MODULE 9: Spatial Statistics in Epidemiology and Public Health
Lecture 7: Point Processes

Jon Wakefield and **Lance Waller**

References

- Baddeley, A., Rubak, E., and Turner. R. (2015) *Spatial Point Patterns: Methodology and Applications in R*. Boca Raton, FL: CRC/Chapman & Hall.
- Diggle, P.J. (1983) *Statistical Analysis of Spatial Point Patterns.* London: Academic Press.
- Diggle, P.J. (2013) *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Third Edition* CRC/Chapman & Hall.
- Waller and Gotway (2004, Chapter 5) *Applied Spatial Statistics for Public Health Data*. New York: Wiley.
- Møller, J. and Waagepetersen (2004) *Statistical Inference and Simulation for Spatial Point Processes*. Boca Raton, FL: CRC/Chapman & Hall.

Goals

- Describe basic types of spatial point patterns.
- Introduce mathematical models for random patterns (stochastic processes) of point-location events.
- Introduce analytic methods for describing patterns in observed collections of events.
- We model the *location* of each event as a random variable in space.
- *NOTE:* These probability models often motivate the model structures we use for disease mapping, spatial (count) regression, etc.

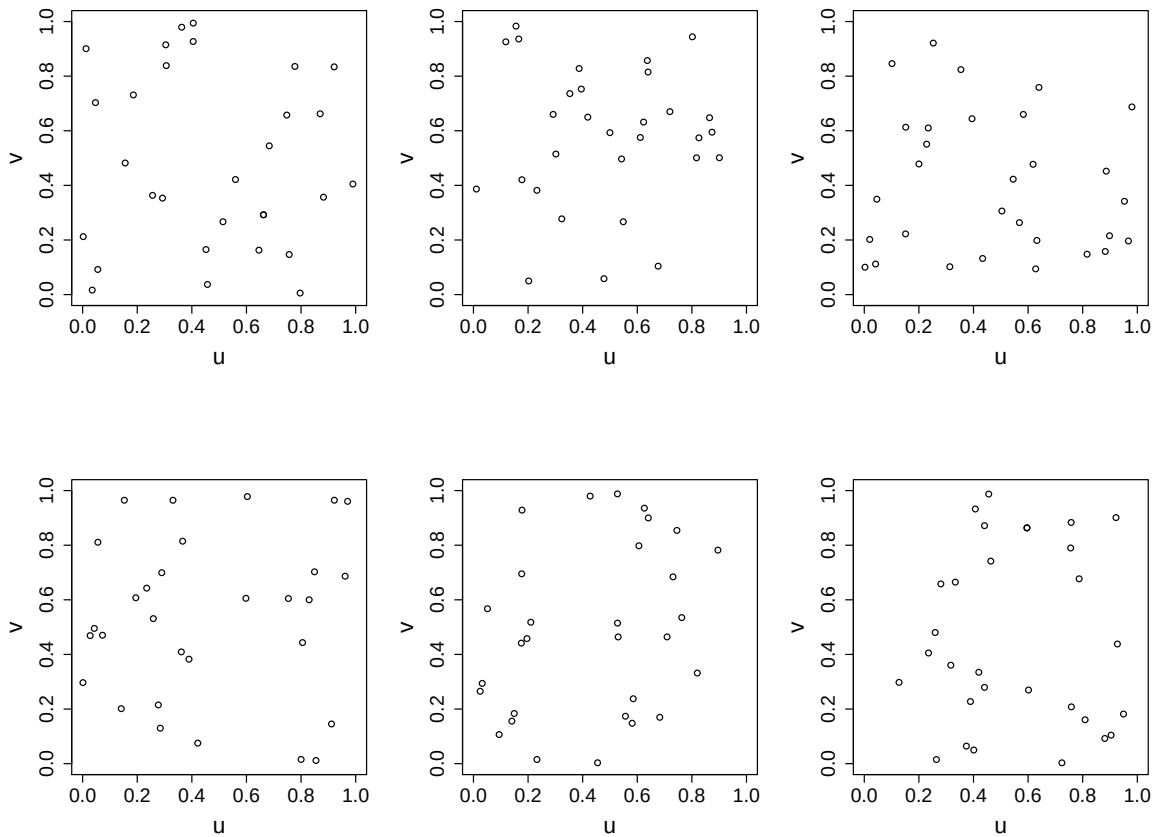
Terminology

- *Realization*: An observed set of event locations (a data set).
- *Point*: Where an event *could* occur.
- *Event*: Where an event *did* occur.

Complete Spatial Randomness (CSR)

- Start with a model of "lack of pattern".
- Events equally likely to occur anywhere in the study area (uniform distribution).
- Event locations independent of each other.

Six realizations of CSR

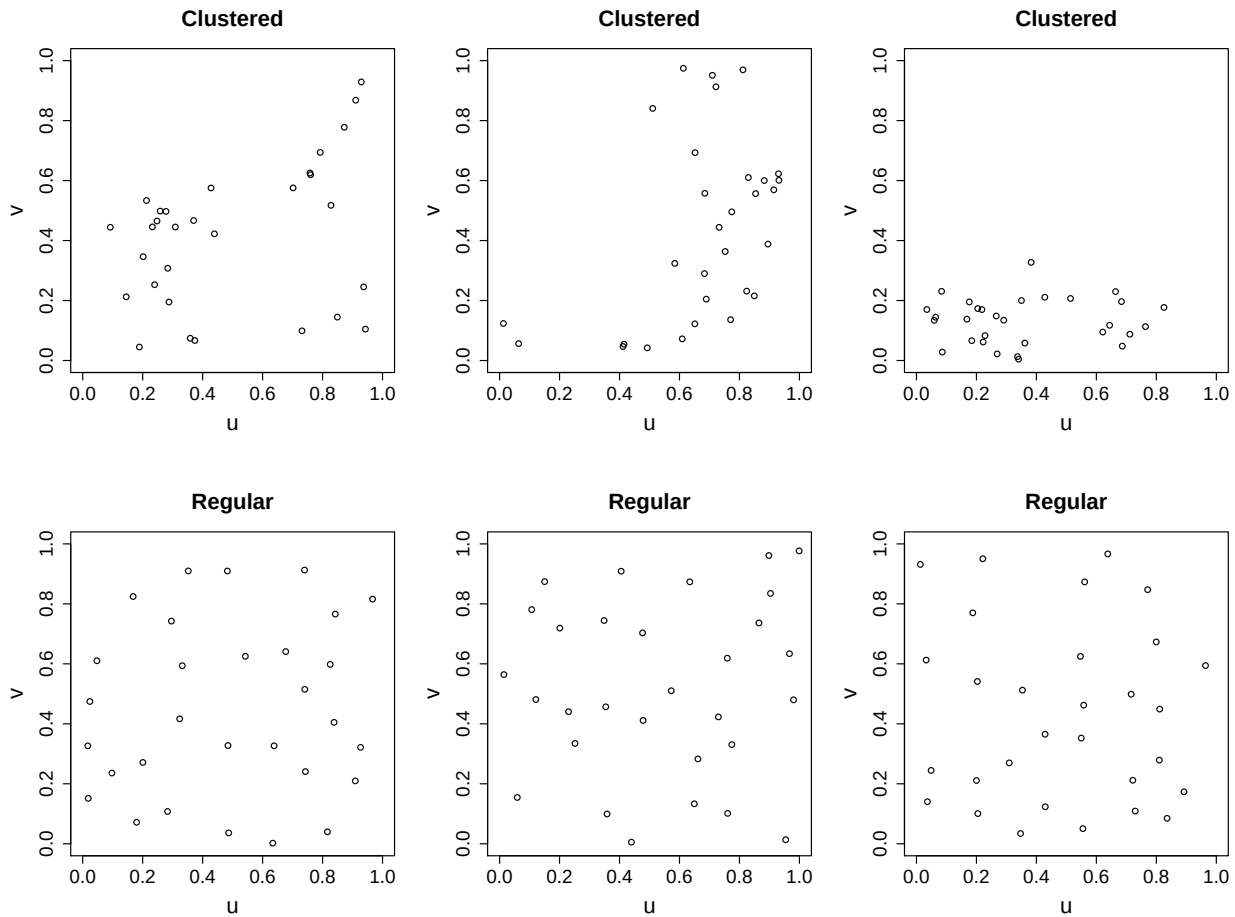


CSR as a boundary condition

CSR serves as a boundary between:

- Patterns that are more "clustered" than CSR.
- Patterns that are more "regular" than CSR.

Too Clustered (top), Too Regular (bottom)



Spatial Point Processes

- Mathematically, we treat our point patterns as realizations of a *spatial stochastic process*.
- A stochastic process is a collection of random variables $X_1, X_2, \ldots, X_N$.
- Examples: Number of people in line at grocery store.
- For us, each random variable represents an event location.

CSR as a Stochastic Process

Let $N(A)$ = number of events observed in region A, and λ = a positive constant.

A *homogenous spatial Poisson point process* is defined by:

(a) $N(A) \sim \text{Pois}(\lambda|A|)$

(b) given $N(A) = n$, the locations of the events are uniformly distributed over A.

λ is the *intensity* of the process (mean number of events expected per unit area).

Is this CSR?

- Criteria (a) and (b) give a "recipe" for simulating realizations of this process:
  * Generate a Poisson random number of events.
  * Distribute that many events uniformly across the study area.

```
runif(n,min(x),max(x))
runif(n,min(y),max(y))
```

Monte Carlo testing

Let $T =$ a random variable representing a test statistic (some numerical summary of the observed data).

What is the distribution of T under H_0?

1. t_1.
2. simulate $t_2, ..., t_m$ under H_0, these values will follow F_0.
3. p.value $= \frac{\text{rank of } t_1}{m}$.

M.C. tests are useful in spatial statistics since we can simulate spatial patterns and calculate the statistics.

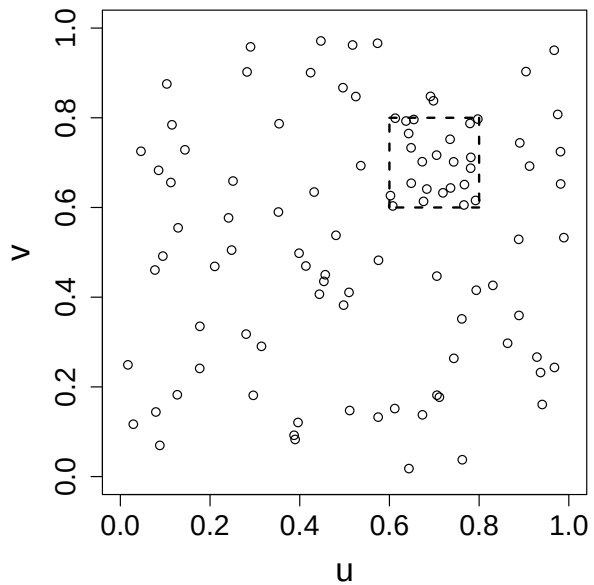
Example: e.g., 592 leukemia cases in $\sim$ 790 regions...

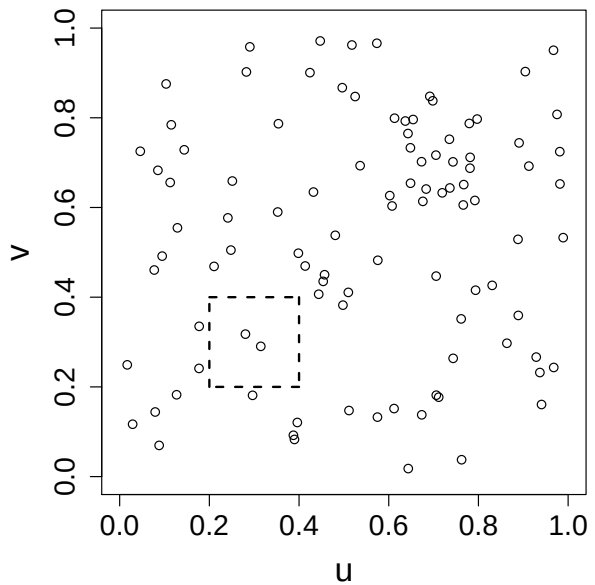
Moving beyond CSR

CSR:

1. is the "white noise" of spatial point processes.
2. characterizes the absence of structure (signal) in data.
3. often the null hypothesis in statistical tests to determine if there is clustering in an observed point pattern.
4. not as useful in public health? Why not?

Heterogeneous population density





Heterogeneous Poisson Process

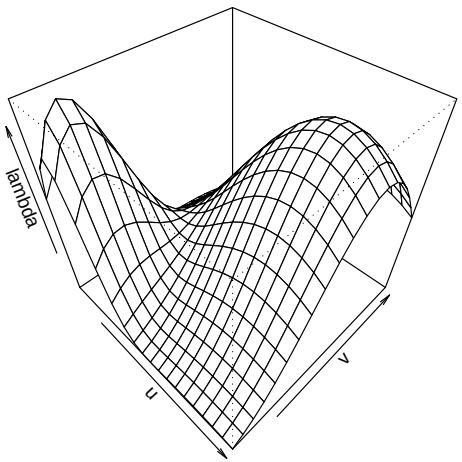
What if λ, the *intensity* of the process (mean number of events expected per unit area), varies by location?

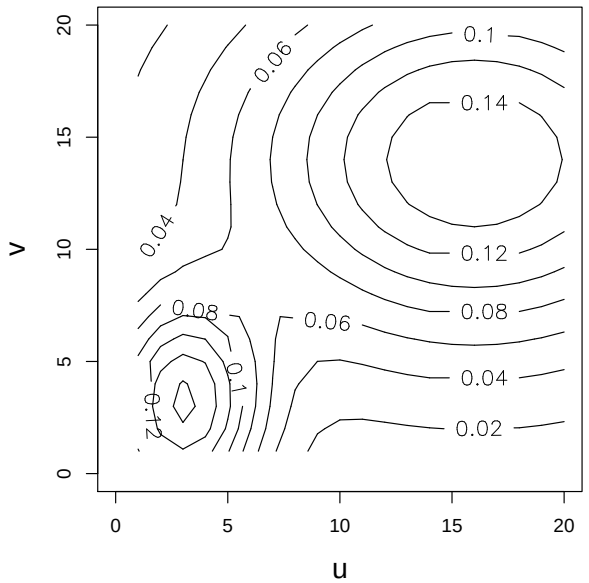
1. $N(A) = Pois\left(\int_{(\boldsymbol{s})\in A} \lambda(\boldsymbol{s}) d\boldsymbol{s}\right)$

   $(|A| = \int_{(\boldsymbol{s})\in A} d\boldsymbol{s})$
2. Given $N(A) = n$, events distributed in A as an independent sample from a distribution on A with p.d.f. proportional to $\lambda(\boldsymbol{s})$.

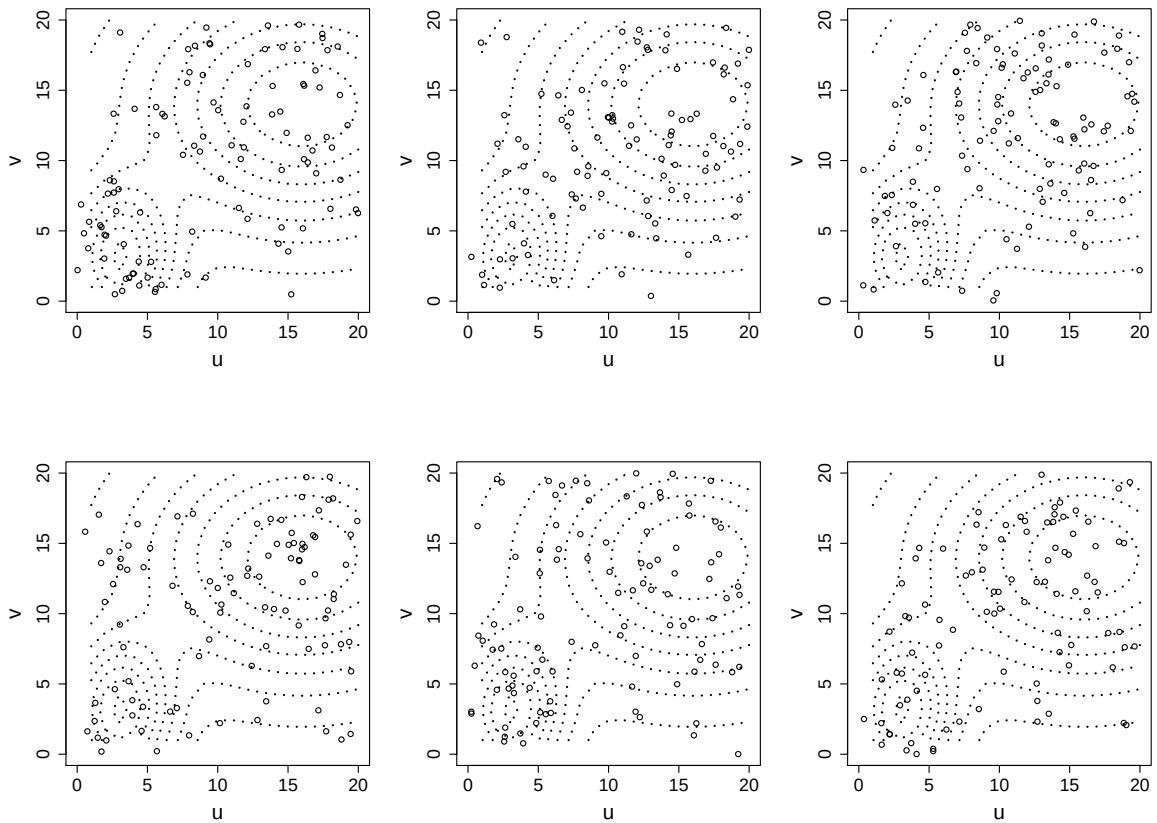
We still have *counts from areas* $\sim$ *Poisson* and *events are distributed proportional to the intensity*.

Example intensity function





Six realizations



IMPORTANT FACT!

Without additional information, no analysis can differentiate between:

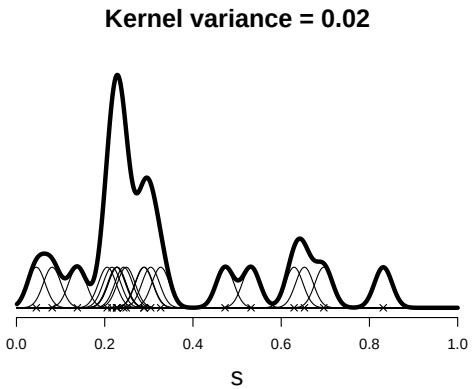
1. *independent* events in a *heterogeneous (non-stationary)* environment
2. *dependent* events in a *homogeneous (stationary)* environment

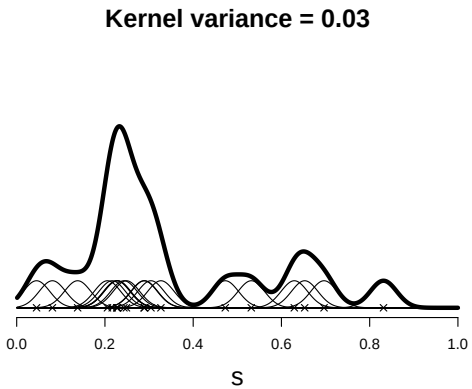
How do we estimate intensities?

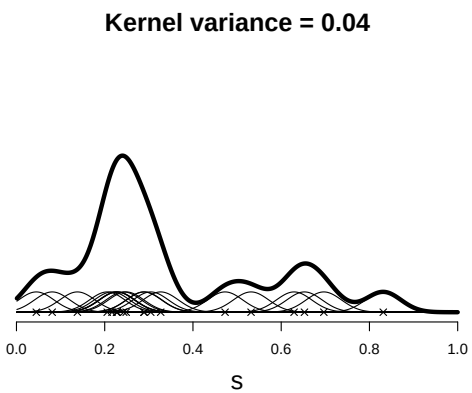
Kernel estimators provide a natural approach (Silverman (1986) and Wand and Jones (1995, `KernSmooth` R library)).

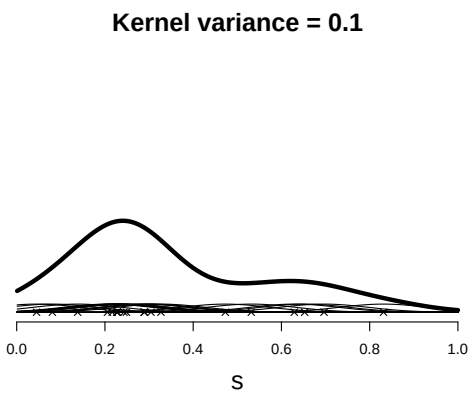
Main idea: Put a little "kernel" of density at each data point, then sum to give the estimate of the overall density function.

Kernels and bandwidths









Kernel estimation in R

`base`

- `density()` one-dimensional kernel

`library(KernSmooth)`

- `bkde2D(x, bandwidth, gridsize=c(51, 51), range.x=<<see below>>, truncate=TRUE)` block kernel density estimation

`library(splancs)`

- `kernel2d(pts,poly,h0,nx=20, ny=20,kernel='quartic')`

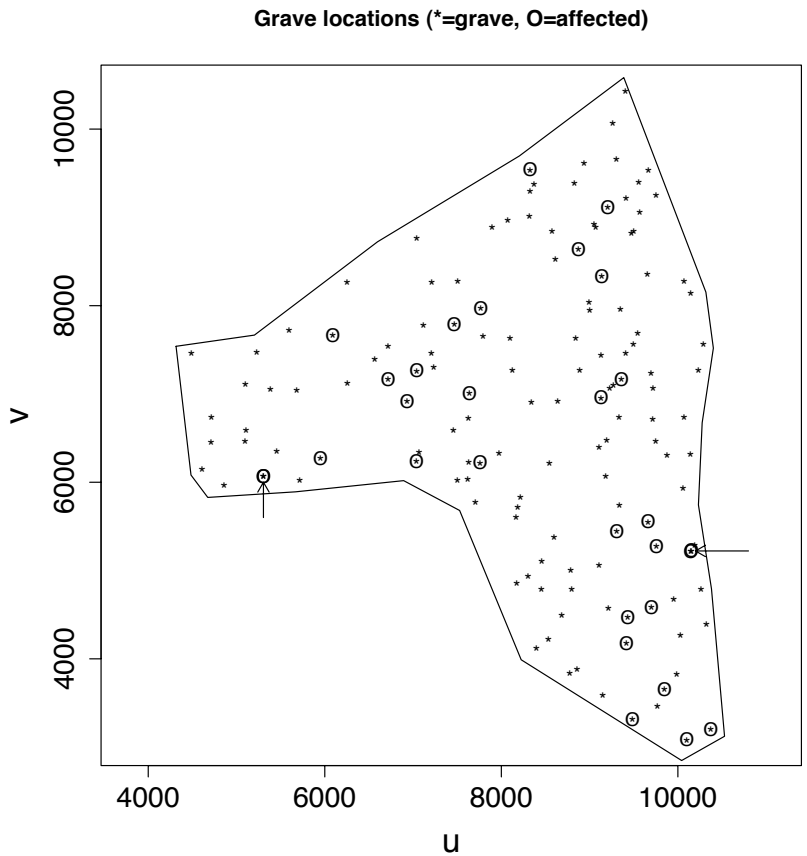
`library(spatstat)`

- `ksmooth.ppp(x, sigma, weights, edge=TRUE)`

Data Break: Early Medieval Grave Sites

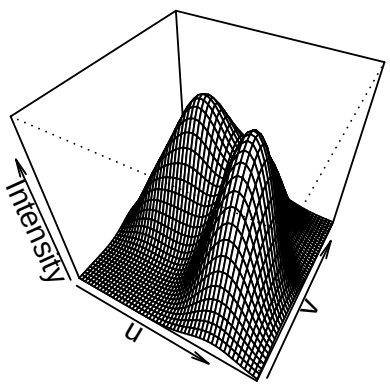
- Alt and Vach (1991). (Data from Richard Wright Emeritus Professor, School of Archaeology, University of Sydney.)
- Archeological dig in Neresheim, Baden-Württemberg, Germany.
- Question: are graves placed according to family units?
- 143 grave sites, 30 with missing or reduced wisdom teeth.
- Could intensity estimates for grave sites with and without wisdom teeth help answer this question?

Plot of the data

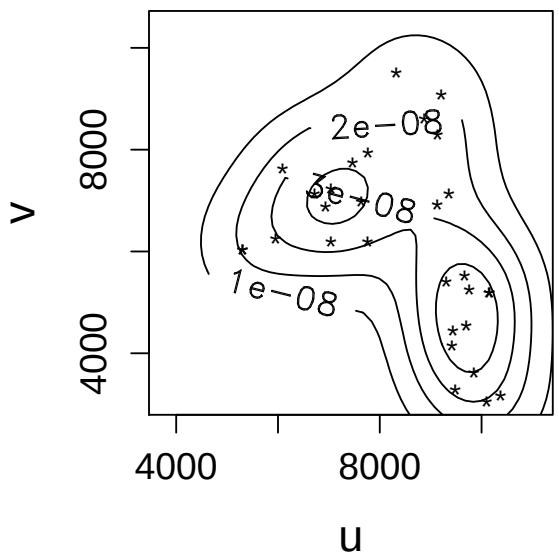


Case intensity

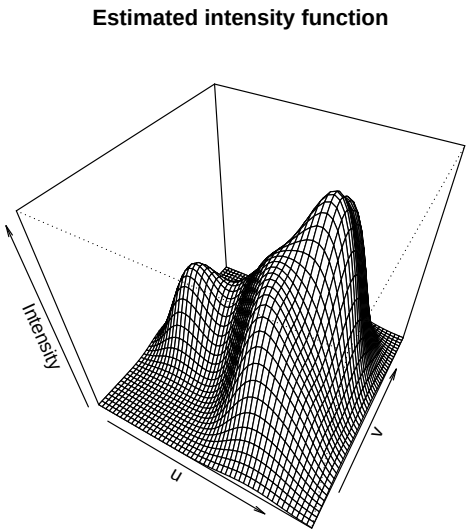
Estimated intensity function

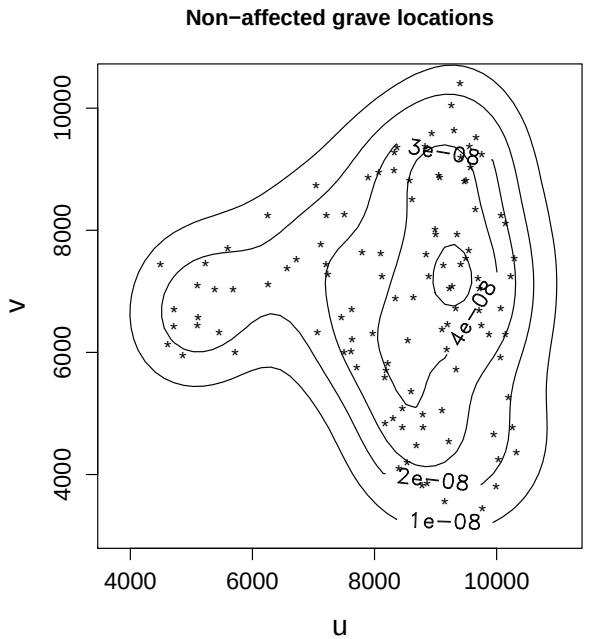


Affected grave locations



Control intensity





What we have/don't have

- Kernel estimates suggest *where* there might be differences.
- No significance testing (yet!)

First and Second Order Properties

- The intensity function describes the *mean* number of events per unit area, a *first order* property of the underlying process.
- What about *second order* properties relating to the variance/covariance/correlation between event locations (if events non independent...)?

Ripley's K function

Ripley (1976, 1977 introduced) the *reduced second moment measure* or K *function*

$$K(h) = \frac{E[\#\text{ events within } h \text{ of a } \textit{randomly} \text{ chosen event}]}{\lambda},$$

for any positive *spatial lag* h.

- Under CSR, $K(h) = \pi h^2$ (area of circle of with radius h).
- Clustered? $K(h) > \pi h^2$.
- Regular? $K(h) < \pi h^2$.

Calculating $K(h)$ in R

```
library(splancs)
```

- `khat(pts,poly,s,newstyle=FALSE)`
- `poly` defines polygon boundary (important!!!).

```
library(spatstat)
```

- `Kest(X, r, correction=c("border", "isotropic", "Ripley", "translate"))`
- Boundary part of `X` (point process "object").

Plots with $K(h)$

- Plotting $(h, K(h))$ for CSR is a parabola.
- $K(h) = \pi h^2$ implies

$$\left(\frac{K(h)}{\pi}\right)^{1/2} = h.$$

- Besag (1977) suggests plotting

$$h \text{ versus } \widehat{L}(h)$$

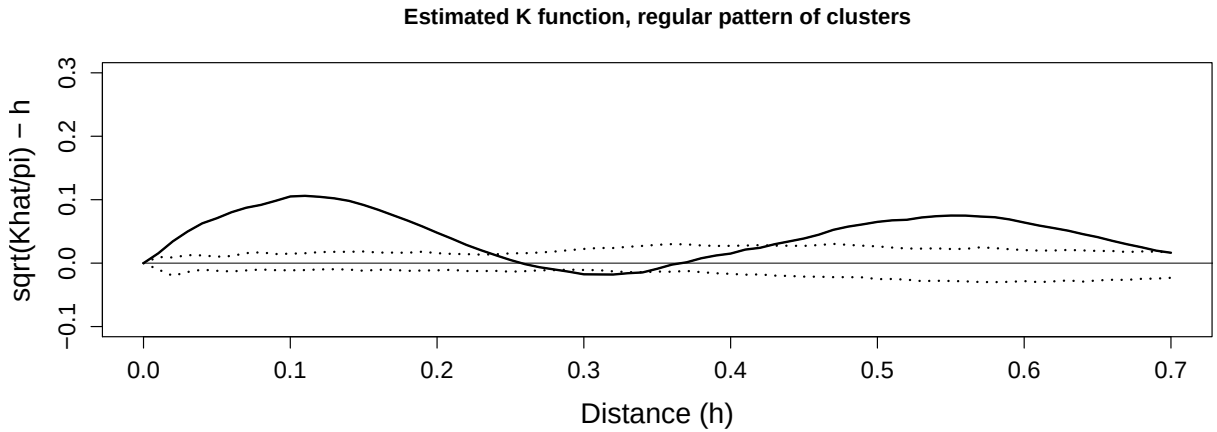
where

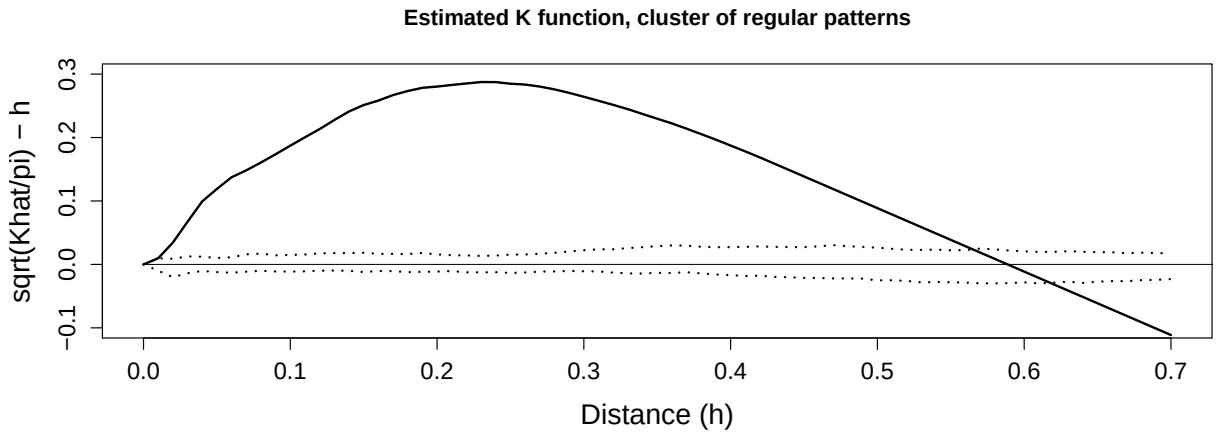
$$\widehat{L}(h) = \left(\frac{\widehat{K}_{ec}(h)}{\pi}\right)^{1/2} - h$$

Monte Carlo Variability and Envelopes

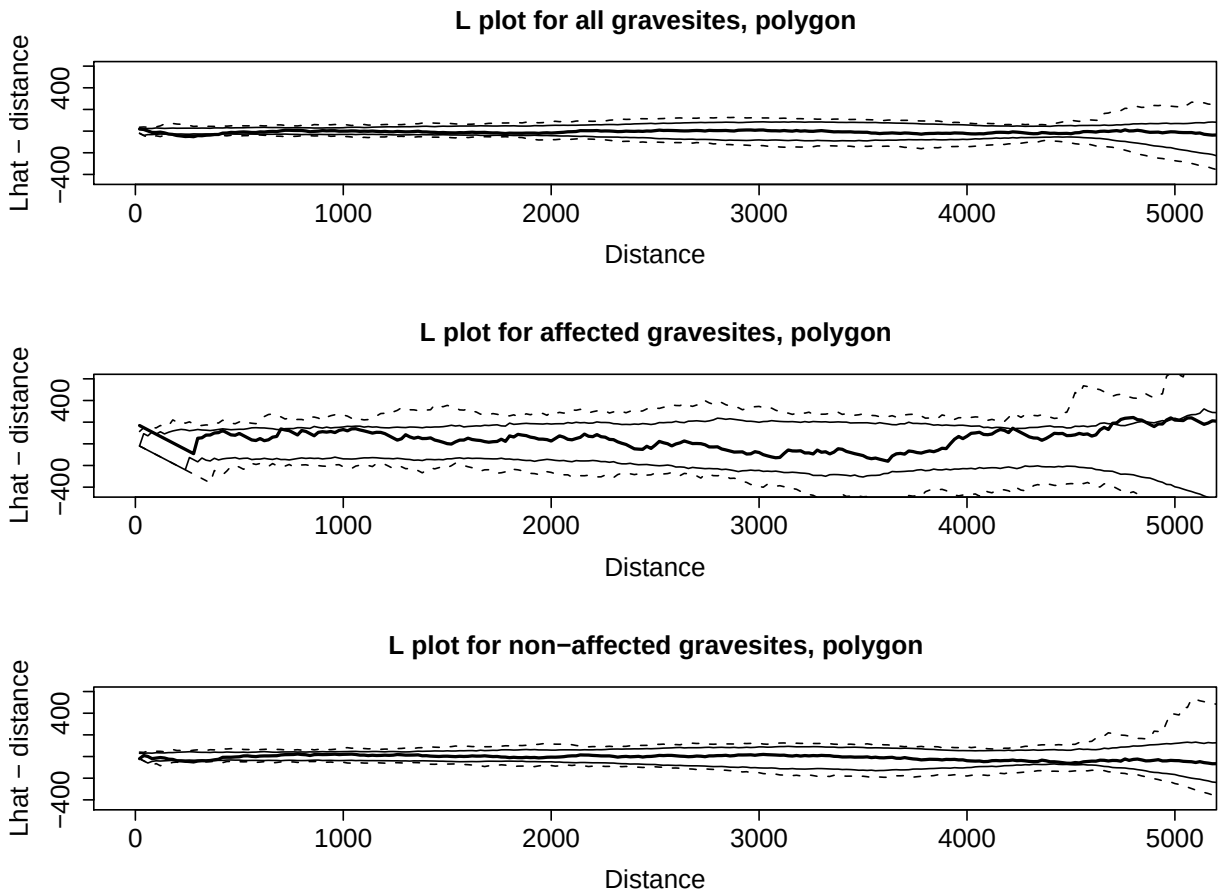
- Observe $\widehat{K}(h)$ from data.
- Simulate a realization of events from CSR.
- Find $\widehat{K}(h)$ for the simulated data.
- Repeat simulations many times.
- Create simulation "envelopes" from simulation-based $\widehat{K}(h)$'s.

Example: Regular clusters and clusters of regularity





Data break: Medieval graves: K functions with polygon adjustment



Clustering?

- Clustering of cases at very shortest distances.
- Likely due to two coincident-pair sites (both cases in both pairs).
- Envelopes based on random samples of 30 "cases" from set of 143 locations.

Notes

- First and second moments do not uniquely define a distribution, and $\lambda(\boldsymbol{s})$ and $K(h)$ do not *uniquely* define a spatial point pattern (Baddeley and Silverman 1984, and in Section 5.3.4).
- Analyses based on $\lambda(\boldsymbol{s})$ typically assume independent events.
- Analyses based on $K(h)$ typically assume a stationary process (with constant λ).
- Remember IMPORTANT FACT! above.

What questions can we answer?

- Are events uniformly distributed in space?
  - Test CSR.
- If not, where are events more or less likely?
  - Intensity estimation.
- Do events tend to occur near other events, and, if so, at what scale?
  - K functions with Monte Carlo envelopes.