GEOG574/Math Introduction to Geostatistics

Point Pattern Analysis

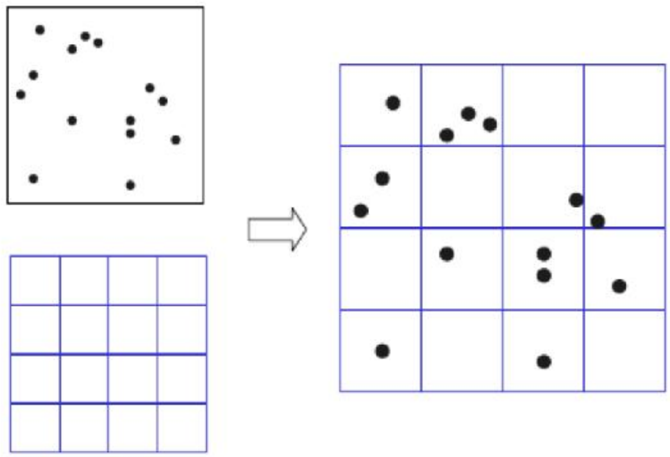
Daoqin Tong

School of Geography &Development
408 Harvill Building
Email: daoqin@email.arizona.edu

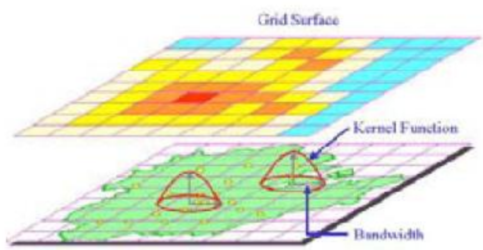
Point Patterns

- A spatial point process is a spatial stochastic process $\{Y(s), s \in \Re\}$, where $\Re$ is random
- A spatial point pattern $\{y(s_1), y(s_2),..., y(s_n); s_i \in R\}$ is a realization of a spatial point process
- Exploring Spatial Point Patterns
  - First order effects
    - ⇒ Quadrat methods
    - ⇒ Kernel estimation
  - Second order effects
    - ⇒ Nearest neighbor distances
    - ⇒ K function

Quadrat Methods

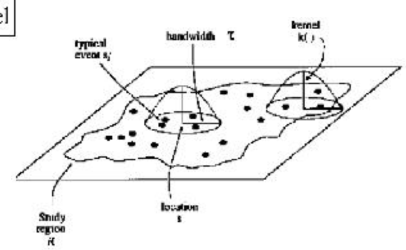


Kernel Estimation



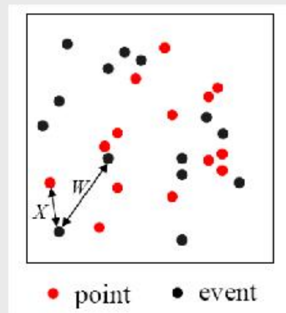
$$\hat{\lambda}_\tau(s) = \frac{1}{\delta_\tau(s)} \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{(s - s_i)}{\tau}\right)$$

where $\delta_\tau(s)$ is the edge correction term;
$k(\bullet)$ is kernel function; τ is bandwidth of kernel



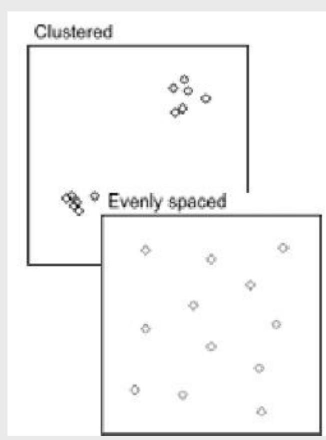
Nearest Neighbor Distances

- The distribution of inter-event distances is closely related to the second order effects
- Two types of distances
  - W: event-event nearest neighbor distances (distances between a randomly chosen event and its nearest neighbor event)
  - X: point-event nearest neighbor distances (distances between a randomly chosen point in the study area and its nearest neighbor event)



Expected behavior of mean nearest neighbor distance

- For a *clustered* pattern
  - All nearest neighbor distances are short
  - So the mean is small
- An *evenly-spaced* pattern
  - Minimum distances are longer
  - Mean is higher



Distance-based Measures

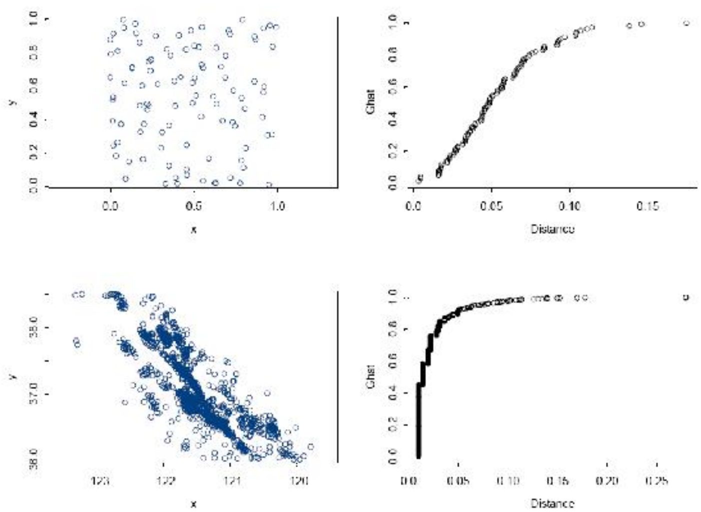
- *W* vs. *X*
  - To get *W*, a complete enumeration of all events is required
  - *X* can be used with random sampling
- Spatial dependence in point patterns can be explored through the observed distributions of *W* or *X*
- Two empirical cumulative probability distribution functions

$$\hat{G}(w) = \frac{\#(w_i \le w)}{n}; \qquad \hat{F}(x) = \frac{\#(x_i \le x)}{m}$$

n: the number of events; m: the number of random points

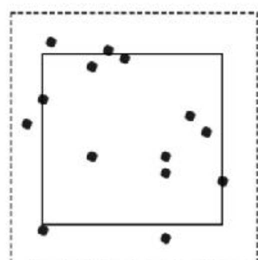
Distance-based Exploratory Analysis

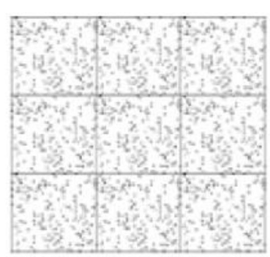
- Plot of *G(w)* against *w* or *F(x)* against *x*
  - If the empirical cumulative distribution function climbs very sharply in the early part before flattening out, then it indicates clustering
  - If it climbs very slowly at the beginning followed by steeply climbing, then it indicates a repulsion or regular pattern
- Plot of *G(w)* against *F(x)*
  - If no interaction, the two values should be very similar
- How to quantify these visual inspections?



Distance-based Exploratory Analysis

- Edge effects
  - Guard area
  - Toroidal edge correction
  - Modified formula



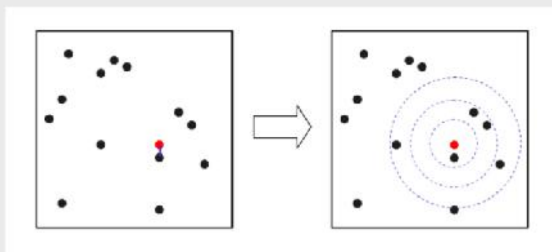


$$\hat{G}(w) = \frac{\#(b_i > w \geq w_i)}{\#(b_i > w)}$$

b_i is the distance from event I to the nearest point on the boundary of R

The K Function

- Nearest neighbor distance analysis
  - Only consider nearest neighbor distances
- The K function
  - Reduced second order moment measure
  - Consider a range of distances



The K Function

- Need to be sure that it is valid to examine second order effects at which scale
- Need to make some assumption
  - Homogeneous or isotropic at the considered scale
  - Otherwise, it is not possible to estimate the second order effects directly from the observed patterns
  - Also, any second order effect can be due to the variation in the first order effect

Definition

- a$K(h)$=E(# of events within distance *h* of an arbitrary event)
- The expected # of events in $\Re$ is a$|\Re|$
- The expected # of ordered pairs within distance *h* is a$^2|\Re|K(h)$
- The K can be estimated by

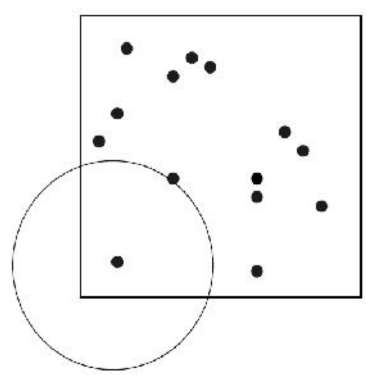
$$\hat{K}(h) = \frac{1}{\lambda^2 |\Re|} \sum \sum_{i \neq j} I_h(d_{ij})$$

Edge Correction

$$\hat{K}(h) = \frac{1}{\lambda^2 |\Re|} \sum \sum_{i \neq j} \frac{I_h(d_{ij})}{w_{ij}}$$

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

w_{ij} is the proportion of circumference of circle centered on the *i*th event and passing through *j*th event



Estimation of Intensity

$$\lambda = \frac{n}{|\Re|}$$

$$\hat{K}(h) = \frac{|\Re|}{n^2} \sum \sum_{i \neq j} \frac{I_h(d_{ij})}{w_{ij}}$$

Modeling Spatial Point Patterns

- Exploratory analysis of spatial point patterns is rather informal and may not be sufficient
- Need more formal analysis
  - Hypothesis testing: comparing summary measures calculated from an observed point pattern with the expected observations under various hypothesized models
  - Statistical modeling: constructing specific models to explain observed patterns

Complete Spatial Randomness

- Complete Spatial Randomness (CSR)-events follows a homogenous Poisson process over the study area
- Consider $\{Y(A), A \in \Re\}$
  - The probability distribution of $Y(A)$ is a Poisson distribution with mean $a|A|$, where a is a constant

$$f_{Y(A)}(y) = \frac{(a|A|)^y}{y!} e^{-a|A|}$$

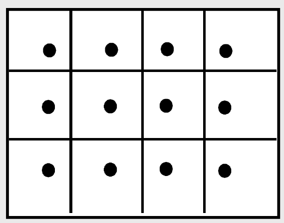
  - $Y(A_i), Y(A_j)$ are independent for any A_i and A_j

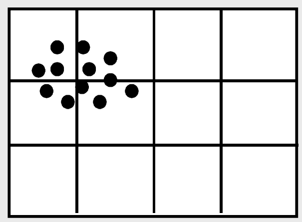
CSR

- CSR implies that conditional on n, events are independently and uniformly distributed over R
  - Any event has an equal probability of occurring at any location in R
  - The location of any event is independent of that of any other
  - CSR can be simulated by generating random events from uniform distribution over R
- CSR provides a baseline hypothesis for testing
  - Regular, clustered, or random

Quadrat count method

- *Equally-spaced* patterns will have most quadrats with similar counts
- *Clustered* patterns will have a few high count quadrats and many which are empty





Statistics of CSR

- The number of events that fall in any quadrat conforms to a *binomial distribution*

$$P(k \text{ events in quadrat}) = \binom{n}{k} \cdot \left(\frac{1}{m}\right)^{k} \left(1 - \frac{1}{m}\right)^{n-k}$$

where

k is the number of events in a quadrat,
n is the total number of events, and
m is the total number of quadrats
1/*m* is the fraction of the region occupied by a quadrat

$$Exp(\text{the number of quadrats that contain } k \text{ events}) = P \times m$$

Poisson distribution

- The binomial distribution is difficult to calculate because *n*! is often very large

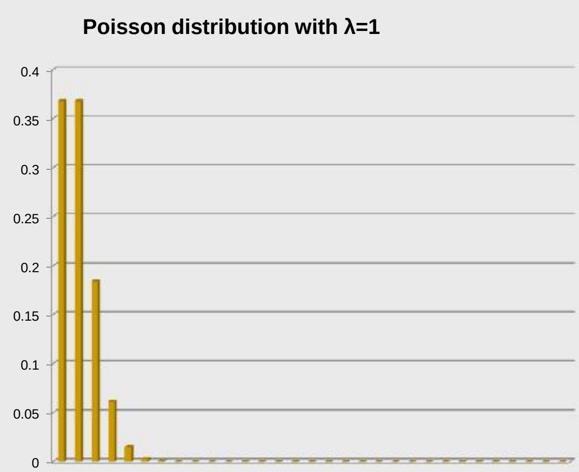
$$50! \approx 30{,}000$$

- So in practice we use the *Poisson distribution* as an approximation:

$$P(k \text{ events in quadrat}) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad a = \frac{n}{m}$$

Expected results for 30 events in 30 quadrats

k	p(k)	p(k)*30
0	0.367879	11.03638
1	0.367879	11.03638
2	0.18394	5.518192
3	0.061313	1.839397
4	0.015328	0.459849
5	0.003066	0.09197
6	0.000511	0.015328
7	7.3E-05	0.00219
8	9.12E-06	0.000274
9	1.01E-06	3.04E-05
10	1.01E-07	3.04E-06
11	9.22E-09	2.76E-07
12	7.68E-10	2.3E-08
13	5.91E-11	1.77E-09
14	4.22E-12	1.27E-10
15	2.81E-13	8.44E-12
16	0	0
17	0	0
18	0	0
19	0	0
20	0	0



Quadrat Tests for CSR

- Divide the study region into m quadrats of equal size
- Find the mean number of points per quadrat $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$
- Find the variance of the number of points per quadrat $s^2 = \frac{1}{m-1}\sum_{i=1}^{m} (xi - \bar{x})^2$
- Calculate the variance-mean ratio (VMR) $VMR = \frac{s^2}{\bar{x}}$

Interpretation

- If VMR ≈1 → expected variation
  - Points approximate random distribution across the study region
- If VMR <1 → less variation than expected
  - Points are spread out or uniform across the study region
- If VMR >1 → more variation than expected
  - Point are more clustered distributed across the study region

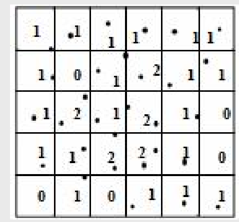
Variance/mean Ratio (VMR)

- Poisson distribution: mean = variance
- *variance/mean ratio* (VMR)

  VMR<1 -> little variation (uniform)

  VMR=1 -> random

  VMR>1 -> good deal of variation (cluster)



$$\bar{x} = \frac{n}{m} = \frac{30}{30} = 1$$

n - # of events

m - # of quadrats

x_i - # of events in the i^{th} quadrat

$$s^2 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{m-1}$$

$$= \frac{5\times(0-1)^2 + 20\times(1-1)^2 + 5\times(2-1)^2}{30-1}$$

$$= 0.345$$

VMR = 0.345 <1 => significantly uniform

Hypothesis Testing

- Hypotheses
  - H_0: Point pattern is random
  - H_A: Point pattern is not random (clustered, or regular)
- Statistical test
  - Chi-square test (df=m-1)
- Test statistic

$$\chi^2_{test} = (m-1)VMR = \frac{(m-1)s^2}{\bar{x}} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{\bar{x}}$$

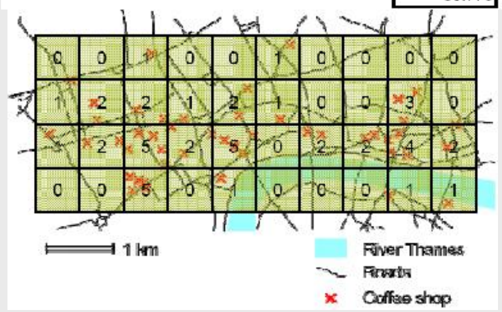
$\chi^2_{1-\alpha,m-1} < \chi^2_{test} < \chi^2_{\alpha,m-1} \Rightarrow$ random

$\chi^2_{test} < \chi^2_{1-\alpha,m-1} \Rightarrow$ regular

$\chi^2_{test} > \chi^2_{\alpha,m-1} \Rightarrow$ clustered

Example on VMR

# events	# quadrats	Difference from expected	Difference squared	Total difference squared
0	18	-1.175	1.380625	24.85125
1	9	-0.175	0.030625	0.275625
2	8	0.825	0.680625	5.445
3	1	1.825	3.330625	3.330625
4	1	2.825	7.980625	7.980625
5	3	3.825	14.630625	43.891875
				85.775



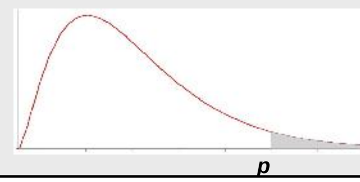
- There are 47 events in 40 quadrats, giving an expected 1.175 per quadrat

$$\chi^2 = \frac{\sum (x_i - \bar{x})^2}{\bar{x}} = \frac{85.775}{1.175} = 73$$

$$df = 40 - 1 = 39$$

$$p(\chi^2) < 0.01$$

Therefore, the coffee shops are clustered



Indices

- Index of Dispersion $\frac{s^2}{\bar{x}}$
- Index of Cluster Size (ICS): $\frac{s^2}{\bar{x}} - 1$
  - Under CSR, E(ICS) =0
  - If ICS>0, the clustering is implied. ICS is the number of extra events
  - If ICS <0, the regularity is implied, ICS is the deficiency in events

Sampling

- Quadrat tests can be used in conjunction with the sampling of point patterns
  - Only need to count m quadrats which are randomly scattered over the study region
- Use sampled quadrats to estimate the intensity of the study region and its associated confidence interval

Estimated intensity

$$\hat{\lambda} = \frac{\bar{x}}{|Q|}$$

$$\operatorname{var}(\hat{\lambda}) = \frac{\hat{\lambda}}{m|Q|}$$

95% confidence interval

$$\hat{\lambda} \pm 2\sqrt{\frac{\hat{\lambda}}{m|Q|}}$$

Limitation of Quadrat Tests

- The relative location of quadrats (if sampling is used) or the relative position of events within a quadrat is not take into account

Nearest Neighbor Tests

- Under CSR, events are independent and $Y(A)$,the # of events in any area A is Poisson distributed

$$f_{Y(A)}(y) = \frac{(a|A|)^y}{y!} e^{-a|A|}$$

- Let x be the radius of a circle, the probability that no events fall within the circle is
- The distribution function F(x) of nearest neighbor point-event distances for CSR is $F(x) = P(X \leq x) = 1 - e^{-a \circledR x^2}$

Nearest Neighbor Tests for CSR

- Mean and variance of X

$$E(X) = \frac{1}{2\sqrt{a}}; \quad Var(X) = \frac{4-\circledR}{4a\circledR}$$

- Similarly, the distribution function G(w) of nearest neighbor event-event distances for CSR is

$$G(w) = P(W \leq w) = 1 - e^{-a \circledR w^2}$$

$$E(W) = \frac{1}{2\sqrt{a}}; \quad Var(W) = \frac{4-\circledR}{4a\circledR}$$

Nearest Neighbor Tests: Issues

- The theoretical distributions of *X* and *W* allow us to derive sampling distributions under CSR of summary statistics for observed nearest neighbor distances
  - The sampling distributions require independent sample point-event or event-event distances
  - Theoretical distributions do not consider edge effects
- Issues
  - Independence assumption may not be valid when large portions of events are used
  - Edge effects will introduce bias

- Tests are based on summary statistics of m randomly sampled nearest neighbor event-event distances $(w_1,\ldots,w_m)$ or point-event distances $(x_1,\ldots,x_m)$
  - Sample size m can be chosen to be less then 0.1n
- Three common tests
  - Clark-Evans
  - Hopkins
  - Byth & Ripley

Clark-Evans

- Randomly sampled m nearest neighbor event-event distances ($w_1,\ldots,w_m$)

- Calculate the sample mean $\overline{w} = \frac{\sum w_i}{m}$
- Compare with the sampling distribution of $\overline{w}$

$$N(\frac{1}{2\sqrt{a}}, \frac{4-\circledR}{4a\circledR m})$$

  – Need to enumerate point pattern completely
  – Need to estimate intensity $\hat{a} = \frac{n}{|\Re|}$

Clark-Evens Test

- It is desirable to use all the nearest neighbor event-event distances ($w_1,\ldots,w_n$)
- Correction to mean and variance

$$E(\overline{w}) = 0.5\sqrt{\frac{|R|}{n}} + 0.051\frac{P}{n} + 0.041\frac{P}{n^{\frac{3}{2}}}$$

$$Var(\overline{w}) = 0.07\frac{|R|}{n^2} + 0.037P\sqrt{\frac{|R|}{n^5}}$$

- Compare with the sampling distribution of $\overline{w} = \frac{\sum w_i}{m}$

$$N(E(\overline{w}), Var(\overline{w}))$$

Hopkins

- Randomly sampled m nearest neighbor event-event distances ($w_1,\ldots,w_m$) and point-event distances ($x_1,\ldots,x_m$)
- Calculate the ratio of the two sample means

$$H = \frac{\sum x_i^2}{\sum w_i^2}$$

- Compare with the sampling distribution of *H*

$$F_{2m,2m}$$

Byth & Ripley

- Randomly sampled m nearest neighbor event-event distances ($w_1,\ldots,w_m$) and point-event distances ($x_1,\ldots,x_m$)
- Calculate the statistic

$$BR = \frac{1}{m}\frac{\sum x_i^2}{\sum (x_i^2 + w_i^2)}$$

- Compare with the sampling distribution of *BR*

$$N(\frac{1}{2}, \frac{1}{12m})$$

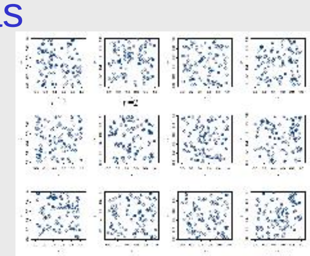
Nearest Neighbor Tests for CSR

- The previous tests only utilize one single summary statistic
- With mapped point patterns, it is possible to compare the estimated distribution function $\hat{G}(w)$ or $\hat{F}(x)$ with their theoretic values $G(w)$ or $F(x)$ under CSR
  - Need to make edge correction for estimated distribution functions
- However, it is hard to assess the significant of the deviation between the estimated from theoretic

Simulation

- Simulate m point patterns with n points under CSR
- Calculate $\hat{G}_i(w)$ for each point pattern without edge correction
- The theoretic value of $G(w)$ without edge correction is estimated as

$$\overline{G}(w) = \sum \hat{G}_i(w) / m$$



- The observed *G(w)*

$$\overline{G}(w)$$

Nearest Neighbor Tests: Simulation

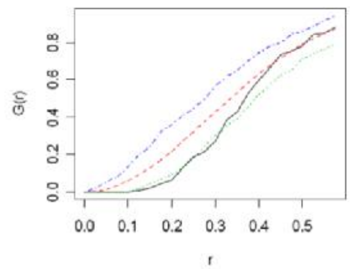
- Constructing the confidence envelopes

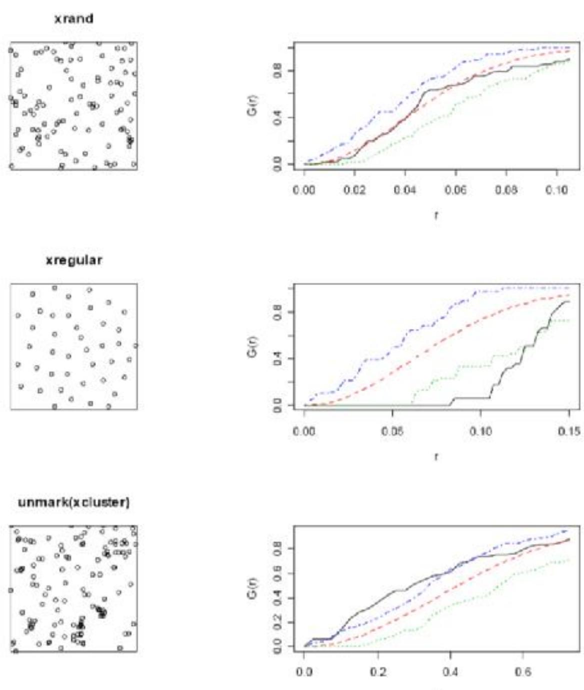
$$U(w) = \max_{i=1,\dots,m} \left\{ \hat{G}_i(w) \right\}$$

$$l(w) = \min_{i=1,\dots,m} \left\{ \hat{G}_i(w) \right\}$$

$$P(\hat{G}(w) > U(w)) = P(\hat{G}(w) < l(w)) = \frac{1}{m+1}$$

- Plot $G(w), \hat{G}(w), U(w), l(w)$





K Function Tests for CSR

- Under CSR, the expected # of events within distance h of a randomly chosen event is $ª®h^2$
- So under CSR, $K(h) = ®h^2$
- Then the empirical K function $\hat{K}(h) = \frac{1}{\lambda^2 |\Re|} \sum \sum_{i \neq j} I_h(d_{ij})$ can be compared with $K(h) = ®h^2$
- Or compared with the transformed form, *L(h)*=0

$$\hat{L}(h) = \sqrt{\frac{\hat{K}(h)}{®}} - h$$

- How to evaluate the significance?

K Function Tests: Simulation

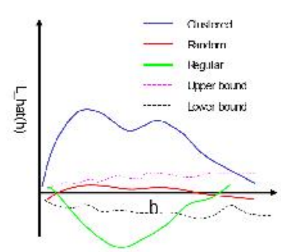
- Simulate m point patterns with n points under CSR
- Constructing the confidence envelopes

$$U(h) \quad \max_{i=1,\dots,m} \{\hat{L}_i(h)\}$$

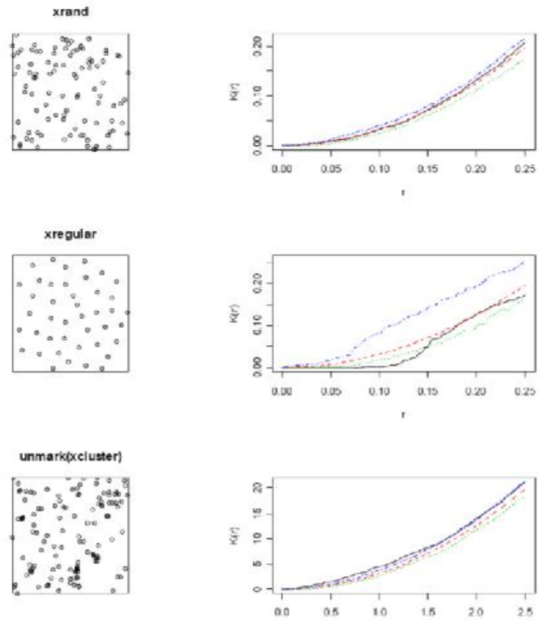
$$l(h) = \min_{i=1,\dots,m} \{\hat{L}_i(h)\}$$

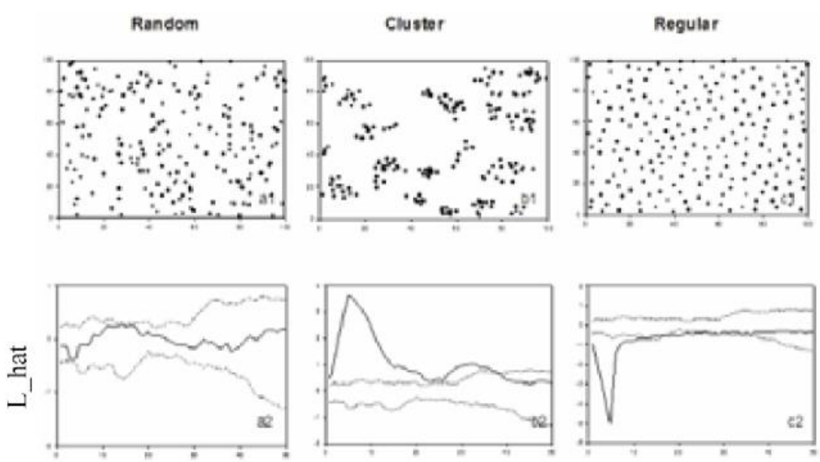
$$P(\hat{L}(h) > U(h)) = P(\hat{L}(h) < l(h)) = \frac{1}{m+1}$$

- Plot $\hat{L}(h), U(h), L(h)$



K Function Tests: Simulation





Population Problem

- A serious difficulty with cluster detection is something we’ve been ignoring so far
  - The background population is not randomly distributed
  - In epidemiology the term used is the ‘at-risk’ population
- What other processes might we use?
  - Heterogeneous Poisson process
  - The Cox process