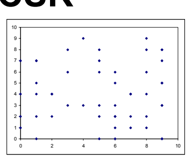
Point Pattern Analysis
Part I

Outline

- Revisit IRP/CSR, first- and second order effects
- What is point pattern analysis (PPA)?
- Density-based point pattern measures
- Distance-based point pattern measures

Revisit IRP/CSR

- Equal probability: any point has equal probability of being in any position or, equivalently, each small sub-area of the map has an equal chance of receiving a point
- Independence: the positioning of any point is independent of the positioning of any other point



$$P(k,n,x) = \binom{n}{k}\left(\frac{1}{x}\right)^{k}\left(\frac{x-1}{x}\right)^{n-k}$$

$$P(k) = \frac{\lambda^{k} e^{-\lambda}}{k!} \text{ and } \lambda = \frac{n}{x}$$

First- & Second Order Effects

- IRP/CSR is not realistic in geography
  - First-Order effect: variations in the density of a process across space due to variations in environment properties: **No** equal probability
  - Second-Order effect: interaction between locations: **NO** independence

What Is Point Pattern Analysis (PPA)

- Point patterns, where the only data are the locations of a set of point objects
  - Represent the simplest possible spatial data
- Examples
  - Hot-spot detection (crime, disease, geological activities)
  - Disease & environmental relations
  - Freeway accidents
  - Vegetation, archaeological studies

What Is PPA (Cont.)

- In a point process the basic properties of the process are set by a single parameter, the probability that any small area will receive a point

What Is PPA (Cont.)

- Requirements for a set of events to constitute a point pattern
  - Mapable pattern on the plane
  - Objectively determined study area
  - An enumeration or census of all entities of interest, not a sample
  - A one-to-one correspondence between objects in the study area and events in the pattern
  - Event locations must be proper (not be the centroids of polygons)

Describing a Point Pattern

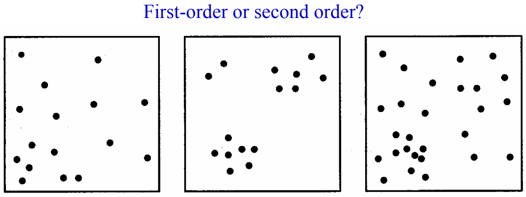
- Point density (first-order or second-order?)
- Point separation (first-order or second-order?)

- When first-order effects are marked, absolute location is an important determinant of observations, and a point pattern has clear variations across space in the number of events per unit area are observed
- When second-order effects are strong, there is interaction between locations, depending on the distance between them, and thus relative location is important

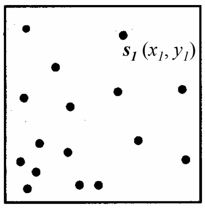
Describing a Point Pattern

- In practice it is close to impossible to distinguish from variation in the environment or interaction
  - Low density → larger separation
  - High density → smaller separation

First-order or second order?



Some Notations



A set of locations S with n events

$$S = \{s_1, s_2, \ldots, s_i, \ldots, s_n\}$$

Each event s_i has two coordinates:

$$s_i = (x_i, y_i)$$

The study region A has an area a

Describing a Point Pattern

Mean Center: the mean X coordinate and the mean Y coordinate

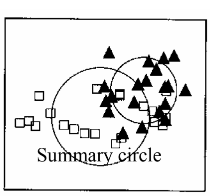
$$\bar{s} = (\mu_x, \mu_y) = \left(\frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n} \right)$$

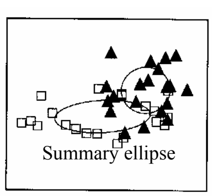
Standard Distance: a measure of how dispersed the events are around their mean center

$$d = \sqrt{\frac{\sum_{i=1}^{n} \left[(x_i - \mu_x)^2 + (y_i - \mu_y)^2 \right]}{n}}$$

Describing a Point Pattern

- A **summary circle** can then be plotted for the point pattern, centered at the mean center with radius *d* *(SD)*
- If the standard distance is computed separately for each axis, a **summary ellipse** can be obtained (2 *d*s)



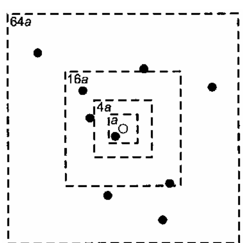


Measuring Point Pattern

- Two basic families
  - Density based measures: 1st order effect
  - Distance based measures: 2nd order effect

Density-Based Measures

- First-order effect
- Three methods
  - Crude density: Overall
  - Quadrat count: counts of events in each quadrat
  - Density estimation: continuous
- Sensitive to the definition of the study area: MAUP & Edge Effect



Crude Density

- Overall intensity of events in the study area

$$\lambda = \frac{n}{a} = \frac{no.(S \in A)}{a}$$

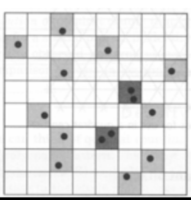
  - Descriptive statistics for summary purpose → too simple

Quadrat Counts

- Count the numbers of events in a set of equal-area quadrats and record these counts as a frequency distribution
  - Equally-spaced patterns will have most quadrats with similar counts
  - Clustered pattern will have a few high count quadrates and many empty ones
- Two approaches:
  - Exhaustive census
  - Random sampling

Quadrat Counts

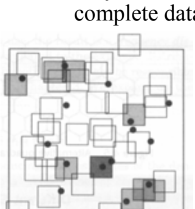
- Approach I: Exhaustive *census* of quadrats
  - Completely fill the study region with no overlaps
  - The choice of origin, quadrat orientation, and quadrat size affects the observed frequency distribution → **MAUP**



Number of events in quadrat	*Census, n = 64*		*Sampling, n = 38*	
	Count	*Proportion*	*Count*	*Proportion*
0	51	0.797	29	0.763
1	11	0.172	8	0.211
2	2	0.031	1	0.026
3	0	0.000	0	0.000

Quadrat Counts (Cont.)

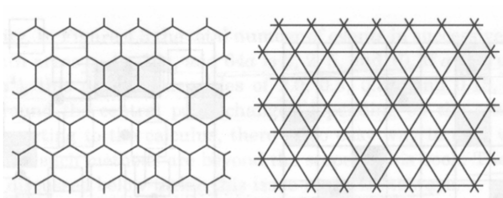
- Approach II: Random sampling approach
  - Possible to increase the sample size simply by adding more quadrats (for sparse patterns)
  - May describe a point pattern without having complete data on the entire pattern



Number of events in quadrat	*Census, n = 64*		*Sampling, n = 38*	
	Count	*Proportion*	*Count*	*Proportion*
0	51	0.797	29	0.763
1	11	0.172	8	0.211
2	2	0.031	1	0.026
3	0	0.000	0	0.000

Quadrat Counts (Cont.)

- Other shapes of quadrats

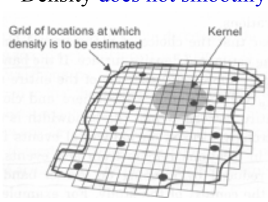


Density Estimation

- Assume: the pattern has a density anywhere in the study area
  - not just locations where there is an event
- Estimated by counting the number of events in a region, or kernel, centered at the location where the estimate is to be made
  - Variation of Quadrat Counts
- Two approaches
  - Simple density estimation
  - Kernel density estimation

Simple Density Estimation

- Simply count the number of events of a pattern in a region (usually, circle) then calculate the density by dividing it with the area of the region
- Bandwidth ***r***
  - Too large or too small? what is the proper size?
- Density does not smoothly vary across space

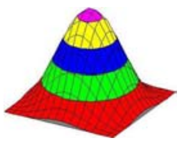


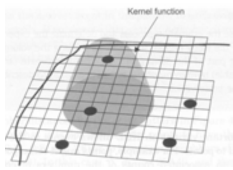
$$\hat{\lambda}_p = \frac{no.[S \in C(p,r)]}{\pi r^2}$$

C(p,r) is a circle of radius centered at the location of interest ***p***

Kernel Density Estimation

- Introduce a distance effect with a kernel function
  - Weight nearby events more heavily than distant ones in estimating the local density
  - e.g. a point at half of bandwidth r only counts for ½ point



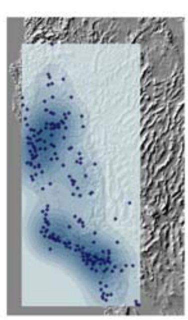


Kernel Density Estimation (Cont.)

- Bandwidth ***r*** → impact the smoothness of the resulting density field depends
  - Narrow kernels produce bumpy surfaces
  - Wide kernels produce smooth surfaces

Applications of Density-Based Methods

- Visualize a point pattern to detect hot spots
- Check whether or not that process is first-order stationary from the local intensity variations
- Link point objects to other geographic data (e.g. disease and pollution)



Summary

- Point pattern is the only pattern
- 1^{st} order vs. 2^{nd} order SV
- Descriptive summary of point pattern
  - Mean center and standard distance
- Density measures
  - Quadrat Counts
  - Density Estimation

Distance-Based Measures

- Second-order effect
  - Look at patterns within the distances among point events
- Two approaches:
  - Nearest Neighbor Distance (NND)
  - Distance functions
    - G function
    - F function
    - K function

Nearest Neighbor Distance

- NND: the distance from an event to the nearest event
  - Usually Euclidean distance → Pythagorean Theorem:

$$d(s_i, s_j) = \sqrt{(x_i - x_j) + (y_i - y_j)} \qquad d_{\min}(s_i) = \min(d(s_i, s_j))$$

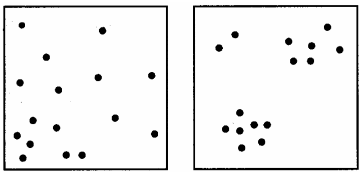
- Mean NND:
  - Summarizes all the nearest-neighbor distances by a single mean value:

$$\bar{d}_{\min} = \frac{\sum_{i=1}^{n} d_{\min}(s_i)}{n}$$

  - Way too simple → lose a lot useful information

Expected Behavior of Mean NND

- Clustered: All NND are short → small mean NND
- Equally-spaced: Minimum distances are longer → larger mean NND

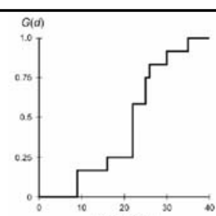


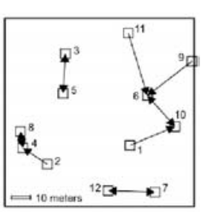
G Function

- **G(*d*)** examines the cumulative frequency distribution of the NNDs btw the event points
  - What fraction of all the NNDs in the pattern are less than a specified distance (d)
  - Simplest distance function

$$G(d) = \frac{\#[d_{\min}(\mathbf{s}_i) < d]}{n}$$

Illustration

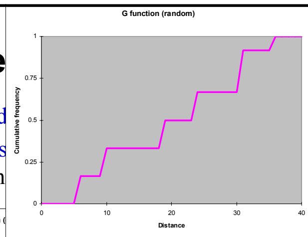


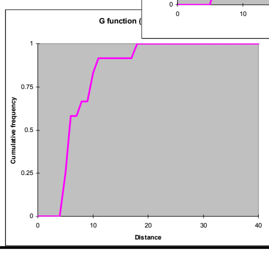


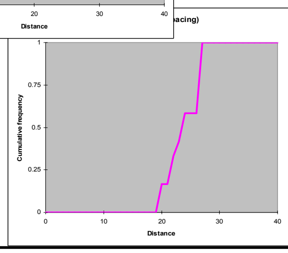
Point	x	y	d_min
1	66.22	32.54	25.59
2	22.52	22.39	15.64
3	31.91	81.21	21.14
4	9.47	31.02	9.00
5	30.78	60.10	21.14
6	75.21	58.93	21.94
7	79.26	7.68	24.81
8	8.23	39.93	9.00
9	98.73	77.17	29.76
10	89.78	42.53	21.94
11	65.19	92.08	34.63
12	54.46	8.48	24.81

Expe … G(d)

- Clustered … distance
- Equally-s … arts at a much lon…







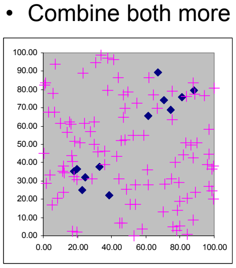
F Function

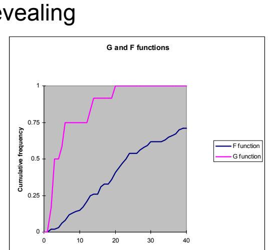
- **F(*d*)** is also the cumulative frequency distribution of the certain NNDs
  - BUT, NNDs are calculated btw a group of random points in the study area and event points
  - Increased sample size → smoother curve

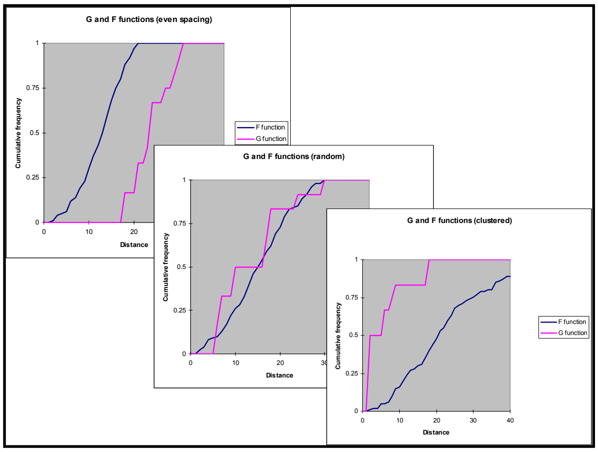
$$F(d)=\frac{\#[d_{\min}(\mathbf{p}_i,S)<d]}{n}$$

G(*d*) vs. F(*d*)

- Both use the idea of NNDs
  - Only NNDs are used
- F(d) is smoother than G(d)
- Combine both more revealing





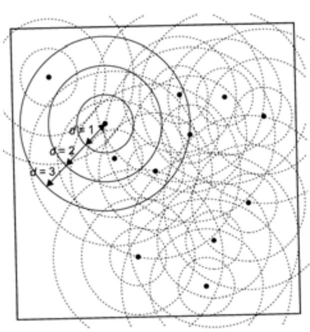


Ripley’s K Function

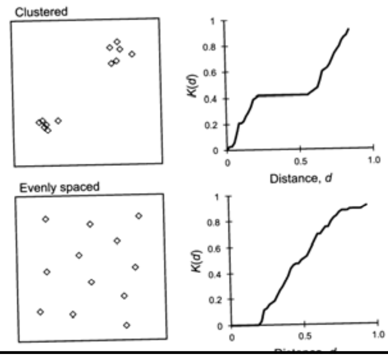
- **K(*d*) makes use of all inter-event distance**
  – Makes it a complex function to calculate
  – But, it includes all possible distance information about the pattern instead of NNDs

Calculating the K(d)

- For each distance *d*:
  1. Determine how many events lie within that distance (*d*) of each event
  2. Calculate an average number of events at the distance *d*
  3. Divide by overall study area event density



Interpretation of K(*d*)



Review (PPT Statistics)

- Density-based :
  - Quadrat Counts
  - Kernel density estimation
- Distance-based:
  - Mean NND
  - G & F functions
  - K function