MODULE 16: Spatial Statistics in Epidemiology and Public Health
Lecture 3: Point Processes

Jon Wakefield and **Lance Waller**

References

- Diggle, P.J. (1983) *Statistical Analysis of Spatial Point Patterns.* London: Academic Press.
- Diggle, P.J. (2014) *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns.* 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 of events.
- Introduce analytic methods for describing patterns in observed collections of events.
- Illustrate the approaches using examples from archaeology, conservation biology, and epidermal neurology.

Terminology

- *Event*: An occurrence of interest (e.g., disease case).
- *Event location*: Where an event occurs.
- *Realization*: An observed set of event locations (a data set).
- *Point*: Any location in the study area.

- *Point*: Where an event *could* occur.
- *Event*: Where an event *did* occur.

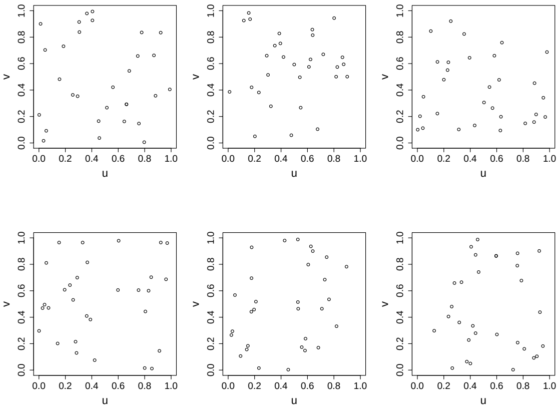
Random patterns

- We use probability models to generate patterns so, in effect, all of the patterns we consider are "random".
- Usually, "random pattern" refers to a pattern not influenced by the factors under investigation.

Complete spatial randomness (CSR)

- Start with a model of “lack of pattern”.
- Events equally likely to occur anywhere in the study area.
- Event locations independent of each other.

Six realizations of CSR

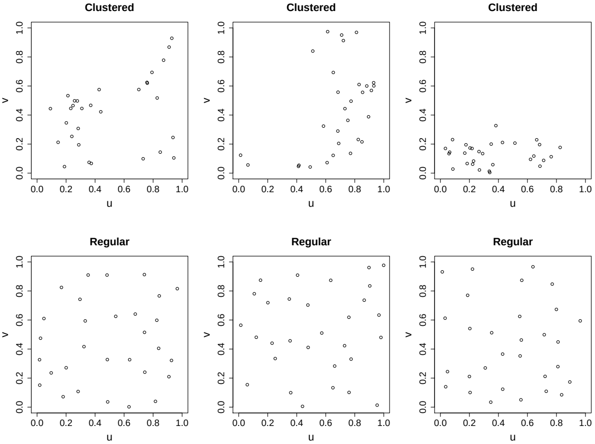


CSR as a boundary condition

CSR serves as a boundary between:

- Patterns that are more “clustered” than CSR.
- Patterns that are more “regular” than CSR.

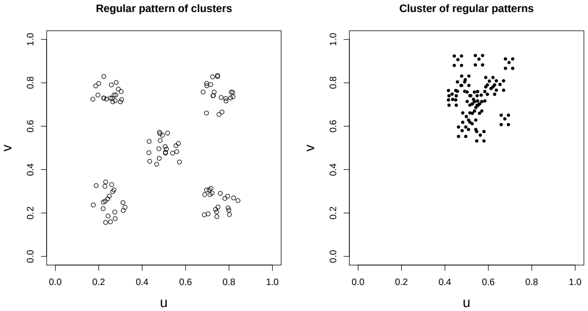
Too Clustered (top), Too Regular (bottom)



The role of scale

- "Eyeballing" clustered/regular sometimes difficult.
- In fact, an observed pattern may be clustered at one spatial *scale*, and regular at another.
- *Scale* is an important idea and represents the distances at which the underlying process generating the data operates.
- Many ecology papers on estimating scale of a process (e.g., plant disease, animal territories) but little in public health literature (so far).

Clusters of regular patterns/Regular clusters



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 Kroger.
- For us, each random variable represents an event location.

Stationarity/Isotropy

- *Stationarity*: Properties of process invariant to translation.
- *Isotropy*: Properties of process invariant to rotation around an origin.
- Why do we need these? They provide a sort of replication in the data that allows statistical estimation.
- Not required, but development easier.

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?

- Criterion (b) describes our notion of uniform and independently distributed in space.
- 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))
```

Compared to temporal Poisson process

Recall the three "magic" features of a Poisson process in time:

- Number of events in non-overlapping intervals are Poisson distributed,
- Conditional on the number of events, events are uniformly distributed within a fixed interval, and
- Interevent times are exponentially distributed.

In space, no ordering so no (uniquely defined) interevent distances.

Spatial Poisson process equivalent

Criteria (a) and (b) equivalent to

1. # events in non-overlapping areas is independent
2. Let $A =$ region, $|A| =$ area of A

$$\lim_{|A| \to 0} \frac{\Pr[\text{ exactly one event in } A]}{|A|} = \lambda > 0$$

3. $$\lim_{|A| \to 0} \frac{\Pr[2 \text{ or more in } A]}{|A|} = 0$$

Interesting questions

- Test for CSR
- Simulate CSR
- Estimate λ (or $\lambda(\mathbf{s})$)
- Compare $\lambda_1(\mathbf{s})$ and $\lambda_2(\mathbf{s}); \mathbf{s} \in D$ for two point processes. (same underlying intensity?

$$\begin{aligned} \text{e.g. } Z_1(\mathbf{s}) &= \text{location of disease cases} \\ Z_2(\mathbf{s}) &= \text{location of population at risk} \\ &\quad \text{or environmental exposure levels.})) \end{aligned}$$

Monte Carlo hypothesis testing

Review of basic hypothesis testing framework:

- $T =$ a random variable representing the test statistic.
- Under $H_0 : U \sim F_0$.
- From the data, observe $T = t_{obs}$ (the observed value of the test statistic).
  p-value $= 1 - F_0(t_{obs})$
- Sometimes it is easier to simulate $F_0(\cdot)$ than to calculate $F_0(\cdot)$ exactly.

Besag and Diggle (1977).

Steps in Monte Carlo testing

1. observe u_1.
2. simulate $u_2, \ldots, u_m$ from F_0.
3. p.value $= \frac{\text{rank of } u_1}{m}$.

For tests of CSR (complete spatial randomness), M.C. tests are very useful since CSR is easy to simulate but distribution of test statistic may be complex.

Monte Carlo tests very helpful

1. when distribution of U is complex but the spatial distribution associated with H_0 is easy to simulate.
2. for permutation tests

   e.g., 592 cases of leukemia in ~ 1 million people in 790 regions

   - permutation test requires all possible permutations.
   - Monte Carlo assigns cases to regions under H_0.

Testing for CSR

A good place to start analysis because:

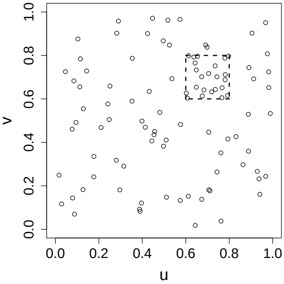
1. rejecting CSR a minimal prerequisite to model observed pattern (i.e. if not CSR, then there is a pattern to model)
2. tests aid in formulation of possible alternatives to CSR (e.g. regular, clustered).
3. attained significance levels measure strength of evidence against CSR.
4. informal combination of several complementary tests to show *how* pattern departs from CSR (although this has not been explored in depth).

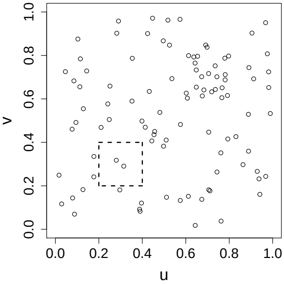
Final CSR notes

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 structure in an observed point pattern.
4. not as useful in public health? Why not?

Heterogeneous population density





What if intensity not constant?

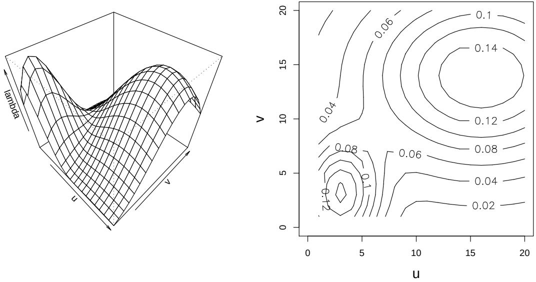
$$
\begin{aligned}
\text{Let } d(x, y) &= \text{tiny region around } (x, y) \\
\lambda(x, y) &= \lim_{|d(x,y)| \to 0} \frac{E[N(d(x, y))]}{|d(x, y)|} \\
&\approx \lim_{|d(x,y)| \to 0} \frac{\Pr[N(d(x, y)) = 1]}{|d(x, y)|}
\end{aligned}
$$

If we use $\lambda(x, y)$ instead of λ, we get a *Heterogeneous (Inhomogeneous) Poisson Process*

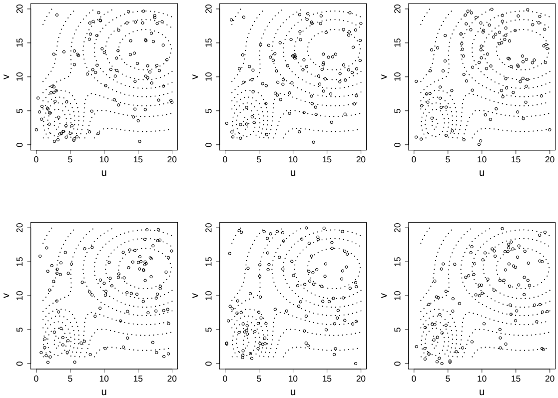
Heterogeneous Poisson Process

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

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?

For an heterogeneous Poisson process, $\lambda(\mathbf{s})$ is closely related to the density of events over A ($\lambda(\mathbf{s}) \propto$ density).

So density estimators provide a natural approach (for details on density estimation see Silverman (1986) and Wand and Jones (1995, `KernSmooth` R library!)).

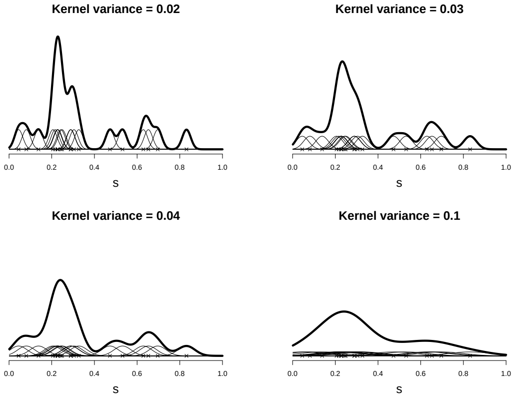
Main idea: Put a little "kernel" of density at each data point, then sum to give the estimate of the overall density function.

What you need

In order to do kernel estimation, you need to choose:

1. kernel (shape) - Epeuechnikov (1986) shows any reasonable kernel gives near optimal results.
2. "bandwidth" (range of influence). The larger *b*, the bandwidth, the smoother the estimated function.

Kernels and bandwidths



1-dim kernel estimation

If we have observations $u_1, u_2, \ldots, u_N$ (in one dimension), the kernel density estimate is

$$\tilde{f}(u) = \frac{1}{Nb} \sum_{i=1}^{N} Kern\left(\frac{u - u_i}{b}\right) \quad (1)$$

where $Kern(\cdot)$ is a kernel function satisfying

$$\int_D Kern(s)ds = 1$$

and b the bandwidth. To estimate the intensity function, replace N^{-1} by $|D|^{-1}$.

What do we do with this?

- Evaluate intensity (density) at each of a grid of locations.
- Make surface or contour plot.

How do we pick the bandwidth?

- Circular question...
- Minimize the Mean Integrated Squared Error (MISE)
- Cross validation
- Scott's rule (good rule of thumb).

$$\widehat{b}_u = \widehat{\sigma}_u N^{-1/(dim+4)} \tag{2}$$

where $\widehat{\sigma}_u$ is the sample standard deviation of the u-coordinates, N represents the number of events in the data set (the sample size), and *dim* denotes the dimension of the study area.

Kernel estimation in R

`base`

- `density()` one-dimensional kernel

`library(MASS)`

- `kde2d(x, y, h, n = 25, lims = c(range(x), range(y)))`

`library(KernSmooth)`

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

More kernel estimation in R

`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

What do we have?

- Alt and Vach (1991). (Data sent from Richard Wright Emeritus Professor, School of Archaeology, University of Sydney.)
- Archeaological dig in Neresheim, Baden-Württemberg, Germany.
- The anthropologists and archaeologists involved wonder if this particular culture tended to place grave sites according to family units.
- 143 grave sites.

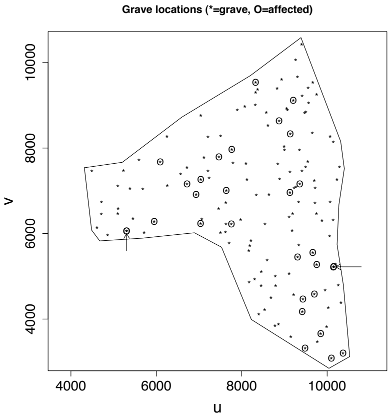
What do we want?

- 30 with missing or reduced wisdom teeth ("affected").
- Does the pattern of "affected" graves (cases) differ from the pattern of the 113 non-affected graves (controls)?
- How could estimates of the intensity functions for the affected and non-affected grave sites, respectively, help answer this question?

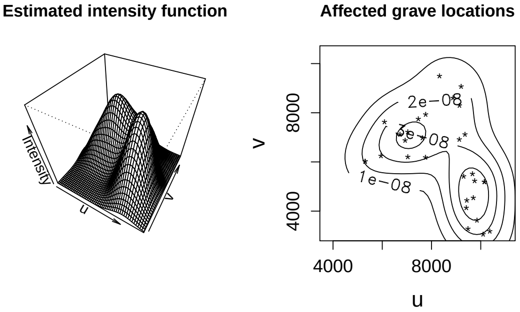
Outline of analysis

- Read in data.
- Plot data (axis intervals important!).
- Call 2-dimensional kernel smoothing functions (choose kernel and bandwidth).
- Plot surface (`persp()`) and contour (`contour()`) plots.
- Visual comparison of two intensities.

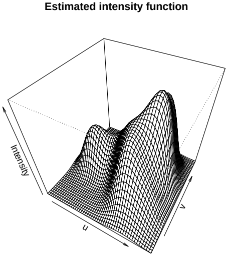
Plot of the data

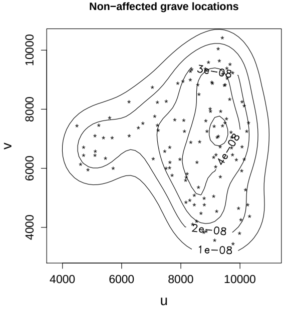


Case intensity



Control intensity





What we have/don't have

- Kernel estimates suggest similarities and differences.
- Suggest *locations* 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*.

NOTE: Use of λ implies assumption of stationary process!

Properties of $K(h)$

- Ripley (1977) shows specifying $K(h)$ for all $h > 0$, equivalent to specifying $\text{Var}[N(A)]$ for any subregion A.
- Under CSR, $K(h) = \pi h^2$ (area of circle of with radius h).
- Clustered? $K(h) > \pi h^2$.
- Regular? $K(h) < \pi h^2$.

Second order intensity?

- $K(h)$ not universally hailed as *the* way (or even a good way) to describe second order properties.
- Provides a nice introduction for us, but another related property is...
- The *second order intensity*, $\lambda_2(\mathbf{s}, \mathbf{u})$,

$$\lambda_2(\mathbf{s}, \mathbf{u}) = \lim_{\substack{|d(\mathbf{s})| \to 0 \\ |d(\mathbf{u})| \to 0}} \frac{E(N(d(\mathbf{s}))N(d(\mathbf{u}))}{|d(\mathbf{s})||d(\mathbf{u})|}$$

Relationship to $K(h)$

- How does $K(\cdot)$ relate to $\lambda_2(\mathbf{s}, \mathbf{u})$?

  In $\Re^2$, for a stationary, isotropic process $\lambda_2(\mathbf{s}, \mathbf{u}) = \lambda_2(\|\mathbf{s} - \mathbf{u}\|)$. Then,

$$\lambda K(h) = \frac{2\pi}{\lambda} \int_0^h u\lambda_2(u)du$$

  So

$$\lambda_2(h) = \frac{\lambda^2 K'(h)}{2\pi h}.$$

- Which to use ($K(\cdot)$ or $\lambda_2(\cdot)$)?
- In theory, $\lambda_2(h)$ often used but $K(h)$ is easier to estimate from a set of data.

Estimating $K(h)$

Start with definition, replacing expectation with average yields

$$\widehat{K}(h) = \frac{1}{\widehat{\lambda}N} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \delta(d(i,j) < h),$$

for N events, where

- $d(i,j)$ = distance between events i and j
- $\delta(d(i,j) < h) = 1$ if $d(i,j) < h$, 0 otherwise.

Edges

- Think about events near "edges" of study area.
- How do we count events within h if distance between edge and observed event is $< h$?
- Unobservable data drive increasing.
- More of a problem as h increases...
- "Edge effects" call for "edge correction".

Edge correction

- Add "guard area" around study area and only calculate $\widehat{K}(h)$ for $h <$ width of guard area.
- Toroidal correction.
- Ripley's (1976) edge correction

$$\widehat{K}_{ec}(h) = \widehat{\lambda}^{-1} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} w_{ij}^{-1} \delta(d(i,j) < h)$$

where $w_{ij} =$ proportion of the circumference of the circle centered at event i with radius $d(i,j)$ within the study area.

Ripley's weights

- Conceptually, conditional probability of observing an event at distance $d(i, j)$ given an event occurs $d(i, j)$ from event i.
- Works for "holes" in study area too. (Astronomy application).
- Requires definition of study boundary (and way of calculating w_{ij}.

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

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

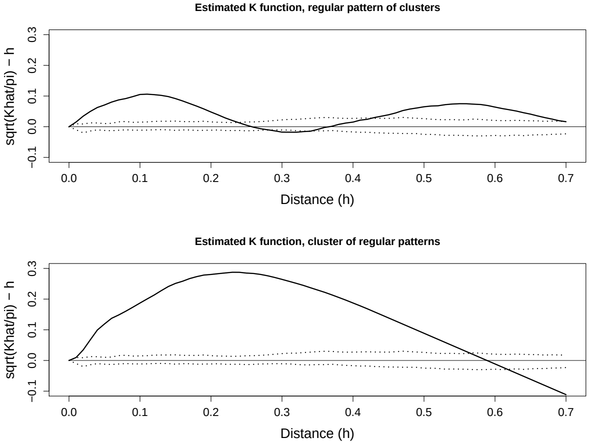
Variability and Envelopes

Are there any distributional results for $\widehat{K}(h)$?

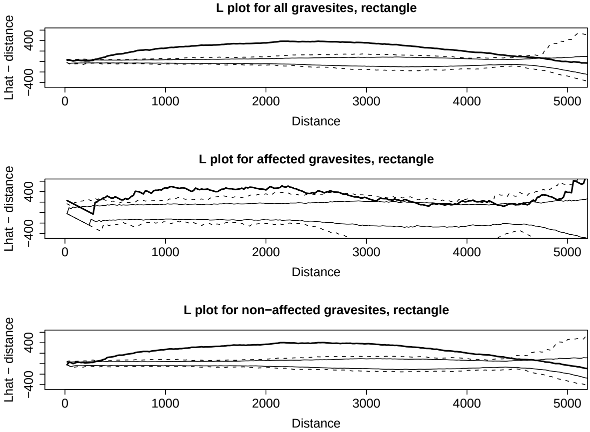
Some, but mostly for particular region shapes. Monte Carlo approaches more general.

- 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



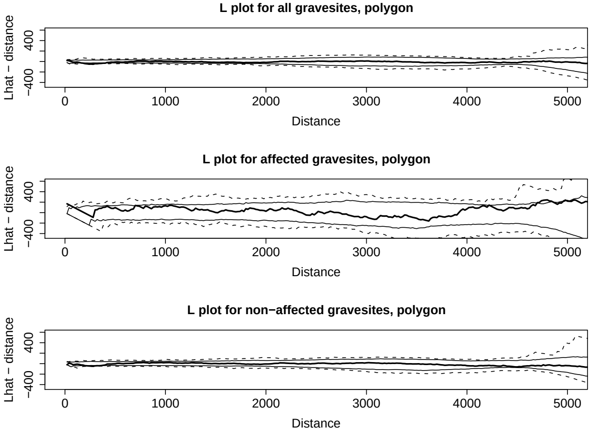
Data break: Medieval gravesites



Clustering!

- Looks like strong clustering, but wait...
- Compares all, cases, controls to CSR.
- Oops, forgot to adjust for polygon, clustering inside square area!
- Significant clustering, but *interesting* clustering?
- Let's try again with polygon adjustment.

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).
- Could we try random labelling here?
- Construct envelopes based on random samples of 30 "cases" from set of 143 locations.

A few notes

- Since $K(h)$ is a function of distance (h), it can indicate clustering at one scale and regularity at another.
- $K(h)$ measures *cumulative* aggregation (# events *up to h*). May result in delayed response to change from clustering to regularity.
- λ_2 (or similar measures) based on $K'(h)$ may respond more instantaneously.
- Often see "envelopes" based on extremes, but what does this mean with respect to increasing numbers of simulations? Quantiles may be better (at least they will converge to something).

Testing

- Envelopes provide *pointwise* inference at particular values of h, but not correct to find deviation then assess whether it is significant.
- Can conduct test of $H_0 : L(h) = h$ (equivalent to $H_0 : K(h) = \pi h^2$) for a *predefined* interval for h (Stoyan et al., 1995, p. 51).
- Monte Carlo test based on

$$T = \max_{0 \leq h \leq h_{max}} \left| \widehat{L}(h) - h \right| .$$

Notes

- Just as the first and second moments do not uniquely define a distribution, $\lambda(\mathbf{s})$ and $K(h)$ do not *uniquely* define a spatial point pattern (Baddeley and Silverman 1984, and in Section 5.3.4).
- Analyses based on $\lambda(\mathbf{s})$ typically assume independent events, describing the observed pattern entirely through a heterogeneous intensity.
- Analyses based on $K(h)$ typically assume a stationary process (with constnat λ), describing the observed pattern entirely through correlations between events.
- Remember Barlett's (1964) result (IMPORTANT FACT! above).

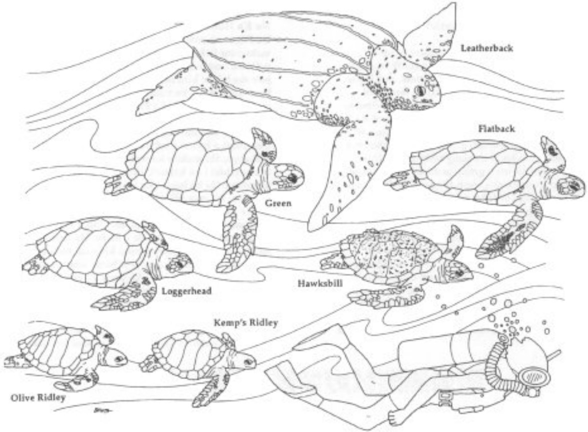
What questions can we answer?

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

Sea turtle biology

- Existed (at least) since Jurassic period ($\sim$ 200-120 mya).
- Air breathing reptile.
- Spends life in ocean, returns to natal beach to lay eggs.
- Lays eggs on land ($\sim$ 100 eggs per clutch, buried in sand).
- Hatchlings appear *en masse*, head to the ocean.

Sea Turtle Species



Consider the hatchling

- Life fraught with peril.
- On menu for crabs, gulls, etc.
- Goal 1: Get to the ocean. Follow the light!

Sea Turtles and Humans: Regulatory implications

- All species “threatened” or “endangered”.
- Regulations in place for fisheries (TEDs), beach-front lighting, beach-front development.
- East coast of Florida (USA), prime nesting site for loggerhead turtles.
- Also popular among green turtles (biannual cycle).
- We consider impact of two different construction projects on Juno Beach, Florida.

Two construction projects

- Construction of 990-foot fishing pier in 1998.
- Environmental impact statement approved if collect data to determine impact on sea turtle nesting.
- Beach nourishment project 2000.
- Dredge up sand offshore and rebuild beach.
- Past studies on impact of beach nourishment on turtle nesting (reduced for 2 nesting seasons, off-shore slope impact?)

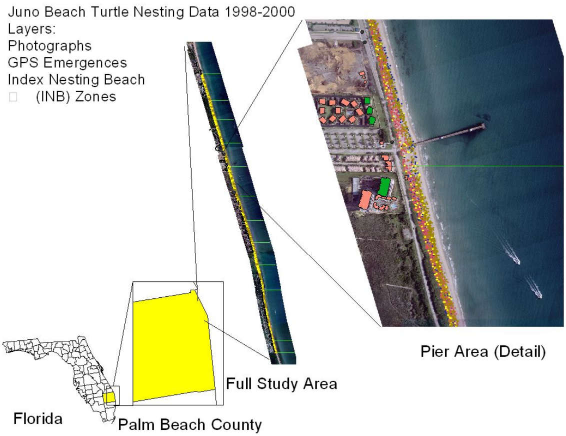
"Standard" approach

- Patrol beach each morning during nesting season.
- Count trackways in each Index Nesting Beach Zone.
- Compare distribution of counts between years.

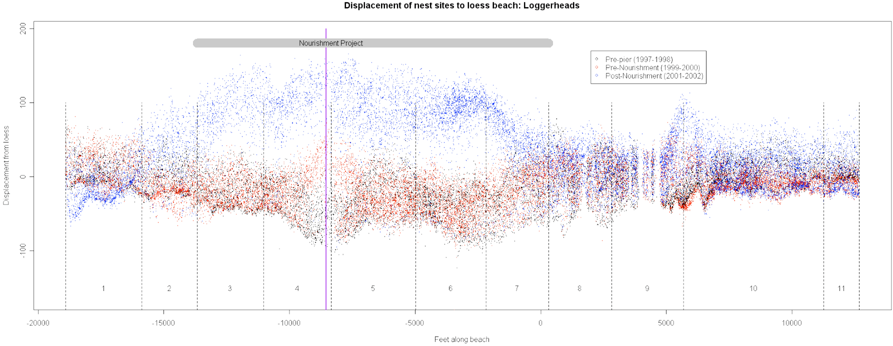
Juno Beach data

- Marinelife Center of Juno Beach volunteers patrol beach each morning.
- Differential GPS coordinates for apex of each “crawl”.
- Store: Date. Nest? Species. Predation?
- Approximately 10,000 points per year.

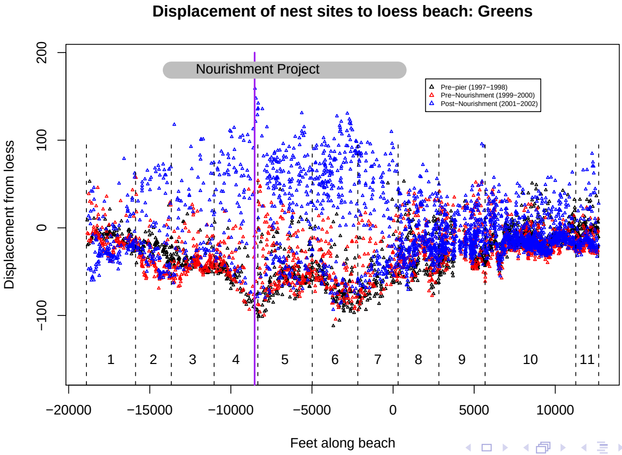
Juno Beach Data



Loggerhead Displacement from Beach



Displacement of Green Emergences



"Linearizing" the beach

- While we have sub-meter accurate GPS locations, depth on beach of secondary interest.
- We want to know *where* differences in nesting pattern occur and whether these observed differences are significant.
- Treat emergence locations as realization of a spatial point process in one dimension (events on a line).
- Ignore depth on beach by projecting all points to loess curve representing beach.
- Three time periods of interest: Pre-pier (1997-1998), pre-nourishment (post-pier) (1999-2000), post-nourishment (2001-2002).

Kernel estimates for each time period

- Let $\lambda(\mathbf{s})$ denote *intensity function* of point process (mean number of events per unit distance).
- Intensity proportional to pdf.
- Data: locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$.
- Estimate via kernel estimation:
$\widehat{\lambda}(\mathbf{s}) = \frac{1}{nb} \sum_{i=1}^{n} \text{Kern}\left[\frac{(\mathbf{s}-\mathbf{s}_i)}{b}\right]$
- $\text{Kern}(\cdot)$ is kernel function, b bandwidth (governing smoothness).

Ratio of kernel estimates

- Kelsall and Diggle (1995, *Stat in Med*) consider ratio of two intensity estimates as spatial measure of relative risk.
- Suppose we have n_1 "before" events, n_2 "after" events.
- $\lambda_1(\mathbf{s})$, $\lambda_2(\mathbf{s})$ associated before/after intensity functions.
- Conditional on n_1 and n_2, data equivalent to two random samples from pdfs:
  $f_1(\mathbf{s}) = \lambda_1(\mathbf{s}) / \int \lambda_1(\mathbf{u}) d\mathbf{u}$ and
  $f_2(\mathbf{s}) = \lambda_2(\mathbf{s}) / \int \lambda_2(\mathbf{u}) d\mathbf{u}$.

Ratio of kernel estimates

- Consider *log relative risk*

$$r(\mathbf{s}) = \log \frac{f_1(\mathbf{s})}{f_2(\mathbf{s})}.$$

- Some algebra yields:

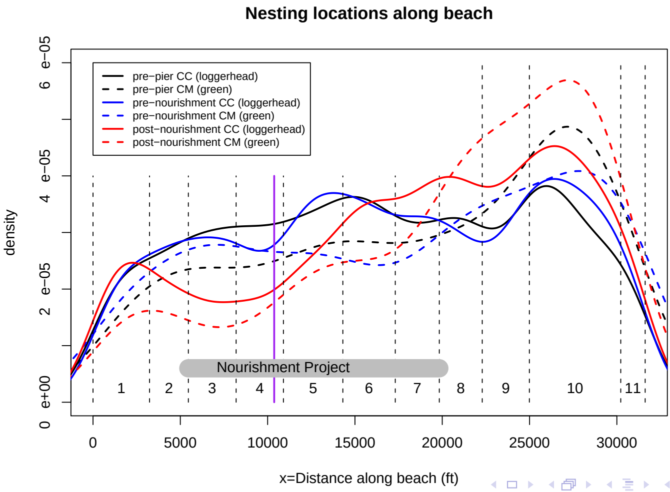
$$r(\mathbf{s}) = \log \left[\frac{\lambda_1(\mathbf{s})}{\lambda_2(\mathbf{s})} \right] - \log \left[\frac{\int \lambda_1(\mathbf{u}) d\mathbf{u}}{\int \lambda_2(\mathbf{u}) d\mathbf{u}} \right].$$

- $H_0 : r(\mathbf{s}) = 0$ for all $\mathbf{s}$.
- How do we test this?

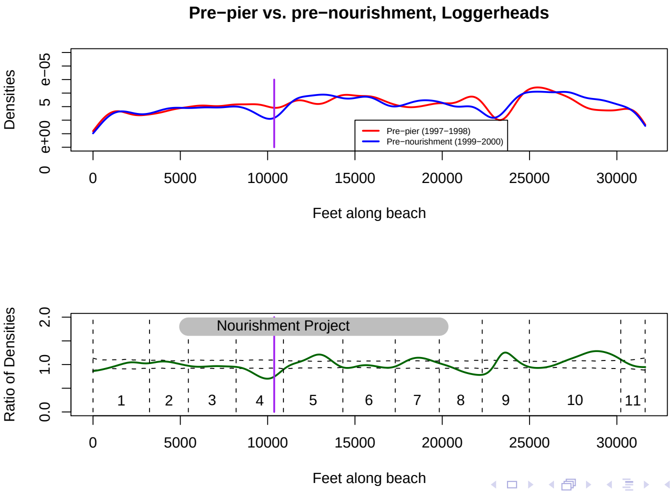
Monte Carlo inference

- Conditional on all $n_1 + n_2$ locations, randomize n_1 to "before", n_2 to "after".
- Calculate $\widehat{f}^*(\mathbf{s})$ and $\widehat{g}^*(\mathbf{s})$, get ratio $r^*(\mathbf{s})$.
- Repeat many (999) times.
- Calculate pointwise 95% tolerance region.

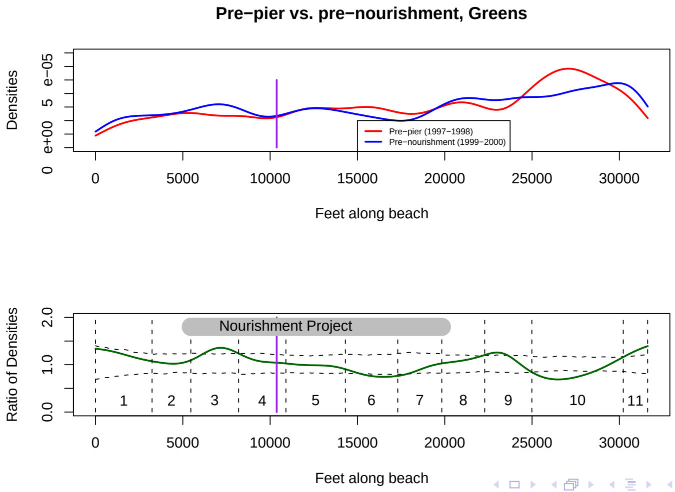
Densities


Nesting locations along beach
pre−pier CC (loggerhead)
pre−pier CM (green)
pre−nourishment CC (loggerhead)
pre−nourishment CM (green)
post−nourishment CC (loggerhead)
post−nourishment CM (green)
Nourishment Project
1
2
3
4
5
6
7
8
9
10
11
density
0 e+00
2 e−05
4 e−05
6 e−05
0
5000
10000
15000
20000
25000
30000
x=Distance along beach (ft)

Pre- vs. post-pier, Loggerheads



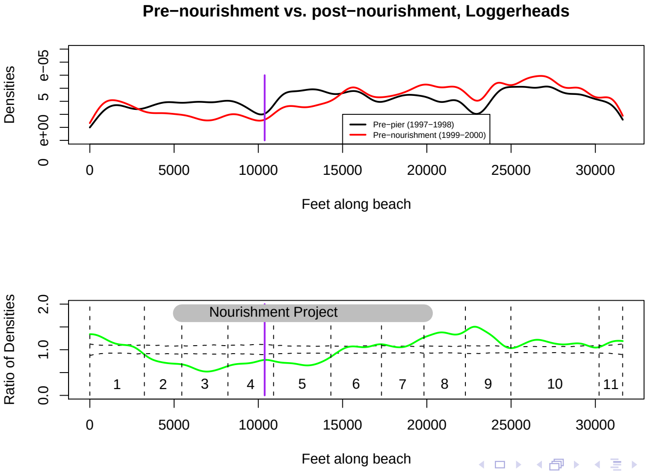
Pre- vs. post-pier, Greens



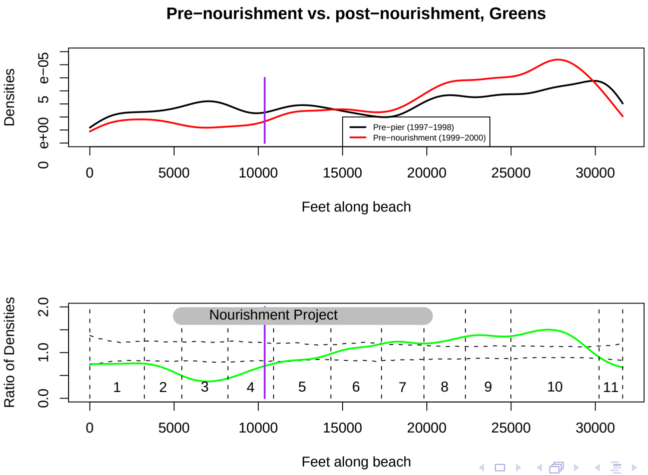
Differences

- Green densities smoother (wider bandwidth due to smaller sample size).
- Clear impact of pier on loggerheads (but very local).
- Conservation strategy? Maintain protected areas nearby.
- Note: impact on *all* crawls (nesting and non-nesting).

Pre- vs. post-nourishment, Loggerheads



Pre- vs. post-nourishment, Greens



Differences

- Reduced nesting in nourishment zone, increases just to south.
- Offshore current suggests approach from north (left on plot).
- Impact on *all* crawls (nesting and non-nesting).
- Suggests impact of nourishment *before* turtle emerges. Slope?

Conclusions

Pairwise comparisons suggest:

- Local impact of pier for loggerheads, but not for greens.
- Shift to south in emergences (nesting and non-nesting) due to nourishment for both loggerheads and greens.

Questions?