Point Processes 2: Spatial Correlations

Michael Noonan

DATA 589: Spatial Statistics

Table of contents

Review

Review

Last lecture we introduced the concept of a spatial point pattern, and how the spatial arrangement of the 'points' is the focus of investigation in point pattern analysis.

We saw how the formal analysis of a point process usually begins with a descriptive analysis of the intensity (first moment quantities).

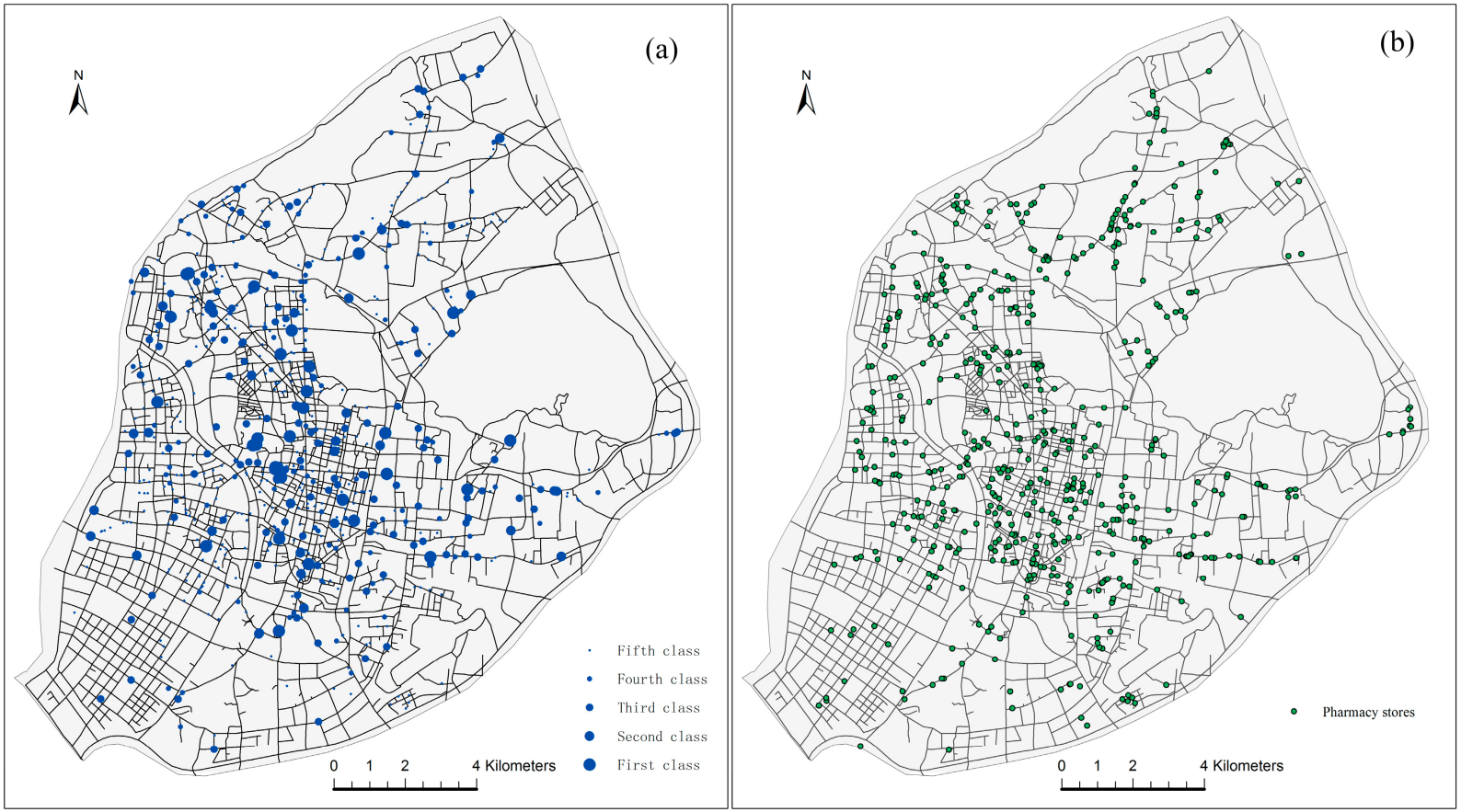
We learned how the spatial patterns in intensity can provide valuable information on a point process, but many of the tests (e.g., kernel estimation, quadrat counting) are sensitive to they way in which they are setup, so analyses should be applied with care.

Today we will focus on how to describe the relationships between points (second moment quantities).

Applied Points Pattern Analysis

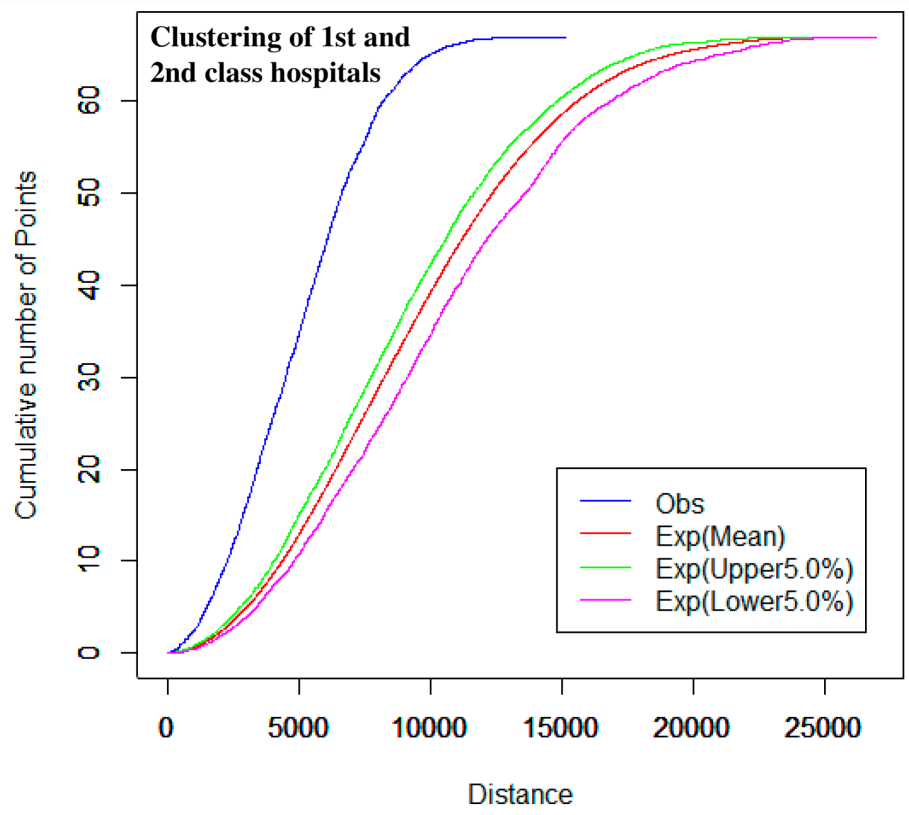
Health care access

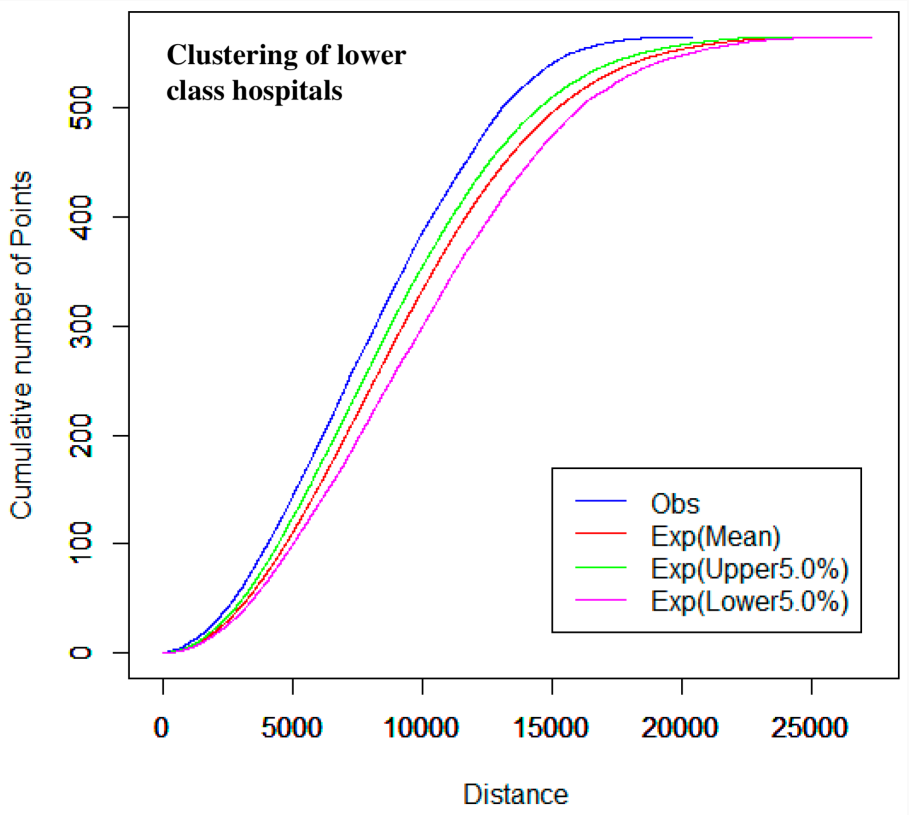
Ni *et al.* (2016) used point pattern analyses to study the characteristics of the spatial distribution of healthcare facilities in Nanjing.



Health care access cont.

Using the K-function, they found that higher-tier hospitals were clustered in space and concentrated in older parts of the city, whereas lower-tier hospitals were randomly distributed.

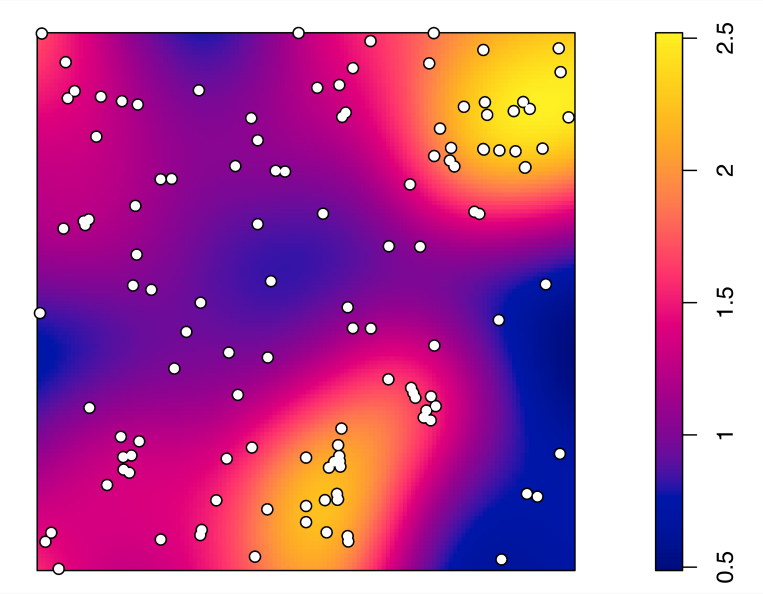




Relationships Between Points

Relationships between points

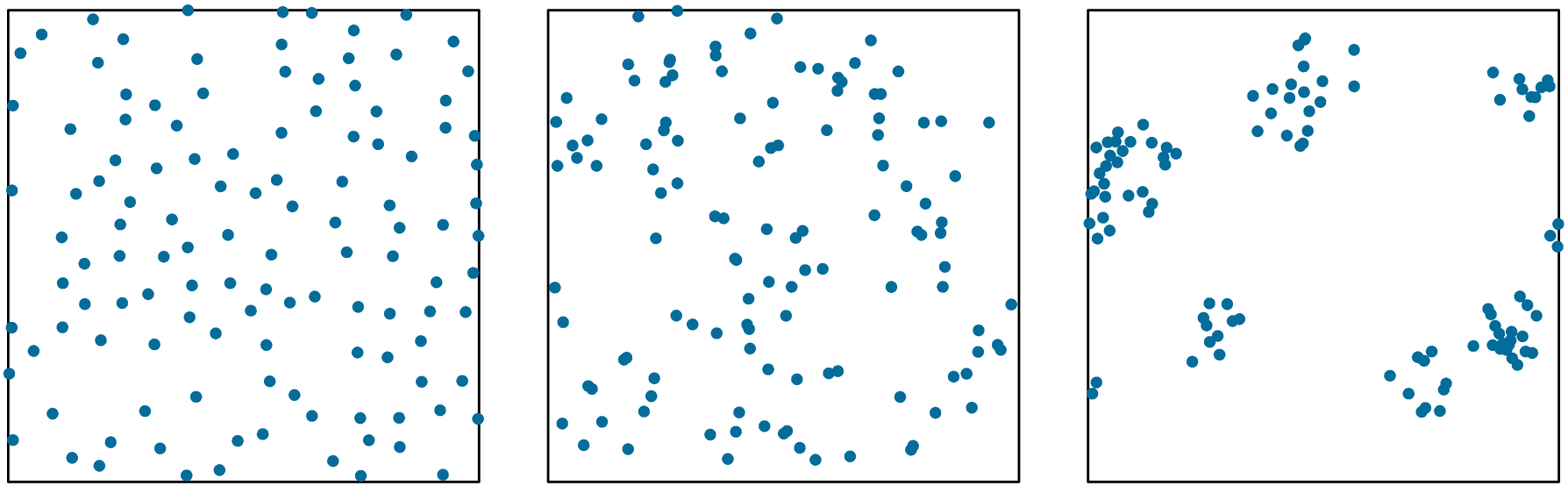
The spatial intensity of a process provides us with information on the number of points we can expect to find at any location u



... but says nothing about the relationships between points.

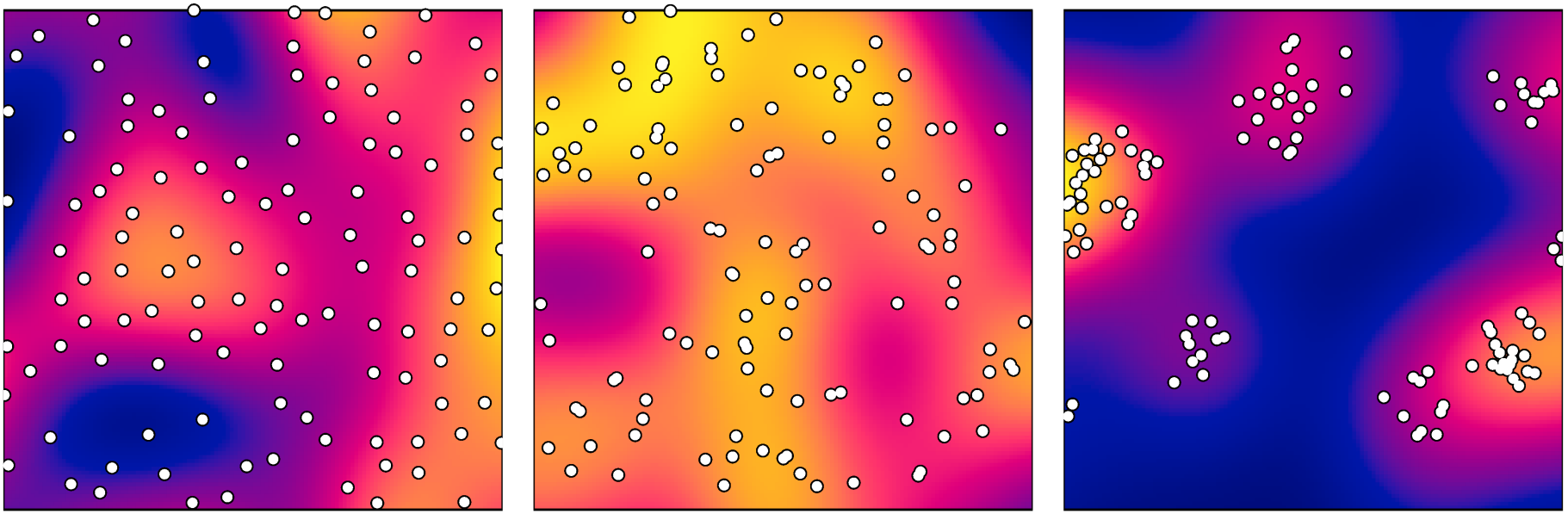
Relationships between points cont.

Points can have a tendency to avoid one another, be independent, or cluster.



Relationships between points cont.

This can generate patterns in the intensity but we don't know if this is caused by environmental factors or relationships between points.



Correlations

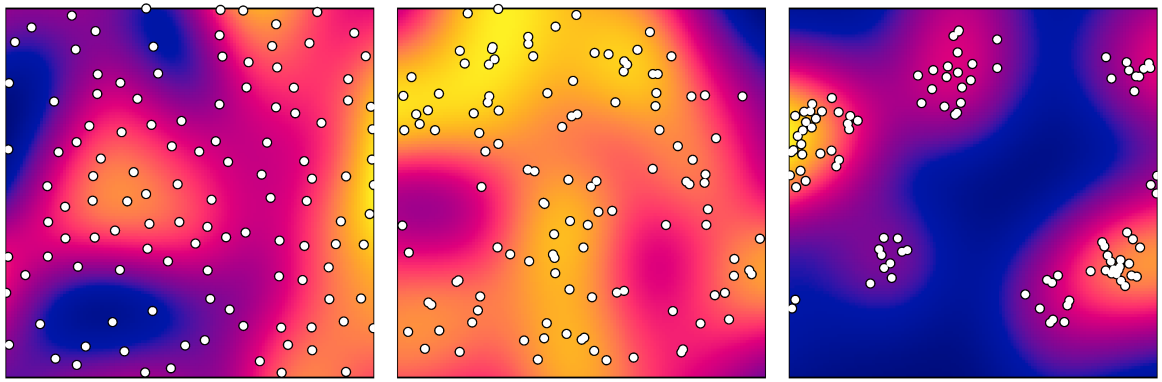
In order to fully understand a point process we need to be able to describe the correlation between points.

In statistics, first moment quantities describe the mean value of a random variable X, second moment quantities describe the mean of X^2 (variance, standard deviation correlation, etc.).

For a point process X, the second moment of $n(\mathrm{X} \cap B)$ can be interpreted as describing patterns in the *pairs* of points x_i, x_j falling in set B.

Correlation caveats

1. Accurately describing correlations, $n(\mathrm{X} \cap B)^2$, requires a an accurate description of the mean, $n(\mathrm{X} \cap B)$.



2. Correlations are summary statistics and do not inherently imply causation.

Morisita's Index

Describing correlations

If we're interested in describing correlations, an easy place to start is with a simple descriptive statistic.

If we subdivide the window into equally sized, m, quadrats, we can count how often a pair of points falls in the same quadrat.

Formally, for a process with n points, there are $n(n-1)$ ordered pairs of distinct points.

When there are m quadrats, the jth quadrat contains $n_j(n_j-1)$ ordered pairs of distinct points, and the total number of ordered pairs of distinct points which fall inside the same quadrat is $\sum_j n_j(n_j-1)$.

Describing correlations cont.

The ratio

$$\frac{\Sigma_j n_j(n_j-1)}{n(n-1)}$$

describes the fraction of all pairs of points which both fall in the same quadrat.

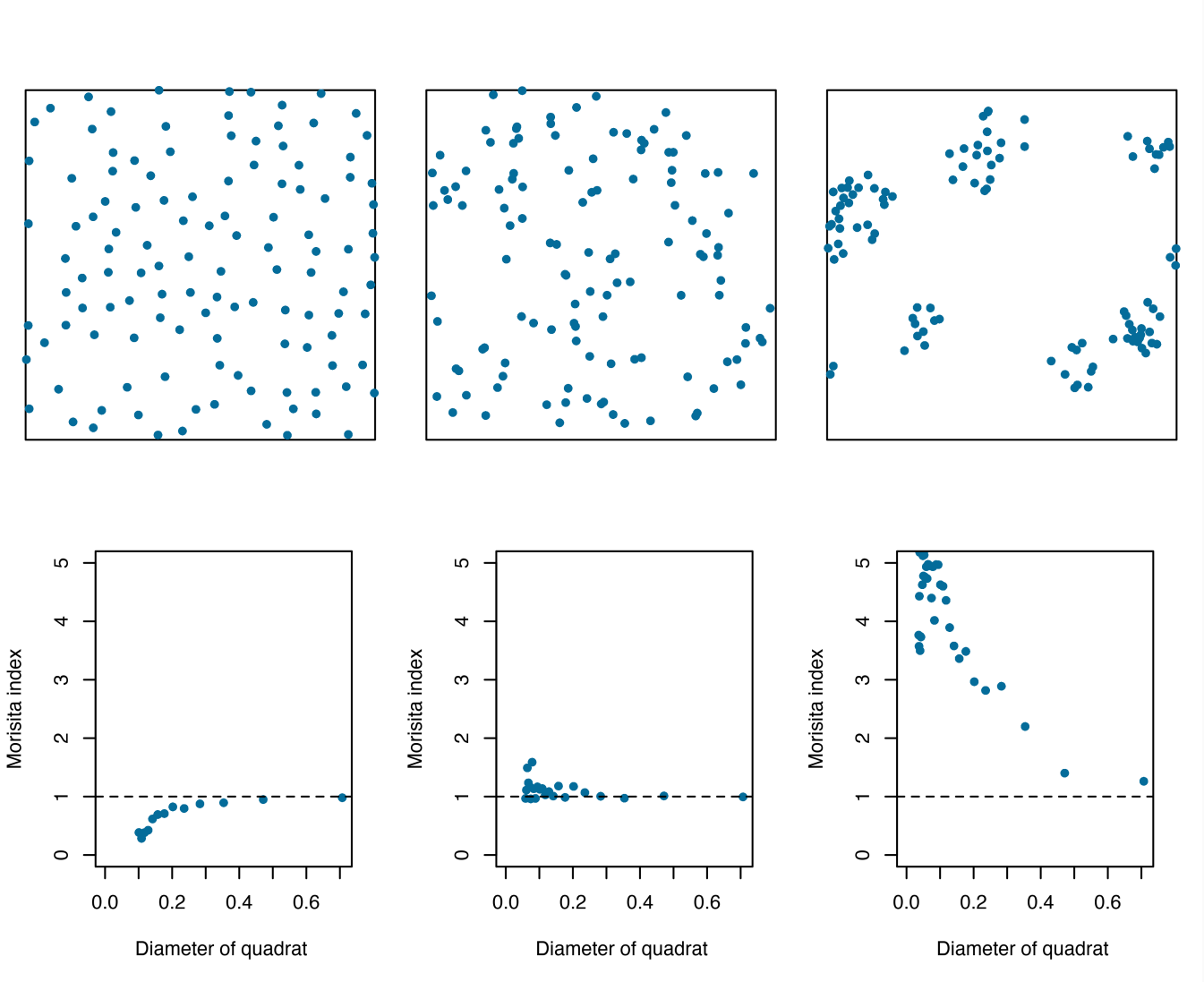
Under an assumption of homogeneity, the probability of a pair of points falling inside equally sized quadrats is just $\frac{1}{m}$, giving us Morisita's Index:

$$M = m\frac{\Sigma_j n_j(n_j-1)}{n(n-1)}.$$

Should be close to 1 if points are independent of one another, lower than 1 if there is avoidance, and greater than 1 if there is attraction.

Morisita's index

Morisita's Index for 3 different point patterns (`miplot()` function).



Morisita's index cont.

Morisita's index can serve as a useful visual diagnostic tool... but the derivation assumed homogeneity.

Large values of M can occur without any underlying attraction between point when intensity is inhomogenous (e.g., your spatial distribution on campus is governed by the need to be in this room, your distribution in this classroom is governed by the arrangement of the seats, Morisita's index would suggest some level of attraction).

If the assumption of homogeneity is broken, the index is not well defined and unlikely to be trustworthy.

The index is based on crude subdivisions, so not sensitive to subtle, fine-scale changes.

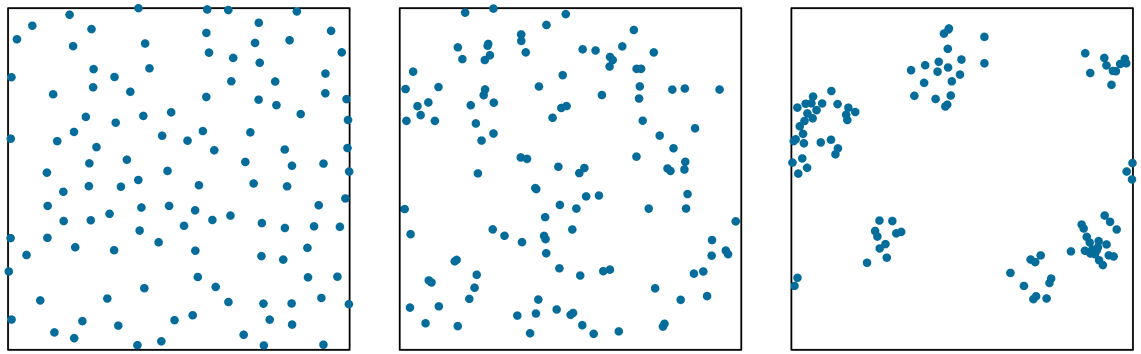
Ripley's K-function

Motivation

Morisita's index describes correlations based on the rate at which pairs of points are found 'close' together

...but if we're interested in the spacing (or distance), why not just build our metric directly off of the separation distances $d_{ij} = ||x_i - x_j||$ between all ordered pairs of distinct points?

Different patterns in clustering should result in different patterns in separation distances.



Pairwise CDF

Let's start by considering the cumulative distribution of pairwise separation distances

$$\begin{aligned}\hat{H}(r) &= \text{fraction of values of } d_{ij} < r \\ &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} 1\{d_{ij} \leq r\}\end{aligned}$$

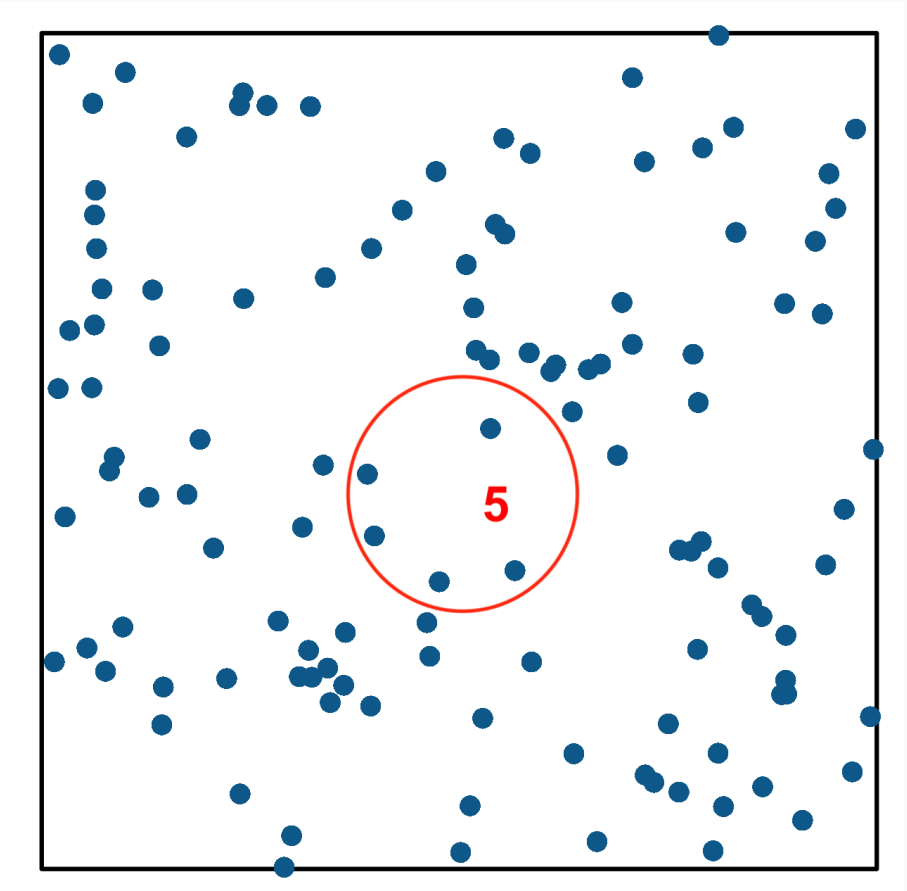
where $1\{d_{ij} \leq r\} = 1$ if true, and 0 if false,
and the sum is taken over all ordered pairs where the indices aren't equal

i.e., $\hat{H}(r)$ is the fraction of pairs of points separated by a distance $\leq r$.

Pairwise CDF cont.

$$\hat{H}(r) = \text{fraction of values of } d_{ij} < r$$

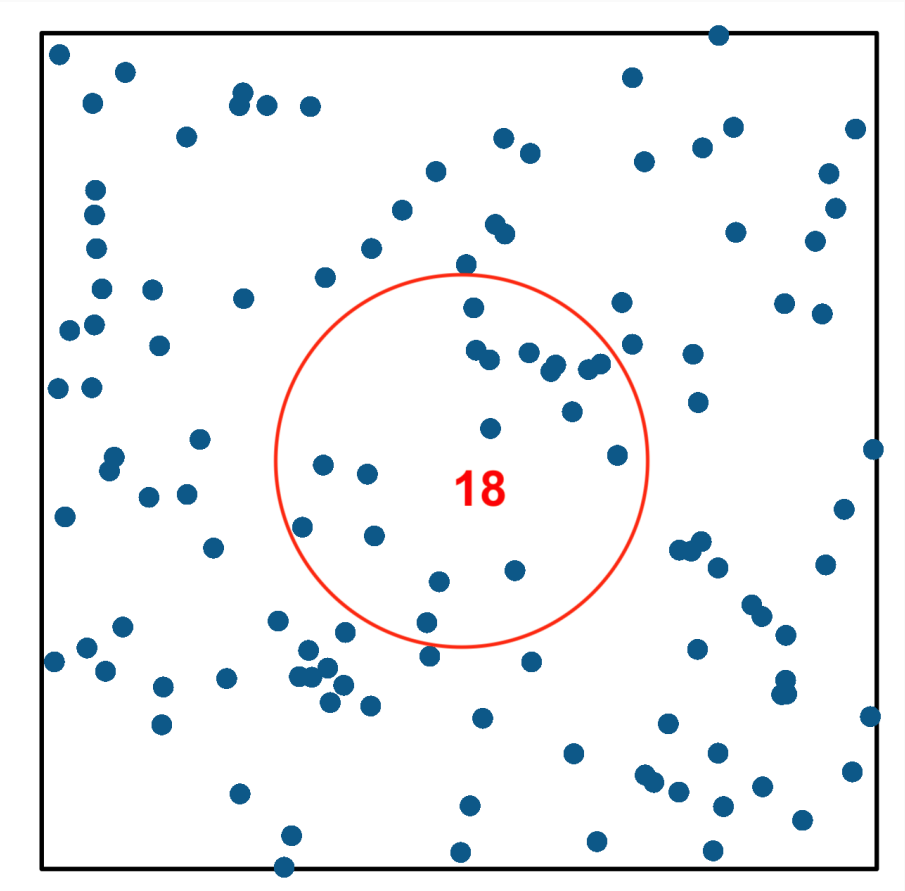
$$= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} 1\{d_{ij} \leq r\}$$



Pairwise CDF cont.

$$\hat{H}(r) = \text{fraction of values of } d_{ij} < r$$

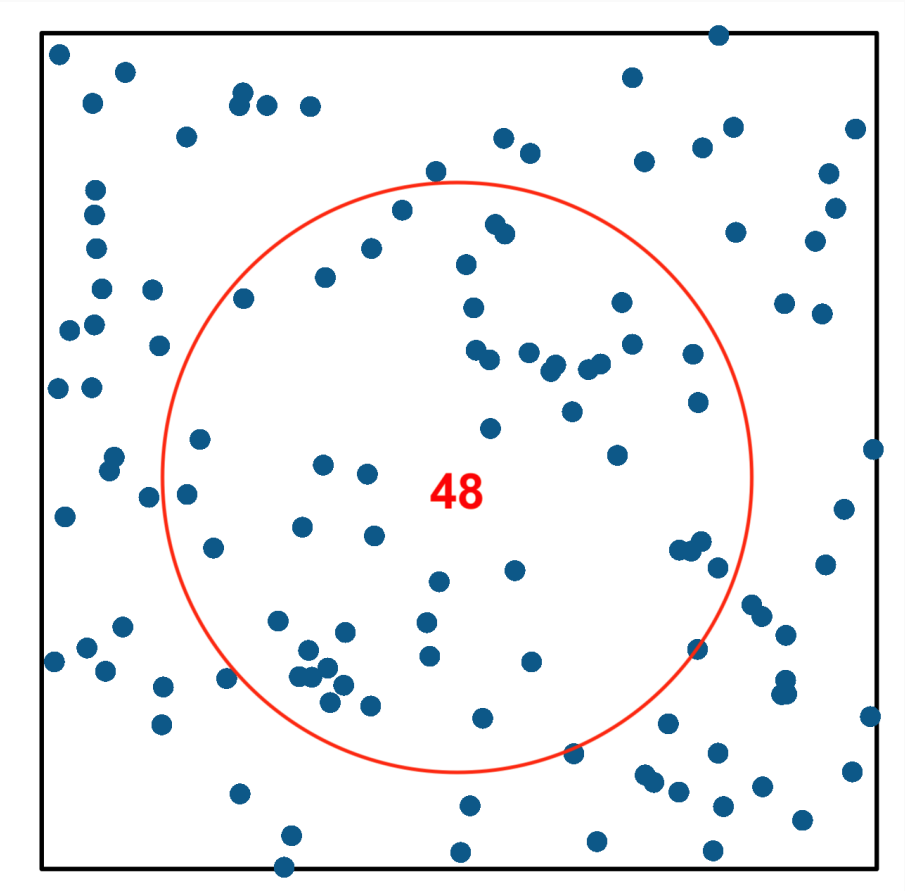
$$= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} 1\{d_{ij} \leq r\}$$



Pairwise CDF cont.

$$\hat{H}(r) = \text{fraction of values of } d_{ij} < r$$

$$= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} 1\{d_{ij} \leq r\}$$



Empirical K-function

$\hat{H}(r)$ is valuable, but is absolute, so it can't be compared between processes with different numbers of points (e.g., clustering of crimes in Kelowna vs. Vancouver, trees in two different forests, etc.).

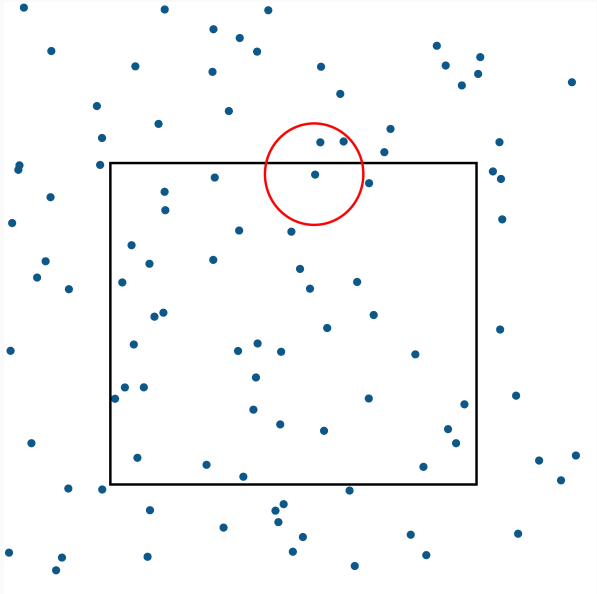
Because the average number of points expected in any radius r is a function of the intensity λ, we can derive a correction for $\hat{H}(r)$ that allows for comparisons between processes (derivations in section 7.3 of Baddeley *et al.* (2015)):

$$\hat{K}(r) = \frac{|W|}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} 1\{d_{ij} \leq r\} e_{ij}(r)$$

where $|W|$ is the observation window,
$e_{ij}(r)$ is an edge correction,
and $\hat{K}(r)$ is the estimated empirical K-function.

Edge corrections

Because our observations are restricted to a window, estimates become biased near the edges.



Two common edge corrections include:

1. **Border correction:** Only use points $> r$ away from border. Computationally fast, but statistically inefficient (only recommended for large datasets).

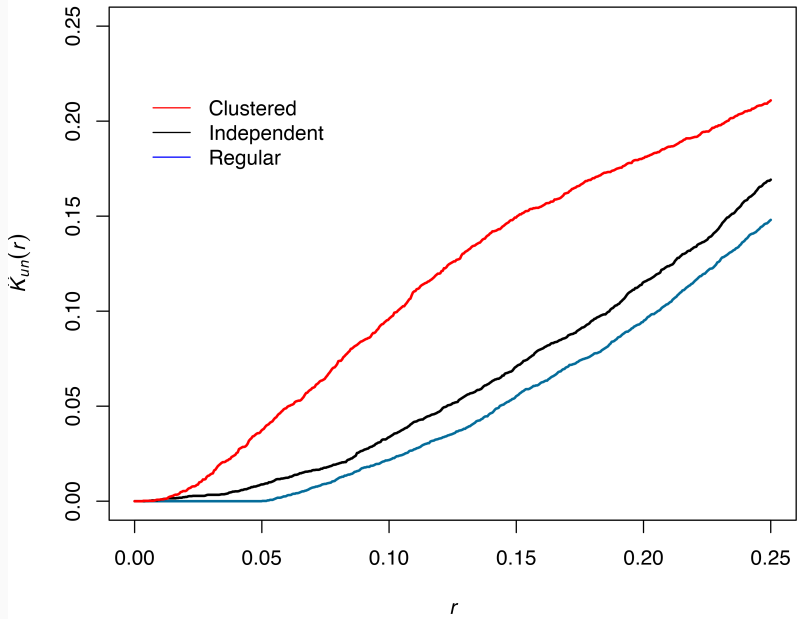
2. **Isotropic correction:** Corrects $\hat{K}(r)$ based on how much of the circle lies outside of the window. Computationally slower, and statistically efficient (but assumes isotropy).

Empirical K-function cont.

The K-function describes the cumulative average number of points falling within distance r of a typical point.

This value is corrected for edge effects.

For our three point processes, the empirical K-functions would look like this:



The patterns are clearly different, but what does this mean for our point processes?

Theoretical K-function

For a homogeneous Poisson point process it can be shown that the expected K-function is given simply by (derivations in section 7.3 of Baddeley *et al.* (2015)):

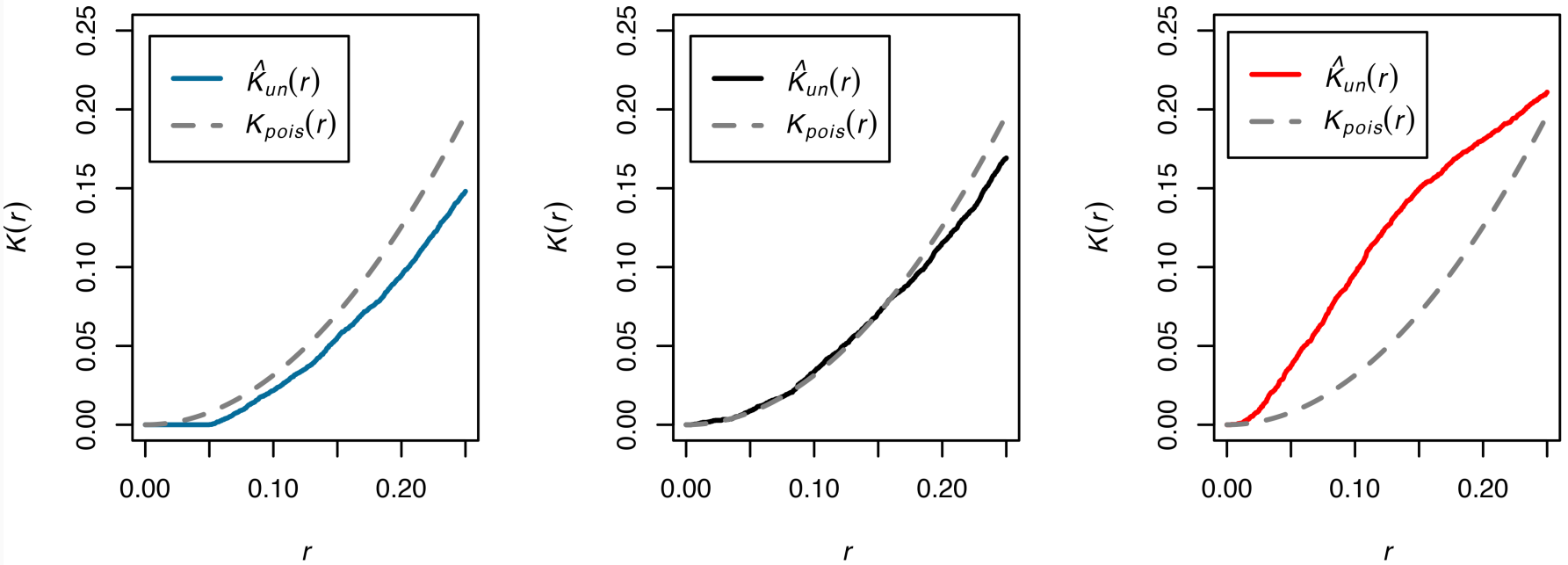
$$K(r) = \pi r^2$$

In other words, it is simply a function of the area of a circle with radius r.

Any deviations between the empirical and theoretical K-functions are an indication of correlations (+ive or -ive).

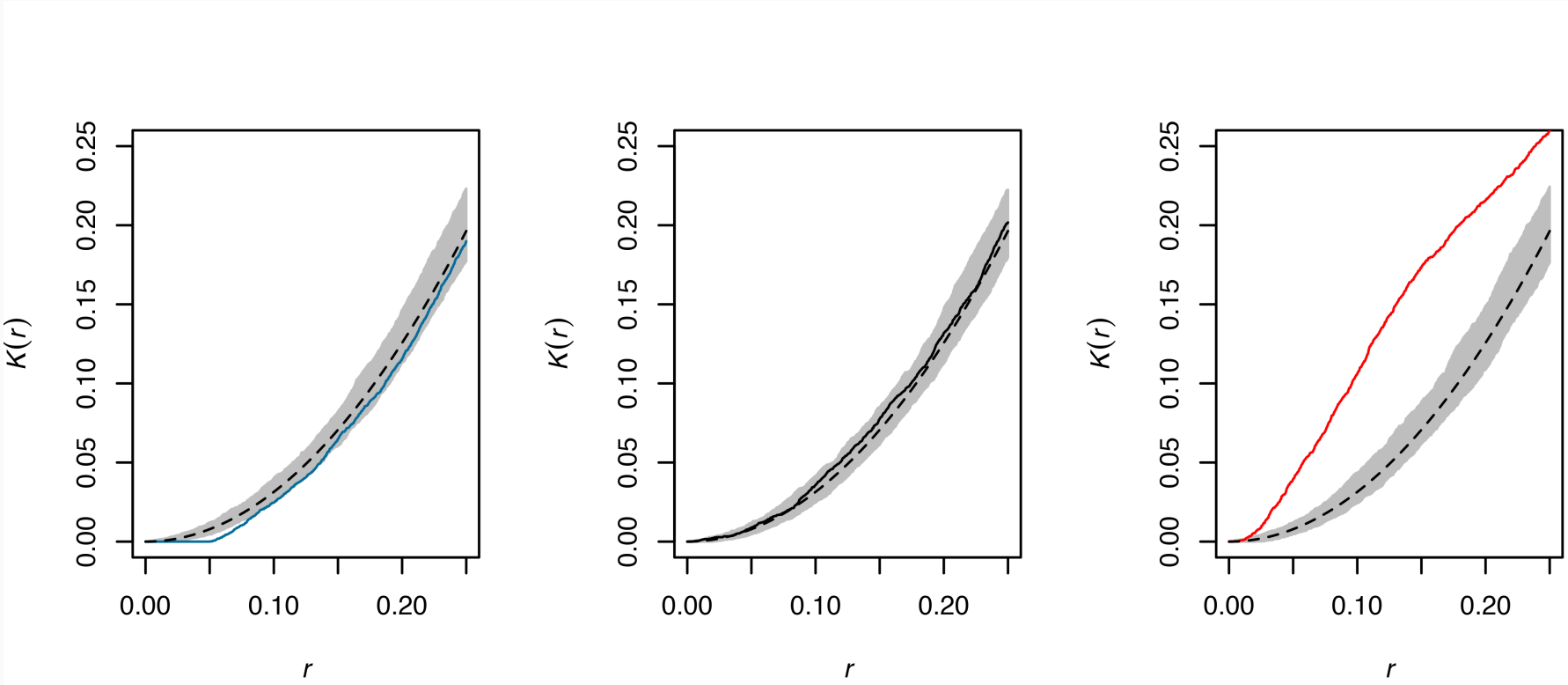
Theoretical K-function cont.

Any deviations between the empirical and theoretical K-functions are an indication of correlations... but what constitutes a meaningful deviation?



Theoretical K-function cont.

We can generate bootstrapped estimates of $\hat{K}(r)$ to obtain confidence intervals (details in section 7.8 of Baddeley *et al.* (2015)).



The K-function overview

Ripley's K-function is a widely used and well respected metric for understanding patterns in spatial correlations.

It describes correlations, not causations (provides no insight on why a perceived correlation may exist).

Tests rely on assumptions of homogeneity, stationarity, and that the process is Poisson (breaking those can produce spurious findings).

Plots of $\hat{K}(r)$ vs r do not provide information on the spatial scale of interactions (often incorrectly used to this end).

Pair Correlation Function

Motivation

Ripley's K-function provides information on whether their are significant deviations from independence between points, but provides limited information on the behaviour of the process (cumulative in nature so contains the contribution of all inter-point distances $\leq r$).

An alternative tool is the pair correlation function $g(r)$, which only contains contributions from inter-point distances $= r$

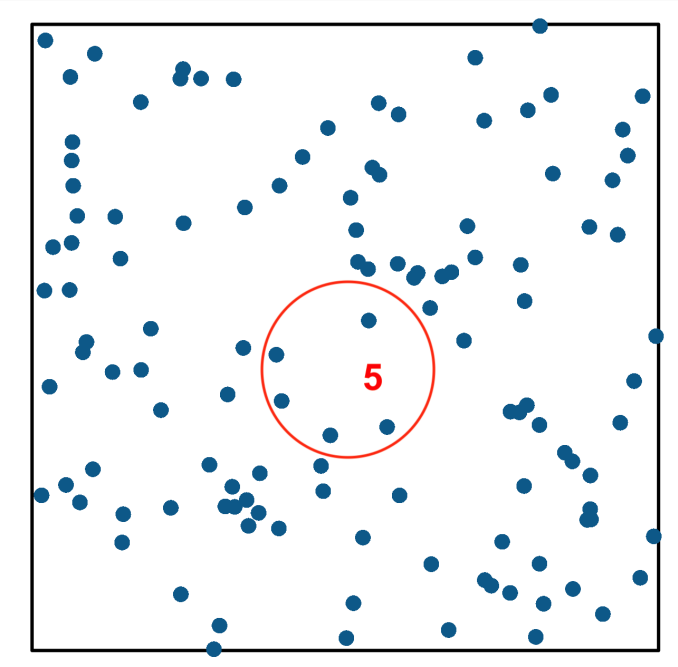
$$g(r) = \frac{K(r)\prime}{2\pi r}$$

i.e., the derivative of the K-function with respect to r.

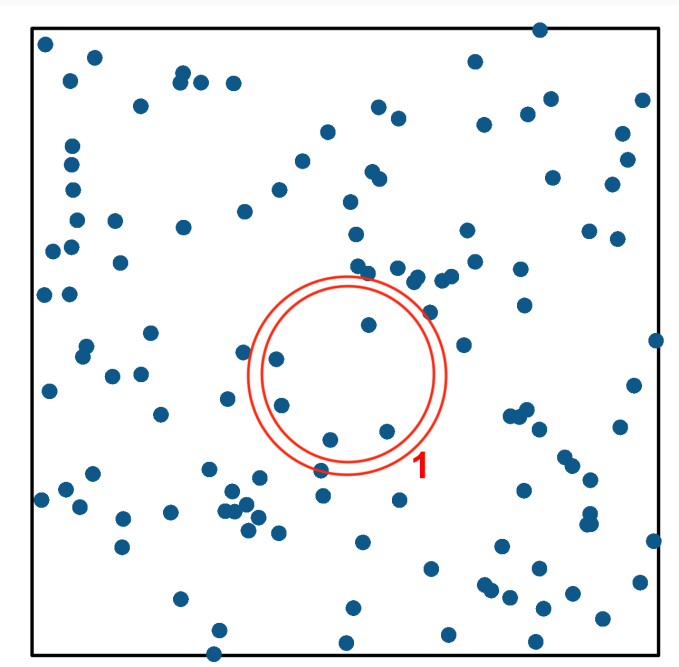
Analogous metrics have arisen independently in other fields of research (e.g., the radial distribution function from physics/chemistry).

Pair correlation function

$K(r)$ counts all points within a circle of radius r.



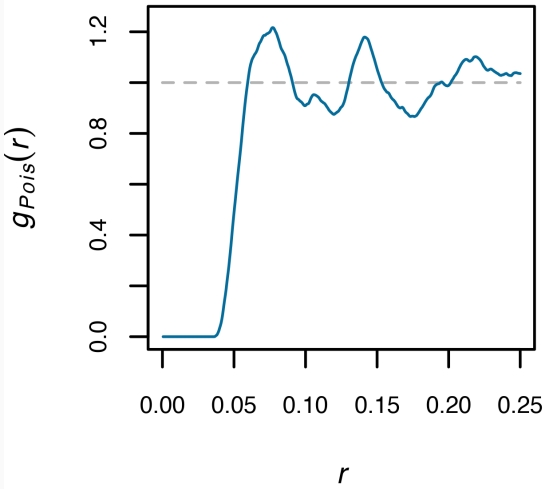
$g(r)$ counts all points within a ring of radii r & $r + h$.

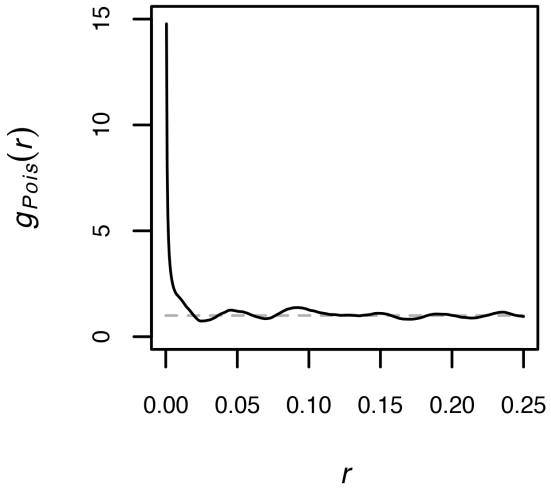


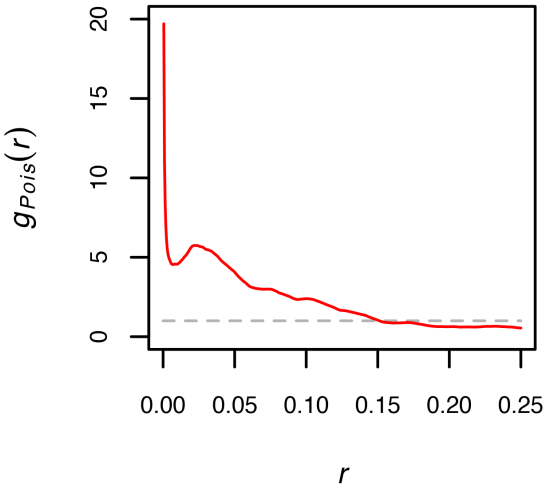
Under CSR, $g(r)$ has an expected value of 1, values < 1 indicate fewer points with separation distance r than expected (i.e., avoidance), and vice versa for $g(r) > 1$.

Empirical PCF

The estimated pair correlation functions for our three case processes would look like this:





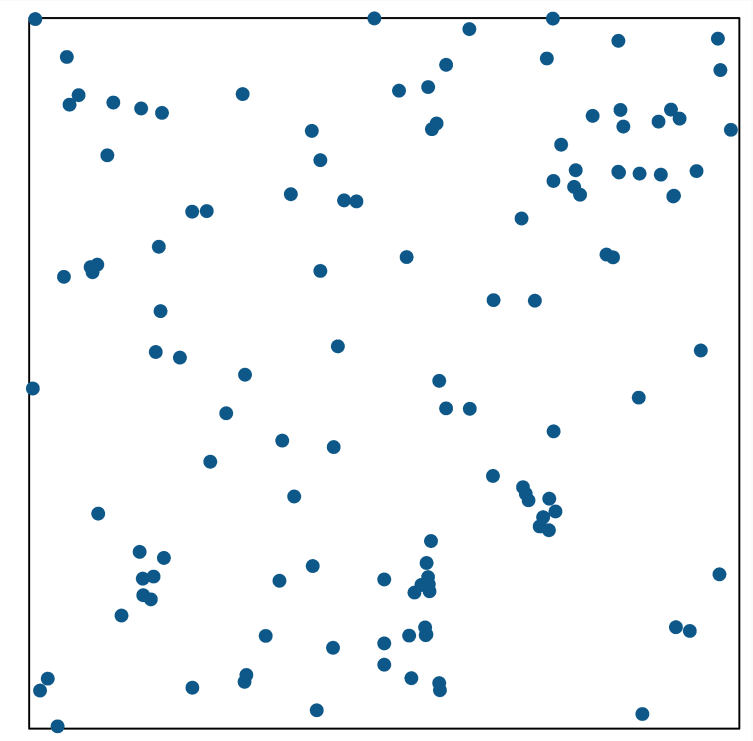


Second Moment Statistics in R

Finnish Pine Dataset

To demonstrate how to apply these tools to real data we will work with the Finnish Pine dataset again.

Do we think the trees are clustered? avoiding each other? independent?

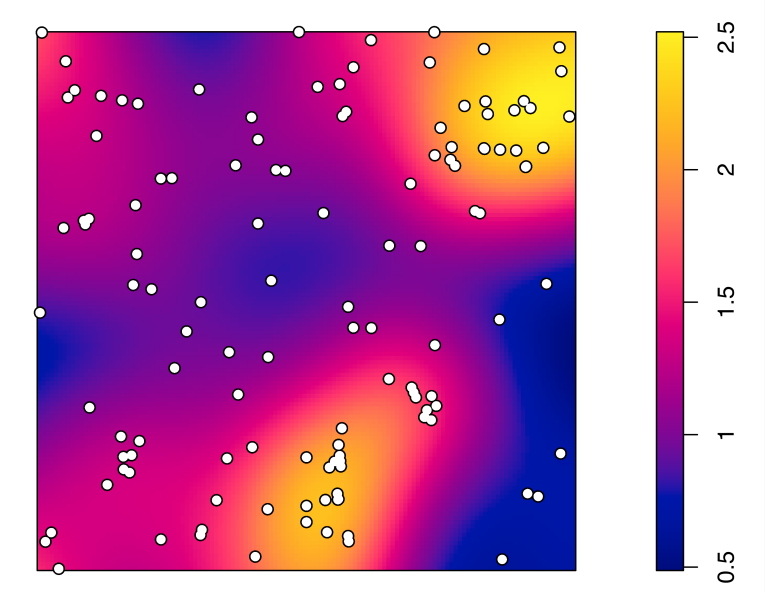


Source: `spatstat` package

Start with the first moment

All analyses should start with the first moment (can't estimate the second moment well if the first isn't understood).

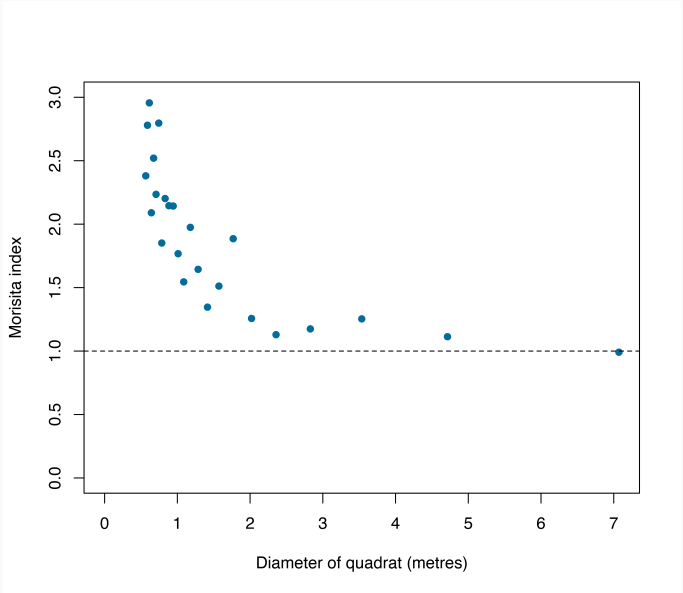
```
#Load in the data
data("finpines")

#Estimate the intensity
lambda_hat <- density(finpines)

#Visualise the first moment
plot(lambda_hat)
points(finpines)
```



Morisita's index

Next we can apply Morisita's index using the `miplot()` function.

```
miplot(finpines,
       ylim = c(0,3),
       main = "",
       pch = 16,
       col = "#046C9A")
```

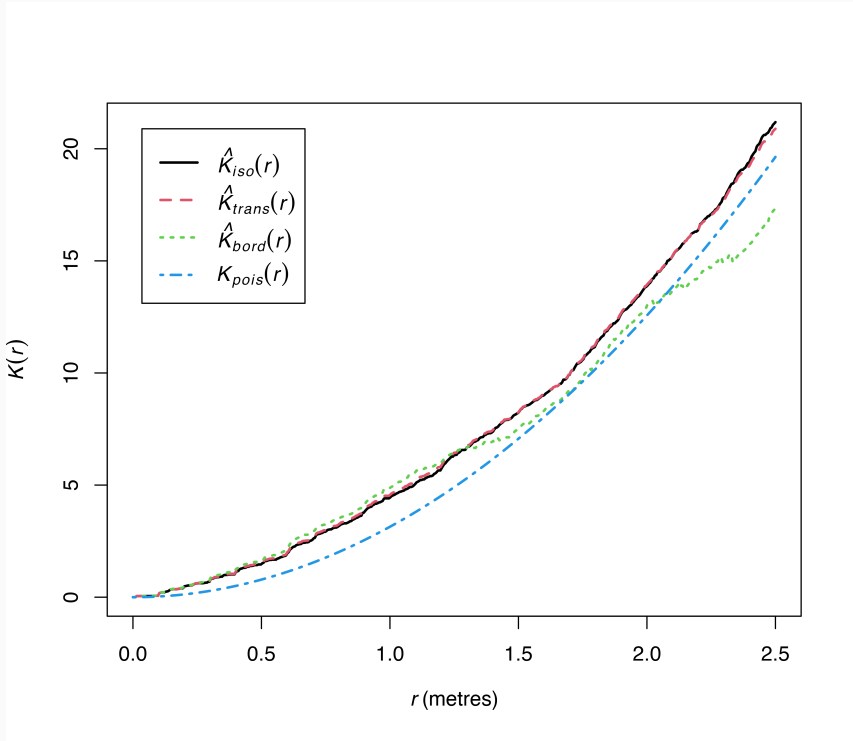


This suggests spatial clustering, but the data appear inhomogeneous, so this might not be trustworthy.

K-function

The K-function can be estimated using the `Kest()` function.

```
#Estimate the k-function
k_finpines <- Kest(finpines)

#visualise the results
plot(k_finpines,
     main = "",
     lwd = 2)
```

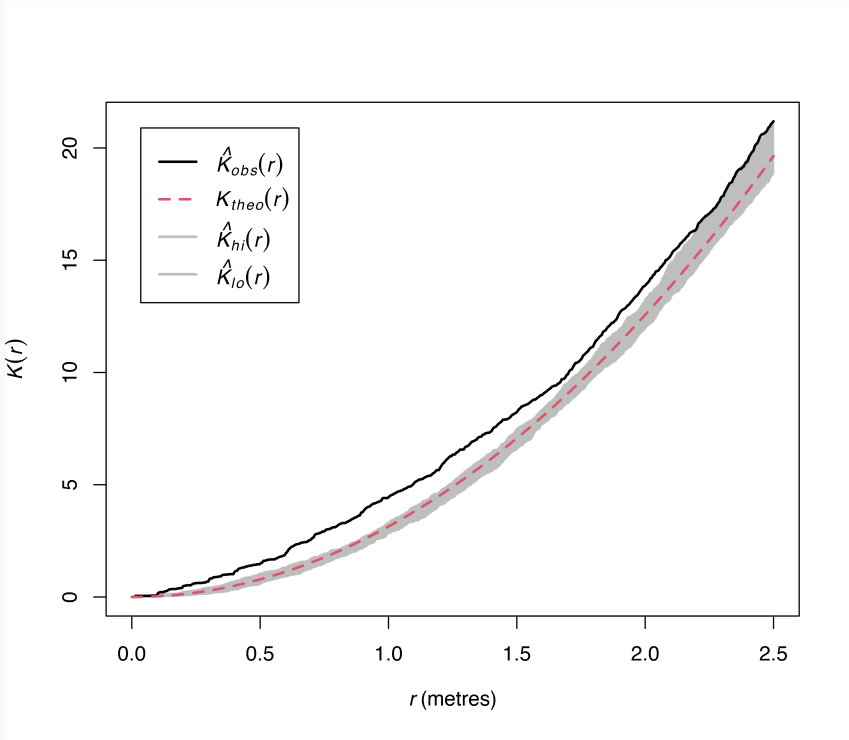


Here again the results suggest clustering... but some confidence intervals would help...

K-function cont.

Bootstrapped confidence intervals for the theoretical K-function can be estimated using the `envelope()` function.

```
# Bootstrapped CIs
# rank = 1 means the max and min
# values will be used for CI
E_finpines <- envelope(finpines,
                       Kest,
                       rank = 1,
                       nsim = 19,
                       fix.n = T)


# visualise the results
plot(E_finpines,
     main = "")
```

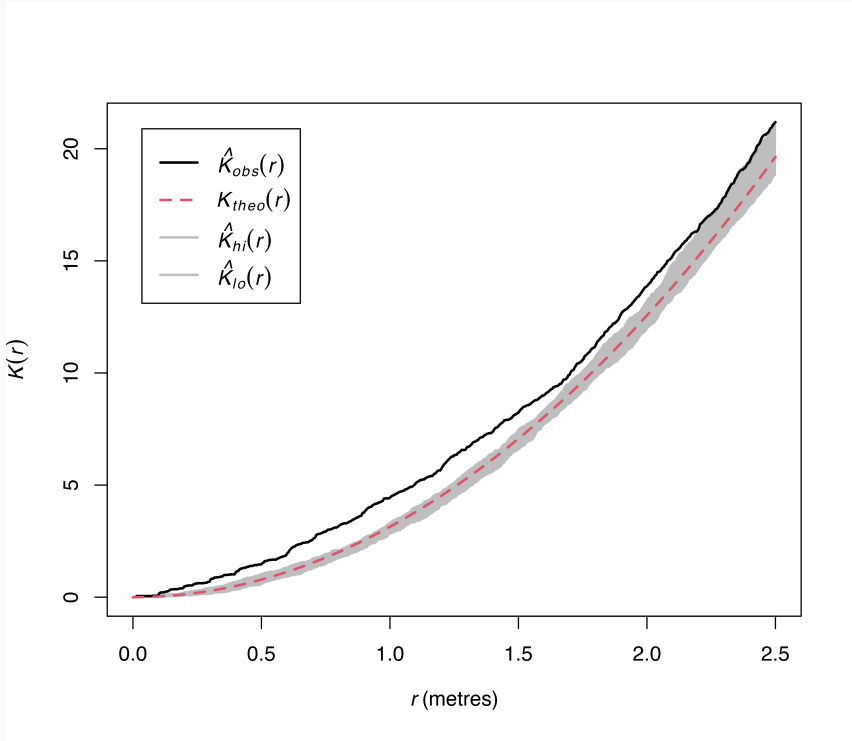


Now we have evidence that suggests *significant* clustering, but these estimates assume homogeneity.

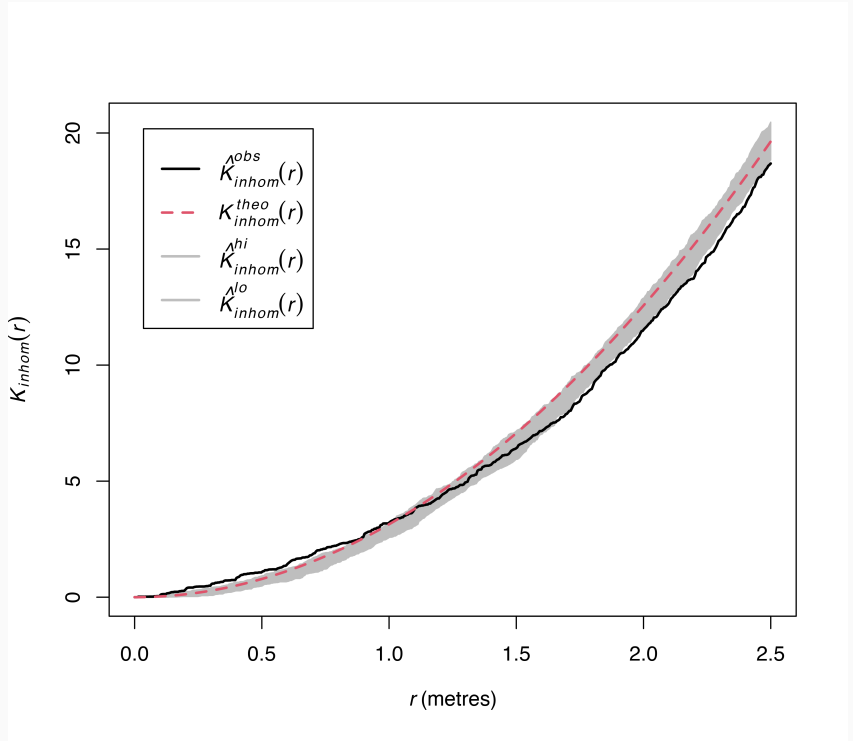
Inhomogeneous K-function

We can correct for inhomogeneity by weighting the data based on $\lambda(u)$ (the `Kinhom()` function).

Homogeneous K-function



Inhomogeneous K-function

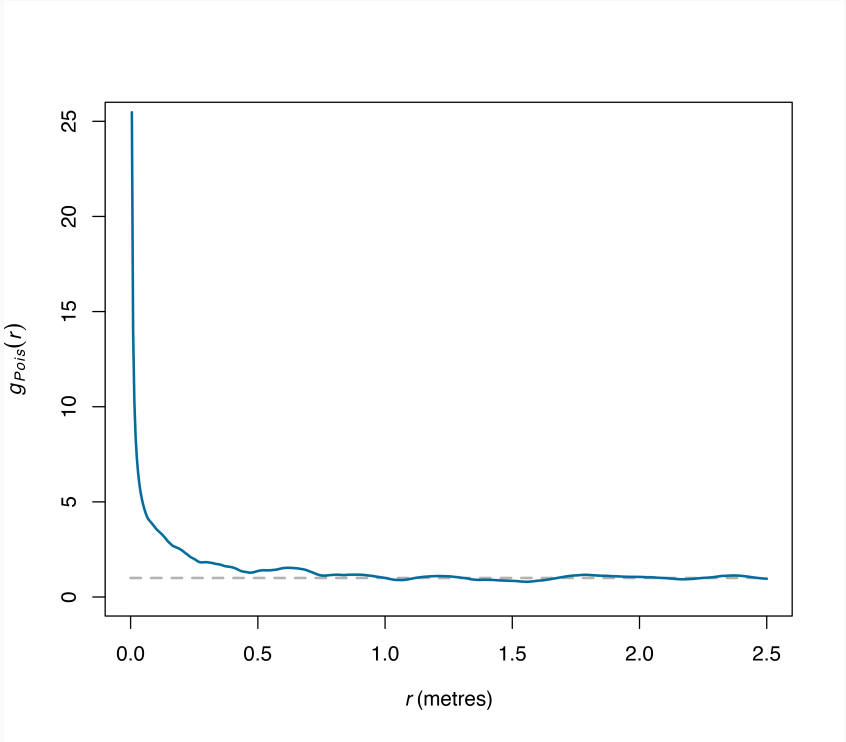


When correcting for inhomogeneity, the clustering is not as strong...

g-function

The g-function can be estimated using the `pcf()` function.

```
# Estimate the g function
pcf_finpines <- pcf(finpines)

# visualise the results
plot(pcf_finpines,
     theo ~ r,
     ylim = c(0,25),
     main = "",
     col = "grey70",
     lty = "dashed")

plot(pcf_finpines,
     iso ~ r,
     col = c("#046C9A"),
     add = T)
```

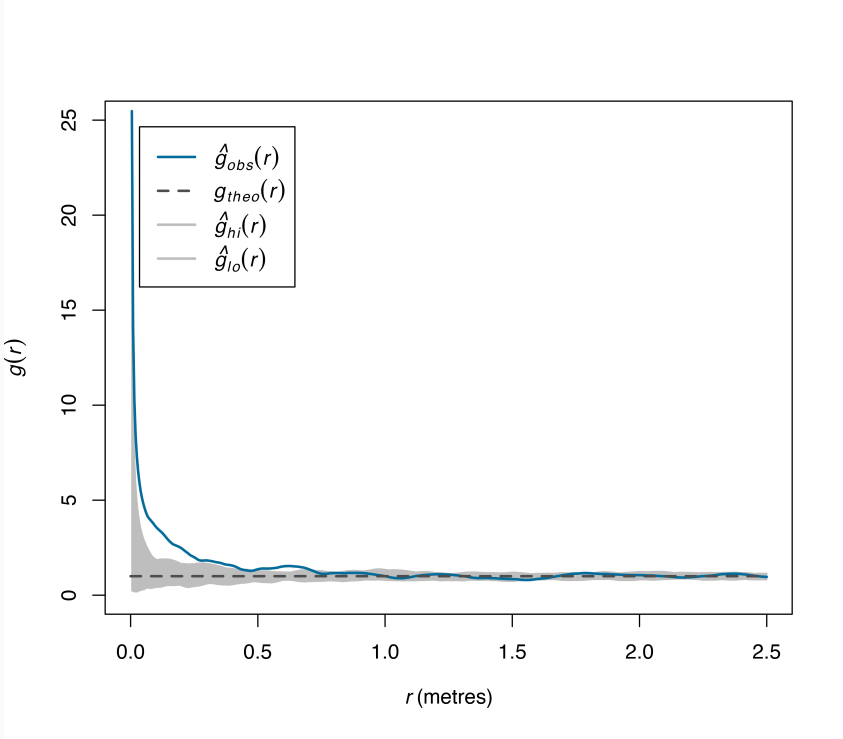


Seems like clustering occurs for ca. 0.75 meters.

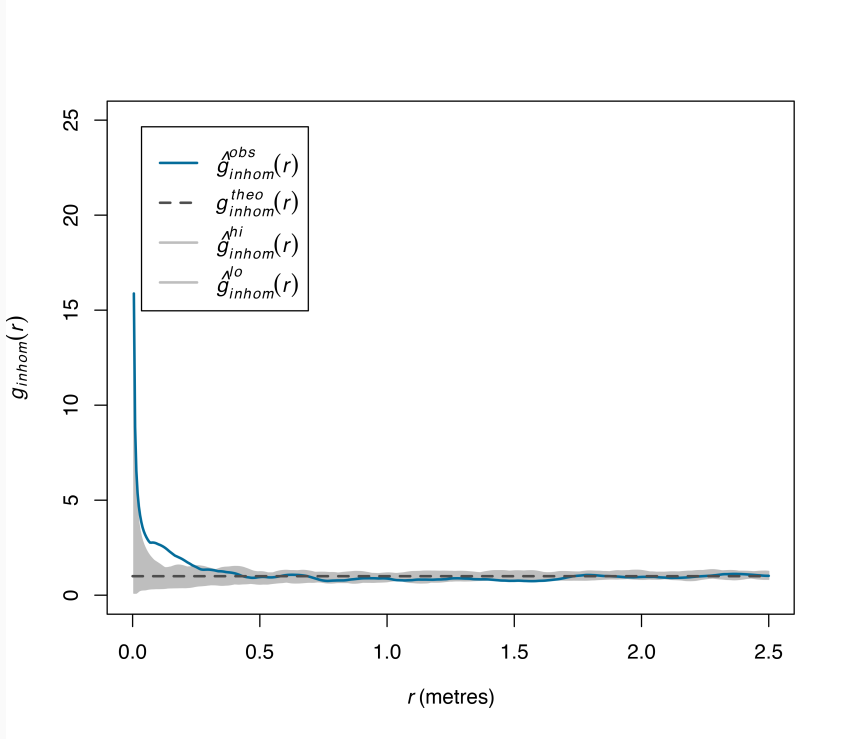
Inhomogeneous g-function

We can also correct for inhomogeneity by weighting the data based on $\lambda(u)$ (the `pcfinhom()` function).

Homogeneous g-function



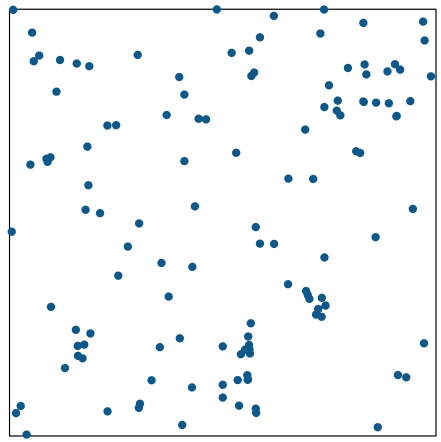
Inhomogeneous g-function

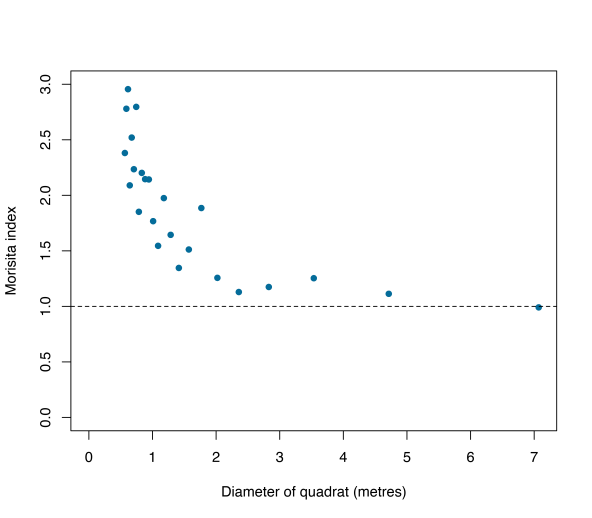


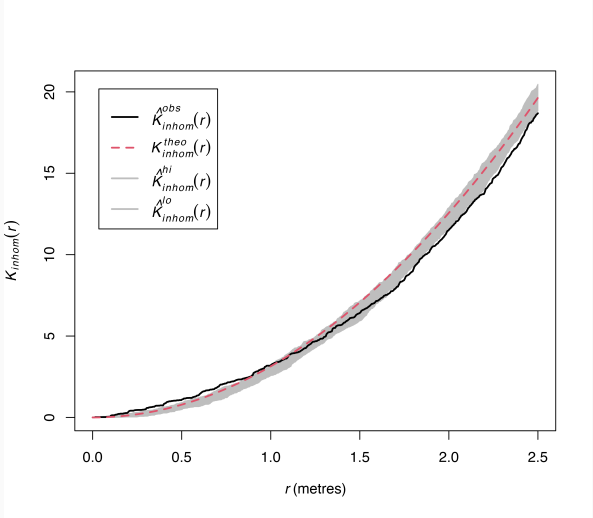
Again, when correcting for inhom., the clustering is not as strong...

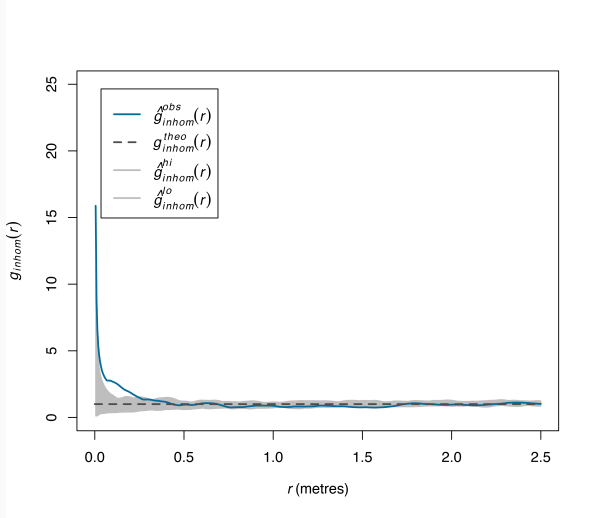
Conclusion

All lines of evidence point towards clustering and non-independence in Finnish pines.









Second moment considerations

The corrections for inhomogeneity all assume the intensity is unbiased.

There are many different corrections for edge effects and anisotropy, and different bootstrapping approaches (also all described in Baddeley *et al.* (2015))

Lots of additional information on the spacing between points can be obtained via metrics that we won't be covering (see chapter 8 of Baddeley *et al.* (2015) if you're interested).

Take home messages

The formal analysis of a point process usually begins with a descriptive analysis of the first moment.

Second moment descriptive statistics can help us further understand the factors at play in a point process (clustering, avoidance, etc...), but these metrics are correlative in nature and do not provide us with information on the underlying cause (e.g., we saw that Finnish pines were clustered, but we weren't able to say anything about why this was the case).

Descriptive statistics are a great place to start when analysing point data, but to fully understand our system we need to be able to model it, which we will cover next lecture.

References

Baddeley, A., Rubak, E. & Turner, R. (2015). *Spatial point patterns: methodology and applications with R*. CRC press.

Ni, J., Qian, T., Xi, C., Rui, Y. & Wang, J. (2016). Spatial distribution characteristics of healthcare facilities in nanjing: Network point pattern analysis and correlation analysis. *International journal of environmental research and public health*, 13, 833.