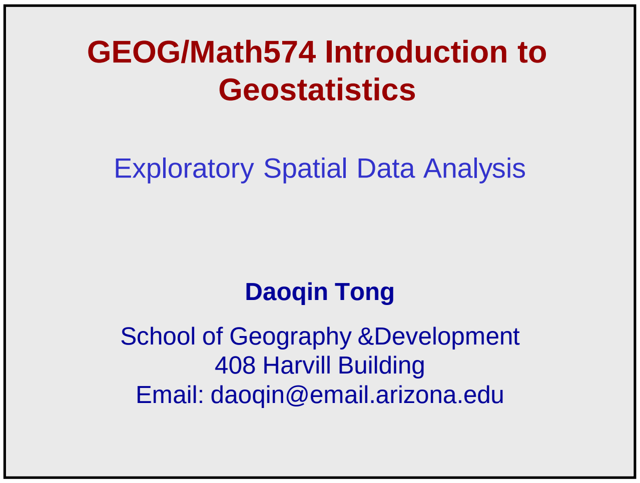
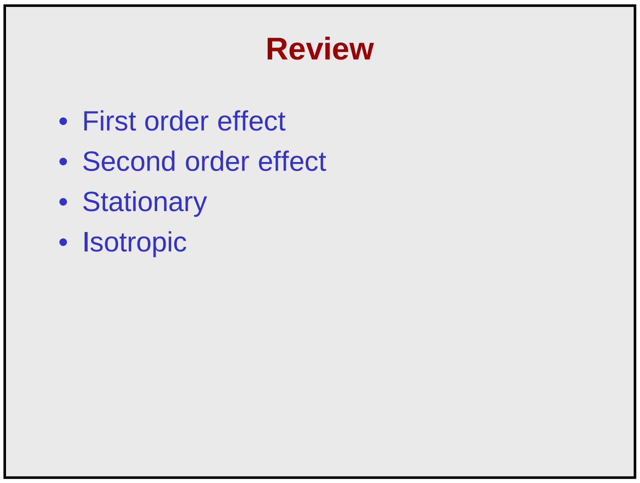


GEOG/Math574 Introduction to Geostatistics

Exploratory Spatial Data Analysis

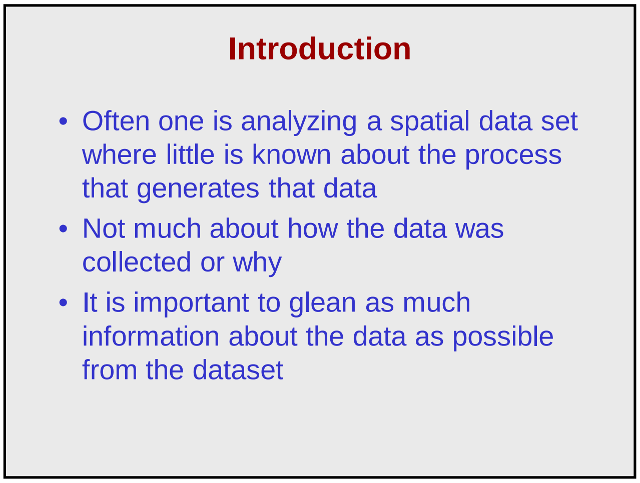
Daoqin Tong

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



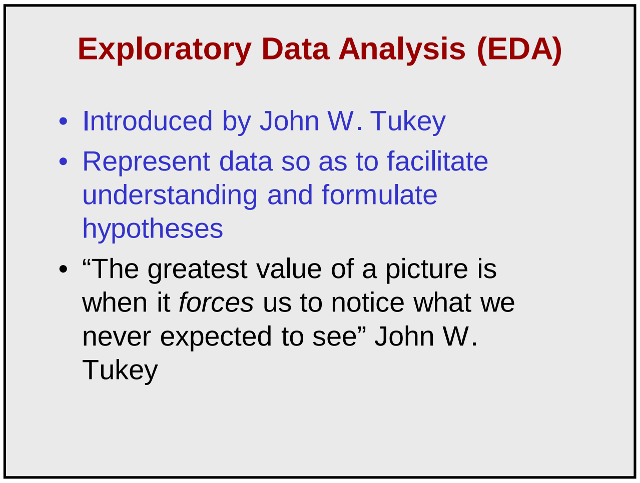
Review

- First order effect
- Second order effect
- Stationary
- Isotropic



Introduction

- Often one is analyzing a spatial data set where little is known about the process that generates that data
- Not much about how the data was collected or why
- It is important to glean as much information about the data as possible from the dataset

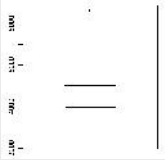


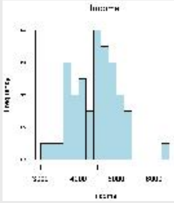
Exploratory Data Analysis (EDA)

- Introduced by John W. Tukey
- Represent data so as to facilitate understanding and formulate hypotheses
- “The greatest value of a picture is when it *forces* us to notice what we never expected to see” John W. Tukey

Graphical Techniques

- Box Plot
  - Five-number summaries(smallest, Q1, Median, Q3, largest observation)
  - By John W. Tukey
- Histogram
  - A graphical display of tabulated frequencies
  - shape

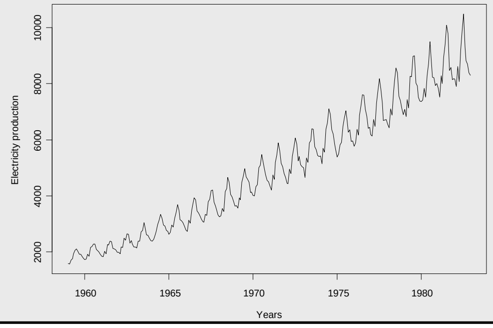




Graphic Techniques

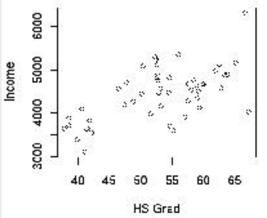
Run chart

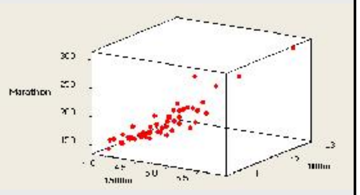
Australia monthly production of electricity from Jan 1956 to Dec 1981 (in unit of m·KWH)



Graphical Techniques

Scatterplot



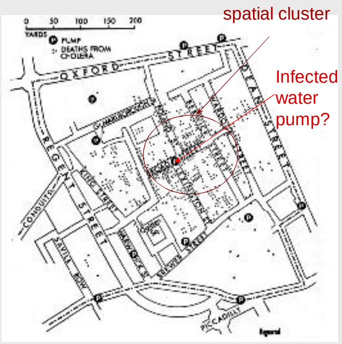


Basic Summary statistics

Mean, standard deviation/variation, skewness, kurtosis, minimum, 25th percentile, median, 75th percentile, maximum

An Example

- Dr. John Snow
- Investigation of deaths from cholera, London,1854



spatial cluster

Infected water pump?

Exploratory Spatial Data Analysis

- Plot of data locations coded by value of variables
- Plot of data value against first coordinate (E-W coordinate)
- Plot of data value against second coordinate (N-S coordinate)
- Plot of data value against third coordinate (vertical coordinate)
- Fit a trend surface to the data (one variable at a time) in R

Point Pattern Analysis

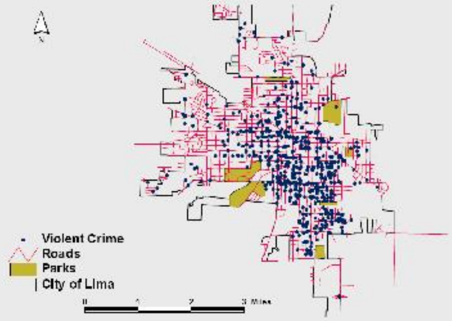
Point Patterns

- A spatial point process is a spatial stochastic process $\{Y(s), s \in \Re\}$, where $\Re$ is random
- A spatial point pattern $\{y(s_1), y(s_2), \ldots, y(s_n); s_i \in R\}$ is a realization of a spatial point process
- Two important terms
  - Events: observations on the observed locations
  - Points: other arbitrary locations

Applications

- Various disease
- Crimes
- Galaxies in space
- Plants
- Etc

Violent Crime in Lima (1999)



(Alan Murray, 2000)

Point Patterns

- Mapped point pattern
  - A complete map of events in the study area
  - All relevant events in the study area have been recorded
- Sampled point pattern
  - A subset of all events are recorded in a sample of different areas of the study area
  - Complete enumeration is not feasible
  - Forestry, ecology

Objectives

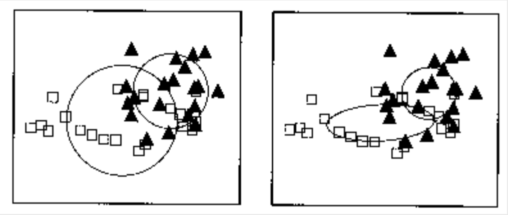
- To determine whether the spatial distribution of data points exhibit a systematic pattern as opposed to randomness
- To estimate the variation of the intensity at a large scale
- To detect the presence of spatial dependence among events
- To find an underlying model generating the observed patterns

Describing a point pattern

- Summary measures
- First order effect
  - – Density-based measures
  - – Quadrat count method
  - – Kernel estimation
- Second order effect
  - – Distance-based measures
  - – Nearest neighbor distance
  - – K/L function

Summary measures

- **Mean center**
- **Standard distance / relative distance**
- **Standard deviational ellipse**



Measures of spatial central tendency - mean center

- Given a set of points $\{(X_1,Y_1),\ (X_2,Y_2),...,(X_n,Y_n)\}$, the mean center is calculated as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\ ,\ \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

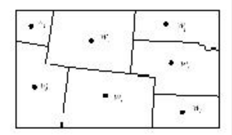
- Minimize the sum of squared distances that individual must travel

$$\min \sum_{i=1}^{n}\left[(X_i - \bar{X})^2 + (Y_i - \bar{Y})^2\right]$$

Measures of spatial central tendency – weighed mean center

- Produced by weighting each X and Y coordinate by another variable (w_i)
- Centroids derived from polygons can be weighted by any characteristic of the polygon

$$\bar{X}_w = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i X_i\ ,\ \bar{Y}_w = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i Y_i$$



Euclidean Median

- Minimize the sum of Euclidean distances from all other points to that central location?

$$\min \sum_{i=1}^{n} \sqrt{(x_i - x_e)^2 + (y_i - y_e)^2}$$

- Weber problem

Minimize $\sum f_i \sqrt{(x_i - \bar{X})^2 + (y_i - \bar{Y})^2}$

  - – Minimize the total costs
- Location-allocation problem
- GIS, marketing analysis

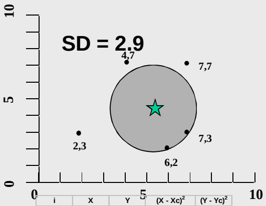
Standard distance

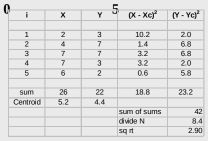
- Spatial equivalent to standard deviation
- Amount of absolute dispersion
- Provides a single unit measure of the spread or dispersion of a spatial distribution

$$S_D = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x}_c)^2 + (y_i - \bar{y}_c)^2}{n}}$$

$$S_D = \sqrt{\left(\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}_c^2\right) + \left(\frac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y}_c^2\right)}$$

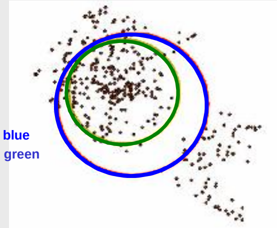
Examples





i	X	Y	(X - Xc)^2	(Y - Yc)^2
1	2	3	10.2	2.0
2	4	7	1.4	6.8
3	7	7	3.2	6.8
4	7	3	3.2	2.0
5	6	2	0.6	5.8
sum	26	22	18.8	23.2
Centroid	5.2	4.4		
			sum of sums	42
			divide N	8.4
			sq rt	2.90

Seasonal effects of childhood respiratory diseases



Winter - blue
Spring - green

First Order Effects

- Intensity
  - – The mean number of events per unit area at one point *s*
  - – Defined as the limit when the size of small region *ds* around point s approaches zero:

$$ª(s) = \lim_{ds \to 0} \left\{ \frac{E(Y(ds))}{|ds|} \right\}$$

- For a stationary point process, $ª(S)$ is constant over the study area
  - – The intensity is ª
  - – For an arbitrary area *A*, $E(Y(A)) = |A|\,ª$

Second Order Effects

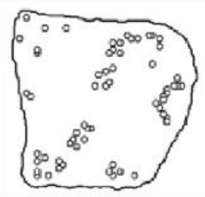
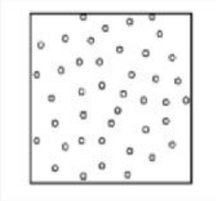
- Second order intensity
  - The relationship between numbers of events in pairs of areas in the study area $\Re$
  - Defined as the limit when the sizes of small regions *ds*$_i$ and *ds*$_j$ around point *s*$_i$ and *s*$_j$ approach zero:

$$\text{ª}(s_i, s_j) = \lim_{ds_i, ds_j \to 0} \{\frac{E(Y(ds_i)Y(ds_j))}{|ds_i| \times |ds_j|}\}$$

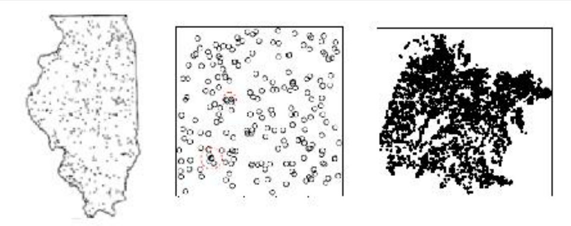
- For a stationary point process
  - Anisotropic $\text{ª}(s_i, s_j) = \text{ª}(s_i - s_j) = \text{ª}(h)$
  - Isotropic $\text{ª}(s_i, s_j) = \text{ª}(|h|)$

Visualizing Spatial Point Patterns

- Dot map
  - Shape of the study area
  - Visual impression of patterns

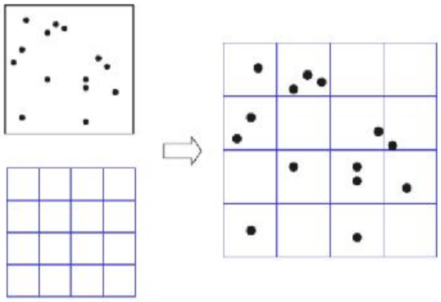
Examples

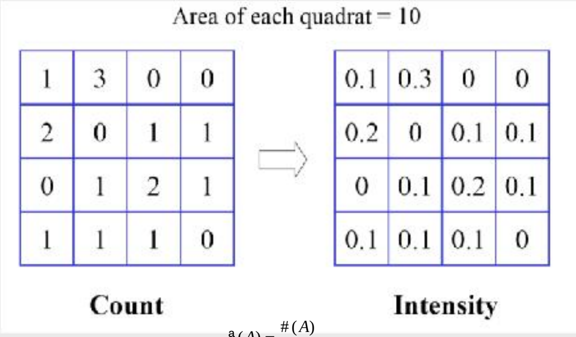


Exploring Spatial Point Patterns

- Summary statistics or plots
- First order effects
  - Quadrat methods
  - Kernel estimation
- Second order effects
  - Nearest neighbor distances
  - K function

Quadrat Methods





Area of each quadrat = 10

1	3	0	0
2	0	1	1
0	1	2	1
1	1	1	0

Count

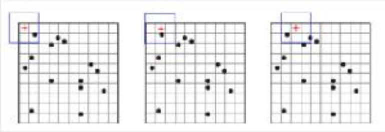
0.1	0.3	0	0
0.2	0	0.1	0.1
0	0.1	0.2	0.1
0.1	0.1	0.1	0

Intensity

$$^{a}(A) = \frac{\#(A)}{|A|}$$

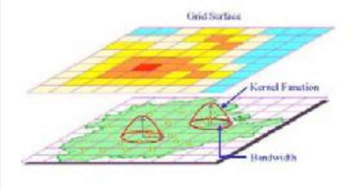
Remarks

- Quadrat can be randomly placed over the study area to gain a rough estimation of the variation in intensity
- No spatial details are considered within each quadrat
- The size of quadrat: large or small
- An alternative approach: moving window?



Kernel Estimation

- An extension of moving window approach
- Originally designed to obtain a smooth estimate of probability density function

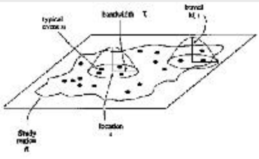


Kernel Estimation

- Notation
  - *s*: an arbitrary location in R
  - $s_1, \ldots, s_n$ are the locations of the *n* observed events
- The intensity at location s is estimated by

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

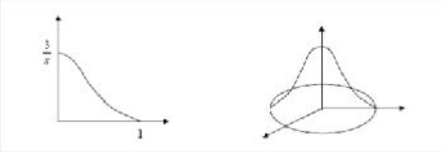
where $\delta_\tau(s)$ is the edge correction term;
$k(\bullet)$ is kernel function; τ is bandwidth of kernel



Kernel Estimation

- Kernel function

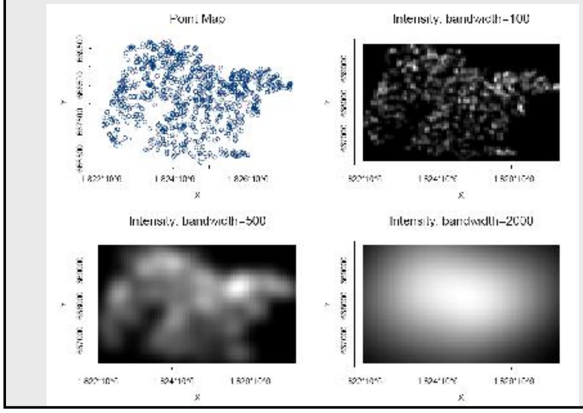
$$k(u) = \begin{cases} \frac{3}{\pi}(1 - u^T u)^2; & u^T u \le 1 \\ 0; & u^T u > 1 \end{cases} \qquad u^T u = |u|^2$$

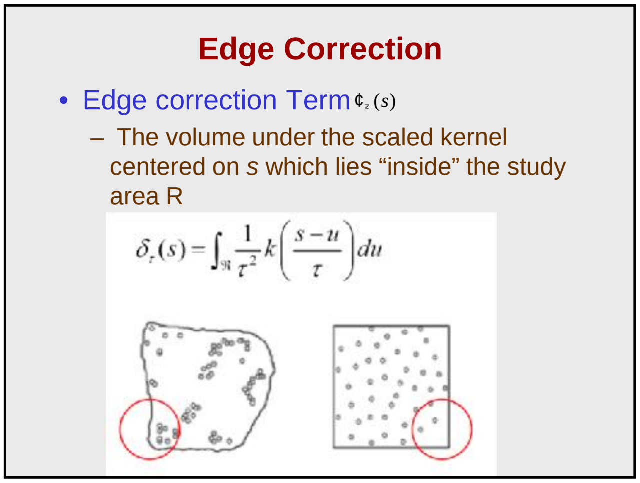


How to choose a bandwidth

- The bandwidth determines the degree of smoothness of the intensity map
- If ^{2} is too large, the intensity will be very smooth and spatial details are averaged out
- If ^{2} is too small, the intensity will be very spiky at the locations of events and flat at the locations with no or few events
- One rule of thumb
- Or, try different bandwidths

Effects of Bandwidth

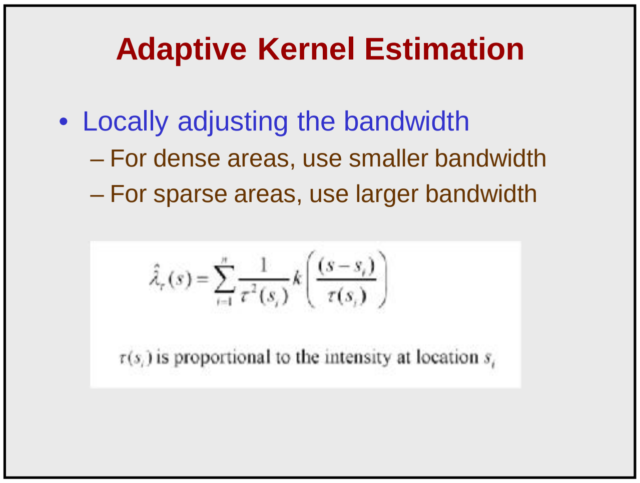




Edge Correction

- Edge correction Term $¢_2(s)$
  - The volume under the scaled kernel centered on *s* which lies “inside” the study area R

$$\delta_\tau(s) = \int_{\Re} \frac{1}{\tau^2} k\left(\frac{s-u}{\tau}\right) du$$



Adaptive Kernel Estimation

- Locally adjusting the bandwidth
  - For dense areas, use smaller bandwidth
  - For sparse areas, use larger bandwidth

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

$\tau(s_i)$ is proportional to the intensity at location s_i