Methods for Spatial Point Patterns

Bailey and Gatrell
Chapter 3

Lecture 9

October 1, 2009

Modeling Spatial Point Patterns

Objective

- To test hypotheses
- To construct specific models to explain observed patterns

Carry out statistical comparison of summary measures calculated from observed events with what we would expect from various hypothesized models

The standard model against which a spatial point pattern is compared is

Complete spatial randomness

Complete Spatial Randomness (CSR)

Reasons for beginning an analysis with a test for CSR:

- Rejection of CSR is a prerequisite for any serious attempt to model an observed pattern
- Tests are used to explore a set of data and assist in the formulations of alternatives to CSR
- CSR operates as a dividing hypothesis between regular and clustered patterns

Tests for complete spatial randomness

A hypothesis of complete spatial randomness for a spatial point pattern $\{Y(A), A \in R\}$ asserts that:

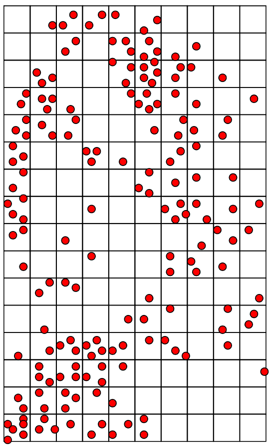
a). The number of events in any planar region with area A follows a Poisson distribution with mean λA

$$f_{y(A)}(y) = \frac{(\lambda A)^y}{y!} e^{-\lambda A}$$

b). Given n events in A, the events are an independent random sample from a uniform distribution on A

a) implies constant intensity – no first order effects

b) implies no spatial interaction- no second order effects



Number of Discoveries per Tract	Number of observed discoveries	Expected Frequency	Expected number of discoveries
0	70	0.3499	56.0
1	42	0.3674	58.8
2	26	0.1929	30.9
3	17	0.0675	10.8
4	3	0.0177	2.8
5	1	0.0372	0.6
6	1	0.0065	0.1
	160		

$$\Pr_{(0)} = e^{(-1.05)}\frac{1.05^0}{0!} = 0.34994$$

$$f_{y(A)}(y) = \frac{(\lambda A)^y}{y!}e^{-\lambda A}$$

Estimated mean= 1.05

Quadrat tests for CSR

Data consist of counts of events ($n_1, n_2, \ldots n_m$) in m quadrats

A simple test statistic uses the Poisson distribution characteristic of equal mean and variance.

One test statistic is the sample variance-to-mean ratio $\frac{s^2}{\bar{x}}$
or *index of dispersion*

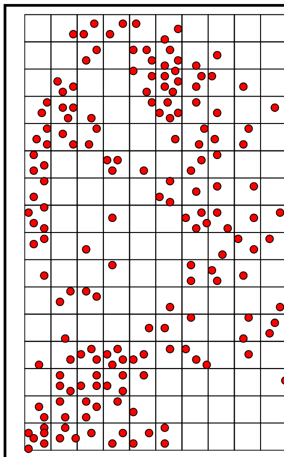
Test statistic

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

$\bar{x}$ is the mean of observed counts, s^2 is the observed variance

Quadrat tests for CSR

- Under CSR the test statistic I is distributed as χ^2_{m-1}
- Compare test statistic I with percentage points of χ^2_{m-1}
- Significantly large values indicate clustering
- Significantly small values indicate regularity



	Number of Discoveries per Tract (n)	Number of tracts with n discoveries (q)	Number of discoveries by quadrat (X)	$n^2q=X_2$
	0	70	0	0
	1	42	42	42
	2	26	52	104
	3	17	51	153
	4	3	12	48
	5	1	5	25
	6	1	6	36
Sum		160	168	408

$$s^2 = \frac{\sum X_2}{\sum X} - \frac{\sum X}{N} \qquad \bar{x} = 1.05$$

$$s^2 = 1.378 \qquad \text{Compare to } \chi^2_{m-1}$$

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

188.33

$$I = 208.66$$

Quadrat tests for CSR

$$\left(\frac{s^2}{\bar{x}}\right) - 1$$ *Index of cluster size (ICS)*

$E(ICS) = 0$ Under CSR

If ICS > 0 then clustering is implied

If ICS < 0 then regularity is implied

Example $s^2 = 1.378$

$\bar{x} = 1.05$

ICS = 0.312

Indicates clustering

9

Quadrat tests for CSR

$$\left(\frac{s^2}{\bar{x}}\right)$$ Index of dispersion

H_0: Index of Dispersion = 1

H_1: Index of Dispersion $\neq$ 1

Can use a z score to test the hypothesis

$$z = \frac{ID_o - ID_e}{\sqrt{2(n-1)}} \qquad z = \frac{1.312 - 1}{\sqrt{2(160-1)}} \qquad z = 2.7819$$

For 95 percent confidence level compare to 1.96

Example $s^2 = 1.378$ $\quad \frac{1.378}{1.05} = 1.312$

$\bar{x} = 1.05$

10

Limitations of Quadrat based tests

Takes no account of position of quadrats or events within quadrats

Loses spatial information

11

Nearest Neighbor tests for CSR

Under CSR events are independent and the number of events in any area is Poisson distributed

$$f_{y(A)}(y) = \frac{(\lambda A)^y}{y!} e^{-\lambda A}$$

Then the probability that no events fall within a circle of radius x around any randomly chosen point is $e^{-\lambda \pi x^2}$

Then the distribution function F(x) of nearest neighbor point-event distances for CSR is

$$F(x) = \Pr(X \leq x) = 1 - e^{-\lambda \pi x^2} \qquad x \geq 0$$

12

Nearest Neighbor tests for CSR

Theoretical distributions for X and W

πX^2 follows an exponential distribution with parameter λ

$2\pi\lambda X^2$ is therefore distributed as χ_2^2

$$E(X) = \frac{1}{2\sqrt{\lambda}} \qquad VAR(X) = \frac{(4-\pi)}{4\lambda\pi}$$

$2\pi\lambda\sum X_i^2$ is distributed as χ_{2n}^2

13

Nearest Neighbor tests for CSR

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

$$G(w) = \Pr(W \le w) = 1 - e^{-\lambda\pi w^2} \qquad w \ge 0$$

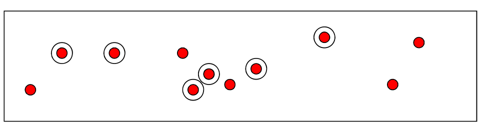
$$E(W) = \frac{1}{2\sqrt{\lambda}} \qquad VAR(W) = \frac{(4-\pi)}{4\lambda\pi}$$

Knowing the theoretical distributions of G and F allows derivation of the sampling distributions under CSR of various summary statistics for observed nearest neighbor distances

14

Nearest Neighbor tests for CSR

Distribution theory for these tests is based on the assumption of independence of *n* nearest neighbor measurements randomly sampled from a region R



This assumption of independence may be violated in case of smalls numbers of events and if the proportion of them used is large.

15

Nearest Neighbor tests for CSR

If the distribution theory used assumes independence and distances are not independent then there will be a tendency to reject the null hypothesis of CSR too often

Blyth and Ripley recommend that m (number of nearest neighbor distances)should be such that m ≤ 0.1n where n is number of events in study area

- Distribution theory for the tests also assumes that the nearest neighbor distances have not been biased by edge effects

Nearest neighbor distances for events near the boundary will be biased – so it becomes important to apply edge corrections

16

Tests of CSR based on Summary Statistics

Use *m* randomly sampled nearest neighbor event–event distances ($w_1 \ldots w_m$) or point-event distances ($x_1 \ldots x_m$)

- **Clarke-Evans**

Expected mean and variance under CSR

Compare $\overline{w} = \sum \frac{w_i}{m}$ with $N\left(\frac{1}{2\sqrt{\lambda}}, \frac{4-\pi}{4\lambda\pi m}\right)$

Normal approximations are good for m>10

Requires a completely enumerated point pattern from which events can be randomly sampled

Unknown λ needs to be replaced by an estimate - usually $\hat{\lambda} = \frac{n}{R}$

Tests of CSR based on Summary Statistics

Correction to allow all event-event distances to be used rather than a sample

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

$$VAR(\overline{W}) = 0.070\frac{R}{n^2} + 0.037P\sqrt{\frac{R}{n^5}}$$

Where P is the perimeter of the study region with area R

Example Computation of Clarke Evans

For Redwood Seedlings $\hat{\lambda} = .00143$

Mean of the nearest neighbor event-event distances 13.2445

Computed as $E(W) = \frac{1}{2\sqrt{\lambda}}$

Variance of the nearest neighbor event-event distances .77309

Computed as $VAR(W) = \frac{(4-\pi)}{4\lambda\pi}$

Compute Z score $z = \frac{\overline{w} - \mu}{\sigma}$ Z = -4.78721

$\overline{w} = 9.035$ $\Pr(Z \leq -4.78721) = .0000$

Tests of CSR based on Summary Statistics

- **Hopkins**

compare $H = \frac{\sum x_i^2}{\sum w_i^2}$ With percentage points of $F_{2m,2m}$

Rationale

In clustered patterns point-event distances x_i will be large relative to event-event distances w_i

In regular patterns point-event distances x_i will be smaller relative to event-event distances w_i

Tests of CSR based on Summary Statistics

- **Blyth & Ripley**

compare $\frac{1}{m}\sum \frac{x_i^2}{\left(x_i^2 + w_i^2\right)}$ with $N\left(\frac{1}{2}, \frac{1}{12m}\right)$

In this case x_i are randomly paired with w_i values

Lower tail of distribution indicates regularity, upper tail indicates clustering

21

Tests of CSR for Nearest Neighbor Distances

For completely enumerated point patterns we can consider the complete estimated distribution function W or X rather than a summary statistic

Goal:

Compare G(W) or F(X) with a theoretical form under CSR

$$G(w) = 1 - e^{-\lambda \pi w^2} \qquad F(x) = 1 - e^{-\lambda \pi x^2}$$

Need to be sure $\hat{G}(w)$ and $\hat{F}(x)$ have been corrected for edge effects since the theoretical distribution assumes no edge effects

22

Tests of CSR for Nearest Neighbor Distances

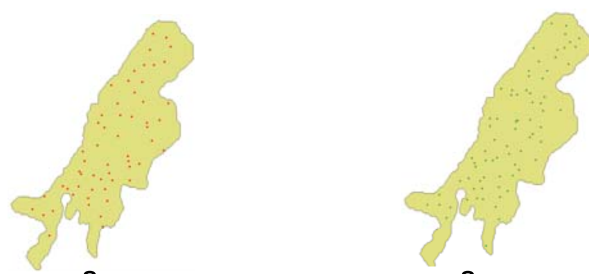
Employ a simulation estimate of the theoretical distribution under conditions of the specific study region R

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

where $\hat{G}_i(w) \quad i = 1, \ldots m$

is an empirical distribution function estimated without edge corrections from one of ***m*** independent simulations of n events under CSR in R

23



S_1 . . . S_m

Create m simulated patterns of n events in R

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

Compare with observed point pattern $\hat{G}(w)$

24

Tests of CSR for Nearest Neighbor Distances

To assess significance of departure, identify upper and lower bounds of the *m* empirical distribution functions

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

$$L(w) = \min_{i=1\cdots,m} \{\hat{G}_i(w)\}$$

Plot $\overline{G}(w)$ $\hat{G}(w)$ *U(w)* and *L(w)*

If compatible with CSR the plots should be linear

25

Tests of CSR for Nearest Neighbor Distances

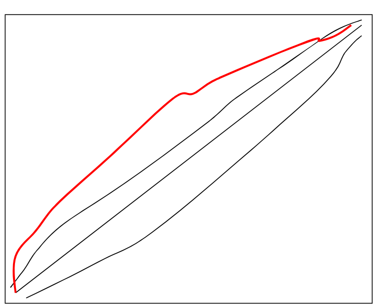
The upper and lower bounds help assess significance of departure from CSR

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

The desired significance level determines the number of simulations that should be run.

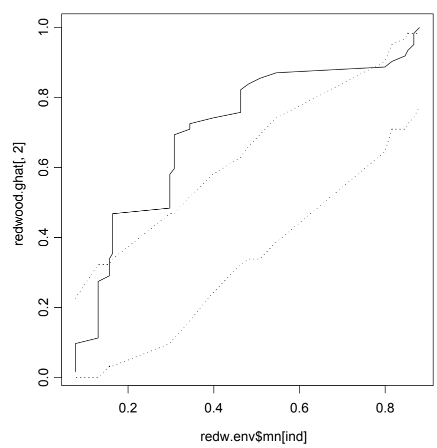
26

$\hat{G}(w)$



Upper and lower simulation envelops around the 45 degree line
Empirical distribution of observed distances falls above line

27



28

K Functions tests for CSR

Under CSR the expected number of events within a distance h of a randomly chosen event is $\lambda\pi h^2$

$\lambda K(h) = \lambda\pi h^2$

so $K(h) = \pi h^2$ under CSR

we then compare $\hat{K}(h)$ with πh^2

29

K Function Tests for CSR

How can we test the significance of departure in terms of the plot?

The sampling distribution of $\hat{K}(h)$ is complex because of the edge corrections

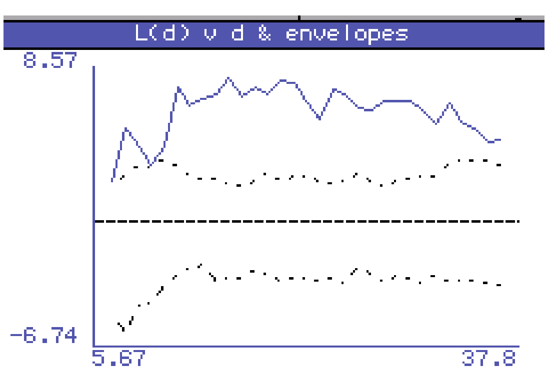
Can use a similar simulation approach to that used for G(w) or F(x)

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

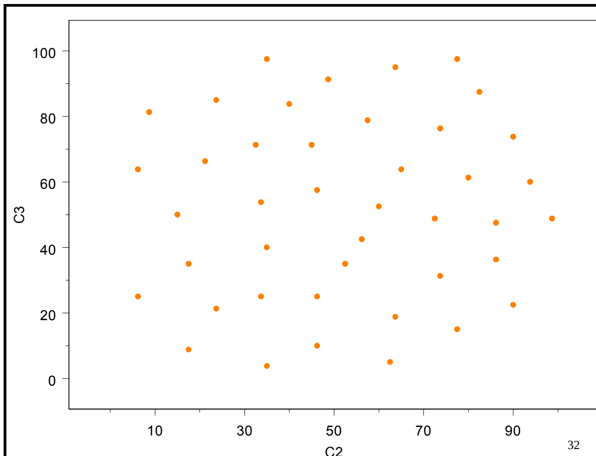
$$L(h) = \min_{i=1\cdots,m} \{\hat{L}_i(h)\}$$

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

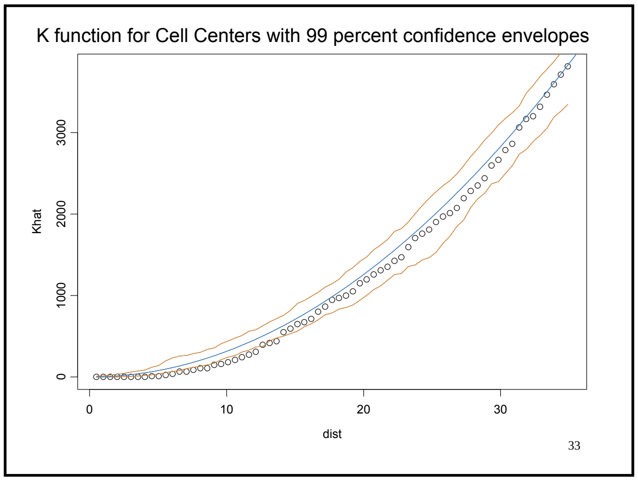
30



31



32


K function for Cell Centers with 99 percent confidence envelopes
Khat
0
1000
2000
3000
0
10
20
30
dist
33