Reminder

- Final Exam on Thursday, lecture time!
- Review the slides/book chapters yourself
- Lab 6 due: Sunday, September 16, 2018, 11:55 PM (NO LATE SUBMISSION!)
- Bonus lab due: Sunday, September 16, 2018, 11:55 PM (NO LATE SUBMISSION!)
- ESCI Evaluation if you haven’t done it yet
- Bring a scantron, a draft paper, a pencil, an eraser and a calculator
- No cheatsheet; formulas will be provided with questions.
- Last office hour today after the lecture!
GEOG 176A: Introduction to Geographic Information Systems

Lecture 21: Final Review

Rui Zhu
What we learned?

GIS Fundamental
- Definition and History of GIS - L01L02 / Chapter 1
- Modeling Locations and scale - Cartographic Roots in GIS - L03L04 / Chapter 2

GIS Core
- Computer Science
  - Obtaining and storing data in GIS - L05L06L07L08 / Chapter 3-4
  - GIS Database Management - L09 / Chapter 5
- Statistics/Cartography
  - Visualization in GIS - L15L16 / Chapter 8
  - Spatial analysis in GIS - L10-L14 / Chapter 6-7

Others
- GIS systems - L17L18 / Chapter 9
- GIS applications - L19L20 / Chapter 10
- Future of GIS - L21
GIS concepts …
Lecture 1 - 2

● Definitions of GIS
  ○ Toolbox
  ○ Information system
  ○ GIScience
  ○ GIS industry
  ○ Society

● Spatial primitives

● Difference between spatial and non-spatial data: spatial data can be referenced to a location on a map
Lecture 1 - 2

● History of GIS
  ○ GIS’s origins lie in thematic cartography (overlay) - Before Computer (BC)
  ○ Early GISs: CGIS, ODYSSEY, ArcView … - After Tomlinson (AT)

● The development of GIS has been significantly facilitated by
  ○ PC
  ○ Workstation
  ○ Graphic user interface (GUI)
Lecture 3 - 4

- GIS’s root in cartography
  - A map is a representation of some geographic phenomenon as a set of symbols and at a particular scale whose representative fraction is less than one to one.
  - How to measure locations on a map is essential of cartography

- Symbols: representation of geospatial information
- Map scales: granularity of geospatial information
- Models of the Earth - locating geospatial information
  - On the Earth
    - Geodesy (sphere, ellipsoid)
    - Geodetic datum
    - Geoid
    - Geographic coordinate system
  - On maps
    - Projected coordinate system
Lecture 3 - 4

Step 1: Mapping data onto the Earth surface
Geographic Coordinate System (GCS)

Step 2: Projecting data onto a flat map
Projected Coordinate System (PCS)

Azimuthal  Cylindrical  Conical
Lecture 5 - 8

GIS: Computational modeling

- Conceptual model:
  - Whether we should consider geographic information as discrete objects or continuous field?

- Logical model
  - Objects: objects associated with coords and attributes?
  - Field: a grid with an array of cells?

- Physical model
  - How are we going to encode data: binary or hexadecimal or …?
  - Whether data are stored in flat files or a relational database?
  - In what format the data should be stored?
    - Objects: coverage, AutoCAD, PostScript, …
    - Filed: Grid, TIFF, PNG, JPEG, …
Lecture 5 - 8

- Discrete → object → vector:
  - points + lines (polylines) + areas (polygons)
- Continuous → field → raster:
  - Cells with resolution + values
- Use vector to model field → TIN
Lecture 5 - 8

● Obtaining and storing geospatial data
  ○ To obtain geospatial data: importing from existing databases + digitizing + scanning + on-screen digitizing + field work
  ○ To store geospatial data:
    ■ Representation: binary + ASCII
    ■ Data format: flat file + geodatabase (relational)
Lecture 9

● Core functions of DBMS
  ○ Database definition
    ■ Data dictionary
  ○ Data entry
  ○ Data organization and management
    ■ Flat files
    ■ Relational database (primary key and foreign key + table joins and relates)
  ○ Data update (e.g., add, delete and change)
  ○ Data query
    ■ Select by attributes
    ■ Select by locations
    ■ SQL (structured query language)
Lecture 10 - 14

- **Non-spatial analysis**
  - **Descriptive analysis**
    - Categorical /text data → word frequency / geospatial semantic analysis / natural language processing ...
    - Numeric data → max, min, range, mean, median, quantile, variance, std, pearson’s correlation coefficient ...
  - **Inferential analysis**
    - Estimation of parameters (degree of freedom)
    - Confidence interval
  - **Data visualization:** histogram, bar/pie chart, line chart, box plot, scatter plot, radar plot …
Lecture 10 - 14

- **Spatial analysis**
  - Descriptive analysis
    - Point patterns analysis: min, max, and range of coordinates, (weighted) mean center, standard distance, average nearest neighbor index
    - Spatial clustering analysis: K-means, DBSCAN, hotspot analysis
  - Inferential analysis
    - Regression models: (multiple) linear regression, geographically weighted regression
    - Spatial interpolation (in terrain analysis)
Lecture 10 - 14

- Terrain analysis
  - Representation of terrain
    - Vector: point samples (regular and irregular), contours
    - Raster: DEM/DTM/DSM
    - TIN
    - Others: voxel, 3D point cloud
  - Terrain analysis
    - Slope and aspect
    - Hillshade
    - Elevation profile
    - Intervisibility and viewshed
    - Curvature
    - Flow accumulation
Lecture 10 - 14

- Point samples → interpolation
  - Deterministic: IDW
  - Stochastic: Kriging
    - $W_i$ is calculated based on semivariogram
    - Semivariogram is calculated based on the variance of attribute change with regard to distance
Lecture 15 - 16

- Cartographer’s paradox
- Elements of a map
- Different types of maps
- How to choose a suitable map? → normalization, human cognition

HSV
- **Hue**: specifies the hue (color) to which the color will be set
- **Saturation**: specifies the intensity of saturation to which the color will be set
- **Value**: specifies the intensity of white in the color
Lecture 17 - 18

- Which GIS?
  - What functions should this GIS have?
    - Data capture: dissolving, mosaicing, rubber sheeting, generalization
    - Storage: data fusion (geo-ontology)
    - Management: masking
    - Retrieval: identifying, buffer
    - Analysis: map algebra (local, focal, zonal, global)
    - Display / visualization
    - Others: documentations, model builder, arcpy
  - Commercial or Open Source GIS? → advantages and disadvantages
  - Do you need to develop your own GIS applications? → algorithms vs applications
Lecture 19

GIS in action:

Using GIS to facilitate information management and decision making in many domains:

- Climate change, transportation, city planning, security, health, disaster responses, resource allocation, agriculture, business management, military and so on…..

John Snow’s 1854 Cholera Map - 1854
Lecture 20

● The future of GIS
  ○ Big Data: different types of sensors (Obtaining and Storing Geographic Data)
  ○ GeoAI (Spatial Analysis)
  ○ CyberGIS (GIS Database Management)
  ○ Web-GIS / Semantic Web (GIS Systems)
  ○ Smart City (GIS applications)
  ○ Geoprivacy
What we learned again?

**GIS Fundamental**
- Definition and History of GIS - L01L02 / Chapter 1
- Modeling Locations and scale - Cartographic Roots in GIS - L03L04 / Chapter 2

**GIS Core**
- Computer Science
  - Obtaining and storing data in GIS - L05L06L07L08 / Chapter 3-4
  - GIS Database Management - L09 / Chapter 5
- Statistics/Cartography
  - Visualization in GIS - L15L16 / Chapter 8
  - Spatial analysis in GIS - L10-L14 / Chapter 6-7

**Others**
- GIS systems - L17L18 / Chapter 9
- GIS applications - L19L20 / Chapter 10
- Future of GIS - L21
Spatial computation ...
Earth as the WGS84 ellipsoid

For the WGS84

\[ a = 6,378,137 \]
\[ b = 6,356,752.3 \]

so \( f = 1/298.257 \)

\[ f = 1 - \frac{b}{a} \]

f: flattening of the ellipsoid
Practice question on UTM:

- Ellison Hall is located at around (23846, 3811950) at Zone 11, Northern hemisphere in a UTM coordinate system:
  - What is the longitude range of the zone?
    - [114° west, 120° west] (6° × 20 - 6° and 6° × 20)
  - What is the longitude of the central meridian?
    - 117° west (114° + 3°)
  - Is Ellison Hall to the west or to the east of the central meridian?
    - West (23846 meters is smaller than 500,000 meters)
  - What is the approx. distance between Ellison Hall and the North Pole?
    - About 10,000,000 meters - 3,811,950 meters = 6,188,050 meters
    - To south pole? About 10,000,000 meters + 3,811,950 meters = 13,811,950 meters
Military Grid Reference System (MGRS)

- Based on the UTM coordinate systems
- Adopted by the U.S. Army in 1947
- Used by many other countries and organizations
- Each south-north strip is divided into 20 cells (every cell covers 8 latitudes and 6 longitudes)
- These cells are named from letter C to letter X (letters A, B, I, O, Y, Z are not used)
Military Grid Reference System (MGRS)
Units in angles

- **Degree, minute, second**
  - 1 degree = 60 minutes; 1 minute = 60 seconds

- **Radian to degree**
  - 360 degrees (rotational measure) equals $2\pi$ radians (distance measure)
  - 180 degrees = $\pi$ radians!
    - 1 degree = $\pi / 180$ radians
    - 1 radian = $180 / \pi$ degree
  - E.g.
    
    \[
    \frac{5\pi}{12} \rightarrow \frac{5\pi}{12} \left( \frac{180^\circ}{\pi} \right) \rightarrow 5(15^\circ) \rightarrow 75^\circ
    \]

    \[
    45^\circ \rightarrow 45^\circ \left( \frac{\pi}{180^\circ} \right) \rightarrow \frac{1}{45} \left( \frac{\pi}{180^\circ} \right) \rightarrow \frac{\pi}{4}
    \]
Analysis on attribute data - numbers

- **Mean** is the sum of the values for one attribute divided by the number of records

\[ \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \]

- Mean is sensitive to outliers
- Examples:
  - A set of elevation values in meters: \( \{7, 15, 36, 39, 40, 41\} \), what is the mean?
  - What about: \( \{7, 15, 36, 39, 40, 2410\} \)?
Analysis on attribute data - numbers

- **Median** is the number separating the attribute values into higher half and lower half
  - Order all values from the minimum to maximum, and median is the number that is in the middle
  - If the total number of values is odd, then median is the middle number; If the total number is even, the median is the average of the two in middle
- Median is not sensitive to outliers
- Examples:
  - A set of elevation values in meters: \{7, 15, 36, 39, 40, 41\}, what is the median?
  - What about \{7, 15, 36, 37, 39, 40, 41\}
  - And \{7, 15, 36, 37, 39, 40, 2410\}?
Analysis on attribute data - numbers

- Both variance and standard deviation measure the spread of the data, i.e., how far a set of numbers spread out?

- Variance: $\sigma^2 = \frac{\sum (x_i - \mu)^2}{n}$

- Standard deviation ($\sigma$): the square root of variance

- Example:
  - A set of elevation values: {7, 15, 36, 39, 40, 41}, what is the variance and standard variance of this data?
Normal distribution

- A normal distribution is determined by two parameters:
  - Mean ($\mu$) and standard deviation ($\sigma$)
- If the data $X$ follows a normal distribution, then:
  \[ X \sim N(\mu, \sigma) \]

- If $X \sim N(0, 1)$, $X$ follows a standard normal distribution
Standard normal distribution

- A standard normal distribution $X \sim N(0, 1)$ allows us to calculate the probability of different value ranges:
  - What is the probability for having values in $[0, 1]$?
  - What is the probability for having values in $[0, 2]$?
  - What is the probability for having values $> 1$?
Standard normal distribution

Practice question: the precipitation (in inches) of all cities in a country follows \( N(40, 10) \). What is the probability that a city may have a precipitation larger than 60 inches?
Inferential analysis on attribute data (Estimation)

- **Degree of freedom (df)**: the amount of information used to estimate a parameter of a population
- **How to calculate df?**
  - The number of sample records minus the number of parameters which have to be estimated during the intermediate steps
  - E.g., if you have \( N \) sample records, the df of estimating population mean is \( N \)
  - E.g., if you have \( N \) sample records, the df of estimating population variance or std is \( (N-1) \)
- **Different numbers of sample records bring different dfs for the estimation**
Examples of Estimation

To estimate the average soil PH value of a big farm, you decide to collect some samples:

**Trial 1:** you collected 50 samples, obtained a sample mean as 6 and a sample std as 0.5, you know the $t = 2.01$ when $df = 49$ with 95% confidence. What is the 95% confidence interval for the average PH?

$$\bar{X} - t_{N-1} \cdot \frac{s}{\sqrt{N}} \leq \mu \leq \bar{X} + t_{N-1} \cdot \frac{s}{\sqrt{N}}$$

$$\frac{6-2.01 \times 0.5}{\sqrt{50}} \leq \mu \leq \frac{6+2.01 \times 0.5}{\sqrt{50}}$$

$$6-2.01 \times 0.07 \leq \mu \leq 6+2.01 \times 0.07$$

$$5.86 \leq \mu \leq 6.14$$

**Conclusion:** You have 95% confidence to say that the PH value is between [5.86, 6.14], when 50 samples were collected.
Examples of Estimation

To estimate the average soil PH value of a big farm, you decide to collect some samples:

**Trial 2:** you collected 5 samples, obtained a sample mean as 6 and a sample std as 0.5, you know the $t = 2.78$ when $df = 4$ with 95% confidence. What is the 95% confidence interval for the average PH?

$$
\bar{X} - t_{n-1, \frac{1}{2}} \cdot \frac{s}{\sqrt{N}} \leq \mu \leq \bar{X} + t_{n-1, \frac{1}{2}} \cdot \frac{s}{\sqrt{N}}
$$

$$
6 - 2.78 \times \frac{0.5}{\sqrt{5}} \leq \mu \leq 6 + 2.78 \times \frac{0.5}{\sqrt{5}}
$$

$$
6 - 2.78 \times 0.22 \leq \mu \leq 6 + 2.78 \times 0.22
$$

$$
5.39 \leq \mu \leq 6.61
$$

**Conclusion:** You have 95% confidence to say that the PH value is between [5.39, 6.61], when only 5 samples were collected.
Descriptive analysis on spatial data

- Min, max and range of coordinates
- MinX, maxX, minY, maxY define the minimum bounding rectangle

(MinX, MinY)  (MaxX, MinY)  (MinX, MaxY)  (MaxX, MaxY)
Descriptive analysis on spatial data

Practice question:

- You have three points whose coordinates are in the form of \((x, y): (1, 8), (4, 7)\) and \((3, 9)\)
- What is the coordinate of the upper left corner of the minimum bounding rectangle of the three points.
Descriptive analysis on spatial data

- **Mean center**: for a set of points, we can calculate their mean for x and y respectively.
Descriptive analysis on spatial data

- **Average nearest neighbor index (ANN)**
  - Could examines the spatial distribution of features (clustered or dispersed)?
  - Compare the average nearest neighbor of observed points with a hypothetical situation in which the same number of points are randomly distributed.

\[
ANN = \frac{\bar{D}_O}{\bar{D}_E}
\]

\[
\bar{D}_O = \frac{\sum_{i=1}^{n} d_i}{n}
\]

\[
\bar{D}_E = \frac{0.5}{\sqrt{n/A}}
\]

This is a theoretical value!
Estimation of linear regression

- Model: \( y = a + bx + \epsilon \)
- Estimation of \( a \) and \( b \) (Least square estimation):
  \[ \epsilon = \hat{y} - y \] where \( \hat{y} = a + bx \) is the estimated value using the model, and \( y \) is the observed value

  minimize \( \epsilon^2 \) to calculate \( a \) and \( b \)

- \( \epsilon \) is named as residuals
- Other estimation approach: maximum likelihood estimation (MLE)
Goodness-of-fit of linear regression

Many approaches:

- Diagnostic on residuals
- R-squared
- Adjusted R-squared
- Chi-square goodness of fit tests
- Split the data into training and testing → cross validation

“All Models are wrong, but some are useful.”
R-squared

● A metric to quantify how good your model fits to your data
● Specifically, it describes **how much variance** the model explained for the data
● Has a value ranges between **[0, 1]**
● 0 means the model does not fit the data at all
● 1 means the model fits the data perfectly
IDW interpolation

- Inverse Distance Weighting
- Closer points have more influences on the target value:

\[ Z = \frac{\sum w_i z_i}{\sum w_i} \text{ where } w_i = \frac{1}{(d_i)^p} \]

- \( p \) is the power parameter which determines the influence of the distance towards weights:
  - \( p=0 \): every point has the same weight regardless of their distance towards the target location
  - \( p=\text{large value} \): points that are further away have less influence on the target location
IDW interpolation

Practice question: you have three points whose elevations are 7, 9 and 11 meters, and their distances to the target locations are 1, 2 and 3 meters respectively. Using a power of 2, what would be your estimation for the elevation at the target locations?

$$Z = \frac{\sum w_i z_i}{\sum w_i} \quad \text{where} \quad w_i = \frac{1}{(d_i)^p}$$

$$Z = \frac{(1/1)^2 \times 7 + (1/2)^2 \times 9 + (1/3)^2 \times 11}{(1/1)^2 + (1/2)^2 + (1/3)^2} = 7.69$$
Reminder

- Final Exam on Thursday, lecture time!
- Review the slides/book chapters yourself
- Lab 6 due: **Sunday, September 16, 2018, 11:55 PM (NO LATE SUBMISSION!)**
- Bonus lab due: **Sunday, September 16, 2018, 11:55 PM (NO LATE SUBMISSION!)**
- **ESCI Evaluation if you haven’t done it yet**
- Bring a scantron, a draft paper, a pencil, an eraser and a calculator
- No cheatsheet; formulas will be provided with questions.
- Last office hour today after the lecture!
KEEP CALM AND GOOD LUCK ON YOUR EXAMS