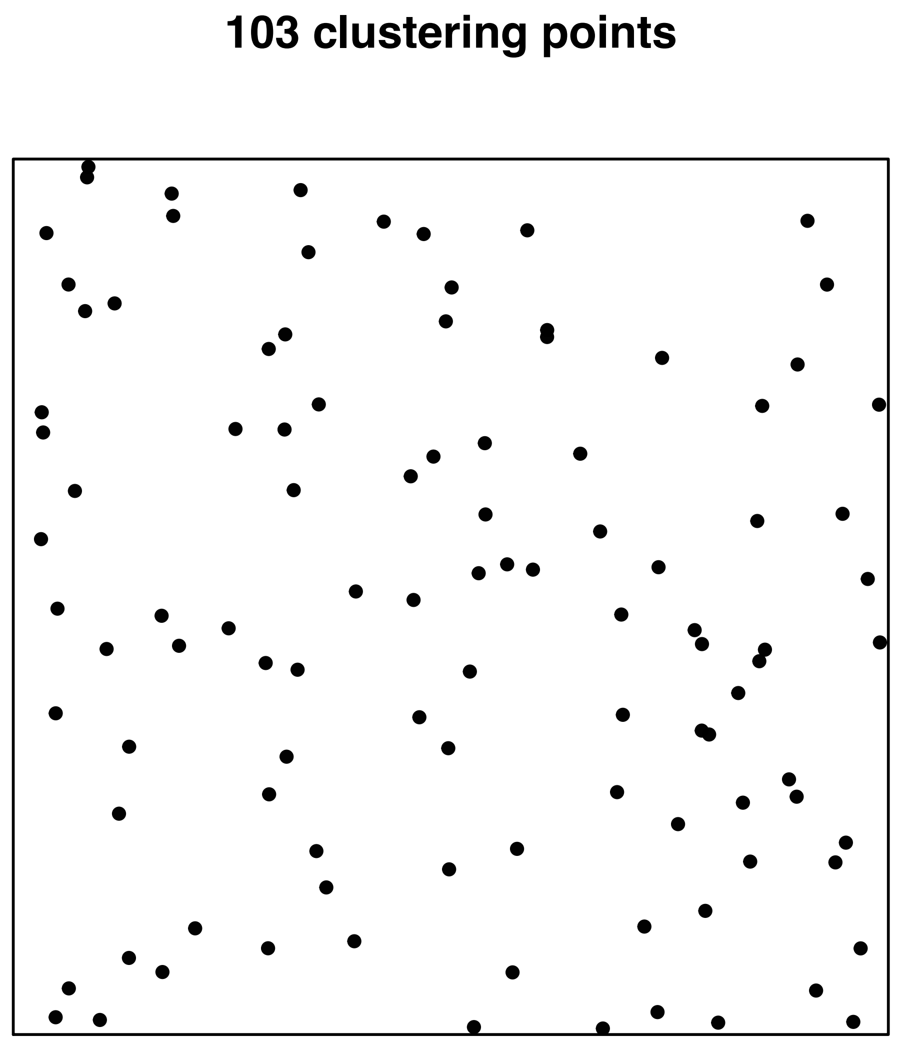
- alternative definition: cumulative distribution function (CDF) of all n event-to-nearest-neighbor distances; instead of computing average $\bar{d}_{min}$ of d_{min} values, compute their CDF
- the G function provides information on event *proximity*

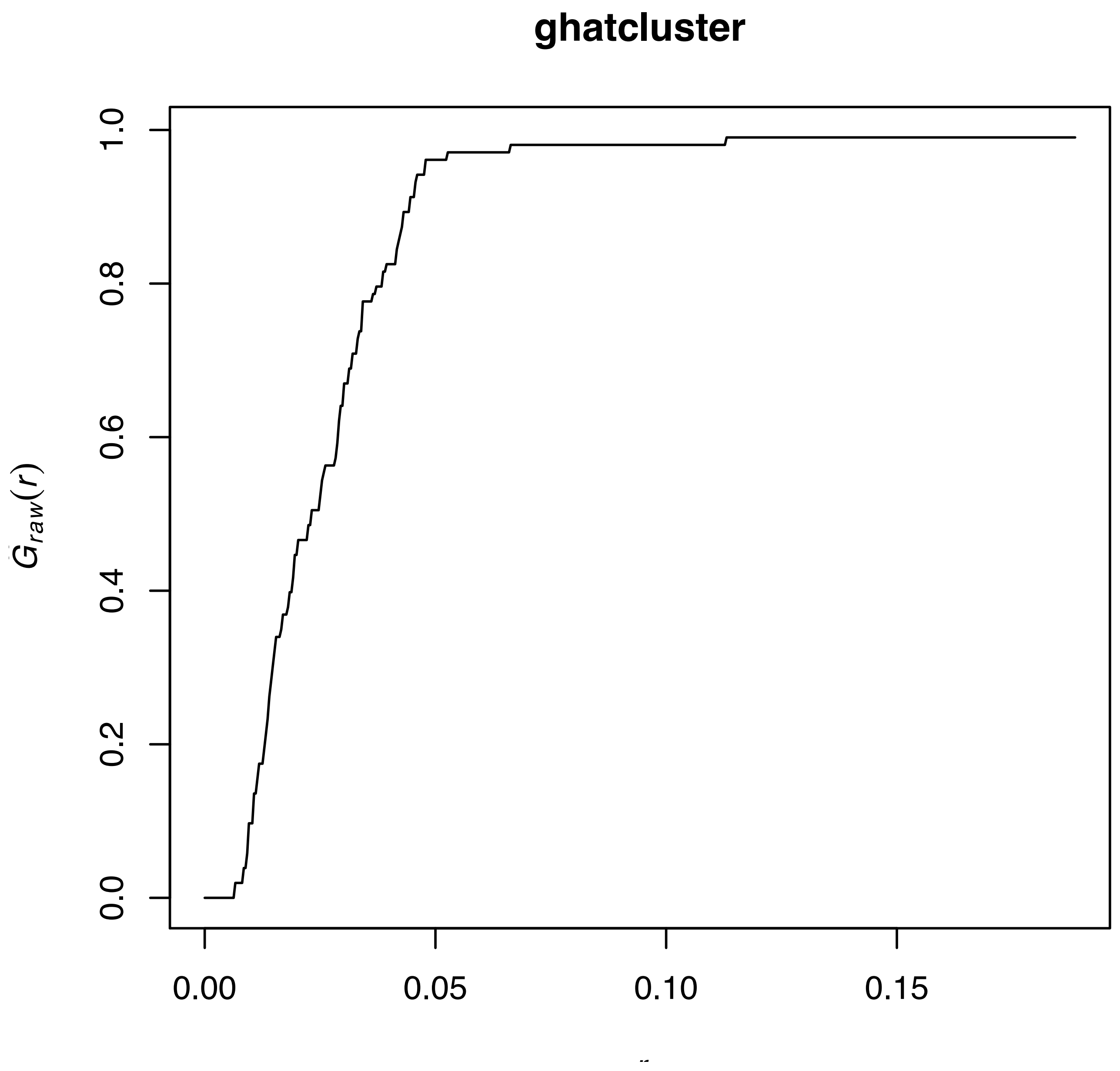
Examples of G function

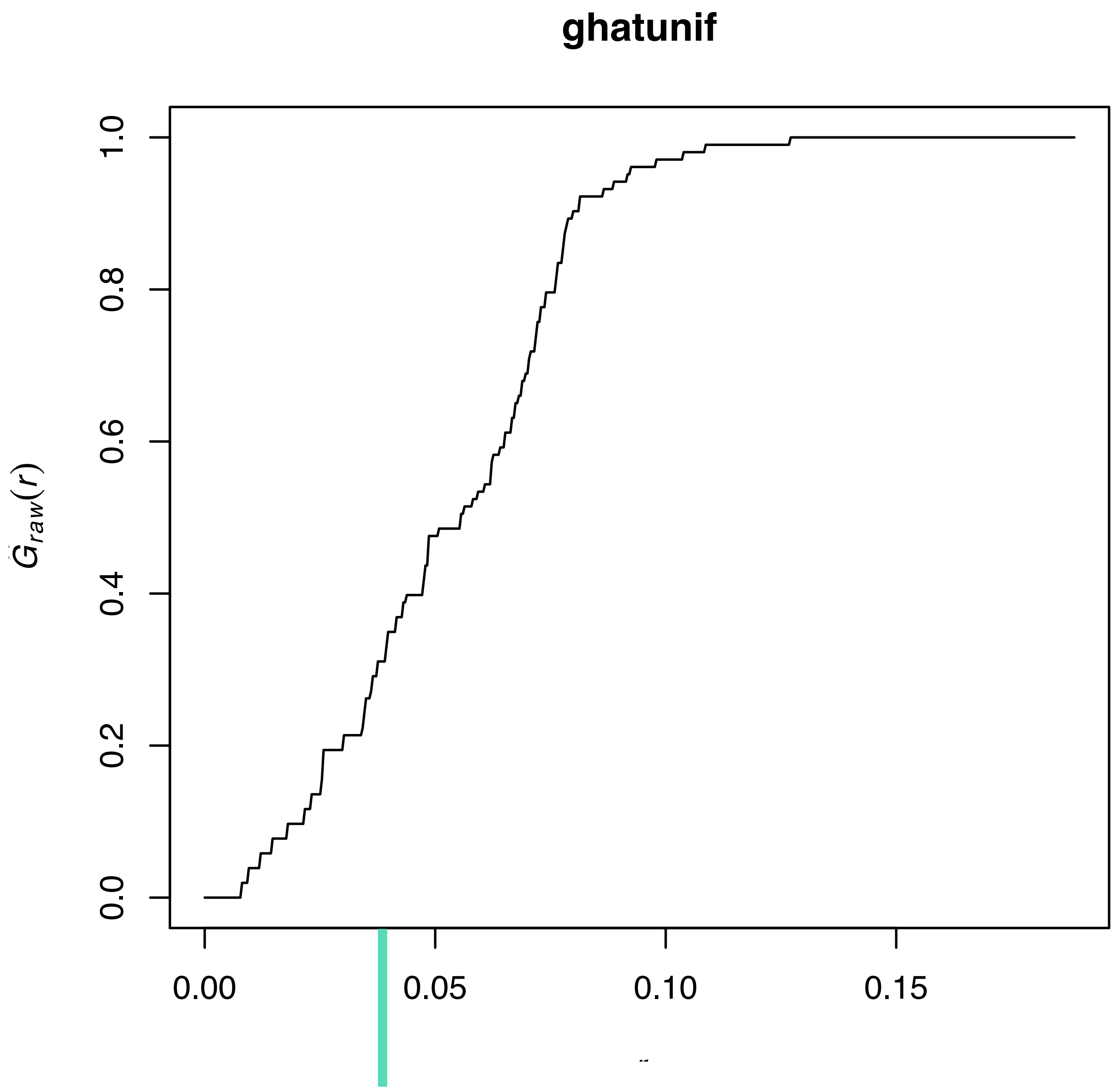
- **for clustered events**, $\hat{G}(d)$ rises at short distances, and then levels off at larger d-values

- **for randomly-spaced events**, $\hat{G}(d)$ rises gradually up to the distance at which most events are spaced, and then increases sharply









F function

- proportion of point-to-nearest-neighbor distances (i.e., *empty space distances*) $\tilde{d}_{min}(s_p)$ no greater than given distance cutoff d, estimated as:

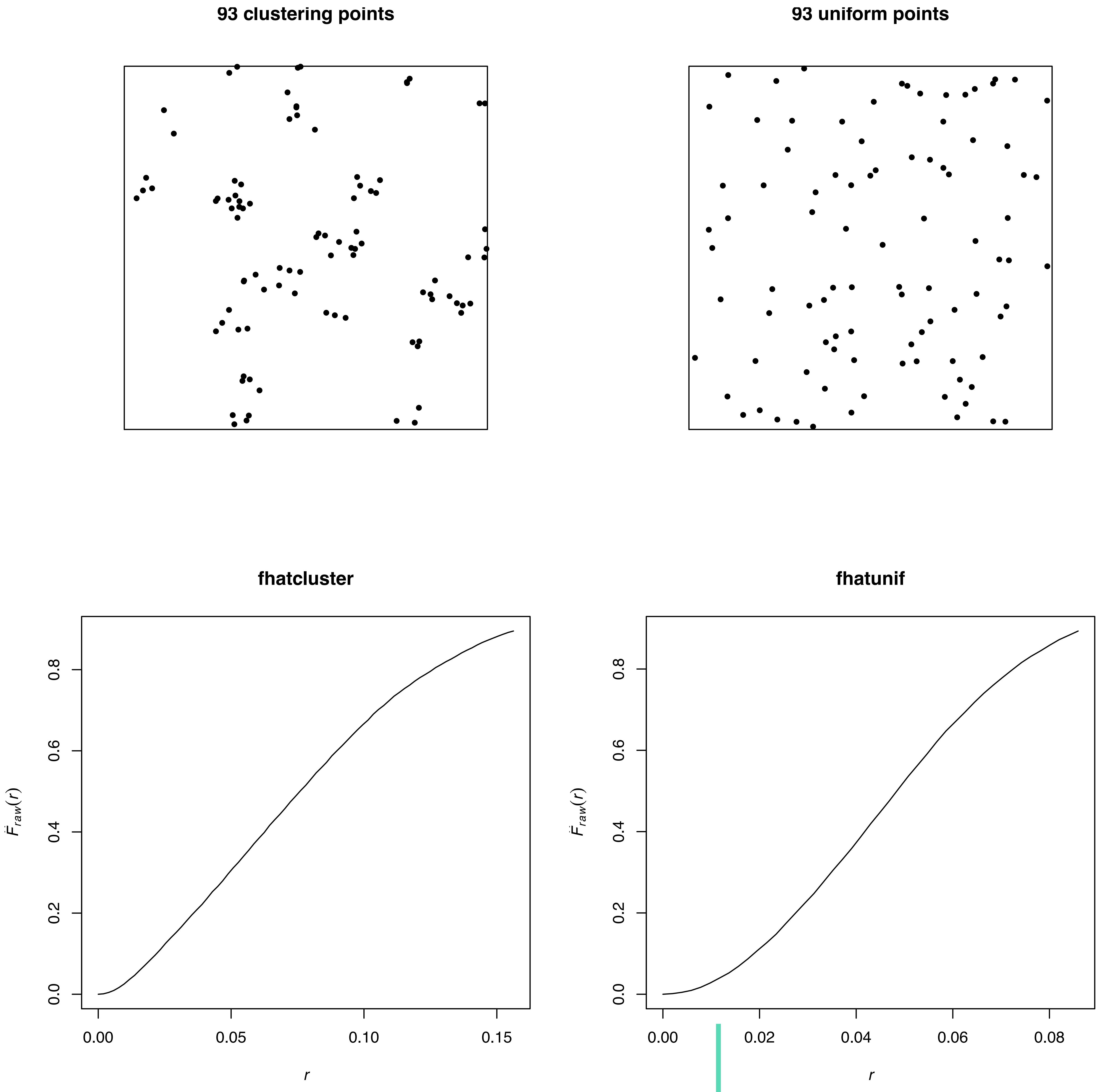
$$\hat{F}(d) = \frac{\#\{\tilde{d}_{min}(\tilde{s}_p) < d, p = 1, \ldots, m\}}{m}$$

- alternative definition: cumulative distribution function (CDF) of all m point-to-nearest-neighbor distances
- the F function provides information on event proximity to voids

Examples of F function

- **for clustered events**, $\hat{F}(d)$ rises sharply at short distances, and the levels off at larger d-values

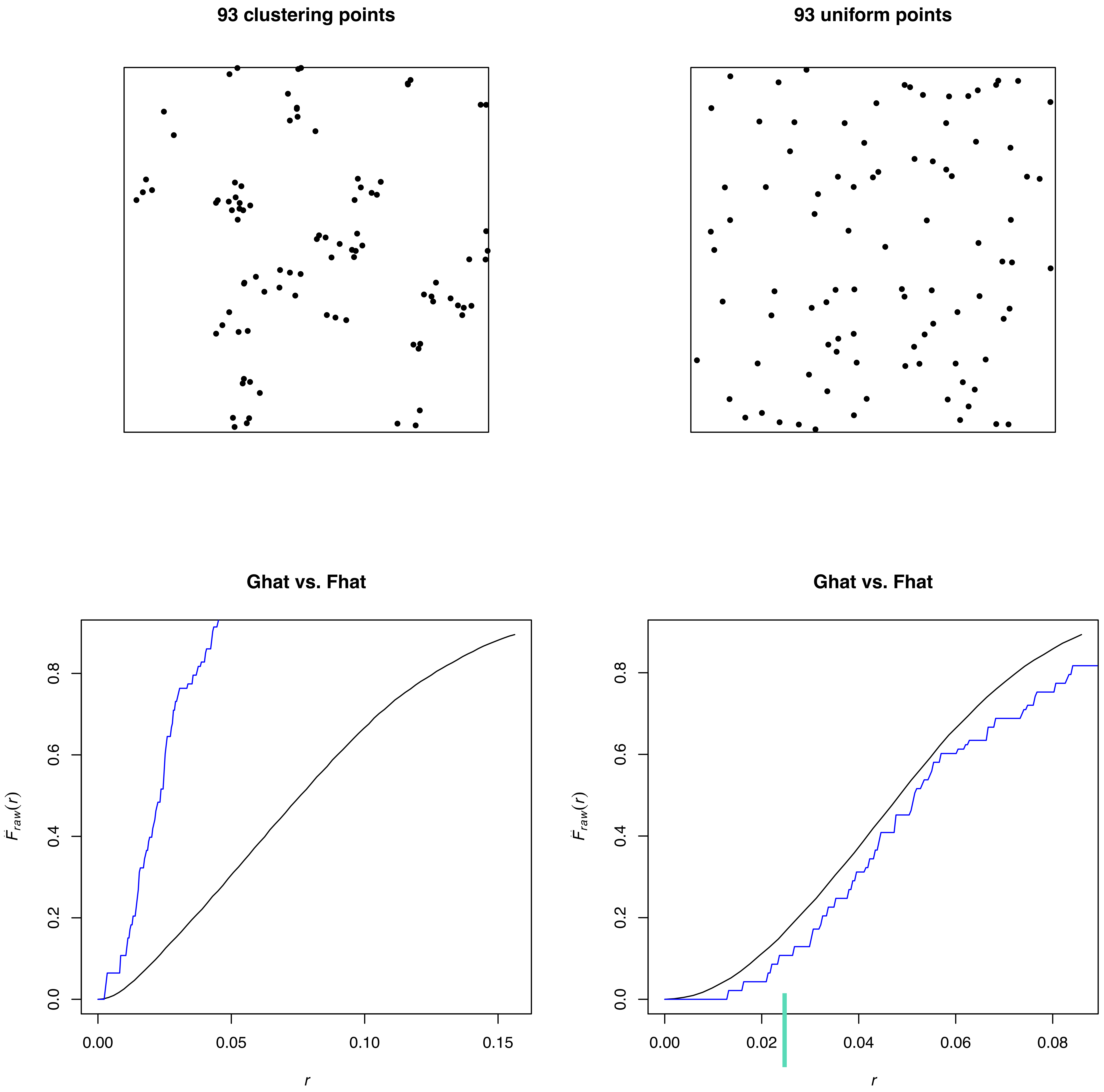
- **for randomly-spaced events**, $\hat{F}(d)$ rises rapidly up to the distance at which most events are spaced, and then levels off (there are more nearest neighbours at small distances from randomly placed points)



comparing G and F functions

- for clustered events, $\hat{G}(d)$ rises faster

- for randomly-spaced events, $\hat{F}(d)$ tends to be close to $\hat{G}(d)$



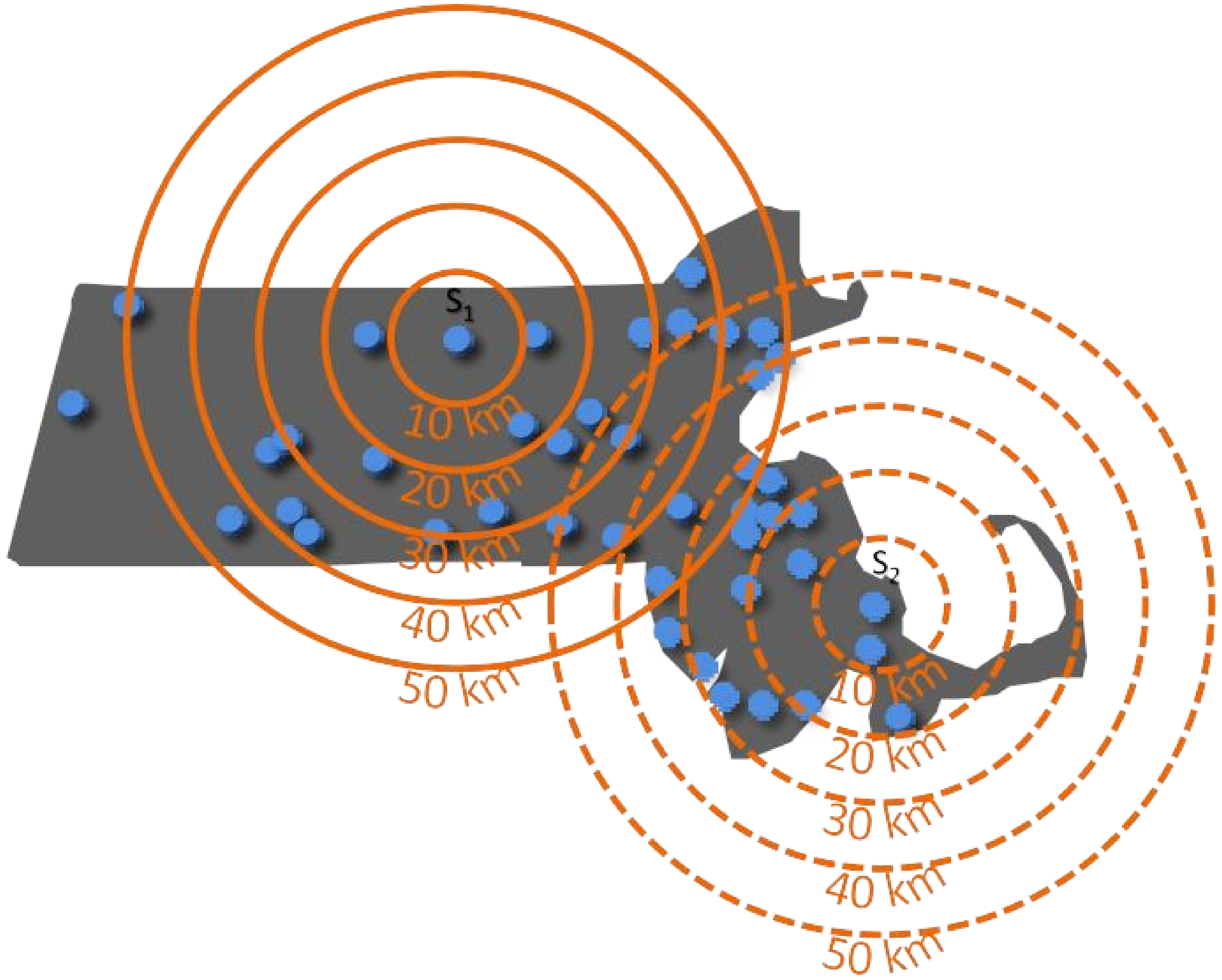
K function

Concept

1. construct set of concentric circles (of increasing radius d) around each event

2. count number of events in each distance "band"

3. cumulative number of events up to radius d around all events becomes the sample K function $\hat{K}(d)$



K function

- Formal definition:

$$
\begin{aligned}
K(d) &= \frac{1}{\lambda}\frac{\#\{d_{ij} \leq d, i, j = 1, \ldots, n\}}{n} \\
&= \frac{|A|}{n}\frac{\#\{d_{ij} \leq d, i, j = 1, \ldots, n\}}{n} \\
&= |A|(\text{proportion of event-to-event distance} \leq d)
\end{aligned}
$$

- In other words, the $\hat{K}(d)$ is the sample cumulative distribution function (CDF) of all n^2 event-to-event distances, scaled by $|A|$

Recap

Spatial point patterns

- set of n point locations with recorded "events"

Describing the first-order effect

- overal intensity
- local intensity (quadrat count and kernel density estimation)

Describing the second-order effect

- nearest neighbour distances
  - the G function
- empty space distances
  - the F function
- pair-wise distances
  - the K function

caveats

- theoretical G, F, K functions are defined and estimated under the *assumption that the point process is stationary (homogeneous)*
- these summary functions *do not completely characterise the process*
- if the process is not stationary, deviations between the empirical and theoretical functions (e.g. $\hat{K}$ and K) are not necessarily evidence of interpoint interaction, since they may also be attributable to variations in intensity

descriptive vs statistical points pattern analysis

Descriptive analysis:

- set of quantitative (and graphical) tools for characterizing spatial point patterns
- different tools are appropriate for investigating first- or second-order effects (e.g., kernel density estimation versus sample G function)
- can shed light onto whether points are clustered or evenly distributed in space

Limitation:

- no assessment of *how* clustered or *how* evenly-spaced is an observed point pattern
- no yardstick against which to compare observed values (or graph) of results

descriptive vs statistical points pattern analysis

Statistical analysis:

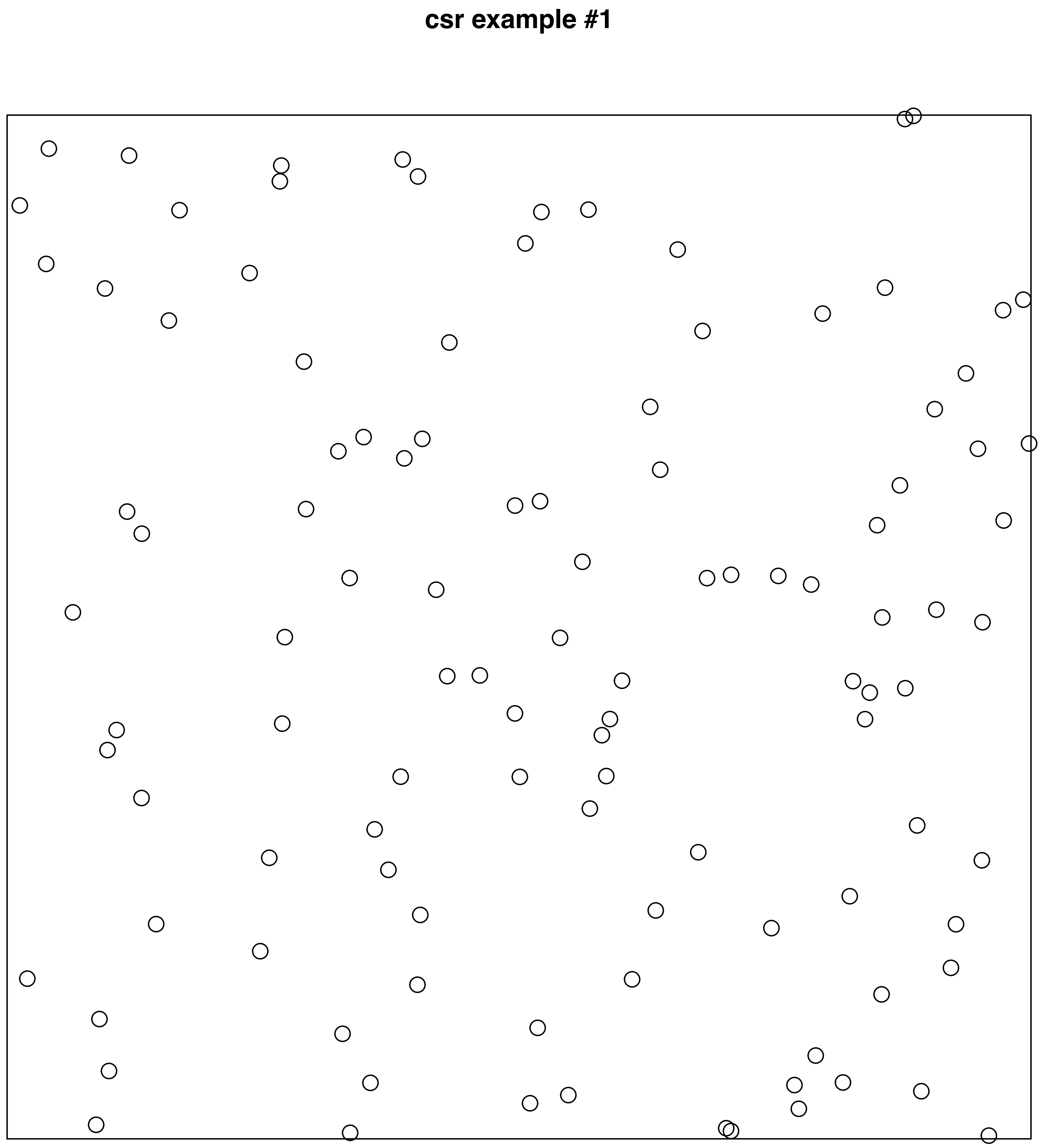
- assessment of whether an observed point pattern can be regarded as one (out of many) realizations from a particular spatial process
- measures of confidence with which the above assessment can be made (how likely is that the observed pattern is a realization of a particular spatial process)

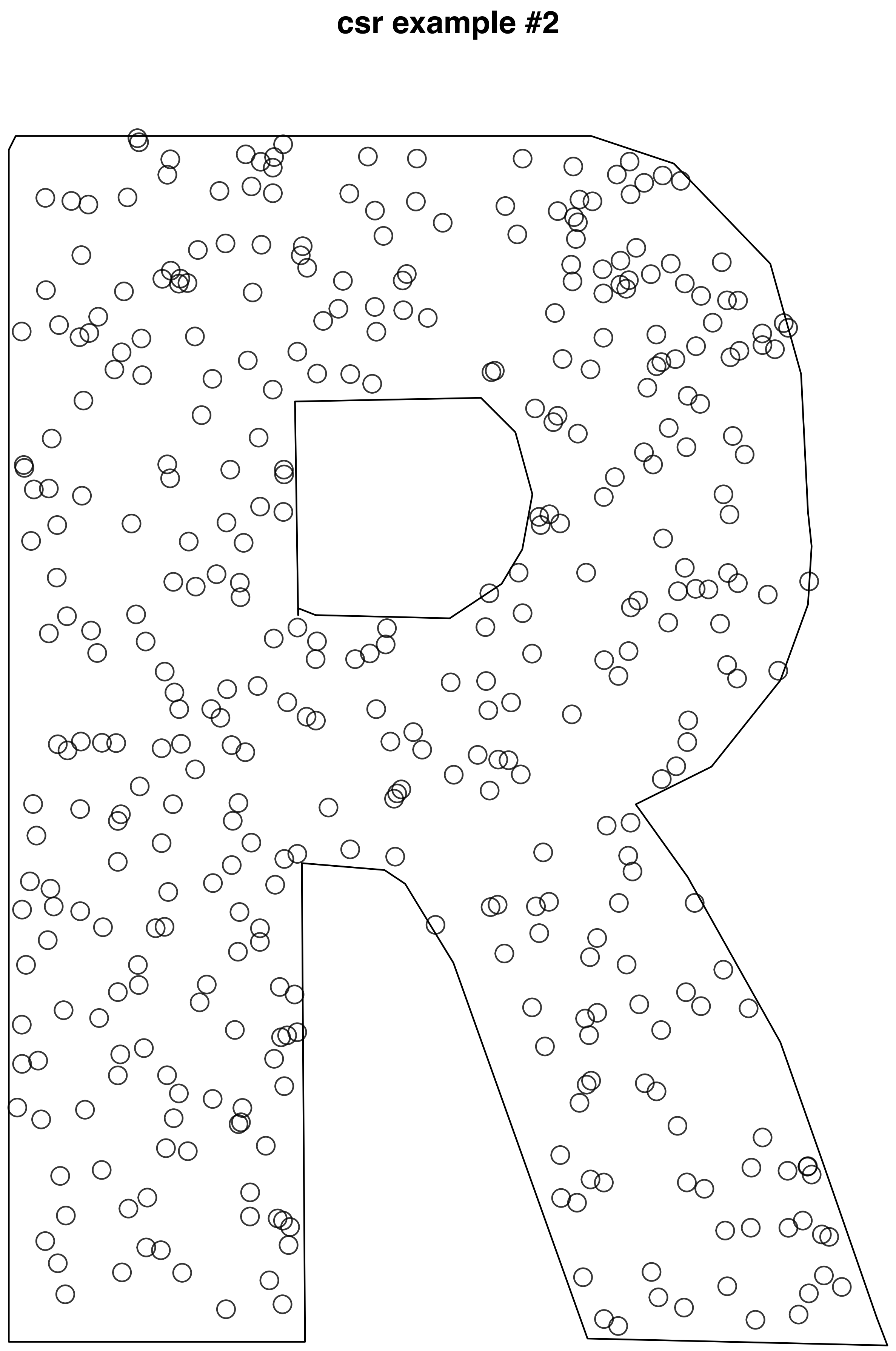
Are daisies randomly distributed in your garden?

complete spatial randomnes (CSR)

- yardstick, reference model that observed point patterns could be compared with, i.e., null hypothesis
- = *homogeneous (uniform) Poisson point process*
- basic properties:
  - the number of points falling in any region A has a Poisson distribution with mean $\lambda|A|$
  - given that there are n points inside region A, the locations of these points are i.i.d. and uniformly distributed inside A
  - the contents of two disjoint regions A and B are independent

complete spatial randomnes (CSR)
example





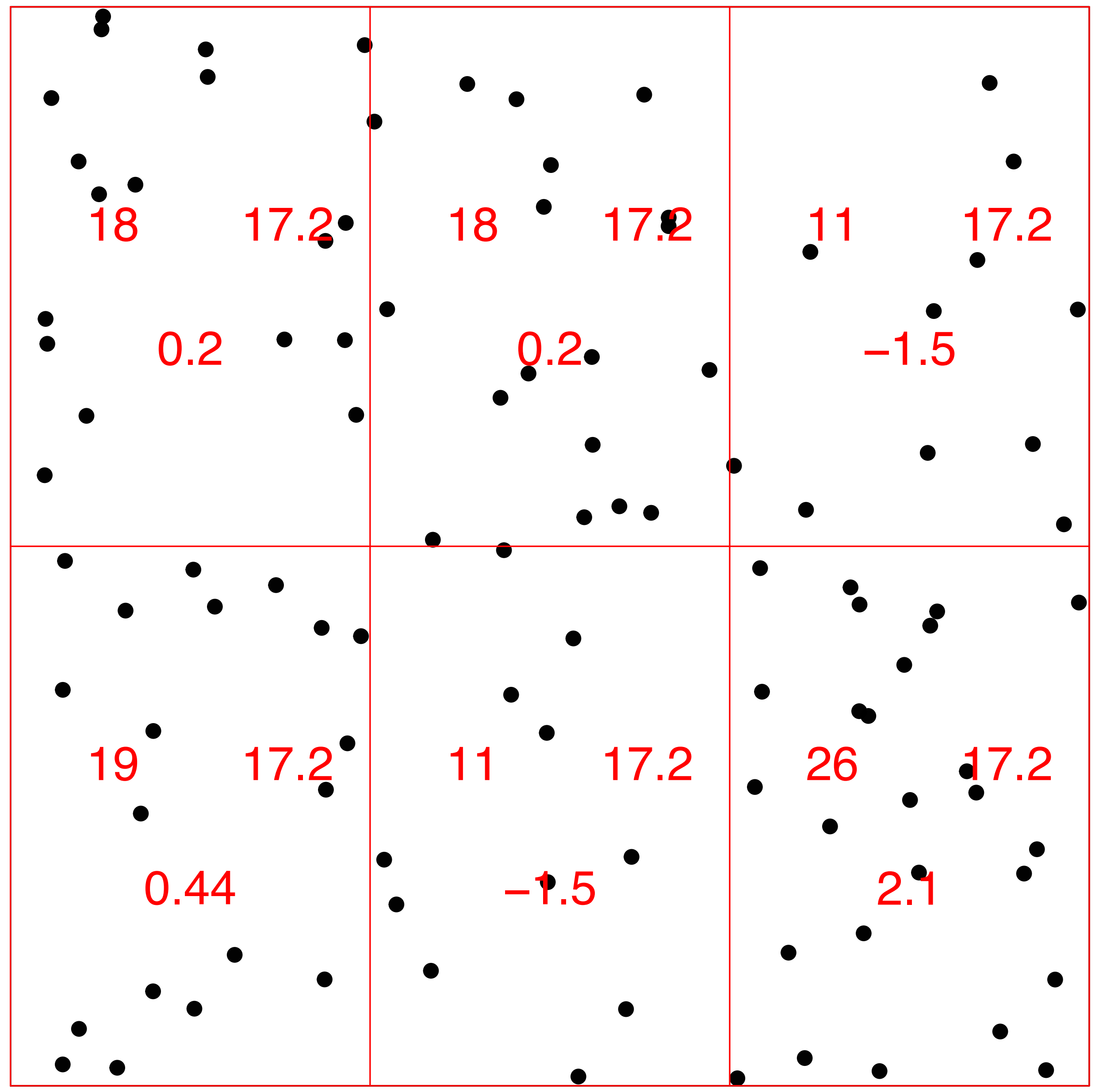
quadrant counting test for CSR

- partition study area A into L sub-regions (quadrats), $A_1, \ldots, A_L$
- count number of events $n(A_l)$ in each sub-region A_l
- Under the null hypothesis of CSR, the $n(A_l)$ are i.i.d. Poisson random variables with the same expected value
- The Pearson χ^2 goodness-of-fit test can be used
    - test statistics: Pearson residual $\sum_l \epsilon(A_l)^2$

$$\epsilon(A_l) = \frac{n(A_l) - \mu(A_l)}{\sqrt{\mu(A_l)}},$$

      where $\mu(A_l)$ indicates the expected number of events in A_l
    - $\sum_{l=1} \epsilon(A_l)^2$ is assumed to follow χ^2 distribution

example



- three values indicate the number of observations, CSR-expected number of observations, and the Pearson residuals
- p-value = 0.617

nearest neighbour index (NNI) test under CSR

Nearest neighbour index

- Compares the mean of the distance observed between each point and its nearest neighbor ($\bar{d}_{min}$) and the expected mean distance under CSR $E(d_{min})$

$$NNI = \frac{\bar{d}_{min}}{E(d_{min})}$$

- Under CSR, we have:

$$E(d_{min}) = \frac{1}{2\sqrt{\lambda}}$$

$$\sigma(d_{min}) = \frac{0.26136}{\sqrt{n^2/A}}$$

nearest neighbour index (NNI) test under CSR

Nearest neighbour index test

- Test statistics:

$$z = \frac{\bar{d}_{min} - E(d_{min})}{\sigma(d_{min})},$$

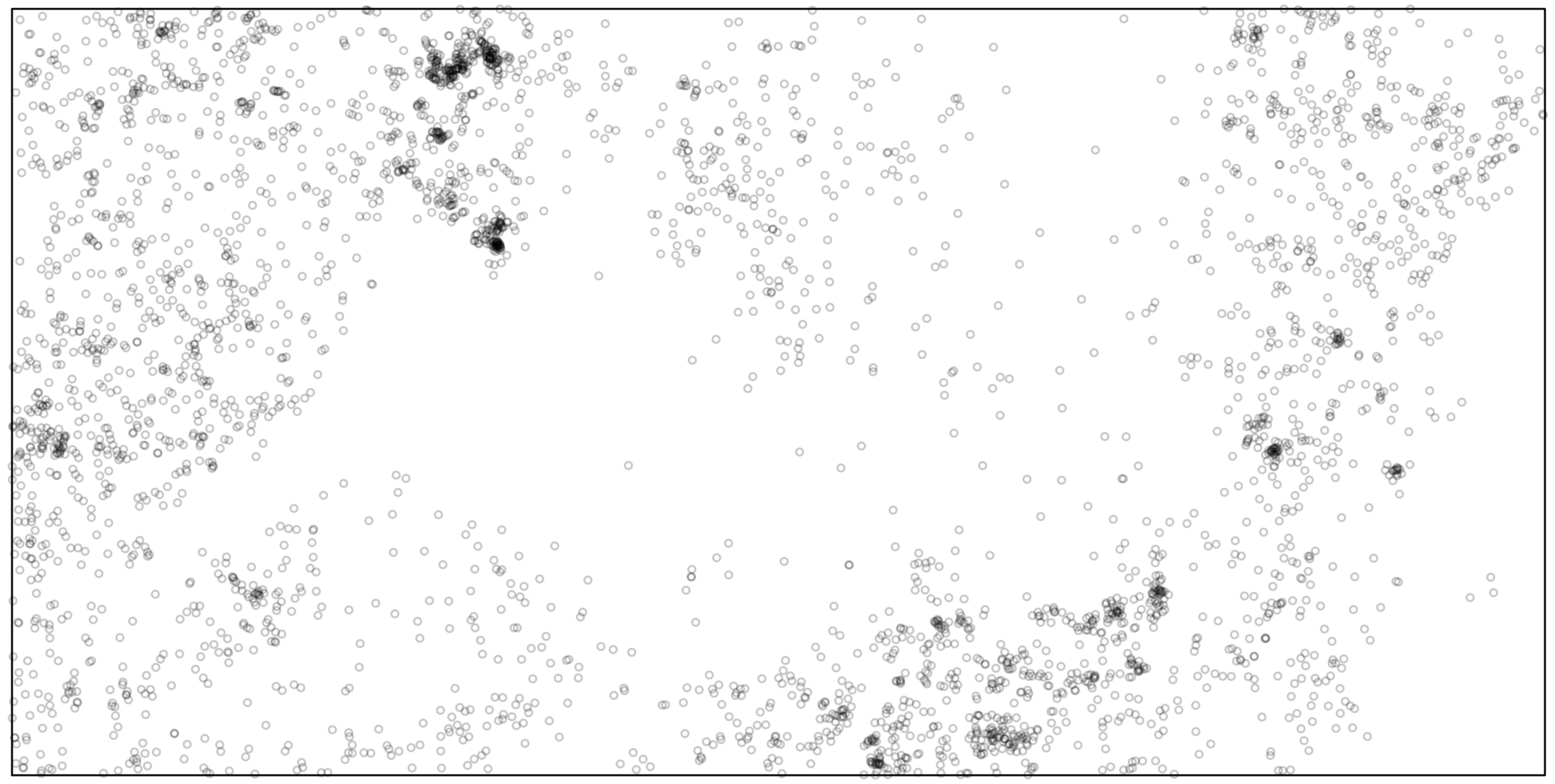
- z is assumed to follow Gaussian distribution, thus, if $z < -1, 96$ or $z > 1.96$, we are 95% confident that the distribution is <u>not</u> randomly distributed

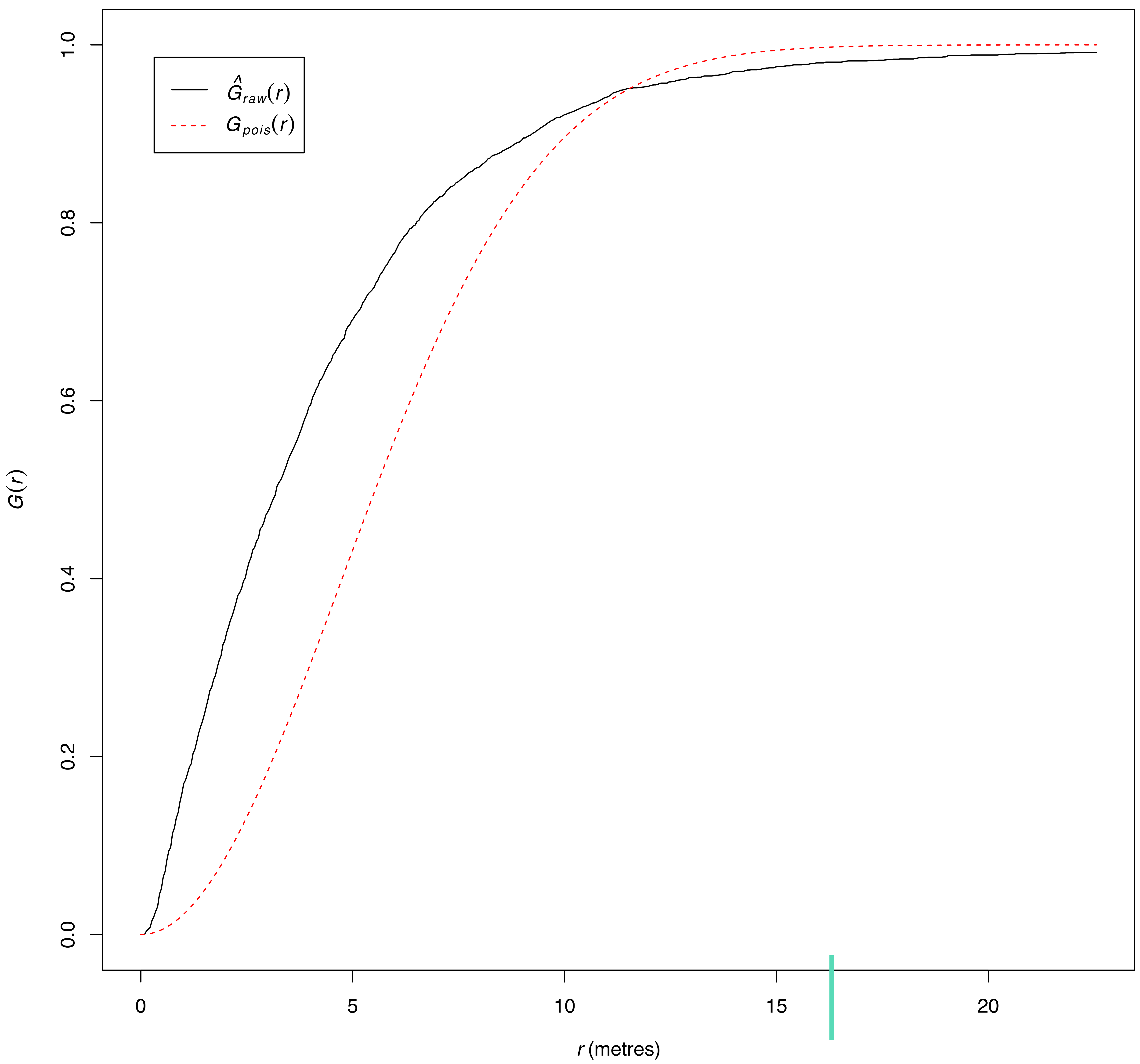
G function under CSR

- The G function is a function of nearest-neighbour distances
- For a homogeneous Poisson point process of intensity λ, the nearest-neighour distance distribution (the G function) is known to be:

$$G(d) = 1 - \exp\{-\lambda \pi d^2\}$$

G function under CSR



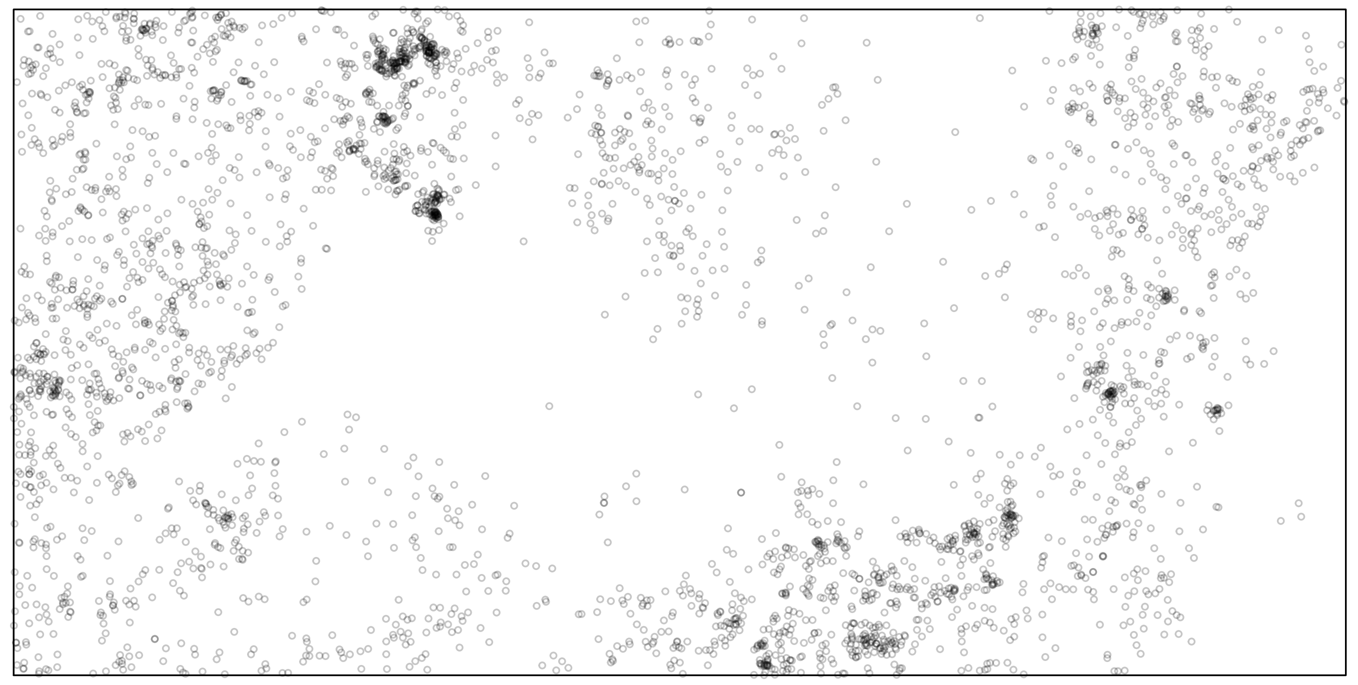


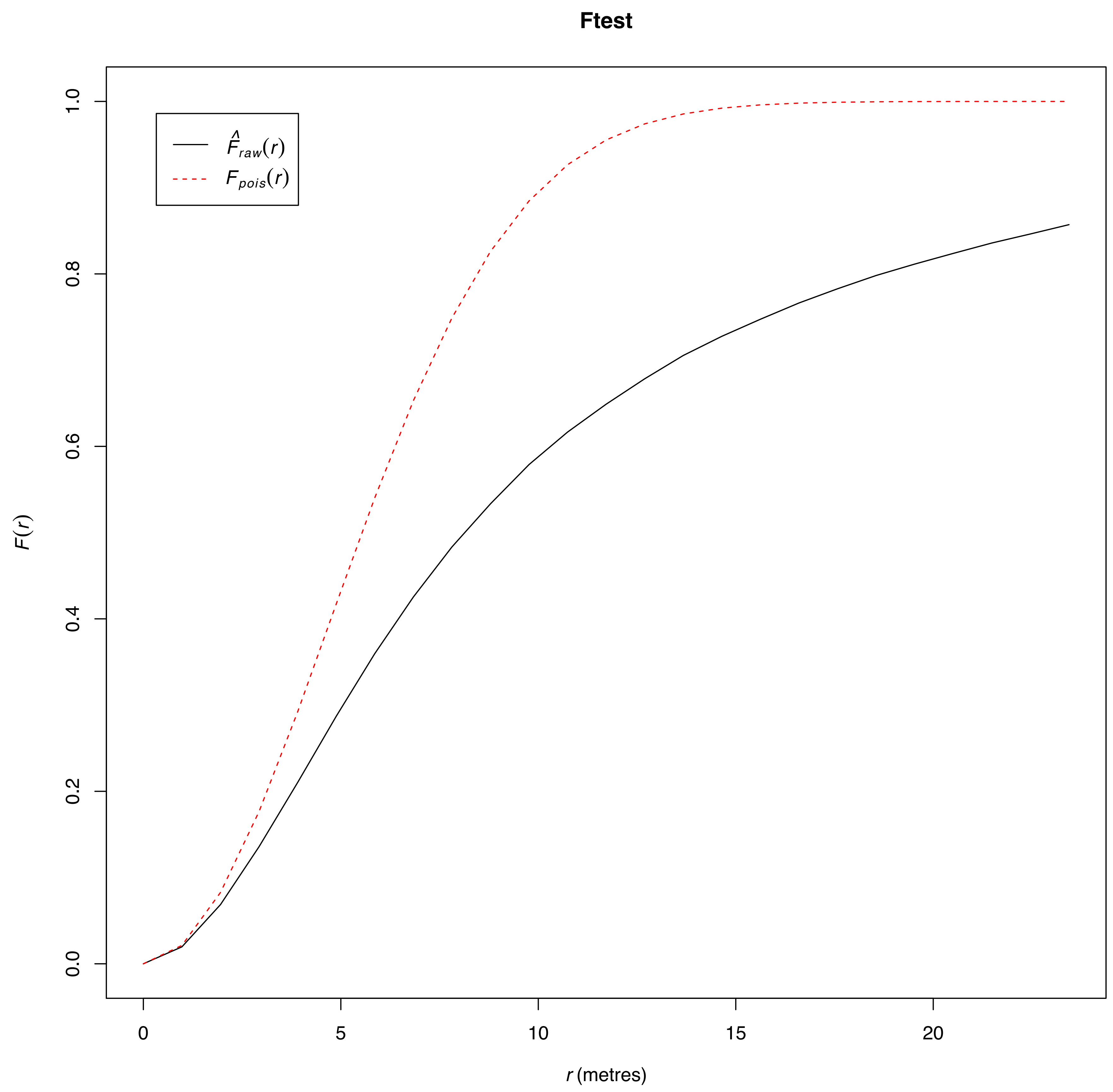
F function under CSR

- The F function is a function of empty space distances
- For a homogeneous Poisson point process of intensity λ, the empty space distance distribution (the F function)is known to be:

$$F(d) = 1 - \exp\{-\lambda \pi d^2\}$$

- Equivalent to the G function
- Intuitively, because points (events) of the Poisson process are independent of each other, the knowledge that a random point is a event of a point pattern does not affect any other event of the process





K function under CSR

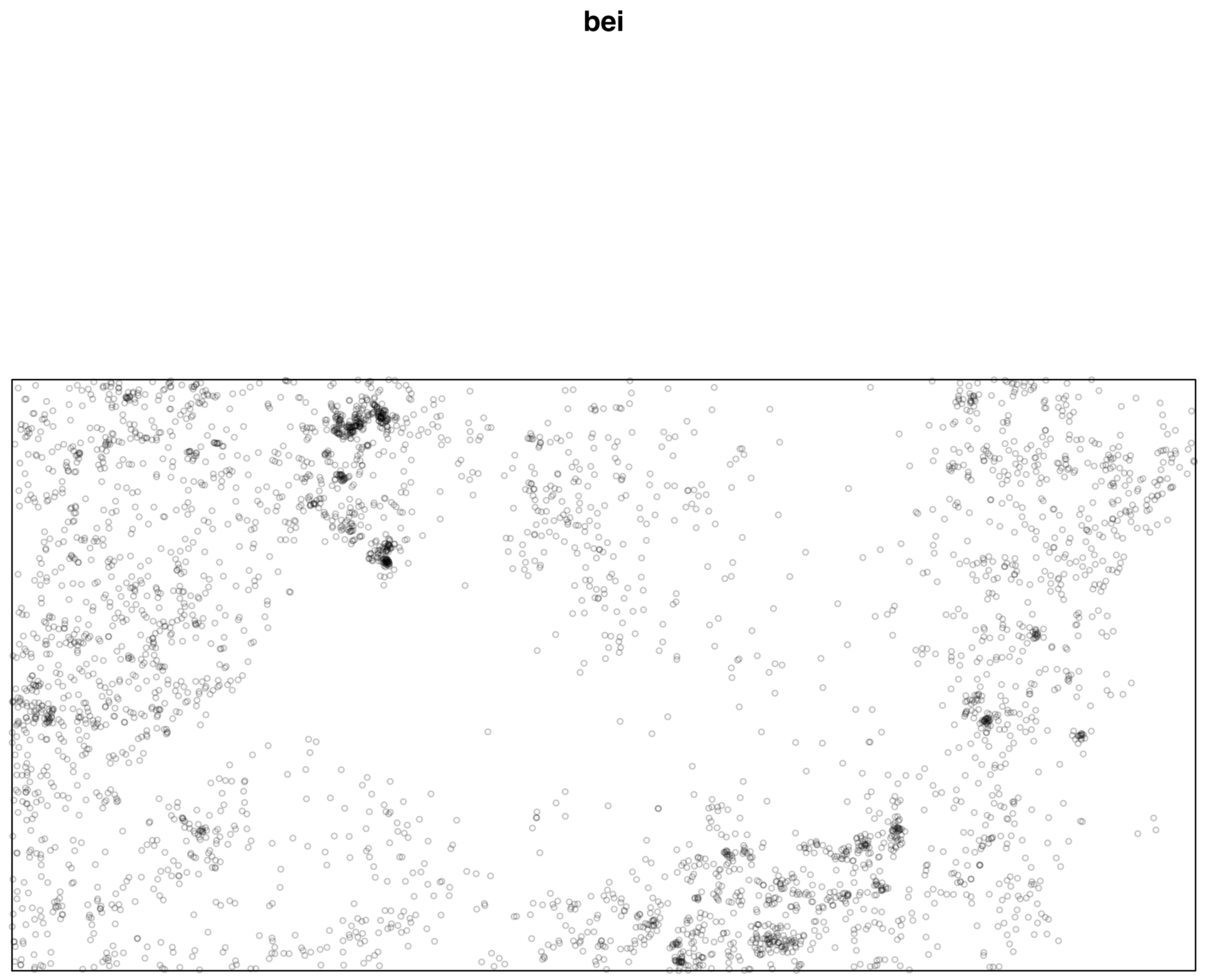
- The K function is a function of pair-wise distances
- For a homogeneous Poisson point process of intensity λ, the pair-wise distance distribution (the K function) is known to be:

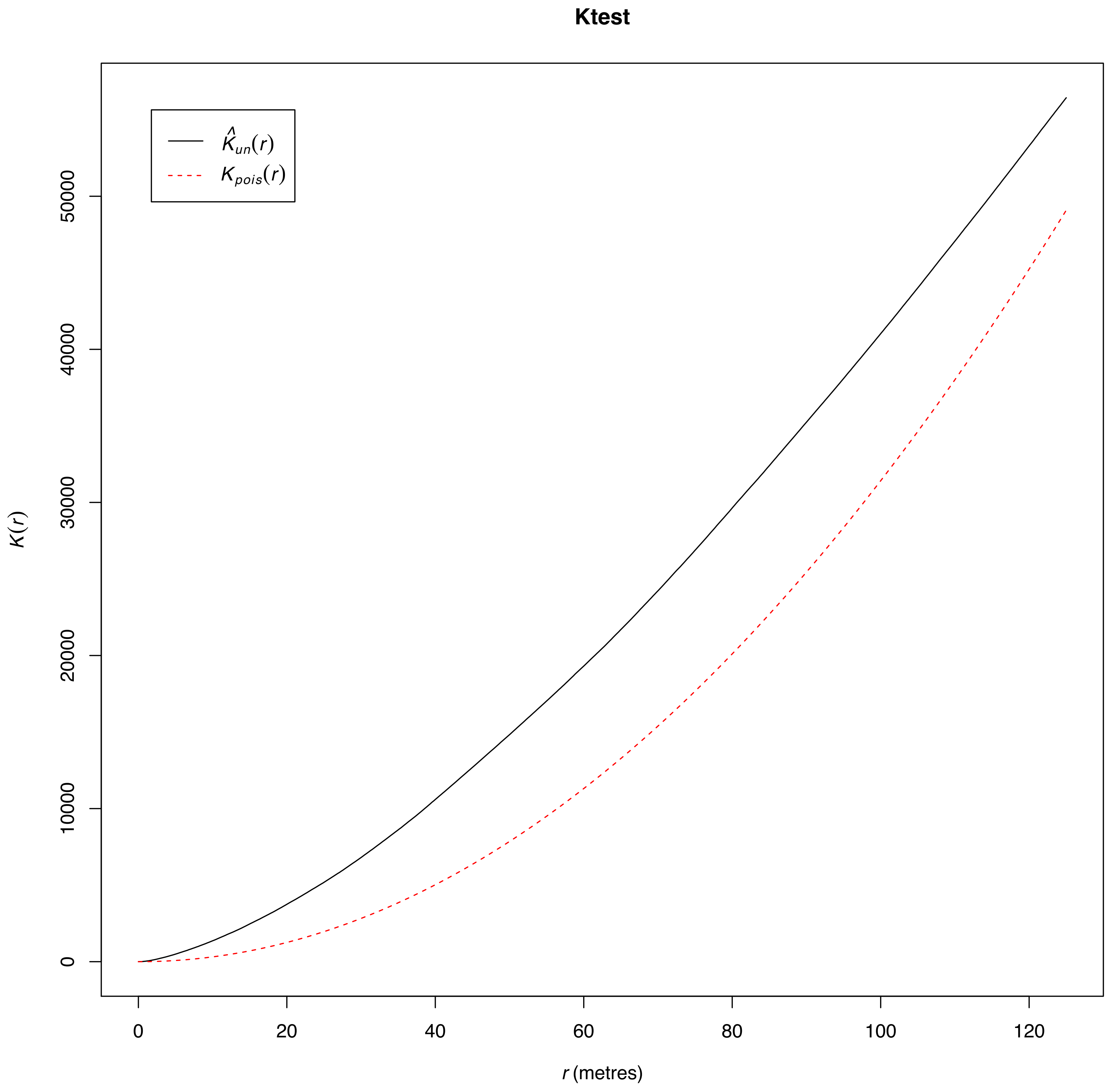
$$K(d) = \pi d^2$$

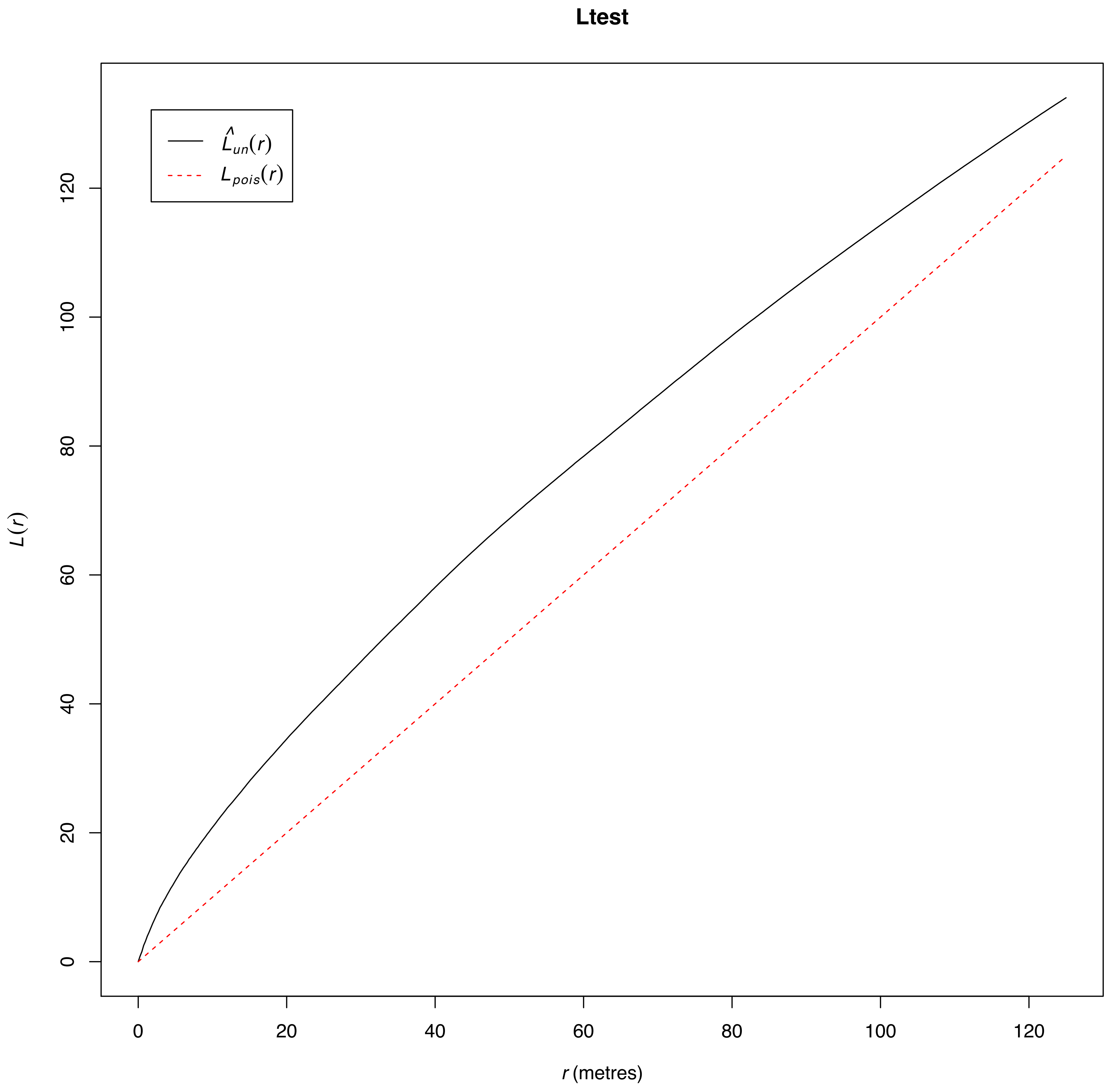
- A commonly-used transformation of K is the L-function:

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

K function under CSR



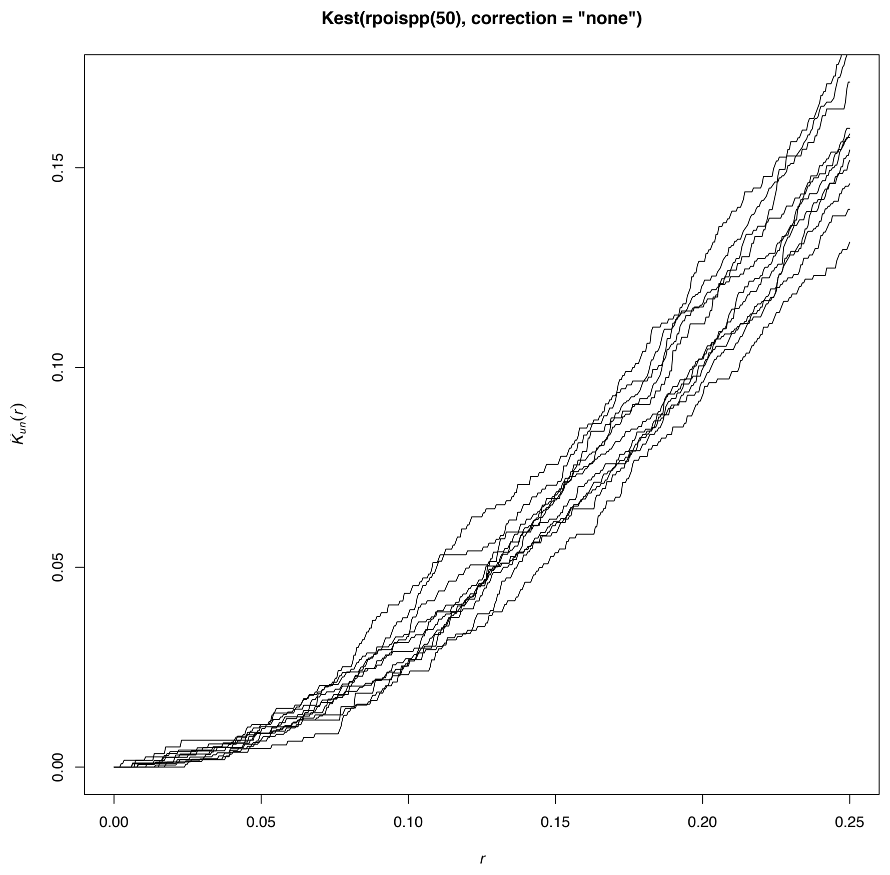




Monte Carlo test

- because of random variability, we will never obtain perfect agreement between sample functions (say the K function) with theoretical functions (the theoretical K functions), even with a completely random pattern

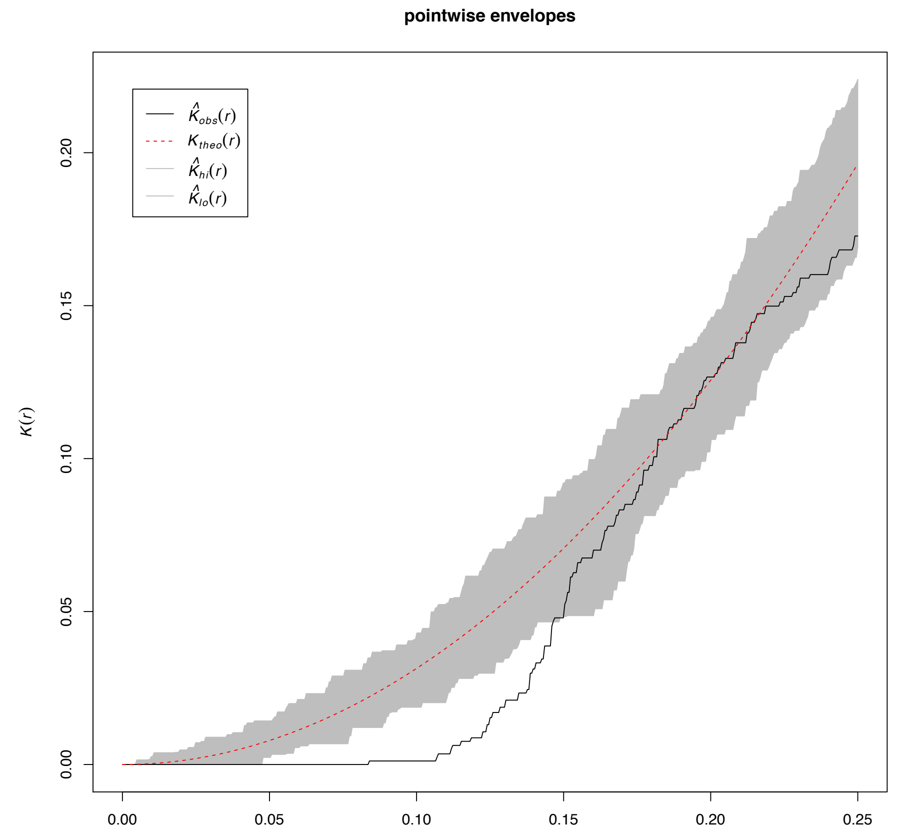
Example



Monte Carlo test

- A *Monte Carlo* test is a test based on simulations from the null hypothesis
- Basic procedures:
  - generate M independent simulations of CSR inside the study region A
  - compute the estimated K functions for each of these realisations, say $\hat{K}^{(j)}(r)$ for $j = 1, \ldots, M$
  - obtain the pointwise upper and lower envelopes of these simulated curves
  - not a confidence interval

Example



recap

Statitsical analysis of spatial point patterns:

- allows to quantify departure of results obtained via exploratory tools, e.g., $\hat{G}(d)$, from expected such results derived under specific null hypotheses, here CSR hypothesis
- can be used to assess to what extent observed point patterns can be regarded as realizations from a particular spatial process (here CSR)
- Same concepts can be applied for hypothesis of other types of point processes (e.g., Poisson cluster process, Cox process)

Sampling distribution of a test statistics

- lies at the heart of any statistical hypothesis testing procedure, and is tied to a particular null hypothesis
- simulation and analytical derivations are two alternative ways of computing such sampling distributions (the latter being increasingly replaced by the former)

@rschifan

schifane@di.unito.it

http://www.di.unito.it/~schifane