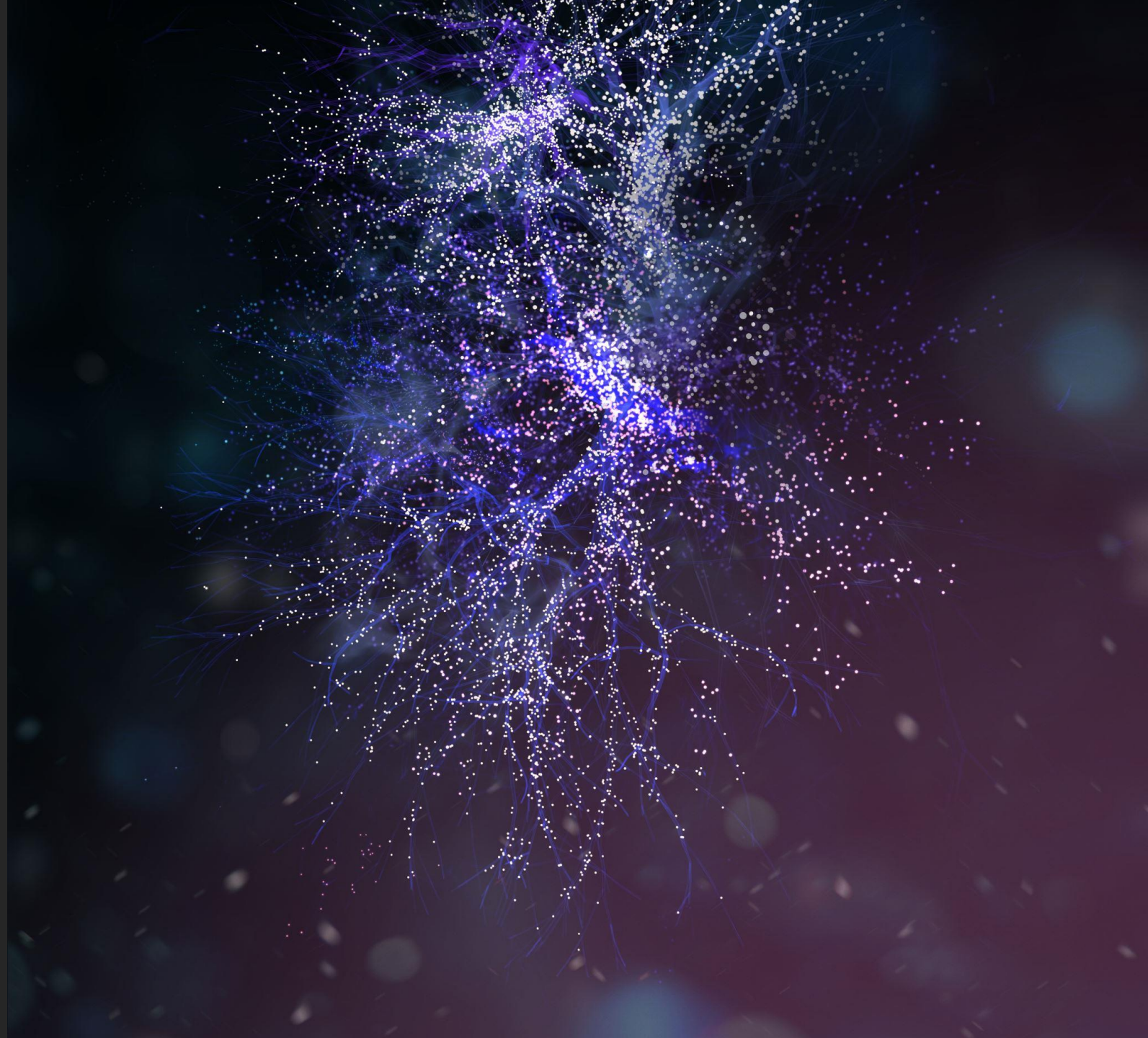
DSA 303 SPATIAL DATA ANALYSIS 7E

K-FUNCTION ANALYSIS



SECOND ORDER ANALYSES

- The basic underlying **assumption** behind second order analysis is, the marginal distributions of points have fixed intensity, but the joint distribution of all points is such that individual distribution of points are not independent! (Basically, saying identically distributed but not independent.)
  - Example: if a disease is contagious, the reporting of an incidence in one location is likely to be accompanied by other reports nearby.
- The **second order analysis** of point patterns, commonly known as *K*-function analysis, is the methodology advocated by Ripley's *K*-function and extended from thereon.

RIPLEY'S *K*-FUNCTION

- The Ripley's *K*-function is a method used to describe how point patterns occur over a given area of interest.
- It is **the expected number of extra events within distance *h*** from an arbitrary event and is *defined* by,

$$K(h) = \frac{E(N_h)}{\lambda}$$

where N_h is the number of events within a distance *h* of a randomly chosen event of all the recorded events and λ is the intensity of the process measured in events per unit area.

K-FUNCTION AND CSR

- If the process is an HPP (or CSR), with intensity λ, $E(N_h)=\lambda\pi h^2$ (same as 5d, slide 4).
- That means for an HPP, the *K*-function simplifies to,

$$K_{HPP}(h) = \pi h^2$$

- $K_{HPP}(h)$ can be used as a benchmark to assess clustering of either processes. If,
  - $K(h) > K_{HPP}(h)$: there is an excess of nearby points, i.e., clustering at the *spatial scale* associated with the distance h .
  - $K(h) < K_{HPP}(h)$: there is spatial dispersion at the spatial scale associated with the distance h . The presence of one point suggest other points are less likely to appear nearby than for an HPP.

THE EMPIRICAL *K*-FUNCTION

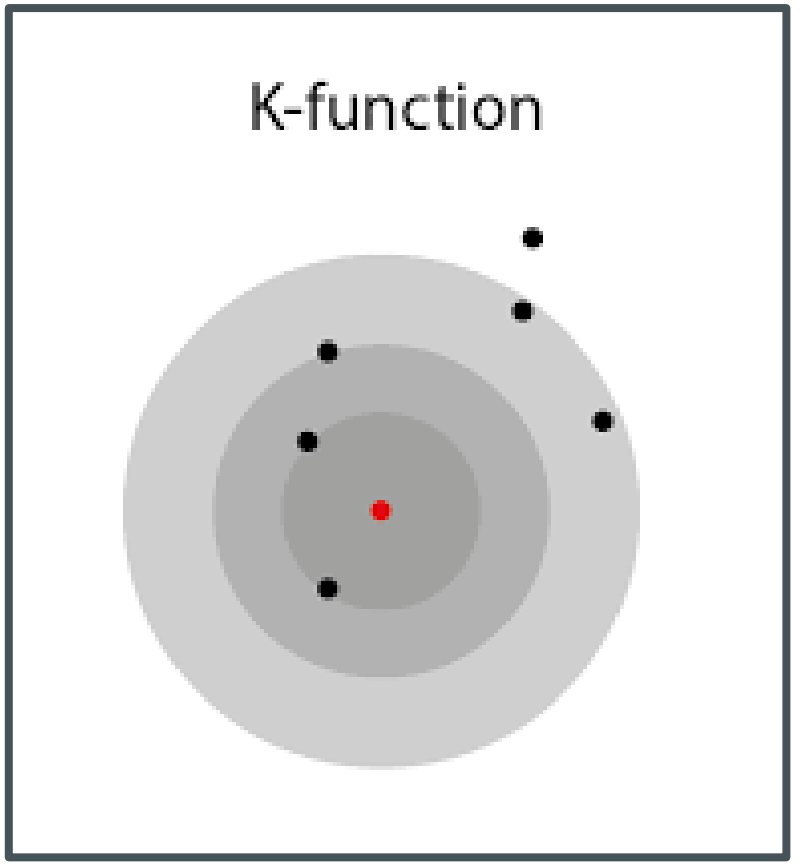
- When working with a sample of data points, the *K*-function for the underlying distribution is usually unknown. Thus, an estimate must be made using the sample.
- A natural estimator of the intensity in region *A* is,

$$\hat{\lambda} = \frac{n}{A}$$

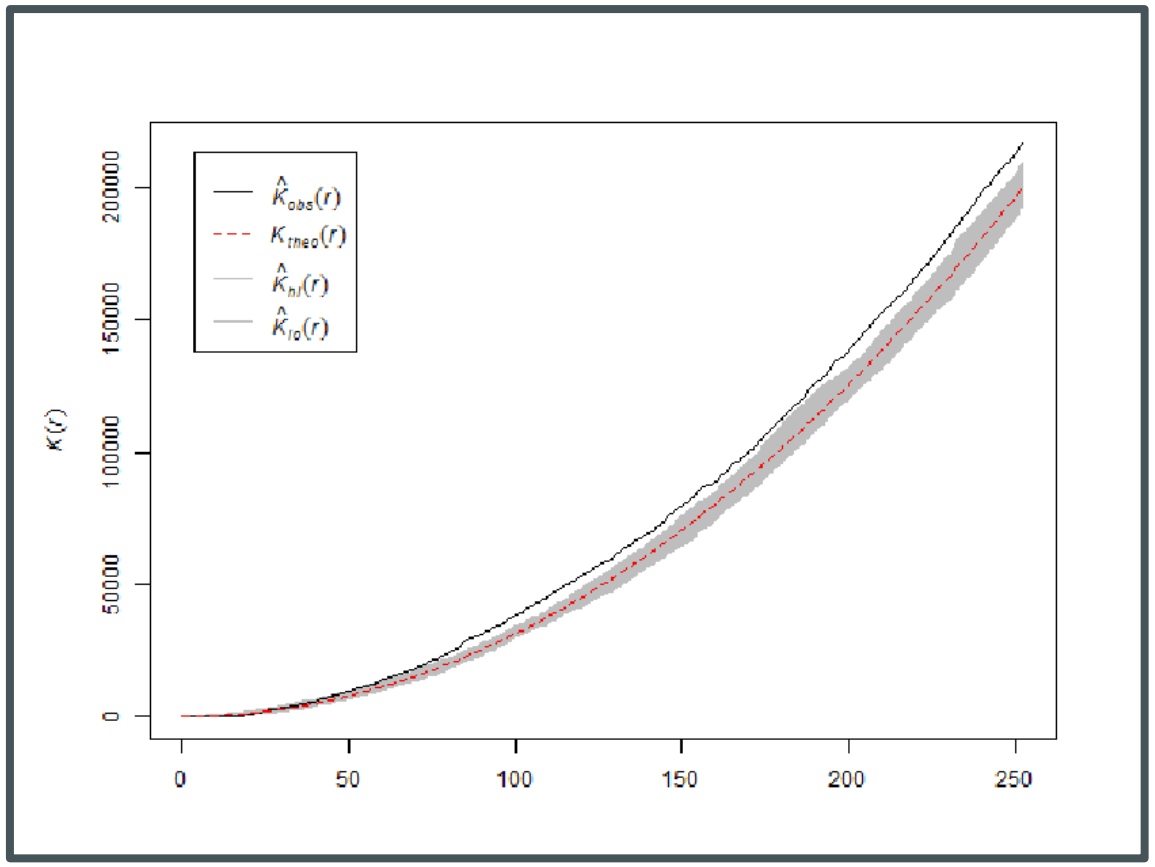
- A naïve moment estimator of $E(N_h)$ is,

$$\hat{E}(N_h) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} I(h_{ij} \leq h)$$

where, $I(h_{ij} \leq h)$ is an indicator function that equals 1 if the distance from a given observation is less than or equal to *h*. The inner sum yields the number of observed extra events within distance *h*, and the outer sum accumulates these counts from one observation to another. The estimate counts the number of neighbouring points found within a given distance *(h)* of each individual point (called inter-event distances).



THE EMPIRICAL *K*-FUNCTION



- Collecting these two, an estimate for the *K*-function is,

$$\widehat{K}(h) = \frac{\widehat{E}(N_h)}{\hat{\lambda}}$$

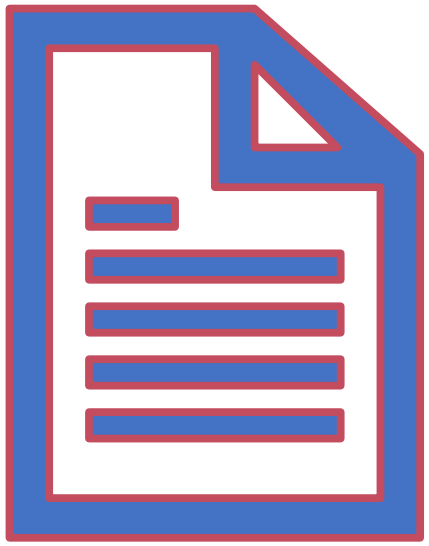
- $\widehat{K}(h)$ is normally plotted against *h* to reveal if any clustering occurs at certain distances.

THE EMPIRICAL *K*-FUNCTION WITH (RIPLEY) EDGE CORRECTION

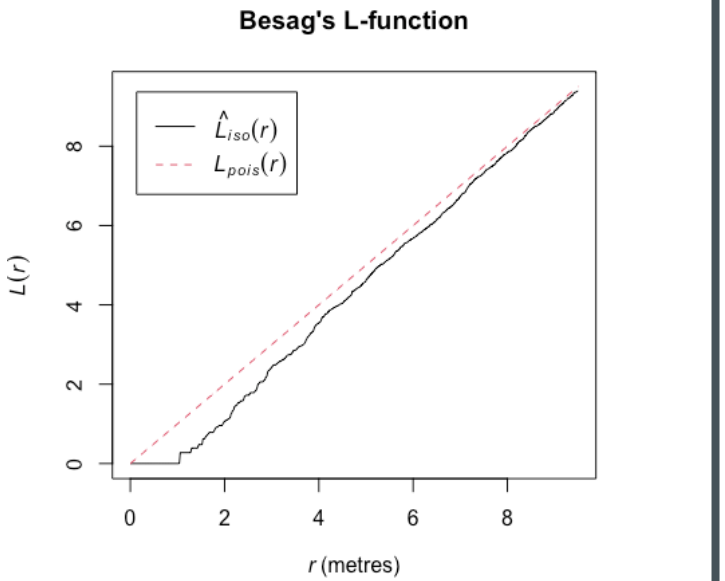
- This original naïve moment estimator is negatively biased because the events outside of the study region are unobserved.
- The theoretical form of *K(h)* assumes points occurring in an infinite plane, however, a real-world sample is taken from a finite study area *A*.
- This mainly affects the events near the boundaries as the counts will be unnaturally low.
- To adjust for this we apply weights, w_{ij} to each pair of observations that correspond to the proportion of circumference of a circle that is within the study region.
- The edge corrected estimate for the *K*-function is given by,

$$\hat{E}(N_h) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{I(h_{ij} \leq h)}{w_{ij}}$$

and $\hat{K}(h) = \frac{\hat{E}(N_h)}{\hat{\lambda}}$



BESAG'S *L*-FUNCTION



- The *L*-function is a transformation of Ripley's *K*-function,

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

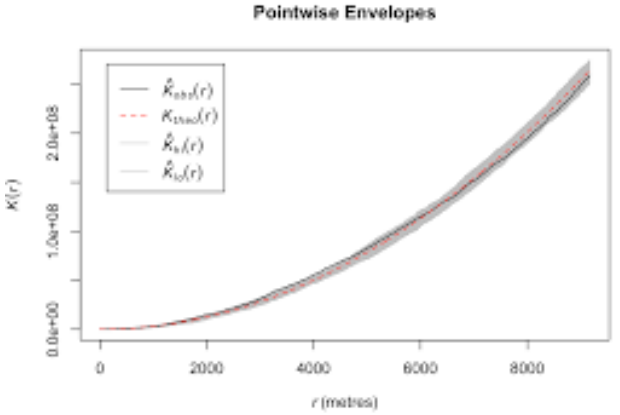
- Under CSR, $L(h) = h$ (prove this!) which has the appealing property of being a "linear function" when it comes to making comparisons.
- Thus, making visual comparisons in a plot of estimated *L*-functions is much easier and the computation is faster.

K-FUNCTION ANALYSIS: REVISITED

- The basis for conducting a *K*-function (or *L*-function) analysis is by comparing the expected and the observed distributions.

SIMULATION ENVELOPES

- An envelope is the highest and the lowest value (usually a 95% interval) for the distribution of *K(h)* [or *L(h)*], for a given value of *h* simulated *around* the $K_{HPP}(h)$ [or $L_{HPP}(h)$].
- These bands are derived using simulations based on randomly sampled estimates under CSR using the expected *K(h)* [or *L(h)*].
- For any distance *h*, if the observed *K*-function is less than or greater than the expected *K*-function, the null hypothesis of CSR is rejected.
  - observed distribution greater than the upper limit of the envelope indicates clustering.
  - observed distribution less than the lower limit of the envelope indicates dispersion.
  - Observed distribution being between the upper and the lower limits of the envelopes indicates CSR.



(MANUAL) STEPS

1. Determine/compare the observed and expected *K*. The observed *K* is obtained through the construction of a circle around each point event (*i*), counting the number of events (*j*), within radius (*h*) of the circle, and repeat the process for all events.
2. Determine the average number of events within successive distance bands. Find overall point density for the study area. The observed *K* is the ratio of numerator to the density of events. This can then be compared to the expected *K*, which is a random pattern, $K(h) = \pi h^2$.
3. Transform $K(h)$ estimates into a square root function to make it linear $L(h)$. This step is optional.
4. Determine the confidence envelope by estimating minimum and maximum values for $K(h)$ (or *L(h)*) from several simulations at 5% level of significance under a null hypothesis of CSR.
5. Plot $K(h)$ (or $L(h)$) estimates on a graph to reveal if any clustering occurs at certain distances.
6. Interpret the results.

PACKAGES YOU NEED

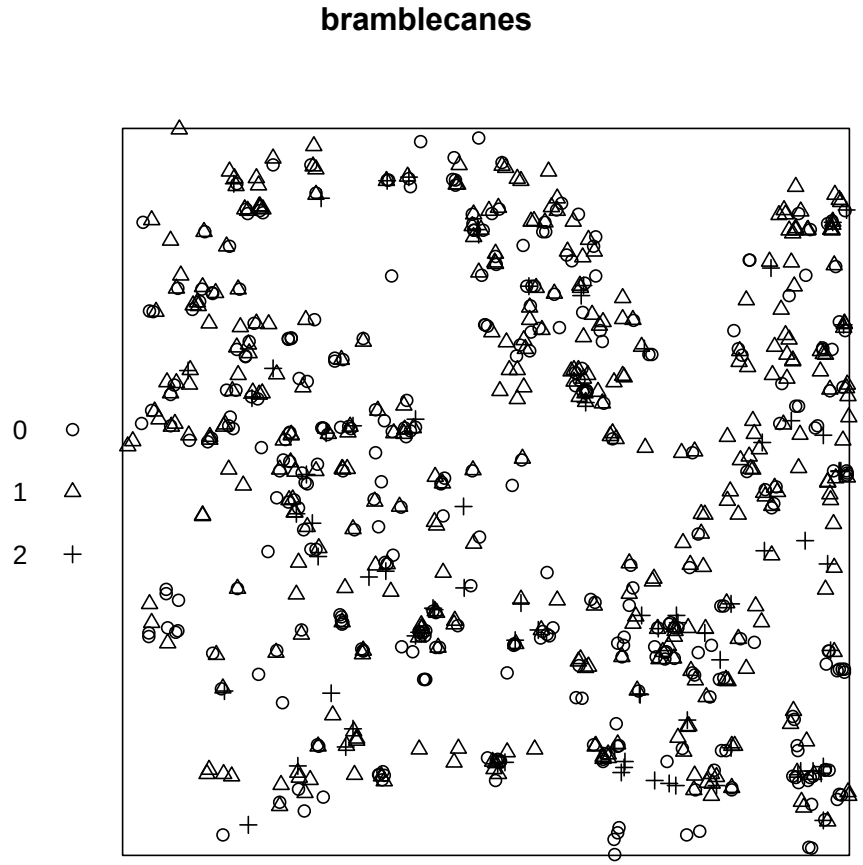
spatstat

DATASET

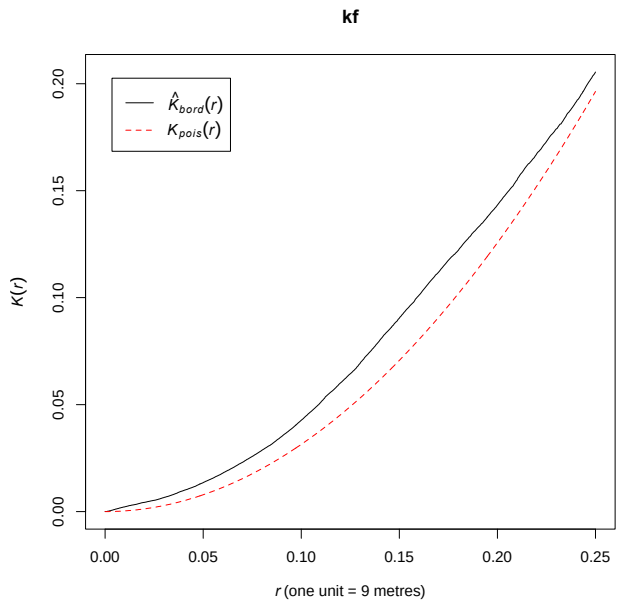
Locations of bramble canes

```
> data(bramblecanes)
> plot(bramblecanes)
```

Different symbols represent different ages (a marked point pattern!)



K-FUNCTION ESTIMATE



"Kest" estimates the *K*-function with an edge correction.

```
> kf = Kest(bramblecanes,
correction="border")
```

```
> plot(kf)
```

Check the construct of this function using ?Kest

An edge correction is needed to reduce bias. The border method or "reduced sample" estimator is the least efficient (statistically) and the fastest to compute. It can be computed for a window of arbitrary shape. Read about other types of edge corrections that are available.

The red dotted lone corresponds to $K_{HPP}(h) = \pi h^2$

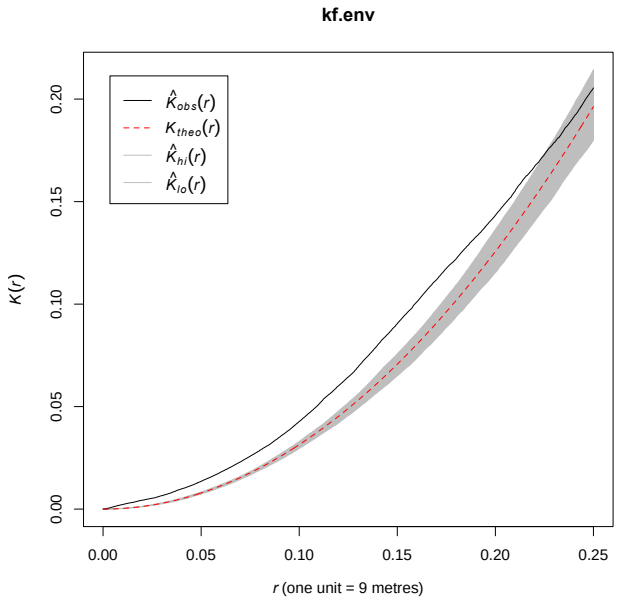
ENVELOPE ANALYSIS

Objective: envelope the highest and the lowest values of $\widehat{K}(d)$ around K_{pois}.

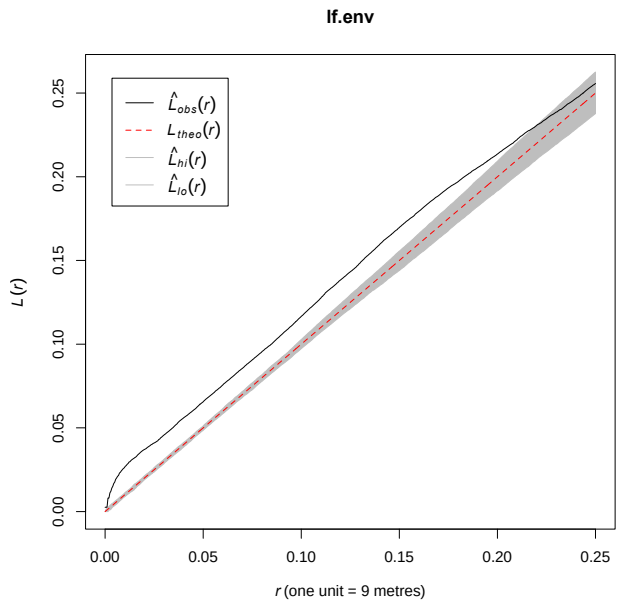
```
> kf.env = envelope(bramblecanes, Kest,
correction="border")
```

```
> plot(kf.env)
```

Observation: estimated *K*-function is higher than the envelope of simulated *K*-functions for CSR until *d* becomes quite large, suggesting strong evidence of clustering of locations of bramble canes.



L-FUNCTION ESTIMATE




```
> lf.env = envelope(bramblecanes, Lest,
correction="border")
```

```
> plot(lf.env)
```

The red dotted lone corresponds to $L_{HPP}(h) = h$

G-FUNCTION (FROM 5C)

- G-function is the cdf of the nearest neighbour (NN) distance for a randomly selected observation. A better alternative to analysing the NN distances (using CE aggregation index) is to examine the *cdf* of the NN distances.
- For a given distance *h*, *G(h)* is the probability that the NN distance for a randomly chosen sample point is less than or equal to *h*.
- The aim is to compare the observed (cumulative) distribution of the distances from an arbitrary point to the NN, against that of an HPP.
- Refer to 7d for details.

EXERCISE A

- Estimate the G-function and the related envelope for the "bramblecanes" dataset.
  - Hint: use Gest
- Try other types of border corrections applicable to Kest and Lest.
- Look for other types of functions. Example: ?Fest, ?Jest

K-FUNCTION ANALYSIS FOR NEWHAVEN DATASET

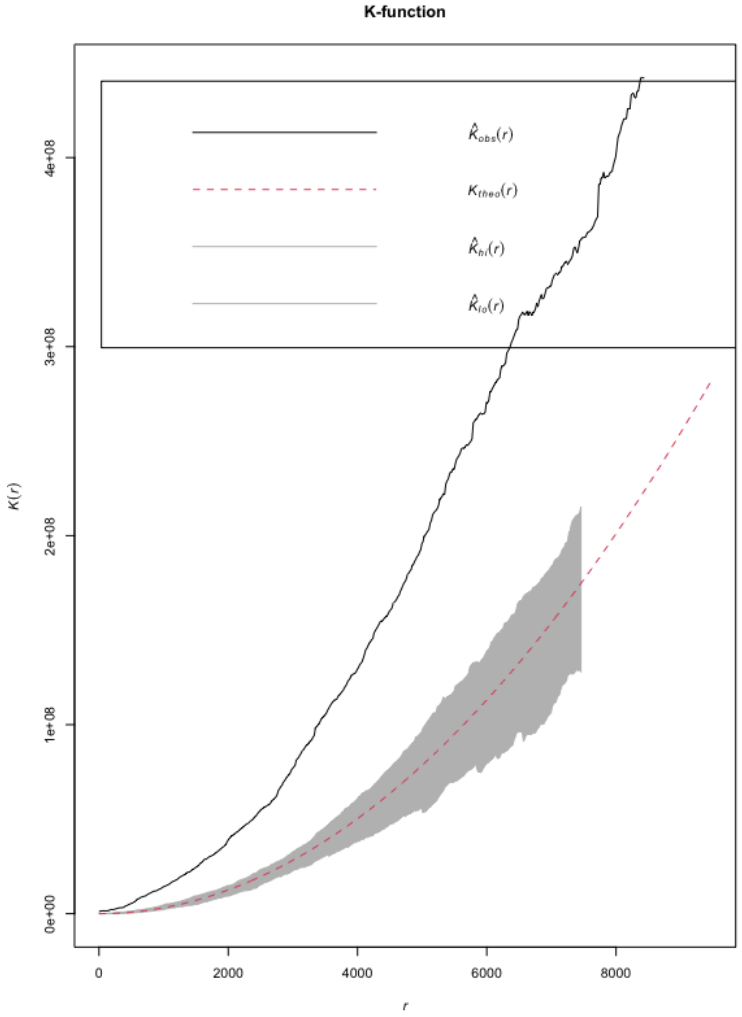
Envelopes:

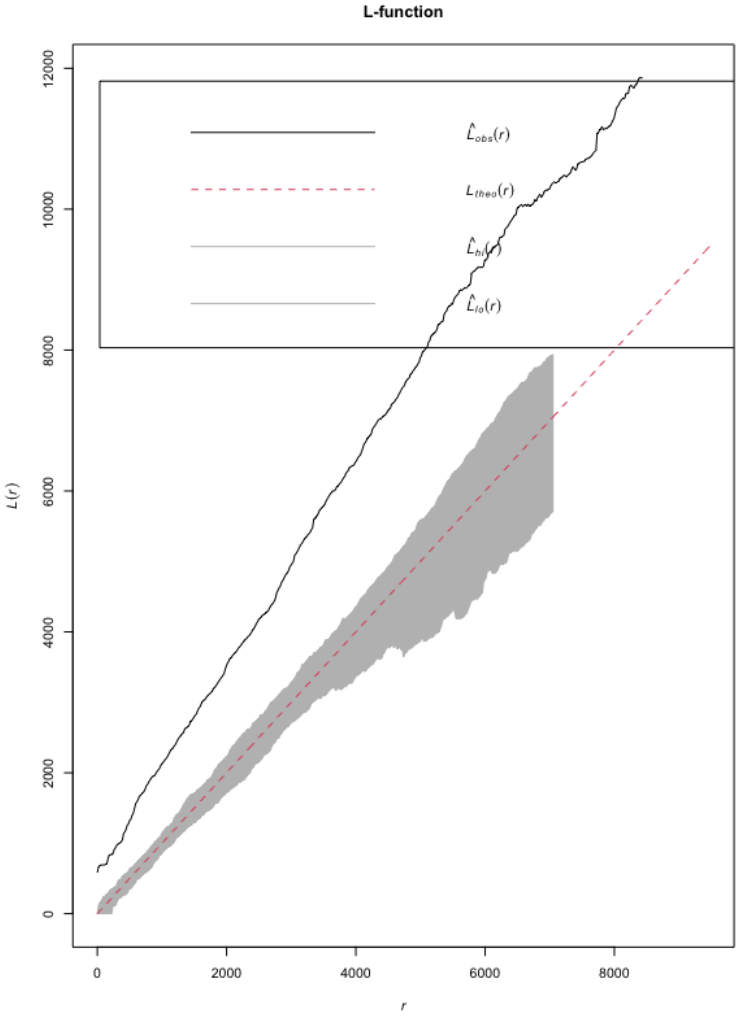
```
> kf_NH =
Kest(bramblecanes,correction="border");
plot(kf_NH)
```

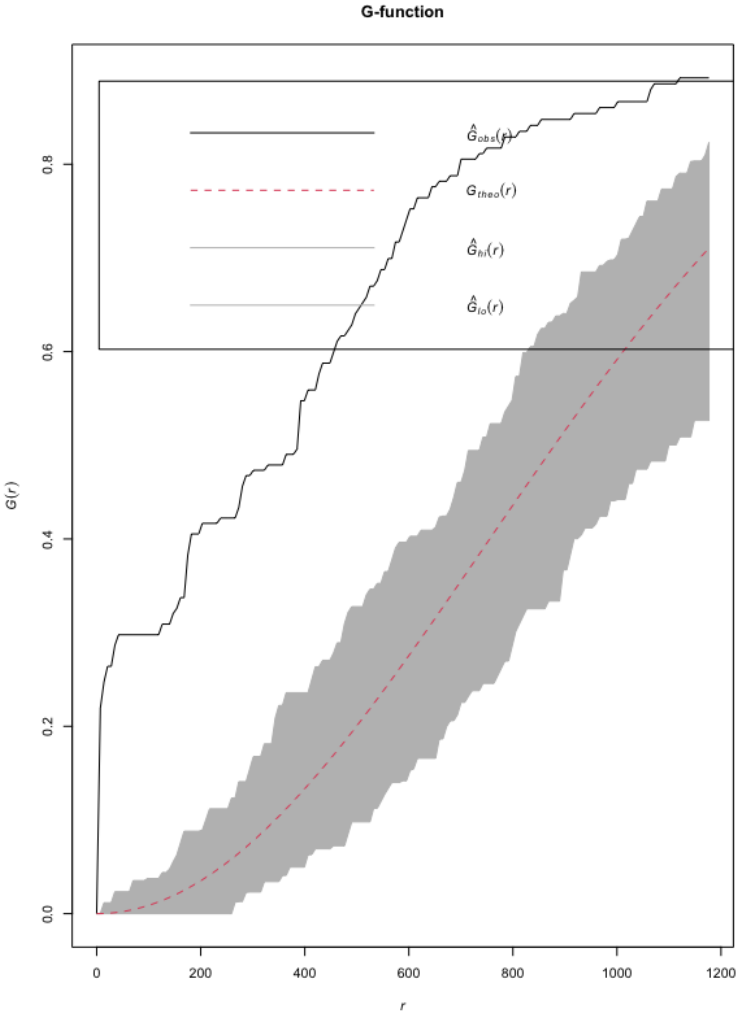
```
> kf.env_NH = envelope(breach.ppp, Kest,
correction="border"); plot(kf.env_NH)
```

```
> lf.env_NH = envelope(breach.ppp, Lest,
correction="border"); plot(lf.env_NH)
```

```
> gf.env_NH = envelope(breach.ppp, Gest,
correction="best"); plot(gf.env_NH)
```


K-function
$\hat{K}_{obs}(r)$
$K_{theo}(r)$
$\hat{K}_{hi}(r)$
$\hat{K}_{lo}(r)$
K(r)
0e+00
1e+08
2e+08
3e+08
4e+08
0
2000
4000
6000
8000
r


L-function
$\hat{L}_{obs}(r)$
$L_{theo}(r)$
$\hat{L}_{hi}(r)$
$\hat{L}_{lo}(r)$
L(r)
0
2000
4000
6000
8000
10000
12000
0
2000
4000
6000
8000
r


G-function
$\hat{G}_{obs}(r)$
$G_{theo}(r)$
$\hat{G}_{hi}(r)$
$\hat{G}_{lo}(r)$
G(r)
0.0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
1200
r

CONCLUSION

- Strong evidence of clustering in occurrences of breaches of peace in Newhaven.
- Plot commands:

```
> par(mfrow=c(1,3))
> plot(kf.env_NH, main="K-function")
> plot(lf.env_NH, main="L-function")
> plot(gf.env_NH, main="G-function")
> par(mfrow=c(1,1))
```

TAKE HOME POINTS

- The theoretical construct behind second order analyses of point patterns using K-, L-, G-functions.
- Estimation of K-, L-, G-functions.
- Edge corrections: the why and the how…

IMPORTANT R FUNCTIONS

- Kest()
- Lest()
- Gest()
- envelope()

REFERENCES

- ***Spatial Analysis*** by Tonny Oyana, 2nd edition, Chapter 6.
- ***Applied Spatial Data Analysis with R*** by Roger S. Bivand, Edzer Pebesma, and Virgilio Gómez-Rubio, 2nd edition, (2013), Chapter 7.
- https://bookdown.org/lexcomber/brunsdoncomber2e/Ch6.html
- https://cran.r-project.org/web/packages/spatstat/spatstat.pdf
- https://cran.r-project.org/web/packages/maptools/maptools.pdf
- Ripley, B.D. (1977) Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 172 – 212.
- Besag, J. (1977) Discussion of Dr Ripley's paper. *Journal of the Royal Statistical Society, Series B*, **39**, 193–195.