GEOG 413/613

LECTURE 10
Course Review

1

Non-spatial Statistics

Measurement Concepts
- Precision
- Accuracy
- Validity
- Reliability

Classification Methods
- Equal intervals
- Quantile breaks
- Natural breaks

Presentation
- Histograms
- Frequency tables
- Scatter Plots
- Line Graphs

2

2

Spatial Statistics

Central Tendency
- Mean Center
- Weighted Mean Center
- Median Center
- Manhattan Center

Dispersion
- Standard distance
- Relative Distance

3

3

Non-spatial statistics

Central Tendency
- Mode
- Median
- Mean

Dispersion
- Deviation
- Average Deviation
- Range
- Standard Deviation
- Variance
- Covariance

Relative Position
- Kurstosis
- Skewness

4

4

Analysis of Spatial Data

- Spatial autocorrelation
- The modifiable area unit problem
- Ecological fallacy
- Scale
- Non-uniformity of space
- Edge/boundary effects

5

5

Geospatial Databases

- A database is an integrated set of data on a particular subject
  - Data are related and represent a specific aspect of the world
  - Data are for a specific purpose
- A geospatial database
  - Has entities (house, river, lake, road…)
  - Has attribute of these entities (location, size, type, name…)
  - Has spatial relationships (distances between entities, adjacency…)

6

6

Geospatial Databases

- Advantages
  - Data stored at a single location reduces redundancy
    - Consider cadastral data needed by different levels of govt or departments
  - Maintenance costs decrease
  - Multiple applications and users can use the same data
    - Data are not dependent on software
  - Data sharing is easier
    - Multiple interfaces and operations
  - Data security and standards
- Disadvantages
  - The cost of acquiring DBMS software can be quite high
  - A DBMS can add unnecessary complexities for data management in small projects
  - Single-user GIS will often be better for files rather DBs

7

7

DBMS

- Functions
  - A data model
  - Data input capability
  - Indexing
  - A query language
  - Security
  - Backup and recovery
  - Database administration tools
  - Applications
  - Application programming interfaces (APIs) for further customization

8

8

DBMS Types and Extensions

- Types:
  - relational (RDBMS)
  - object (ODBMS)
  - object-relational (ORDBMS)
- Extensions
  - ORDBMS have spatial database extension
  - Indexing; Storage management; Transaction services; Query language; DB replication services; Query parser; Query optimizer

9

9

Storing Geospatial Data in DBMS Tables

- Database tables are designed along the following principles
  - There is only one value in each cell at the intersection of a row and column.
  - All values in a column are about the same subject.
  - Each row is unique
  - There is no significance to the sequence of columns
  - There is no significance to the sequence of rows

10

10

SQL

- SQL (Structured Query Language) programming language designed to retrieve sets (row and column combinations) of data from relational databases
- It is the standard database query language it has geographic capabilities
- Some DBMS can have proprietary SQL extensions that are usually only used on their system

11

11

Database Design

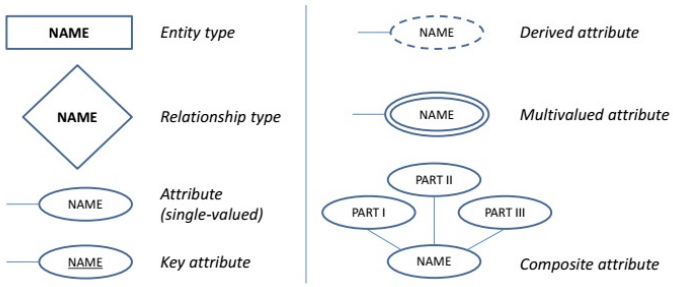
- Database Design
  - Conceptual model
  - Logical model
  - Physical model
- Database Schema and Instances
  - the description of the database is called *database schema*
  - the data items that reside in a database at a specific point in time form the *database instance*

12

12

Entity-Relationship Model

- ER-model: diagrammatic representation of the *miniworld* into a set of entities and their relationships.

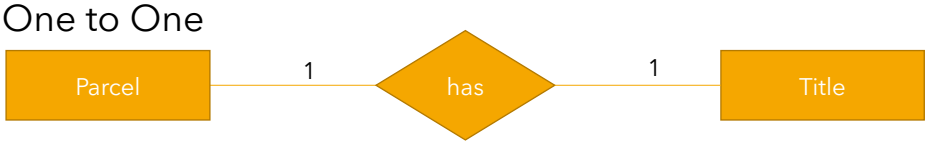


13

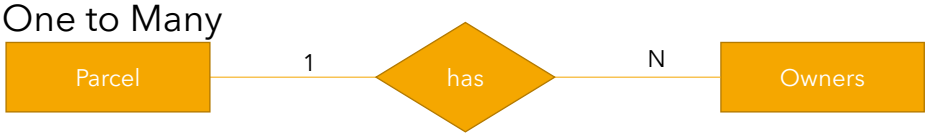
13

Entity and Relationship Types

- One to One



- One to Many



- Many to Many

14

14

Normalization

- The process minimizing redundancy in a database
  - Characterizes the level of redundancy in a relational schema
  - Provides mechanisms for transforming schemas in order to remove redundancy
  - Data normalization follows certain rules which are categorized as "normal forms". There are 6 normal forms but we'll briefly look at only 3
- In principle, any information that can be applied to more than one record should be moved to its own table.
  - Each successive normal form applied must meet the rules of the previous form

15

15

Indexing Geographic Information

- A database index is a special representation of information about objects that improves searching
  - B-tree indexing
  - Grid indexing
  - Quadtree indexing
  - R tree

16

16

Editing and Data Maintenance

- Transaction
- Versioning

17

17

GeoWeb

- Web GIS
  - web service technology
    - Key elements:
      - A server and a client
      - The server performs the requested GIS operations and sends responses to the client via HTTP.
      - The format of the response sent to the client can be in many formats, such as HTML, binary image, XML, JSON, etc

18

18

Web GIS Advantages

- A wide reach
- A wide user base
- Cross-platform capability
- Low cost (relative to potential usage)
- Easy to use
- Unified updates
- Numerous applications

19

19

Essential elements of a web GIS **application**

- A web application
  - Software to visualize and interact with geographic information
- Digital basemaps
  - Geographic context for each application e.g. Transportation, Topographic, Terrain, Imagery
- Operational layers
  - Additional layers for the operation e.g. sensor feeds, editing layers
- Tasks and tools in the web GIS application
  - Client tasks, server tasks
- One or more geospatial databases

20

20

4 Broad Themes of Big Data

- Information
  - Data are created, shared and utilised extensively in recent times
  - The proliferation of personal mobile devices
    - connected to the Internet
    - equipped with digital sensors
  - Expanding variety in form
- Pervasive (Wide impact)
  - Many fields
    - Examples: Elections; Google searches linked to epidemiology and economics

21

21

4 Broad Themes of Big Data

- Technology
  - Needs intensive computational and storage specs
  - Hadoop
    - Open source parallel computing.
    - Google, Yahoo, FaceBook
- Methods of Analysis
  - cluster analysis; genetic algorithms; natural language processing; machine learning; neural networks; predictive modelling; regression models; social network analysis; sentiment analysis; signal processing and data visualisation

22

22

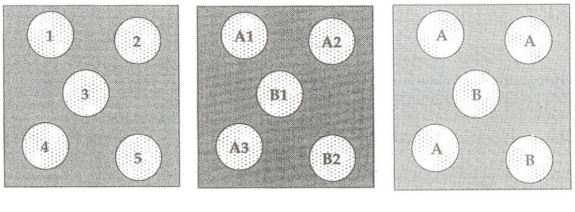
Landscape Metrics

- Pattern and Process
  - Processes in natural systems
    - Spatial pattern (form) can influence process
  - Spatial patterns
    - Formed from processes
    - Patterns tell us about process
  - Quantifying the complexity of nature
    - measure ecological functioning (e.g. biodiversity, connectivity)
    - measure land use processes (land consumption, fragmentation, urban sprawl)

23

23

Landscape Metrics



Patch level | Class level | Landscape level

Farina, A. 2000. Landscape ecology in action. Kluwer Academic Publishers, Netherlands

Patch-level metrics are defined for individual patches, and characterize the spatial character and context of patches.

Class-level metrics are integrated over all the patches of a given type (class).

Landscape-level metrics are integrated over all patch types or classes over the full extent of the data (i.e., the entire landscape).

24

24

Landscape Metrics

Composition

- Proportional Abundance
- Richness
- Diversity
- Evenness

Configuration

- Patch size distribution and density:
- Patch shape complexity:
- Core Area
- Isolation/Proximity:
- Contrast
- Dispersion
- Contagion and Interspersion
- Subdivision
- Connectivity

25

25

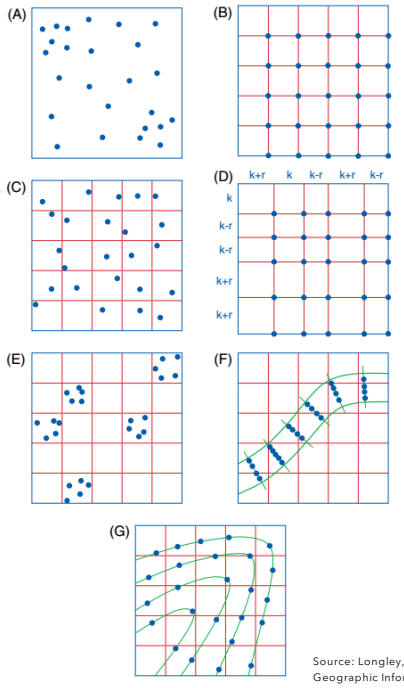
Sampling Methods

- Provides knowledge about a whole population
  - i.e. make inference about a population from the sample data
- Larger sample sizes are more accurate representations of the whole
  - Large samples are costly: time, labour
    - Can be wasteful since we can statistically infer from appropriate samples
- A sampling strategy with the minimum bias is the most statistically valid

26

26

Sampling Methods



Spatial sample designs: (A) simple random sampling, (B) systematic sampling, (C) stratified random sampling, (D) stratified sampling with random variation in grid spacing, (E) clustered sampling, (F) transect sampling, and (G) contour sampling.

Source: Longley, Paul A.; Goodchild, Michael F.; Maguire, David J.; Rhind, David W.. Geographic Information Science and Systems, 4th Edition. Wiley.

27

27

Random Sampling

- Random sampling: each member of the population has an equal chance of being selected
  - Advantages:
    - Can be used with large sample populations
    - Avoids bias
  - Disadvantages:
    - Can disproportionately represent some parts of the population at the expense of others

28

28

Systematic Sampling

- Systematic Sampling: Samples are chosen at regular intervals
  - Sample locations are evenly distributed for example every two metres along a transect line
    - systematic sampling implies a regularly spaced grid
      - Advantages:
        - It is more straight-forward than random sampling
        - Provides a good coverage of the study area
      - Disadvantages:
        - It is more biased: not all points have an equal chance of being selected
        - It may lead to over or under representation if there is periodicity in the data (e.g. sampling at the same interval as the location of erosion barriers along a beach. Or a city road grid)

29

29

Stratified sampling

- Stratified sampling: used when the parent population is made up of sub-groups that of interest.
  - Divide the sampling design into strata(classes), and then select a sample from each stratum
  - The strata are defined so that individuals inside each class are similar based on the characteristic believed to influence the phenomena

30

30

Stratified sampling

- Advantages:
  - If the proportions of the subgroups are known, the results are representative of the whole population
  - Correlations and comparisons can be made between subgroups
- Disadvantages:
  - The proportions of the subgroups must be known

31

31

Levels or Scales of Measurement

- Nominal
  - Categorical data e.g. land use type, religious affiliation
- Ordinal
  - Ranked data , e.g. main, secondary, minor roads
- Interval:
  - Interval between any two units can be measured on scale. Zero value is assigned arbitrarily e.g. Celsius and Fahrenheit scales (80°F is not twice as hot as 40°F)
- Ratio:
  - interval data with an absolute zero value

32

32

Multivariate Exploratory Data Analysis

- the initial investigations on data
  - discover patterns
  - Reduce dimensions
  - identify anomalies/outliers
  - test hypothesis (e.g. observed vs expected)
  - check assumptions
    - Descriptive statistics
    - Visualization

33

33

Multivariate Exploratory Data Analysis

- Common Tools
  - Histogram
  - Box plot
  - Scatter plot matrix
  - Bar graph
  - Parallel Coordinate Plots

34

34

Dimensionality Reduction

- Data dimension is the number of variables for a measured theme/dataset
- Data with a high dimensionality is difficult to visualize
- Reducing the dimensionality of the data helps understand the intrinsic aspects of the dataset
  - Find structure within features
  - Aid in visualization
- The methods maximize information while minimizing differences between the original data and the new lower dimensional representation
- Principal Component Analysis is one such method

35

35

Principal Component Analysis

- Widely used method for dimensionality reduction
- The transformation between original data and the new lower dimensional representation is a linear projection
  - find a linear combination of the original features principal components.
  - The principal components will maintain as much as is possible the same variance as the original data
  - The principal components are uncorrelated (orthogonal)

36

36

Principal Component Analysis

- PCA major steps
  - Standardize the variables
    - Centre (deviation from the mean)
    - Scale (divide the deviation by the standard deviation)
  - Calculate the covariance matrix
    - Covariance – how 2 variables vary with each other
    - If you have more than 2 variables, then you gave more than one covariance (given variables x,y,z –> cov(x,y), cov(x,z), cov(y,z))
  - Calculate the eigenvectors and eigenvalues of the covariance matrix

$$Av = \lambda v$$

$$A - matrix$$
$$v - eigenvector\ for\ \lambda$$
$$\lambda - eigenvalue\ for\ v$$

37

37

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)

38

38

Clustering Analysis

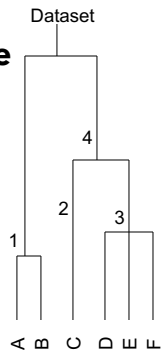
- Searching for groups in the data in such a away that objects belonging to the same cluster resemble each other, whereas objects in different clusters are dissimilar.
- Two methods are partitioning and hierarchical clustering

39

39

Cluster Analysis

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



Dendogram

40

40

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).

41

41

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

42

42

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

43

43

Point Pattern Analysis

- Examining the spatial arrangement of point locations on the landscape
- Two methods are ***nearest neighbour analysis*** and ***quadrat analysis***.
  - Standardized Nearest Neighbour Index
  - Variance-Mean Ratio

44

44

Measures of Spatial Autocorrelation

- Moran's I Statistic
  - A global spatial autocorrelation measure that calculates the relationship between locations of observations (w_{ij}) the similarity between the attributes (c_{ij}) at those locations
- Local Indicators of Spatial Association
  - G-Statistic (Getis-Ord's)for Measuring High/Low Clustering

45

45

Bivariate Regression

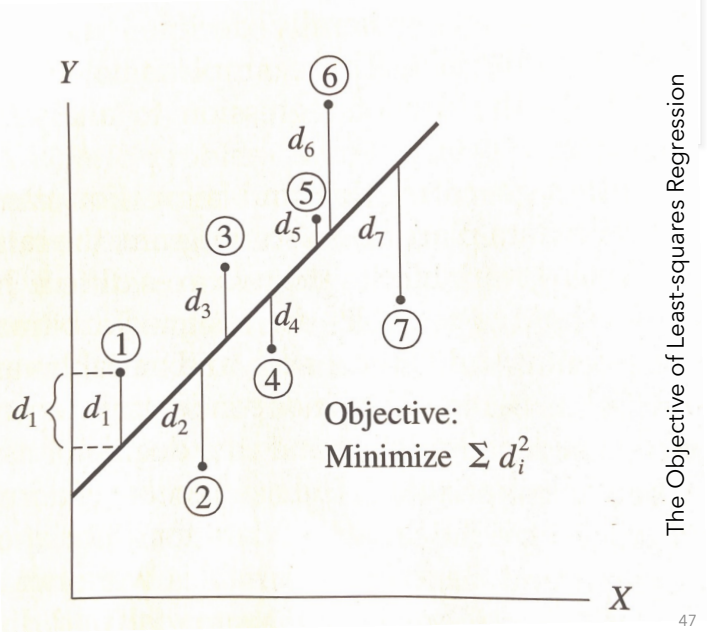
- Bivariate Analysis
  - Two variables are explored in detail
  - Assumption is that one variable influences/affects the other
    - Independent Variable (Explanatory Variable)
      - The variable creating the influence/effect
    - Dependent Variable
      - The variable receiving the influence or effect

46

46

Bivariate Regression Line

- The least-squares regression line is unique
  - Minimizes the sum of the squared vertical distances between each data point and the line
- The regression is given by the straight-line equation
  $Y = a + bX$
- *a* is the intercept on the Y-Axis
  - Represents the value of Y when X is zero
- *b* represents the slope of the line
  - Also, the correlation coefficient



The Objective of Least-squares Regression

47