Spatial Statistics
Point Process Data Unit 1

PM569 Spatial Statistics

Lecture 8: November 4, 2016

Point Pattern Data: A Historical Perspective

In 1946, R.D. Clarke wrote a report about heavily bombed region of South London.

AN APPLICATION OF THE POISSON DISTRIBUTION

BY R. D. CLARKE, F.I.A.
of the Prudential Assurance Company, Ltd.

READERS of Lidstone's *Notes on the Poisson frequency distribution* (*J.I.A.* Vol. LXXI, p. 284) may be interested in an application of this distribution which I recently had occasion to make in the course of a practical investigation.

During the flying-bomb attack on London, frequent assertions were made that the points of impact of the bombs tended to be grouped in clusters. It was accordingly decided to apply a statistical test to discover whether any support could be found for this allegation.

An area was selected comprising 144 square kilometres of south London over which the basic probability function of the distribution was very nearly constant, i.e. the theoretical mean density was not subject to material variation anywhere within the area examined. The selected area was divided into 576 squares of $\frac{1}{4}$ square kilometre each, and a count was made of the numbers of squares containing 0, 1, 2, 3, ..., etc. flying bombs. Over the period considered the total number of bombs within the area involved was 537. The expected numbers of squares corresponding to the actual numbers yielded by the count were then calculated from the Poisson formula:

$$Ne^{-m}(1+m+m^2/2!+m^3/3!+\ldots),$$

where $$N=576 \quad \text{and} \quad m=537/576.$$

The result provided a very neat example of conformity to the Poisson law and might afford material to future writers of statistical text-books.

The actual results were as follows:

No. of flying bombs per square	Expected no. of squares (Poisson)	Actual no. of squares
0	226·74	229
1	211·39	211
2	98·54	93
3	30·62	35
4	7·14	7
5 and over	1·57	1

Point Pattern Data: A Historical Perspective

- During WWII, Germany launched 1,358 V-2 Rockets at London.
- The V-2 had speed and a trajectory that made it invulnerable to interception, but its guidance systems were primitive, so it was thought that it couldn't hit specific targets.
- After strikes began in 1944, bomb damage maps were interpreted as showing that impact sites were clustered.
- If the V-2 strikes were clustered, then the guidance systems were more sophisticated than thought.
- R.D. Clarke set out to analyze these data to determine if the data were clustered or not.

Point Pattern Data: A Historical Perspective

- Clarke took a 12 km x 12 km region and sliced it up in to a grid of 576 squares, (144 km^2, so each grid square is 1/4 km^2).
- For each square, Clark recorded the total number of observed bomb hits. There were 537 total in the study area.
- He then recorded the number of squares with $k = 1, 2, 3, ...$ hits.
- The expected number of squares with k hits was derived from the Poisson distribution $\sum_{k=1}^{n} \frac{e^{-\lambda}\lambda^k}{k!}$ where $\lambda = \frac{537}{576}$ and $n = 576$.

No. of flying bombs per square	Expected no. of squares (Poisson)	Actual no. of squares
0	226·74	229
1	211·39	211
2	98·54	93
3	30·62	35
4	7·14	7
5 and over	1·57	1
	576·00	576

Point Pattern Data: A Historical Perspective

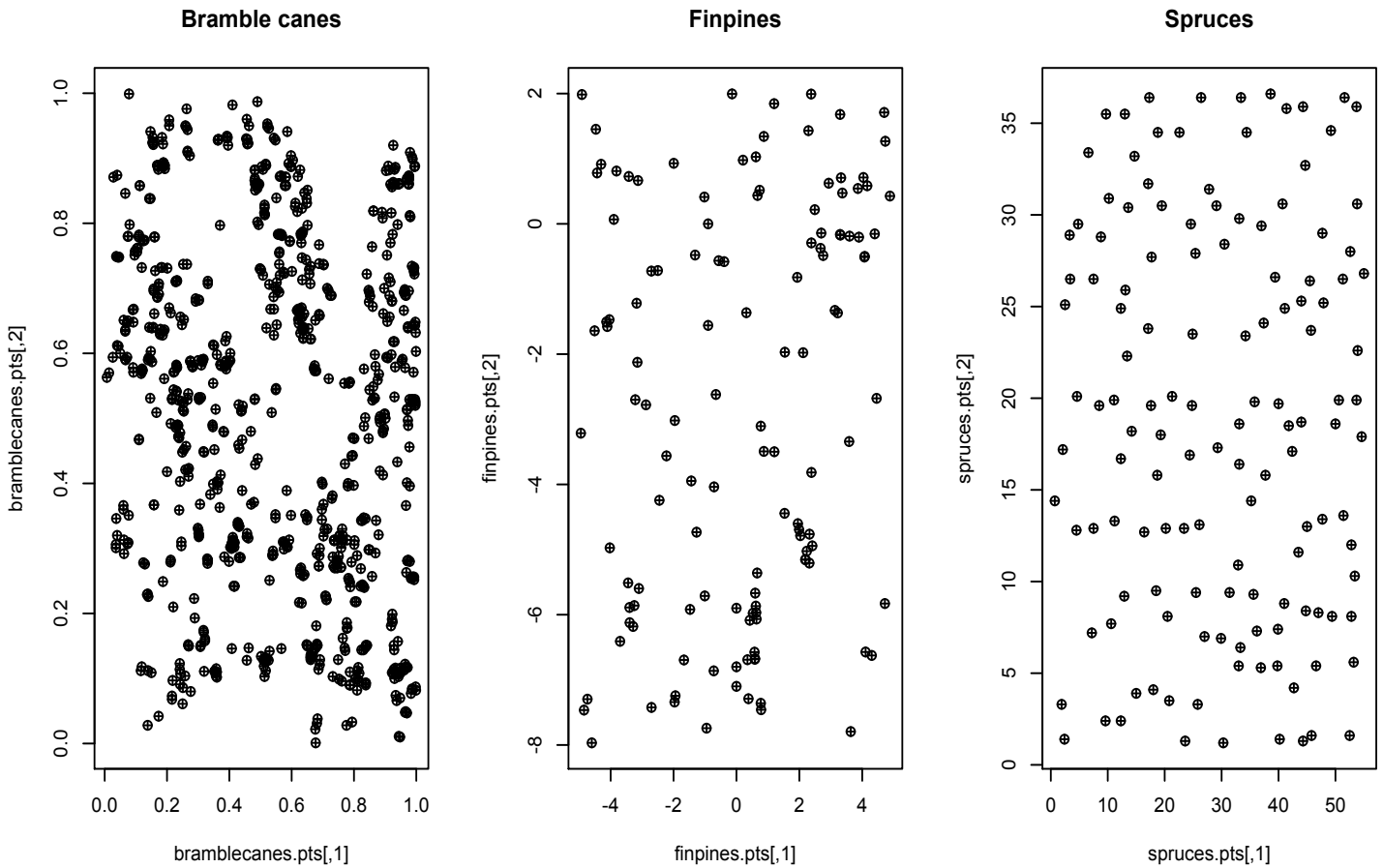
- Doing the cross tabulation of observed vs expected, he found $\chi^2 = 1.17$ which with 4 degrees of freedom (n-1 groups, the 0 group was excluded) has p-value=0.88.
- The occurrence of clustering would have been reflected in an excess of squares with a high number of bombs or none at all.
- The insignificant p-value and the closeness of fit of the data to the Poisson distribution indicates that the V-2 impact sites were random rather than clustered.

Point Pattern Data

- Goal in point pattern data analysis is to assess whether there is a spatial pattern in occurrences of an event
- Distinguish between a point and an event location
- In geostatistics our points were locations in a domain that we made a measurement. These points make up a set of spatial random variables for which we wanted to determine the spatial relationships (via the covariance function)
- Point patterns consist of event locations where we are concerned with the presence/absence of an event rather than the value of the measurement at a point.
- We ask the question: are the events that we observe in our domain from a completely random spatial process or are they exhibiting some type of pattern or clustering?
- For point pattern analysis in R we use the package `spatstat`.

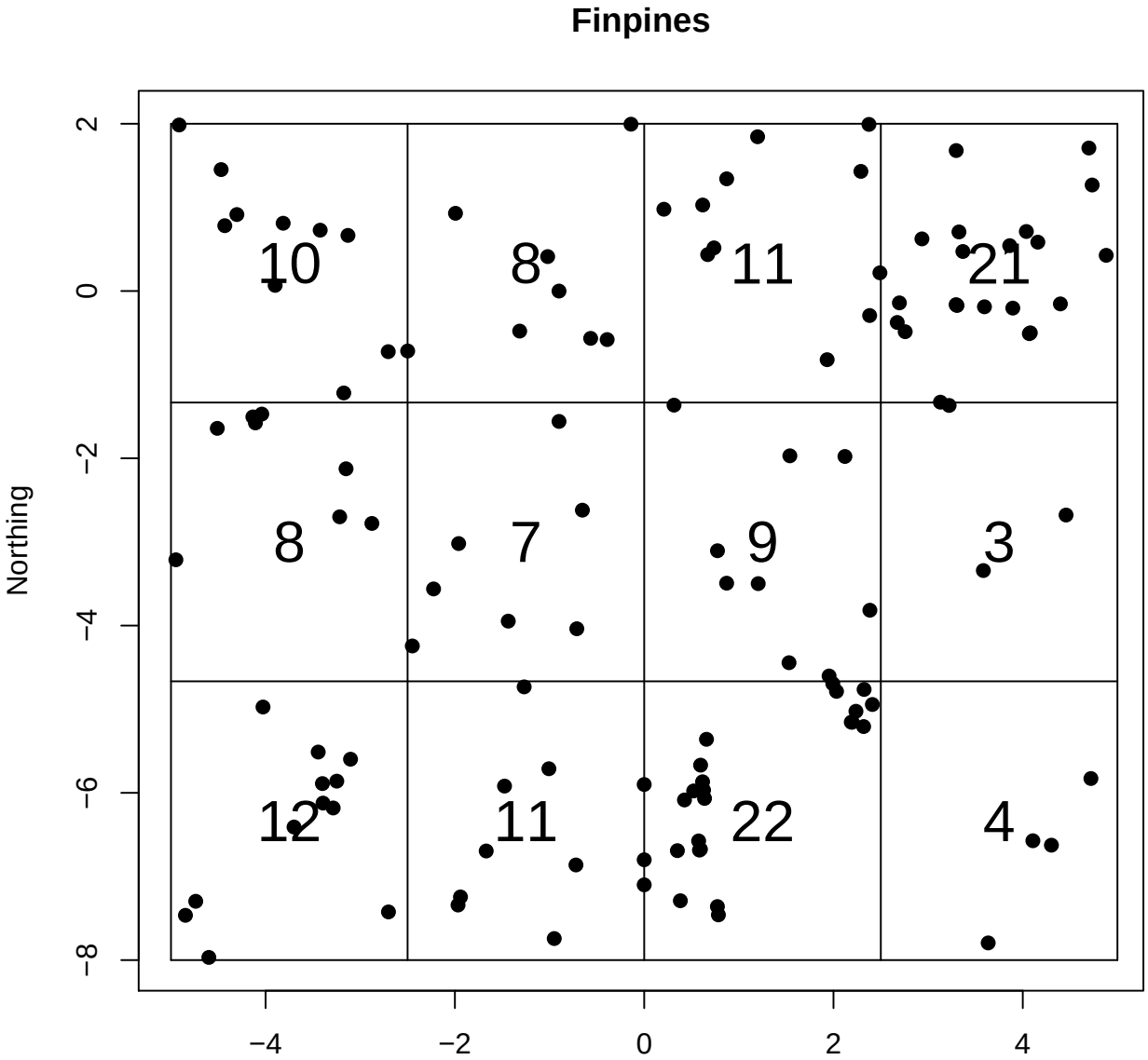
Point Pattern Data: Examples

Here we have the points of locations of three types of trees, each in a rectangular area.



Point Pattern Data: Quadrant Count

A simple quadrant count of the Finnpines points.

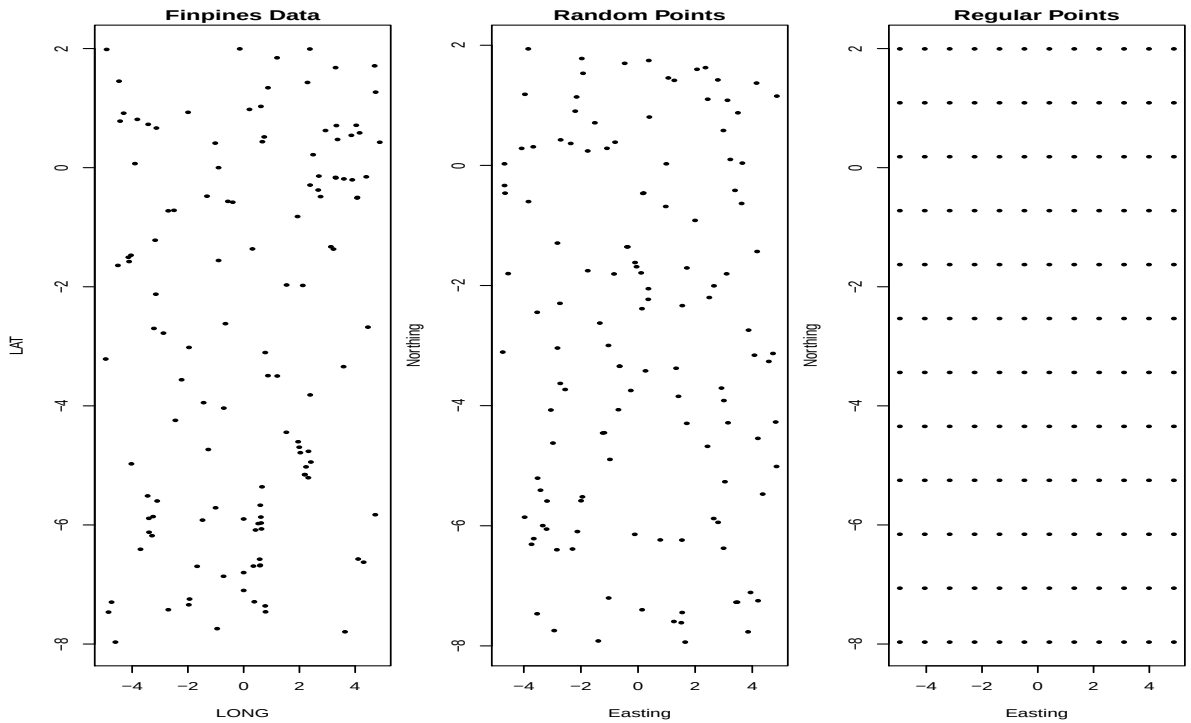


Point Pattern Data

Questions of interest are:

- Are points closer together than they would be by chance?
- Are the points more regularly spaced than they would be by chance?
- What model might reproduce our observed pattern?

Comparing the Finpines data to a spatially random process and a regular pattern:



Point Pattern Data

Point pattern notation:

- Spatial location in (x,y) denoted as s.
- $Y(s)$ represents the presence or absence of Y where $Y(s) = 1$ if there is an observed case at location s, and $Y(s) = 0$ otherwise.
- Spatial domain of observed cases: D, $D = \{s; Y(s) = 1\}$.
- The null hypothesis: no spatial pattern (complete spatial randomness).
- Find a statistic to test whether the data is clustered, or regular.
- Develop a model to generate spatial pattern (PCP, IPP, Cox, SIP).

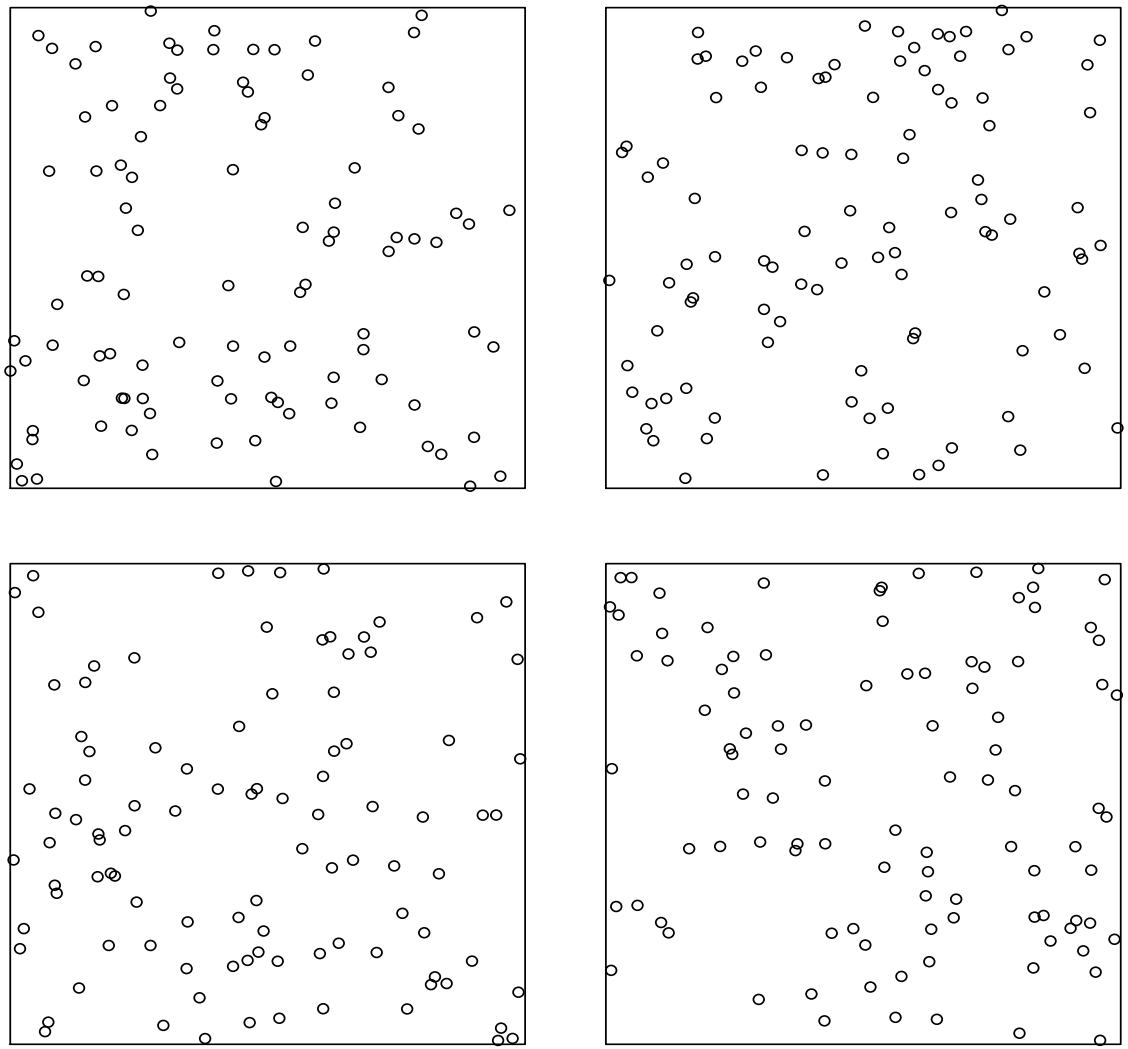
Point Pattern Data: Complete Spatial Randomness

Spatial Randomness

- Typical terms that are used are spatial randomness, random pattern, at random or by chance.
- Complete spatial randomness (CSR): events are uniformly distributed across a domain D and are independent of each other.
- CSR means an event is equally likely to occur at any location or region within D.
- Testing for CSR is the most basic test which can be performed on point pattern data.

Point Pattern Data: Complete Spatial Randomness

Here are 4 realizations of a random uniform process (CSR) in a 1x1 box



Point Pattern Data: Complete Spatial Randomness

- A point process which is CSR pis defined as being a stationary **homogeneous spatial Poisson point process (HPP)**.
- Homogeneous Poisson process have the following characteristics:
  1. For any n events in region D, the events are an independent random sample from a uniform distribution where each point is a location where an event could occur is equally likely to be chosen as an event.
  2. The number of events n occurring within region D is a random variable following a Poisson distribution with mean $\lambda|D|$ where λ is a constant and $|D|$ is the area of the region.

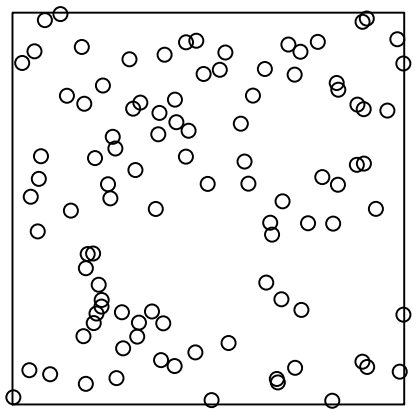
Properties of Homogeneous Poisson Processes

- The Poisson distribution allows the total number of observed events to vary from realization to realization while maintaining a fixed but unknown number of events per unit area.
- The expected number of events per unit area is the intensity, λ where $\frac{n}{|D|} = \lambda$.
- Under a HPP, the location of one point in space does not affect the probabilities of other points appearing nearby. The intensity of the point process in area A is a constnat $\lambda > 0$.

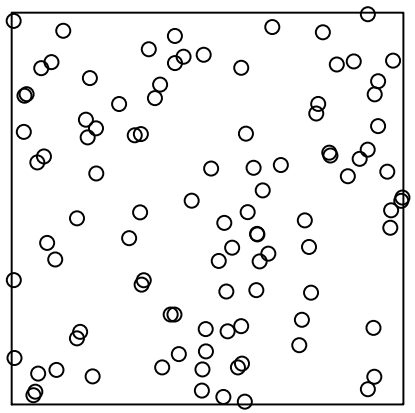
Point Pattern Data: Homogeneous Poisson Process

Here we illustrate two scenarios that result in the same thing:

intensity = 100, unit square



intensity = 1, 10 x 10 square



Properties of Homogeneous Poisson Processes

- The number of events in non-overlapping regions are statistically independent.
- This can be formalized by taking a subregion of a homogeneous Poisson process, $A \in D$
- Can think of this similarly to stationarity

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

This implies that the probability of a single event in an increasingly small area is a constant independent of the location of the region A within D.

$$\lim_{|A| \to 0} \frac{\text{P[two or more events in A]}}{|A|} = 0$$

This implies that the probability of a two or more events in the same location is zero.

Properties of Homogeneous Poisson Processes

- if we let n be the number of events in A, then n following a Poisson distribution means

$$P(n = k) = \frac{exp(-\lambda|A|)(\lambda|A|)^k}{k!}$$

Again, homogeneity is similar to stationarity and isotropy for geostatistical data. The intensity of the point process λ does not vary as a function of spatial location within our domain. Homogeneity means that the intensity is constant across the study area.

Testing for CSR

- Many tests of CSR use Monte Carlo methods.
- Compare the observed value of a test statistic to its distribution under the null hypothesis of CSR.
- Simulate a large number of CSR processes and compare the test statistic from N_{sim} to test statistic from observed.

Testing for homogeneous CSR

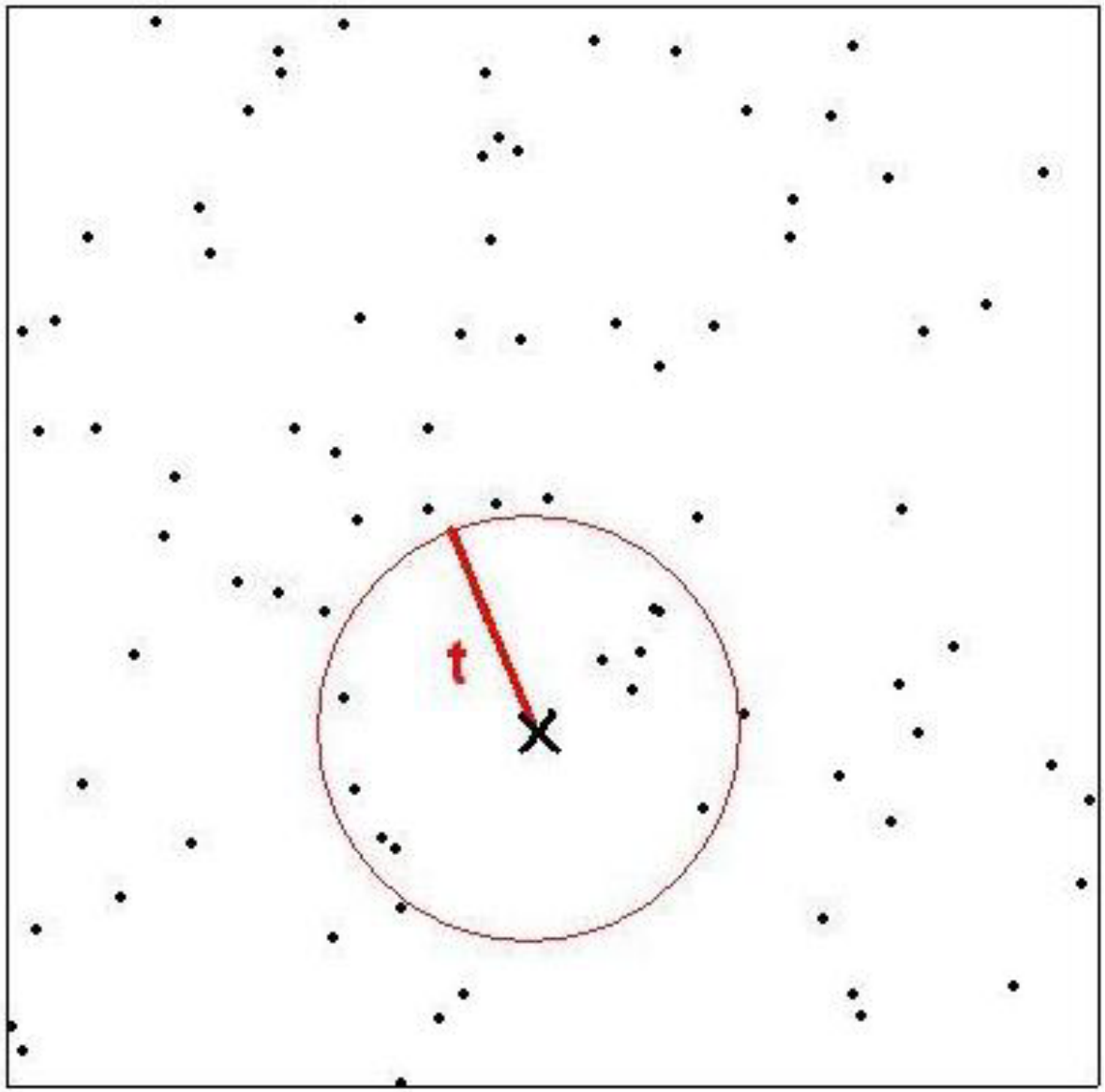
- Ripley's K, $K(h) = \lambda^{-1} E(N_0(h))$
- Where $N_0(h)$ is the number of events within a distance h of an arbitrary event
- $K(h)$ tests the expected number of events within distance h from an arbitrary event (excluding the chosen event itself) divided by the average number of events per unit area
- $K(h)$ is equivalent to showing the variance of the number of events occurring in subregion A (Ripley 1977) so is a second order property of the point process.

Point Pattern Data: Tests of CSR

Testing for homogeneous CSR

- Ripley's K, $K(h) = \lambda^{-1} E(N_0(h))$
- Under CSR $K(h) = \pi h^2$, the area of a circle of radius h

Testing for homogeneous CSR



Testing for homogeneous CSR

- For a process that is more regular than CSR we expect fewer events within distance h of a randomly chosen event
- For a process that is more clustered than CSR we expect more events within distance h of a randomly chosen event
- Estimating $K(h)$:

$$\hat{K}(h) = \hat{\lambda}^{-1} \frac{1}{N} \sum_i \sum_j \delta(d(i,j) < h)$$

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

Testing for homogeneous CSR

- There is an alternate $\hat{K}(h)$ estimator that corrects for edges (boundaries of the region)
- Want to prevent including events that occur outside the boundary but within distance h
- Estimating $K(h)$ accounting for boundaries:

$$\hat{K}_{ec}(h) = \hat{\lambda}^{-1} \frac{1}{N} \sum_i \sum_j w_{ij} \delta(d(i,j) < h)$$

- where $w_{ij} = 1$ if the distance between i and j is less than the distance between event i and the boundary of the region

Point Pattern Data: Tests of CSR

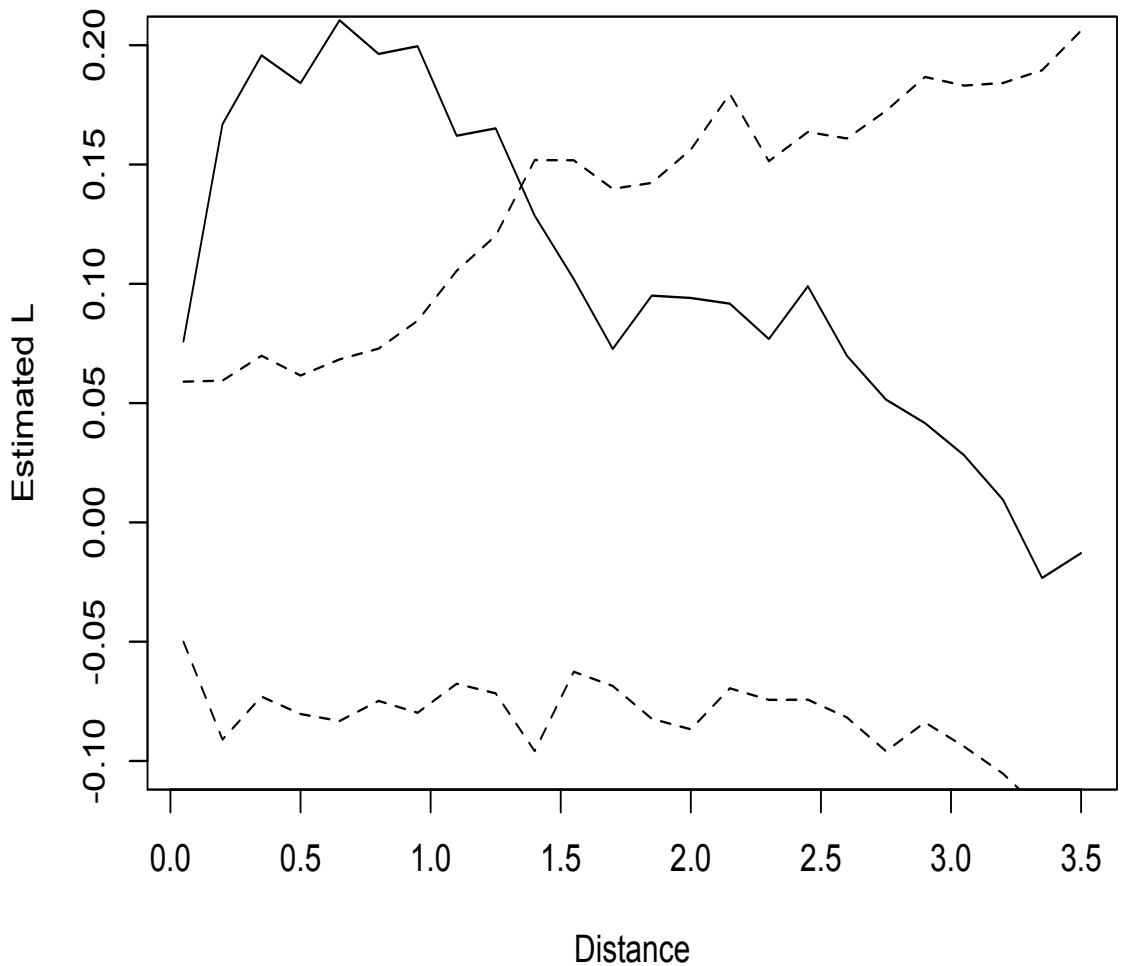
- Using the $K(h)$ function and determining p-values to test CSR.
- Plot $K(h)$; under CSR $K(h) = \pi h^2$ is a parabola.
- $(K(h)/\pi)^{1/2} = h$, so plot h vs $(\hat{K}(h)_{ec}/\pi)^{1/2} - h$.
- Under CSR, $(\hat{K}(h)_{ec}/\pi)^{1/2} - h = 0$.
- Departures from the horizontal line that defines CSR indicate clustering or regularity
- Deviations above the horizontal line indicate clustering because there are more events within distance h than expected.
- Deviations below the horizontal line indicate regularity because there are fewer events within distance h than expected.

Point Pattern Data: Tests of CSR

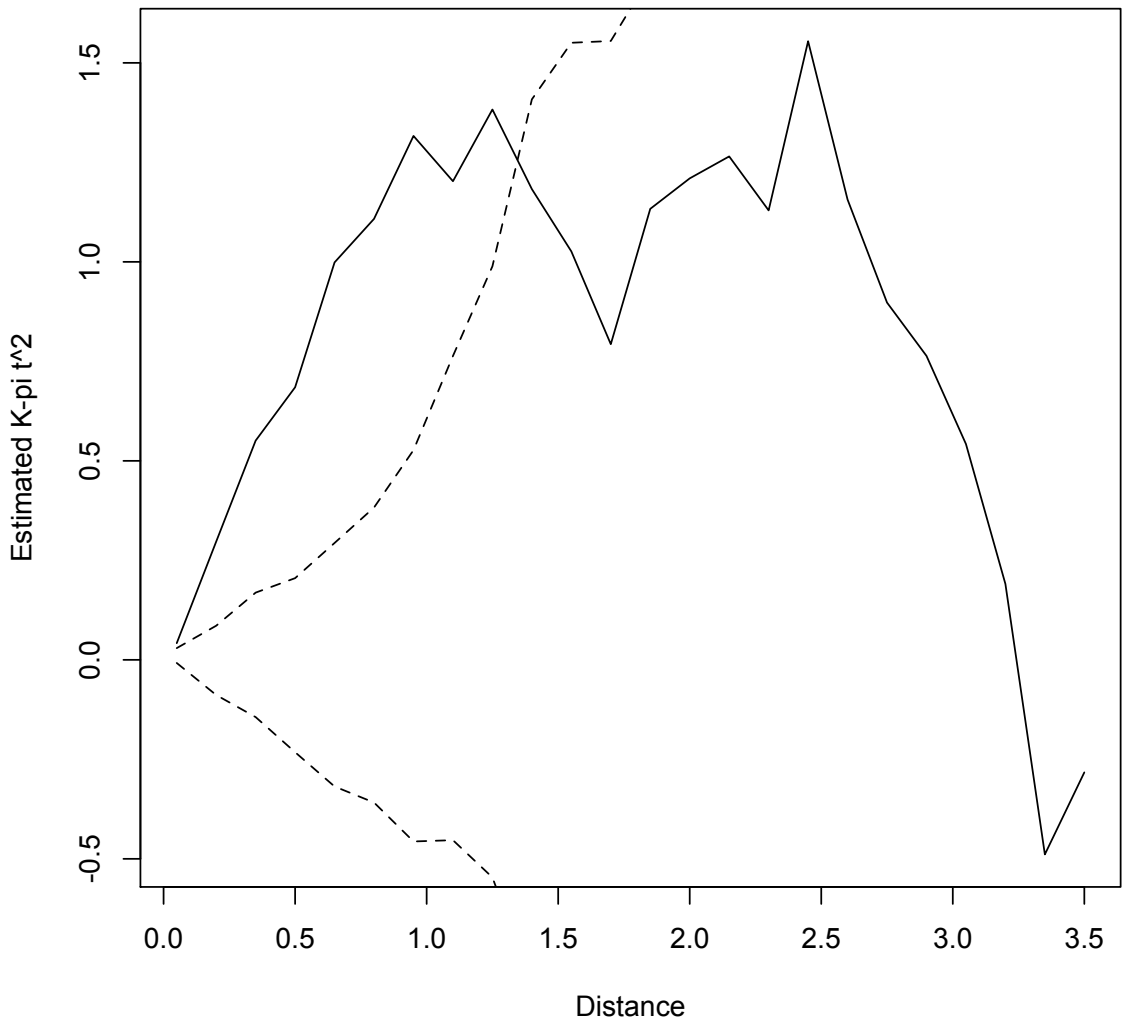
To test CSR based on Riply's K, use Monte Carlo method:

- Simulate k-1 samples (for example, k=100) of n points from a CSR process and compute $\hat{K}_2(h), ..., \hat{K}_n(h)$.
- For each distance h find the upper bound U(h)=max$_i$ $\hat{K}_i(h)$.
- For each distance h find the lower bound L(h)=min$_i$ $\hat{K}_i(h)$.
- For each distance find $\hat{K}_1(h)$ from data.
- Plot distance vs $\hat{K}_i(h) - \pi h^2$.
- Often plot distance (h) vs $\hat{L}(h) - h$ where $\hat{L}(h) = (\hat{K}(h)_{ec}/\pi)^{1/2}$ for better visualization.
- Under CSR, the expected value of $\hat{L}(h) - h = 0$, so we expect a CSR line to be around zero.

Testing for homogeneous CSR



Testing for homogeneous CSR



Point Pattern Data: Tests of CSR

Test statistic

- Order the simulated values $\hat{K}_2(h), ..., \hat{K}_n(h)$, ie for i=1,...,n $\hat{K}_{(1)}(h) \geq ... \geq \hat{K}_{(n)}(h)$
- Let $\hat{K}_{(k)}(h)$ be the kth largest among $\hat{K}_i(h)$
- For each h, the probability that $\hat{K}_1(h)$ is lower than $L(h)$ or higher than $U(h)$ is n^{-1}, i.e.

$$P(\hat{K}_1(h) = \hat{K}_{(k)}(h)) = n^{-1}$$

- We reject the null hypothesis on the basis that $\hat{K}_1(h)$ ranks kth largest or higher. This gives us an exact one-sided test of size k/n

Point Pattern Data: Tests of CSR

- Can also test for CSR based on inter-event distances
- We let $H(h)$ be the theoretical distribution of inter-event distances, $H(h) = P(H \leq h)$
- We estimate $H(h)$ by the empirical distribution function:

$$\hat{H}(h) = \frac{\text{number of paired distances less than h}}{\text{total number of pairs}} = \frac{\#(h_{i,j} \leq h)}{0.5n(n-1)}$$

Point Pattern Data: Tests of CSR

- We know the theoretical $H(h)$ for inter-event distances $h_{i,j}$ if we have an area D which is square or circular.
- Example: Bartlett (1964) described $H(h)$ for a unit circle as

$$H(h) = 1+\pi^{-1}[2(h^2-1)\cos^{-1}(h/2)-h(1+h^2/2)\sqrt{(1-h^2/4)}]$$

- for $0 \leq h \leq 2$

Point Pattern Data: Tests of CSR

- A visual test is to plot $\hat{H}(h)$ vs $H(h)$
- As with Ripley's K we need the distribution of our statistic $\hat{H}(h)$ under CSR
- Simulate k-1 samples of n points from CSR process and compute $\hat{H}_1(h), ..., \hat{H}_s(h)$
- For each inter-event distance h find the upper and lower bounds of $\hat{H}(h)$
- $\hat{H}_U(h) = max_i \hat{H}_i(h)$ and $\hat{H}_L(h) = min_i \hat{H}_i(h)$

Point Pattern Data: Tests of CSR

- Using nearest neighbour distances, let $G(h)$ be the theoretical distribution of NN distances
- Let h_i be the distance from the ith event to the nearest other event in D
- Our estimate of $G(h)$ is

$$\hat{G}(h) = \frac{\#(h_i \leq h)}{n}$$

- Now we need to test our statistic vs the statistic under CSR

Point Pattern Data: Tests of CSR

- Approximation of $G(h)$:
- For any event in our area D, under CSR we have $P(\text{event i is within distance h from j}) = \pi h^2 |D|^{-1}$ where area is represented by $|D|$
- The approximate distribution function of $G(h)$ is $G(h) \approx 1 - (1 - \pi h^2 |D|^{-1})^{n-1}$
- And when you have a large n, $G(h) \approx 1 - exp(-\lambda \pi h^2)$ where $\lambda = n/|D|$ is the intensity as we saw before
- Compare $\hat{G}(h)$ to $G(h)$, find envelope by simulation $\hat{G}_U(h)$ and $\hat{G}_L(h)$
- Plot $G(h)$ vs $\hat{G}(h)$

Point Pattern Data: Inhomogeneous Poisson Process

- A generalization of the HPP is the Inhomogeneous (heterogeneous) Poisson Process (IPP). The IPP occurs when the intensity λ is not constant over the region.
- Many cases homogeneity in intensity is not realistic, for example the locations of trees in a forest may be irregular due to geographic features such as soil, rock, slope or other terrain irregularities.
- In the case of IPP, the intensity is a function that varies spatially, $\lambda(s)$.
- The IPP does not define cluster process, but rather a

Point Pattern Data: Inhomogeneous Poisson Process

Inhomogeneous Poisson process

- We can estimate the intensity function in different ways: parametrically by defining a specific function or non-parametrically using kernel smoothing

$$\hat{\lambda}(s) = \frac{1}{h^2} \sum_i \kappa(\frac{||s - s_i||}{h})/q(||s||)$$

Where $\kappa(s)$ is a kernel function and $q||s||$ is a boundary correction. The distance h is our bandwidth for smoothing

Inhomogeneous Poisson process

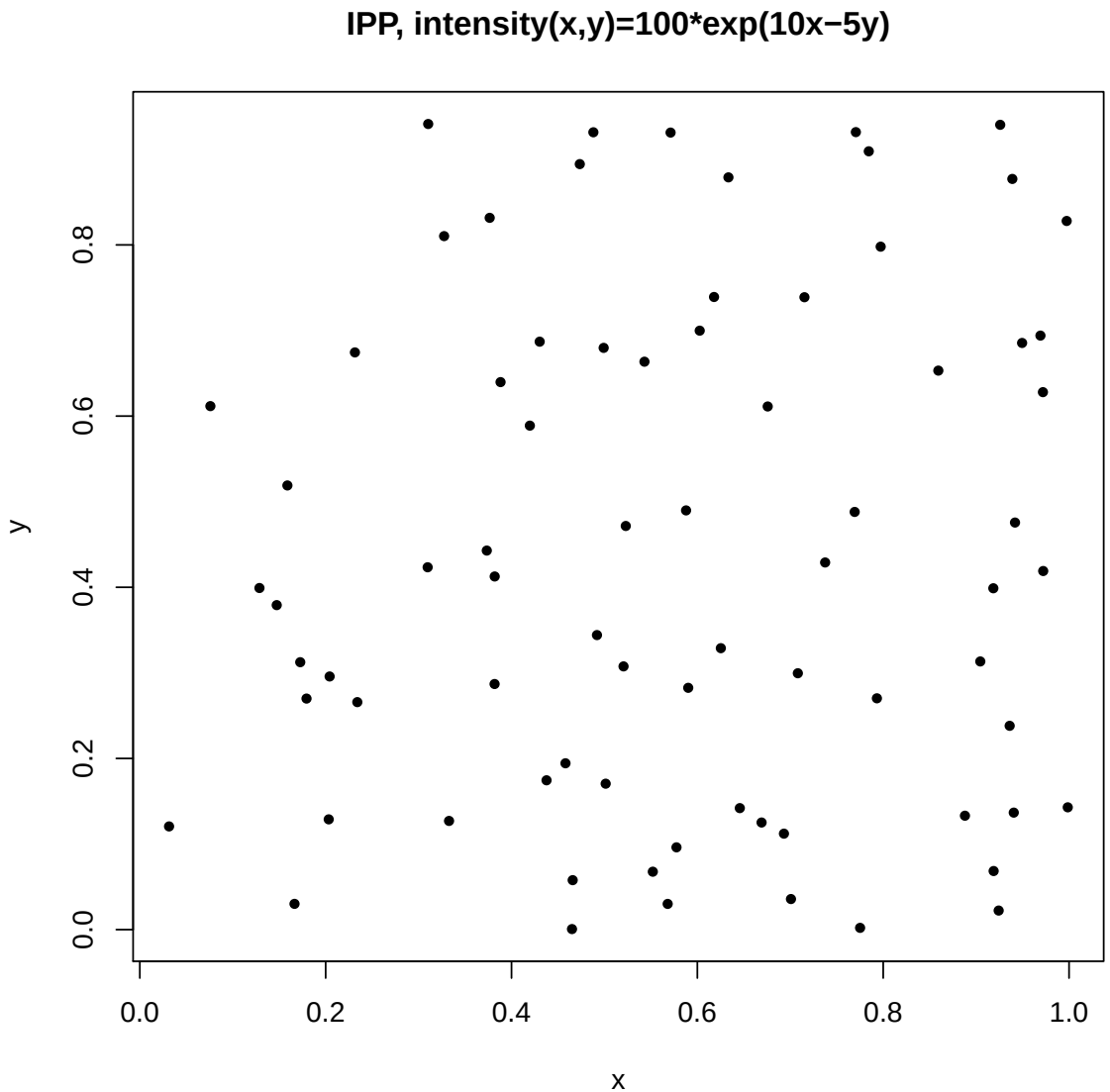
- There are various kernel functions, but a quadratic function is often used:

$$\kappa(s) = \frac{3}{\pi}(1 - ||s||^2)^2$$

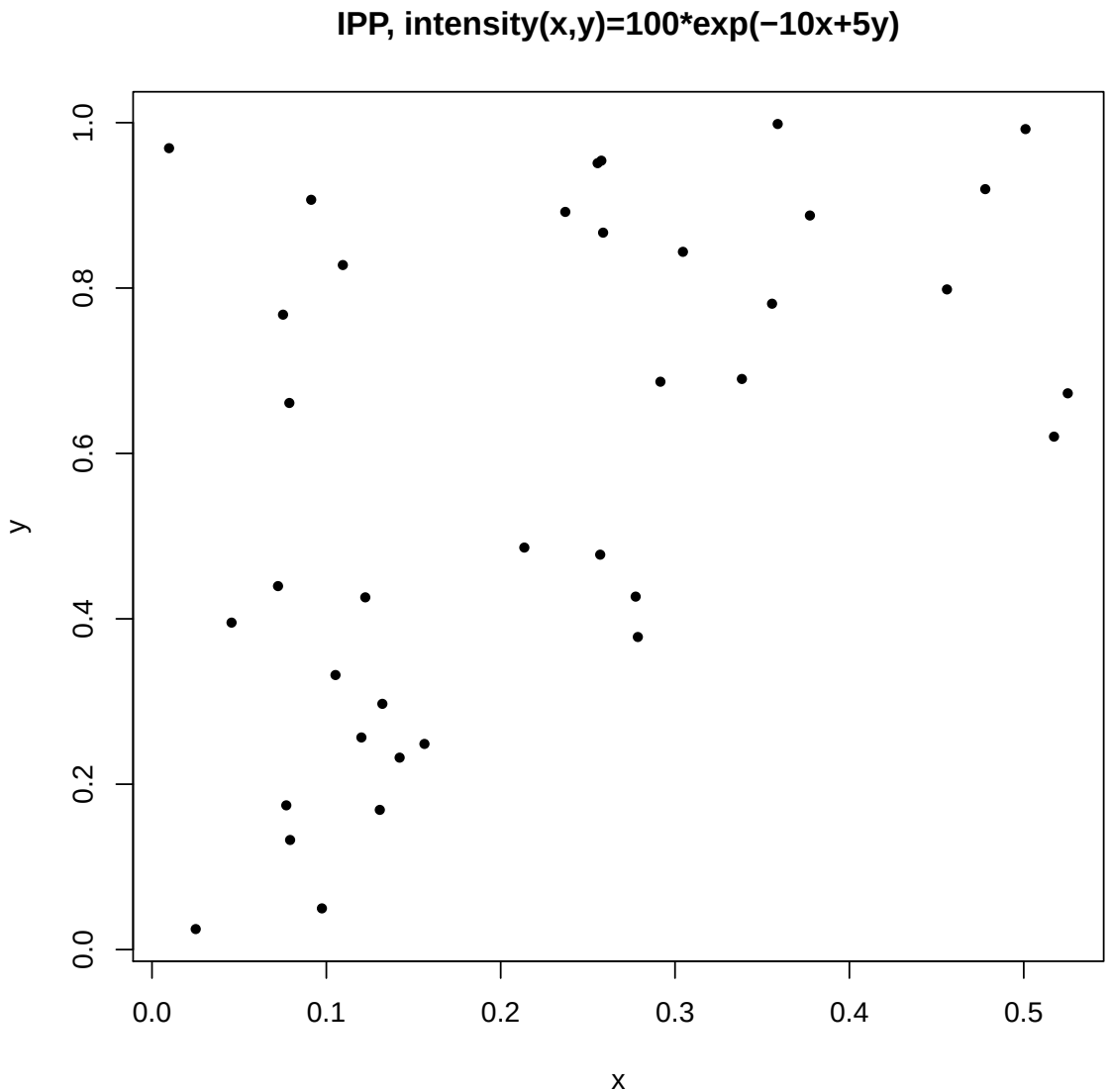
Inhomogeneous Poisson Process

- Example of varying intensity function $\lambda(s)$ could be that intensity varies with location due to environmental heterogeneity
- Example if D i a square unit and N(D)=100
- $\lambda(x, y) = 100 * \exp(10x - 5y)$
- $\lambda(x, y) = 100 * \exp(-10x + 5y)$

Inhomogeneous Poisson Process



Inhomogeneous Poisson Process



Inhomogeneous Poisson Process

- Or we might see that cases of respiratory disease differ with respect to distance from a point source of environmental pollutions_0

$$\lambda(s) = \lambda_0(s) f(||s - s_0||, \theta)$$

- Where $\lambda_0(s)$ models the variation in population density
- f(u,θ) models how the impact of the source varies with distance u (s-s_0) and angle θ

Point Pattern Data: Poisson Cluster Process

Poisson cluster process

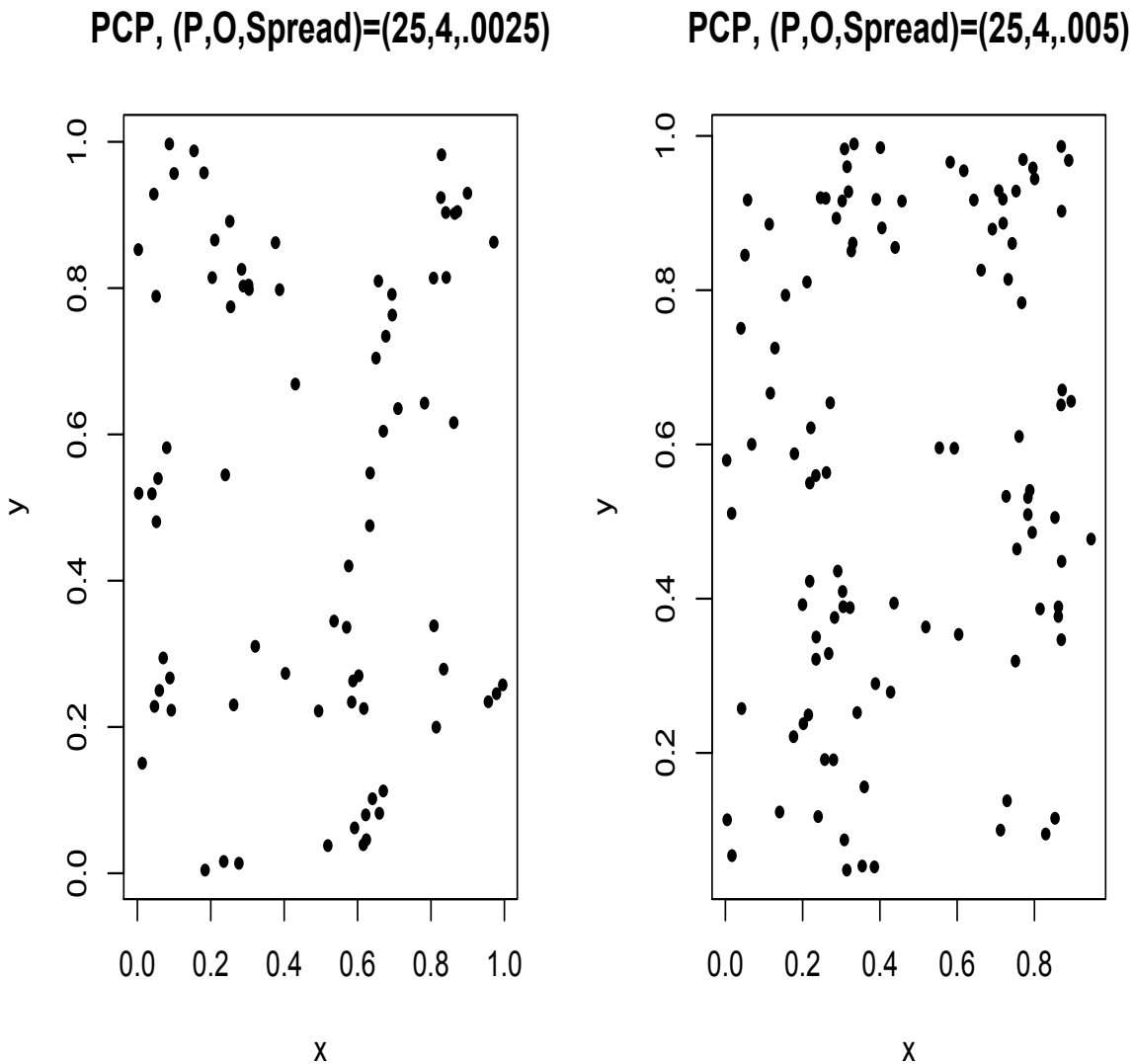
- A spatial point process where each event belongs to a cluster
- There is a parent event that produces a random number of offspring
- Parent events are usually a realization of an Poisson process with intensity $\lambda(s)$
- We have i parents, and each parent produces a random number of offspring, O_i
- The O_i are distributed within h_i of the parent and follow a bivariate probability distribution

Point Pattern Data: Poisson Cluster Process

Poisson cluster process

- Can have homogeneous cluster processes where the intensity of the offspring around a parent is constant λ
- Or an inhomogeneous cluster process where the intensity of the offspring around a parent is not constant across domain $\lambda(s)$
- Parent events are usually a realization of an inhomogeneous Poisson process with intensity $\lambda(s)$ distribution

Poisson cluster process



Poisson cluster process

- Left, we have a unit square as our D with intensity of parents $= 25$ and number of offspring $= 4$ and variation around parents $= 0.00025$
- Right, we have a unit square as our D with intensity of parents $= 25$ and number of offspring $= 4$ and variation around parents $= 0.005$

Poisson cluster process

- Example of PCP: distribution of insect larvae or tree seeds
- Neyman Scott assumptions of homogeneous Poisson cluster process:
  - Parent events are realizations of a Poisson process with intensity ρ
  - Each parent i produces a random number of offspring S_i and the $S - i$ are iid
  - The positions of offspring wrt the parent are iid with bivariate pdf

Poisson cluster process

- Isotropy in our pdf means it must be radially symmetric
- Example is the radially symmetric Gaussian distribution

$$h(x_1, x_2) = \frac{1}{2\pi\sigma^2} \exp(-\frac{x_1^2 + x_2^2}{2\sigma^2})$$

- where σ^2 are the offspring

Poisson cluster process

- Neyman Scott assumptions of homogeneous Poisson cluster process
- Use ρ for intensity of parents, $E(s) = \mu$ is the expected number of offspring
- Then the overall intensity of a clustered process is $\lambda = \rho\mu$ (1st order intensity)
- Need second order properties of this process to derive $K(h)$ under PCP
- Second order intensity $\lambda_2 = \lambda^2 + \rho E[S(S-1)]h_2(s_i - s_j)$