GEOG 413/613

LECTURE 6

1

Mid-Term Exam

1) Which of the following is a measure of dispersion?
   A. Standard deviation
   B. Average deviation
   C. Range
   D. All the above

2) PCA and K-means are methods of reducing the complexity of a dataset
   A. True
   B. False

3) Which of the following is used to show a relationship between variables?
   A. A scatter plot
   B. A Scree Plot
   C. A Bar Graph
   D. A Scatter Plot Matrix

4) Which of the following best describes an outlier?
   A. A stray datapoint
   B. An erroneous datapoint
   C. A datapoint with an extreme value
   D. A datapoint outsize of the range

2

2

1) What is the “modifiable area unit problem”?

2) Why is it important to scale and center a dataset before running the PCA algorithm?

3) How how determine if a dataset contains outliers?

4) An analysis of two separate datasets produced kurtosis values of 2.9 and 1.3 respectively. Briefly describe the meaning of these statistics.

3

3

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

4

4

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

5

5

Spatial Autocorrelation

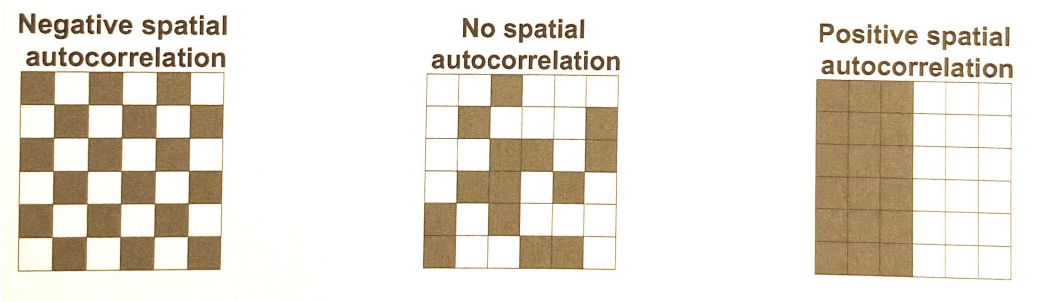
- Most statistical measures are based on the assumption that the values of observations in each sample are independent of one another (this is why we carry out random sampling)
- This independence may be violated by spatial autocorrelation. For example, positive spatial autocorrelation may violate the 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

6

6

Spatial Autocorrelation

- Positive spatial autocorrelation occurs when observations with similar values are closer together (i.e., clustered).
- Negative spatial autocorrelation occurs when observations with dissimilar values are closer together (i.e., dispersed)



7

7

Moran's I Statistic

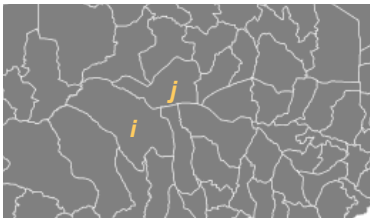
- Moran's I Statistic is a spatial autocorrelation technique that calculates the relationship between locations of observations (w_{ij}) the similarity between the attributes (c_{ij}) at those locations
- Given a set of features and an associated attribute, it evaluates whether the spatial data distribution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

8

8

Moran's I Statistic

If zone *i* is neighboring zone *j*



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

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

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

9

9

- Combines the measurement for attribute similarity and location proximity

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

10

10

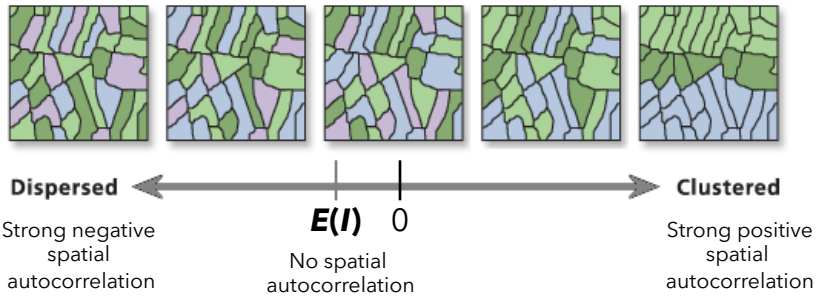
- The numeric scale of Moran's I is anchored at the expected value of

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

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

11

11



12

12

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

13

13

Local Indicators of Spatial Association

- Moran's I index can be disaggregated to provide a series of local indices
- LISA calculates for each feature an index value and a Z score
  - A high negative Z score indicates that the feature is adjacent to features of dissimilar values

14

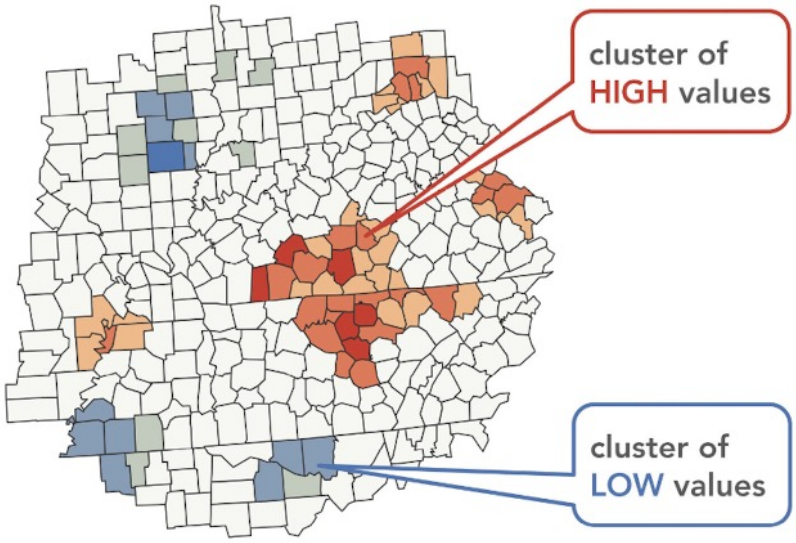
14

Local Indicators of Spatial Association

- Moran's I, Local or Global can only detect the presence of clustering of similar values
  - It cannot tell whether the clustering is made of high values or low values

15

15



16

16

G-Statistic for Measuring High/Low Clustering

- The G-Statistic can separate clusters of high values from clusters of low values.

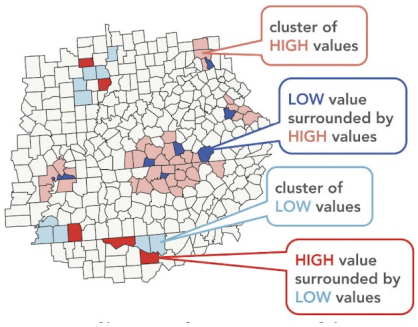
$$G(d) = \frac{\sum\sum w_{ij}(d)x_i x_j}{\sum\sum x_i x_j}, i \neq j \qquad E(G) = \frac{\sum\sum w_{ij}(d)}{n(n-1)}$$

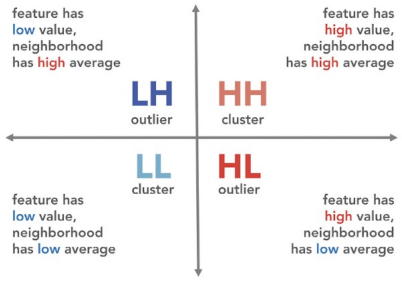
where x_i is the value at location i,
x_j is the value at location j if j is within d of i,
and $w_{ij}(d)$ is the spatial weight
$E(G)$ is the expected value of G

17

17

High/Low Clustering





18

18

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*

19

19

Spatial Interpolation

- The concept of distance decay—that interactions and similarities decline over space in ways that are often systematic—is a fundamental property of geographic data.
- It is associated Tobler's First Law of Geography
- These spatial dependence effects - the ways in which the characteristics of one location are correlated with characteristics of other nearby locations – at the reason we measure spatial correlation

20

20

Spatial Interpolation

- Sometimes we need to estimate the potential for hotspots or anomalies in how a variable is distributed in space.
  - We apply density estimation using a distance decay function
  - Kernel density algorithms are widely used
- On other occasions we need to estimate values at unsampled location

21

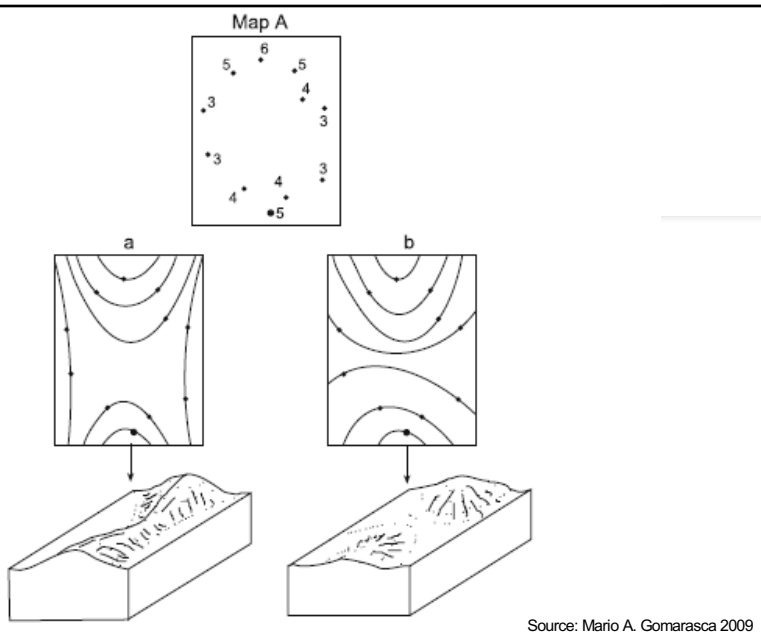
21

Spatial Interpolation

- Point interpolation is essential
  - no interpolation algorithm can be defined as best
  - Factors include:
    - ground morphology or terrain characteristics
    - the application and required accuracy of the output
    - the data structure that the interpolation software can manage

22

22



Source: Mario A. Gomarasca 2009

Two interpolation methods can produce different results

23

23

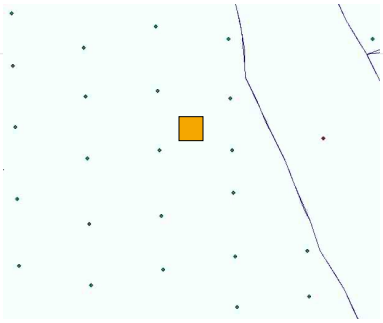
Building Surface Models

- Interpolation Methods
  - among the most popular are:
    - Inverse Distance Weighting
    - Nearest Neighbour
    - Splining
    - Kriging

24

24

DEM creation by interpolation



- **Inverse distance weighted** – simple and quick
- **Nearest neighbour** – honours raw values
- **Spline** – minimizes curvature -> smooth surface
- **Kriging** – uses spatial correlation of points

25

IDW

- IDW weights the value of each point by its distance to the known point being analyzed and averages the values.
- IDW assumes that unknown value is influenced more by nearby than far away points,
- One can control how rapid that distance decay is. Influence diminishes with distance.
- IDW has no method of testing for the quality of predictions, so validity testing requires taking additional observations.
- IDW is sensitive to sampling, with circular patterns often around solitary data points

26

26

Spline

- Fits a curve through the sample data and assigns values to other locations based on their location on the curve
- Thin plate splines create a surface that passes through sample points with the least possible change in slope at all points, that is with a minimum curvature surface.
- Uses piece-wise functions fitted to a small number of data points, but joins are continuous, hence can modify one part of curve without having to recompute whole
- Overall function is continuous with continuous first and second derivatives.

27

27

Spline

- SPLINE has two types: **regularized** and **tension**
- Tension results in a rougher surface that more closely adheres to abrupt changes in sample points
- Regularized results in a smoother surface that smoothes out abruptly changing values somewhat

28

28

Kriging

- Like IDW interpolation:
  - Kriging forms weights from surrounding measured values to predict values at unmeasured locations.
  - the closest measured values usually have the most influence.

29

29

Kriging

- However, the kriging weights for the surrounding measured points are more sophisticated than those of IDW.
- IDW uses a simple algorithm based on distance, but kriging weights come from a semivariogram that is developed by looking at the spatial structure of the data.
- Predictions are made for locations in the study area based on spatial arrangement of measured values that are nearby

30

30

Kriging

- In other words, kriging substitutes the arbitrarily chosen *p* from IDW with a probabilistically-based weighting function that models the spatial dependence of the data.
- The structure of the spatial dependence is quantified in the semi-variogram
- Semivariograms measure the strength of statistical correlation as a function of distance; they quantify spatial autocorrelation
- Kriging associates some probability with each prediction, hence it provides not just a surface, but some measure of the accuracy of that surface
- Kriging equations are estimated through least squares

31

31

DEM creation by interpolation

Inverse Distance weighted - simple

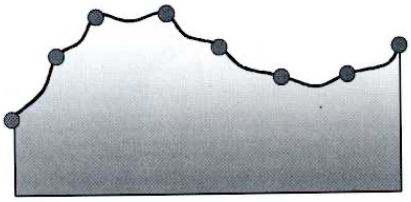


Figure 1.9. Cross section of inverse distance weighted interpolation.

Nearest neighbour – honours raw values

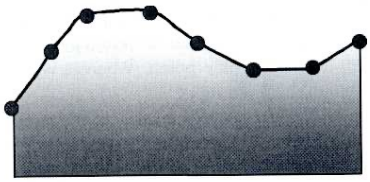


Figure 1.13. Cross section of linear natural neighbors interpolated surface.

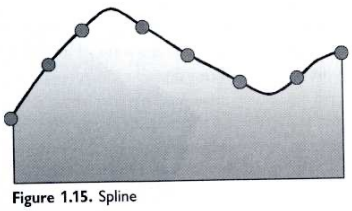


Figure 1.15. Spline

Spline – minimizes curvature -> smooth surface

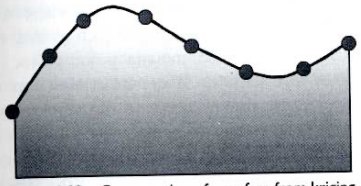


Figure 1.18a. Cross section of a surface from kriging.

Kriging – uses spatial correlation of points (employing semi-variogram of distance v difference)

32

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

33