

GEOG574/Math Introduction to Geostatistics

Point Pattern Analysis

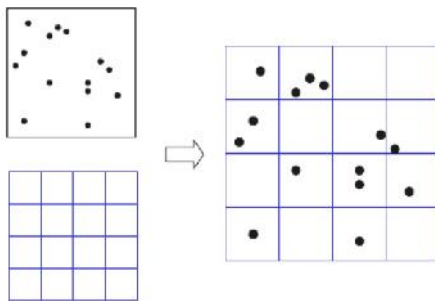
Daoqin Tong

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

Point Patterns

- A spatial point process is a spatial stochastic process $\{Y(s), s \in \mathfrak{R}\}$, where \mathfrak{R} is random
- A spatial point pattern $\{y(s_1), y(s_2), \dots, y(s_n); s_i \in R\}$ 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

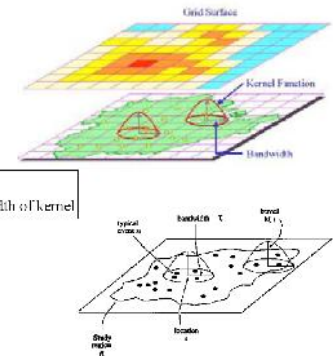


Kernel Estimation

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

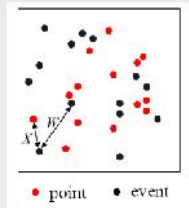
where $\delta_\tau(s)$ is the edge correction term;

$k(\cdot)$ is kernel function; τ is bandwidth of kernel



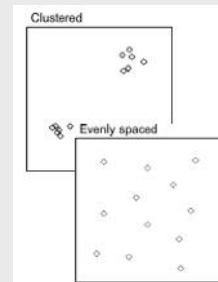
Nearest Neighbor Distances

- 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

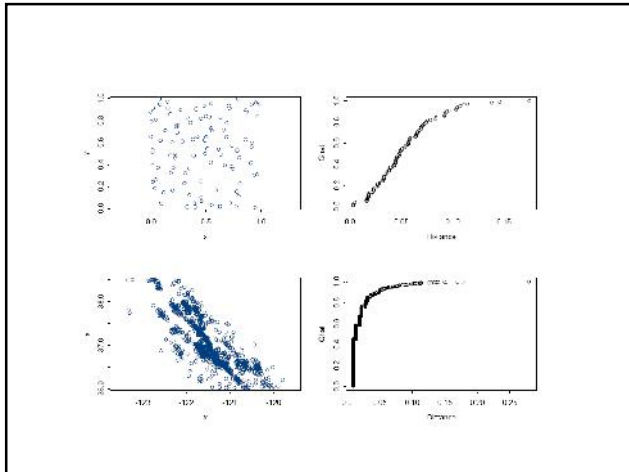
$$\hat{G}(w) = \frac{\#(w_i \leq w)}{n}; \quad \hat{F}(x) = \frac{\#(x_i \leq x)}{m}$$

n : the number of events;

m : the number of random points

Distance-based Exploratory Analysis

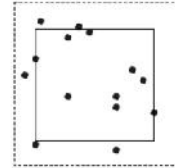
- Plot of $G(w)$ against w or $F(x)$ against x
 - If the empirical cumulative distribution function climbs very sharply in the early part before flattening out, then it indicates clustering
 - If it climbs very slowly at the beginning followed by steeply climbing, then it indicates a repulsion or regular 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

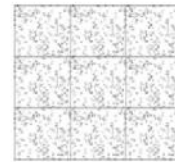
- Edge effects

- Guard area
- Toroidal edge correction
- Modified formula



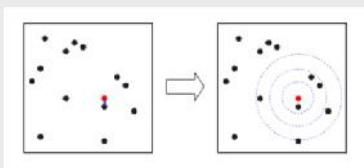
$$\hat{G}(w) = \frac{\#(b_i > w > w_i)}{\#(b_i > w)}$$

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

- $K(h) = E(\# \text{ of events within distance } h \text{ of an arbitrary event})$
- The expected # of events in \mathfrak{R} is $|\mathfrak{R}|$
- The expected # of ordered pairs within distance h is $|\mathfrak{R}|^2 K(h)$
- The K can be estimated by

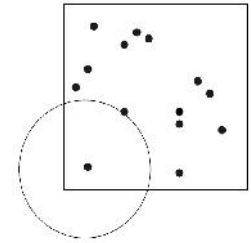
$$\hat{K}(h) = \frac{1}{\lambda^2 |\mathfrak{R}|} \sum_{i \neq j} I_h(d_{ij})$$

Edge Correction

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

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

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



Estimation of Intensity

$$\lambda = \frac{n}{|\mathfrak{R}|}$$

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

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

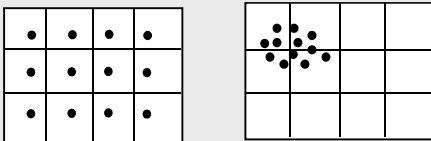
- Complete Spatial Randomness (CSR)- events follows a homogenous Poisson process over the study area
 - Consider $\{Y(A), A \in \mathcal{R}\}$
 - The probability distribution of $Y(A)$ is a Poisson distribution with mean $\lambda|A|$, where λ is a constant
- $$f_{Y(A)}(y) = \frac{(\lambda|A|)^y}{y!} e^{-\lambda|A|}$$
- $Y(A_i), Y(A_j)$ 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

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

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

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

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	95% confidence interval
$\hat{\lambda} = \frac{\bar{x}}{ Q }$	$\hat{\lambda} \pm 2\sqrt{\frac{\hat{\lambda}}{m Q }}$
$\text{var}(\hat{\lambda}) = \frac{\hat{\lambda}}{m Q }$	

Limitation of Quadrat Tests

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

Nearest Neighbor Tests

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

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

- 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

$$F(x) = P(X \leq x) = 1 - e^{-\lambda \pi x^2}$$

Nearest Neighbor Tests for CSR

- Mean and variance of X
- Similarly, the distribution function $G(w)$ of nearest neighbor event-event distances for CSR is

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

$$E(W) = \frac{1}{2\sqrt{\lambda}}; \quad \text{Var}(W) = \frac{4 - \pi}{4\lambda}$$

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, \dots, w_m) or point-event distances (x_1, \dots, x_m)
 - Sample size m can be chosen to be less than $0.1n$
- Three common tests
 - Clark-Evans
 - Hopkins
 - Byth & Ripley

Clark-Evans

- Randomly sampled m nearest neighbor event-event distances (w_1, \dots, w_m)
- Calculate the sample mean $\bar{w} = \frac{\sum w_i}{m}$
- Compare with the sampling distribution of \bar{w}

$$N\left(\frac{1}{2\sqrt{\lambda}}, \frac{4-f}{4\lambda m}\right)$$
 - Need to enumerate point pattern completely
 - Need to estimate intensity $\hat{\lambda} = \frac{n}{|A|}$

Clark-Evens Test

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

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

$$Var(\bar{w}) = 0.07 \frac{|A|}{n^2} - 0.037 P \sqrt{\frac{|A|}{n^3}}$$
- Compare with the sampling distribution of $\bar{w} = \frac{\sum w_i}{m}$

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

Hopkins

- Randomly sampled m nearest neighbor event-event distances (w_1, \dots, w_m) and point-event distances (x_1, \dots, 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, \dots, w_m) and point-event distances (x_1, \dots, 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

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

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

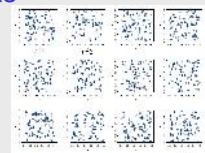
Simulation

- 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

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

- The observed $G(w)$

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



Nearest Neighbor Tests: Simulation

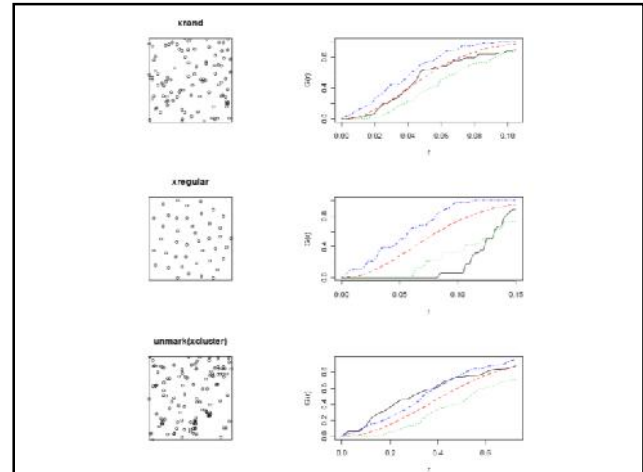
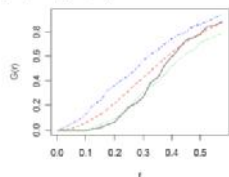
- Constructing the confidence envelopes

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

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

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

- 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 $\pi f h^2$

- So under CSR, $K(h) = \pi f h^2$

- Then the empirical K function can be compared with $K(h) = \pi f h^2$

- Or compared with the transformed form, $L(h) = 0$

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

- How to evaluate the significance?

$$\hat{K}(h) = \frac{1}{\lambda^2} \left[\frac{1}{n(n-1)} \sum_{i \neq j} I_1(d_{ij}) \right]$$

K Function Tests: Simulation

- Simulate m point patterns with n points under CSR

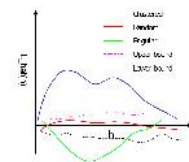
- Constructing the confidence envelopes

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

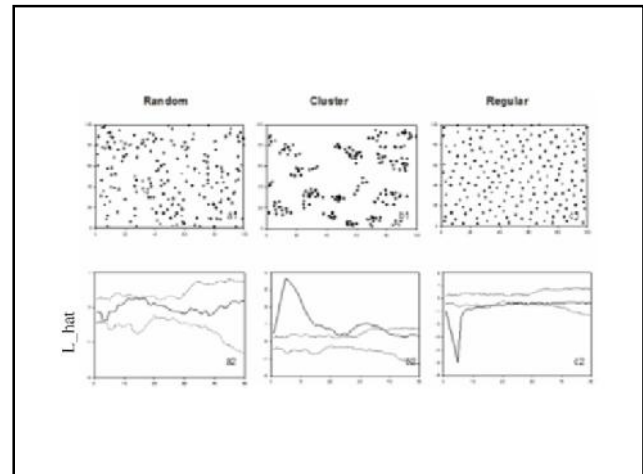
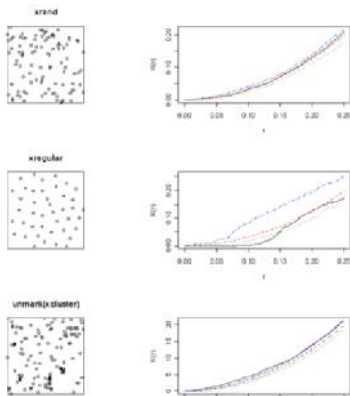
$$l(h) = \min_{i=1, \dots, m} \{ \hat{K}_i(h) \}$$

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

- 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