Lecture 7:
Spatial Analysis II

GEOG413/613

Dr. Anthony Jjumba

1

Spatial Analysis

- Point Pattern Analysis
  - Methods
    - Nearest Neighbor Analysis
    - Quadrat Analysis
  - Spatial Autocorrelation

2

Descriptive Non-Spatial Statistics
- Mean, Standard Deviation, Skewness, Kurtosis

Descriptive Spatial Statistics
- Mean Centre, Weighted Mean Centre, Standard Distance, Relative Distance

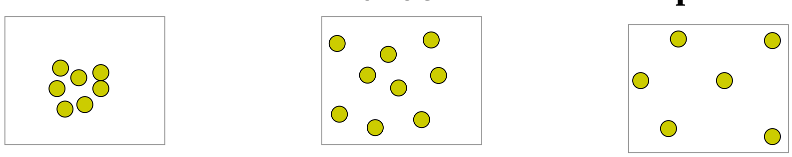
Inferential Spatial Statistics and Analysis
- Point Pattern Analysis (Nearest Neighbor Analysis, Quadrat Analysis)
- Cluster Analysis
- Spatial Autocorrelation
- Factor Analysis
- Spatial Interpolation

3

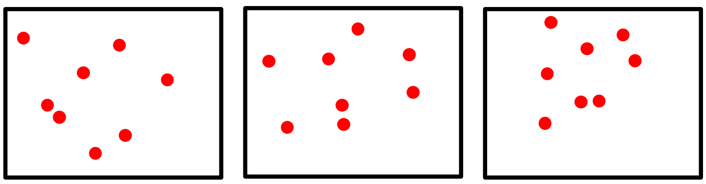
Examining Spatial Data

- How are the data points spatially distributed?

Clustered **Random** **Dispersed**



- How do you know? Always test



4

Point Pattern Analysis

- A set of quantitative tools for examining the spatial arrangement of point locations on the landscape as represented by a conventional map.

- Two methods are ***nearest neighbour analysis*** and ***quadrat analysis***.

5

Nearest Neighbour Analysis

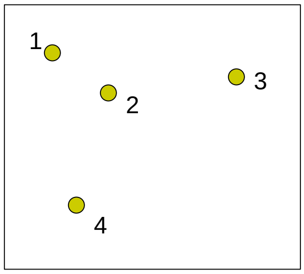
- Distance of each point to its nearest neighbour is measured

- The average nearest distance for all points is then calculated

- Can compare results with expected average for a random distribution

6

Nearest Neighbor Analysis



$d_1 = l_{12}$

$d_2 = l_{21}$

$d_3 = l_{32}$

$d_4 = l_{42}$

$$r_{obs} = \frac{\sum_{1}^{n} d_i}{n}$$

The Average Nearest Neighbor distance = r_{obs}

7

Nearest Neighbor Analysis

- The average nearest neighbour distance is an absolute value
- It is a function of the units in which the distance is measured
- Problem
  - How can we compare data from different regions or studies?
    - Solution: **Standardized Nearest Neighbour Index**

8

Nearest Neighbour Analysis

- The utility of the average nearest neighbour distance comes from comparing the index value for an observed pattern to the results produced from certain distinct point distributions
  - We can compare our results against values for random, clustered and dispersed distributions
  - Problem
    - How can we compare data from different regions or studies?
      - Solution: **Standardized Nearest Neighbour Index**

9

Random Distribution

- For a random distribution, the average nearest neighbour distance is calculated as follows:

$$r_{rnd} = \frac{1}{2\sqrt{n/A}}$$

Where :
A is area of study region
n is number of points

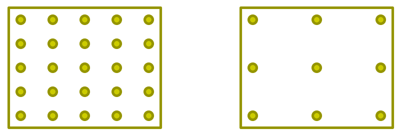
10

Maximum Dispersion Distribution

- If the distribution is perfectly uniform, the average nearest neighbour distance is calculated as follows:

$$r_{dsp} = \frac{1.07453}{\sqrt{n/A}}$$

Question:
Consider the two distributions below, assume that the area is the same.
Do they have different r_{dsp} values?
If so, which one will have a higher r_{dsp} value?



11

Clustered Distribution

- When all points lie in the same position (i.e. maximum clustering) the average nearest neighbour distance is 0

$$r_{cst} = 0$$

12

Standardized Nearest Neighbor Index

- The Standardized Nearest Neighbor Index is computed as a ratio of r_{obs} *to* r_{rnd}, the expected average nearest neighbor distance for a random distribution

$$R = \frac{r_{obs}}{r_{rnd}}$$

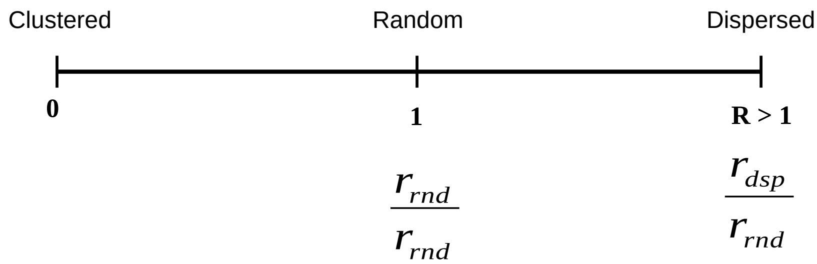
Standardized Nearest Neighbour Index

- To calculate the standardized nearest neighbour index for maximum dispersion:

$$R = \frac{r_{dsp}}{r_{rnd}}$$

Standardized Nearest Neighbour Index

- An actual point pattern can be measured for relative spacing along a continuous scale:



15

Test of Significance

- It is important to test whether a significant difference exists between the observed and random nearest neighbor values.

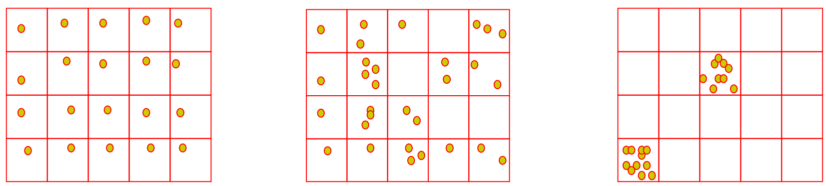
$$Z_r = \frac{r_{obs} - r_{rnd}}{\sigma_{obs}}$$

$$\sigma_{obs} = \frac{0.26136}{\sqrt{n(n/A)}}$$

16

Quadrat Analysis

- Examines the frequency of points occurring in various parts of the study area.
- The point pattern arrangement in the study area is described with the aid of the frequency of points in a cell



Dispersed Random Clustered

Quadrat Analysis

- In quadrat analysis, an index known as the variance-mean ratio (VMR) standardizes the degree of variability in cell frequencies relative to the mean of the cell frequency

$$VMR = \frac{Var}{Mean}$$

$$Mean = \frac{n}{m}$$

where *n* = number of points
m = number of cells
Mean = mean cell frequency
Var = Variance of cell frequencies

Quadrat Analysis

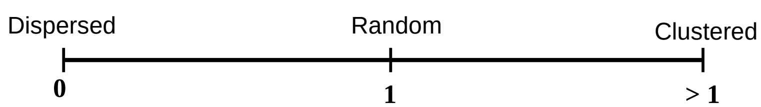
- Variance of Cell Frequencies

$$Var = \frac{\sum f_i x_i^2 - \frac{\left(\sum f_i x_i\right)^2}{m}}{m-1}$$

where f_i = frequency of cells with *i* cases
x_i = number of cases per cell

Quadrat Analysis

- Variance-Mean Ratio (VMR)
  - If each cell contains the same amount of points, then VMR = 0
  - If a point pattern is highly clustered with most cells containing no points, then VMR will be relatively large.
  - If the point pattern is perfectly random, then the mean cell frequency equals the variance of the cell frequency, and VMR = 1



Quadrat Analysis

- Test of Significance
  - Applied to determine if distribution of points is random.
  - The test statistic used is chi-square:

$$X^2 = VMR\ (m - 1)$$

Cluster Analysis

Cluster Analysis

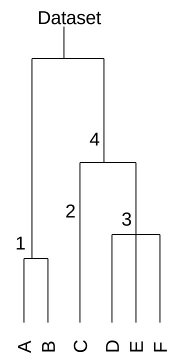
- To reduce data complexity by sorting the data into subsets (clusters) that share some common trait
- Achieve the reduction of observations by minimizing the within-group variation and maximizing the between group variation (i.e. the degree of association between two objects is maximal if they belong to the same group and minimal otherwise)

Clustering Analysis

- Searching for groups in the data in such a away that objects belonging to the same cluster resemble each other, whereas objects in different clusters are dissimilar.
- Two methods are partitioning and hierarchical clustering

Cluster Analysis

- Hierarchical Methods
  - Can be **agglomerative** or **divisive**



Dendogram

25

Cluster Analysis

- Hierarchical Methods
  - Identifies homogeneous groups of variables by using an algorithm that:
    - either starts with each observation in a separate cluster and combines clusters until only one is left (agglomerative),
    - or starts with the whole dataset and proceeds to divide it into successively smaller clusters (divisive).

26

Cluster Analysis

- Partitioning Methods
  - Based on specifying an initial number of groups, and iteratively reallocating observations between groups until some equilibrium is attained
  - The most popular method of partitioning is the *k-means* method

Cluster Analysis – Partitioning methods

- K-Means Clustering
  - Observations are classified as belonging to one of *k* groups
  - Group membership is determined be calculating the centroid for each group, then assigning each observation to the group with the nearest centroid

Spatial Autocorrelation

- Nearest Neighbor Analysis, Quadrat Analysis and Cluster Analysis are limited as they only consider either the location or the attribute of the observations
- Spatial autocorrelation detects spatial patterns of a data distribution by considering both the ***locations*** and the ***attributes*** of the observations

29

Spatial Autocorrelation

- Spatial Autocorrelation – correlation of a variable with itself through geographic space.
  - The first law of geography: "everything is related to everything else, but near things are more related than distant things" – Waldo Tobler
  - If there is a systematic pattern in the spatial distribution of a variable, then there is spatial autocorrelation
    - *positive spatial autocorrelation* – neighbors are more alike
    - *negative autocorrelation* – neighbors are not alike/different
    - *no spatial autocorrelation* – Random distribution

30

Spatial Autocorrelation

- Most statistical measures are based on the assumption that the values of observations in each sample are independent of one another (this is why we carry out random sampling)
- This independence may be violated by spatial autocorrelation. For example, positive spatial autocorrelation may violate the independence if the samples were taken from nearby areas
- Therefore it may lead to incorrect conclusions about relationships between variables
- Goals of spatial autocorrelation are
  - To measure the strength of spatial autocorrelation for a given distribution
  - To test the assumption of independence or randomness

31

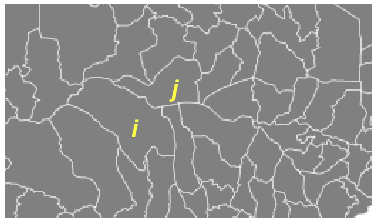
Moran's I Statistic

- Moran's I Statistic is a spatial autocorrelation technique that calculates the relationship between locations of observations (w_{ij}) the similarity between the attributes (c_{ij}) at those locations
- Given a set of features and an associated attribute, it evaluates whether the spatial data distribution is clustered, dispersed or random.
- It is applied to zones or points with continuous variables associated with them.

32

Moran's I Statistic

If zone *i* is neighboring zone *j*



Similarity of Attribute c_{ij} $\quad c_{ij} = (x_i - \overline{x})\,(x_j - \overline{x})$

Proximity of Location w_{ij} $\quad w_{ij} = 1/d_{ij}$

Sample Variance s^2 $\quad s^2 = \sum (x_i - \overline{x})^2$

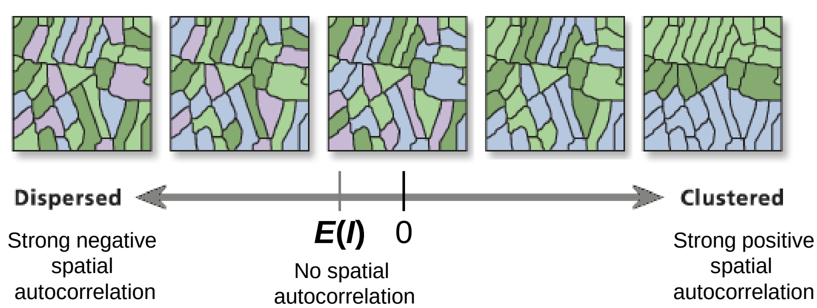
- Combines the measurement for attribute similarity and location proximity

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} c_{ij}}{s^2 \sum_{i=1}^{n}\sum_{i=1}^{n} w_{ij}}$$

- The numeric scale of Moran's I is anchored at the expected value of

$$E(I) = \frac{-1}{n-1}$$

- A Moran's I value equal to *E(I)* demonstrates that the pattern has no spatial autocorrelation and hence can be considered random



Testing for Significance

$$Z = \frac{I - E_I}{S^2}$$

Assuming a two-tailed test at the 0.05 significance level, the observed degree of spatial autocorrelation is significant if it is beyond the critical value of
Z = -1.96 or +1.96

Minimum Sample Size

- In statistics, it is often necessary to determine if our sample size is large enough to consider our results valid
- If the number of observations is too small the sample will not adequately represent the population
- Therefore, it is important to calculate the *minimum sample size* if there is spatial autocorrelation.
  - Generally the sample size will be larger than if spatial autocorrelation is not present
  - Different types of statistics have their methods for way calculating the *minimum sample size*

References

- David O'Sullivan, David Unwin, 2010 *Geographic Information Analysis*. Hoboken, NJ, John Wiley

- David W. S. Wong, Jay Lee, 2005. *Statistical Analysis of Geographic Information with ArcView GIS and ArcGIS*. Hoboken, NJ, John Wiley