GEOG 413/613

LECTURE 10

1

Multivariate Exploratory Data Analysis

- Graphical Methods
- PCA
- **K-Means Cluster Analysis**

2

2

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)

3

3

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.

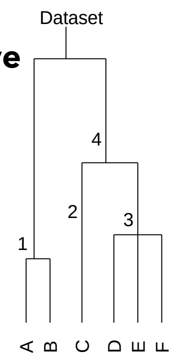
- Two methods are partitioning and hierarchical clustering

4

4

Cluster Analysis

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



Dendogram

5

5

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).

6

6

Cluster Analysis: K-means

- 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
  - commonly used as an unsupervised machine learning algorithm for partitioning a given data set into a set of *k* groups
    - *k* represents the number of non-overlapping groups (clusters) specified by the user

7

7

K-Means Analysis

- K-Means Clustering
  - Group membership is determined be calculating the centroid for each group, then assigning each observation to the group with the nearest centroid
  - The primary objective in k-means clustering is to define clusters so that the total **within-cluster variation** is minimized and the **between group variation** is maximized

8

8

K-Means

$$vc_k = \sum_{x_i \in c_k} (x_i - \mu_k)^2$$

Where:

- vc_k is the sum of the within cluster variation
- x_i is the data point belonging to the cluster c_k
- $\boldsymbol{\mu}_k$ the mean value of the points assigned to cluster c_k

K-Means Algorithm

1. Specify *k* the number of clusters/groups to be created
2. Select randomly *k* objects from the data set as the initial cluster centers or means
3. Assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid
4. For each of the *k* clusters update the cluster centroid by calculating the new mean values of all the data points in the cluster..
5. Iterate through 3 and 4 to minimize the total within sum of squaures

K-Means

- For a multivariate dataset
  - divided into *K* distinct clusters
  - points within a cluster are as close as possible in the multi-dimensional space
  - Points within a given cluster are as far away as possible from points in other clusters.
- The dataset is a set of objects (rows) with each with a set of *n* attributes

11

11

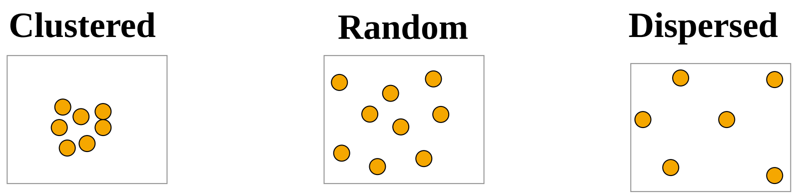
Point Pattern Analysis

12

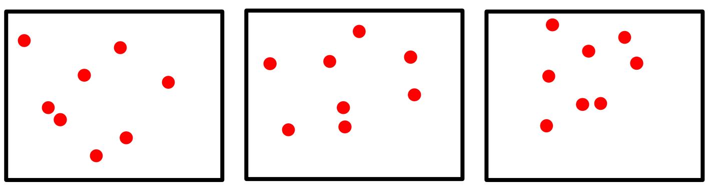
12

Examining Spatial Data

- How are the data points spatially distributed?



- How do you know? Always test



Point Pattern Analysis

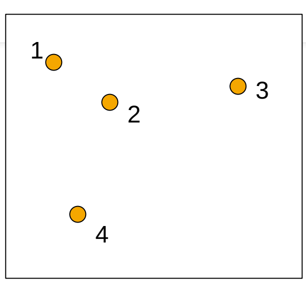
- A set of quantitative tools for examining the spatial arrangement of point locations on the landscape as represented by a conventional map.

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

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

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}

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**

17

17

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

18

18

Random Distribution

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

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

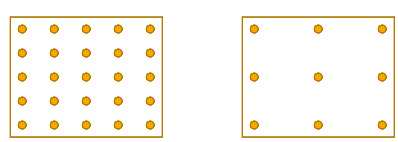
Where :
A is area of study region
n is number of points

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?



Clustered Distribution

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

$$r_{cst} = 0$$

21

21

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

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

22

22

Standardized Nearest Neighbour Index

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

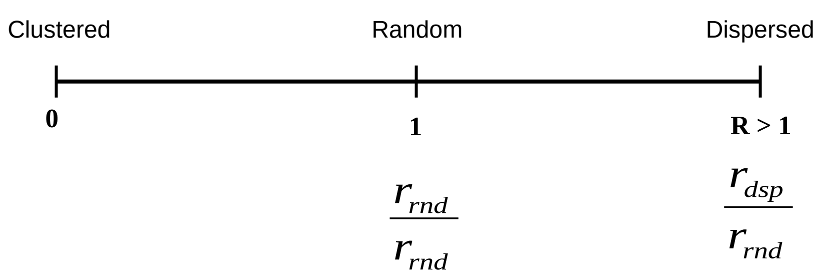
$$R = \frac{r_{dsp}}{r_{rnd}}$$

23

23

Standardized Nearest Neighbour Index

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



24

24

Test of Significance

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

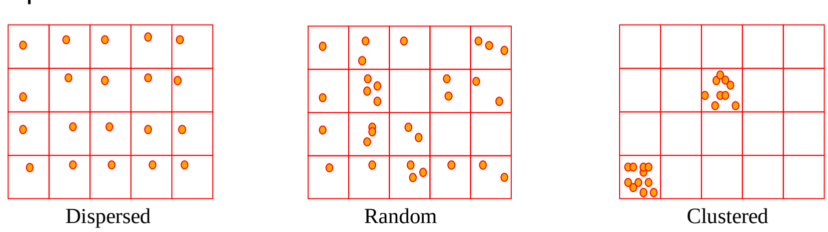
$$Z_r = \frac{r_{obs} - r_{rnd}}{\sigma_{obs}}$$

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

25

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



26

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

27

27

Quadrat Analysis

- Variance of Cell Frequencies

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

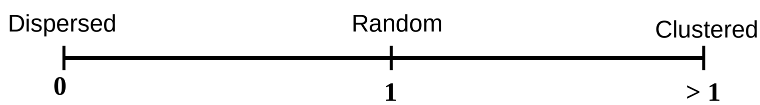
where f_i = frequency of cells with i cases
x_i = number of cases per cell

28

28

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:

$$X^2 = \text{VMR}\,(m - 1)$$