GEOG 176A: Introduction to Geographic Information Systems

Lecture 10: Spatial Analysis I
(chapter 6)

Rui Zhu
Review

GIS Database Management

- Database definition → Data dictionary
- Data entry → manual entering + import from existing databases
- Data organization and management → relational databases (primary/foreign keys, join/relate tables)
- Data update → add, delete change
- Data query → non-spatial attributes and spatial entities
Common spatial operations / queries

- **Buffering**: create zones around points, lines or areas based on distance
Common spatial operations /queries

- **Overly**: calculate values based on multiple map layers
Common spatial operations /queries
Model builder

- Implemented in ArcGIS
- Combine multiple operations/queries together, and save it for next time’s use
Spatial analysis - Why is it there?

- We have a lot of geographic data, but how can we use them?
- Spatial Analysis:
  - Hypothesis testing: I have a hypothesis, could my geographic data help to validate or reject my hypothesis?
  - Knowledge discovery: Could I discover some knowledge which was not known previously from the data?
  - Predicting the future: Could I develop a model based on the data to predict what is likely to happen in the future?
- A lot of GIScience researches are about spatial analysis (or GIS analysis)
Spatial analysis in GIS

ArcGIS Toolboxes
Spatial analysis v.s Non-spatial analysis

- How is spatial analysis different from non-spatial analysis?
  - Spatial analysis absorbs a lot of methods from other domains, e.g., statistics, computer science, economics ...
  - What makes spatial analysis unique is the geospatial data
  - Spatial analysis examine:
    - Coordinates (geospatial locations)
    - Attribute (non-spatial information)
    - Links between them
Spatial analysis

- Analysis on attribute data (non-spatial data)
  - E.g., population, temperature, soil PH value, …
  - Often relates back to the spatial part → where is this outlier?

- Analysis on spatial data (coordinates/locations)
  - E.g., spatial extent covered, distributed patterns, network connectivity, …
  - Often relates back to the attribute part → what makes the spatial data clustered?
Analysis on attribute data

- Often **first** focuses on **one** attribute (one column) of the data table

- What is the type of the data?
  - Categorical / Text data: e.g., Land cover and land use, content of tweets
  - Numeric data: e.g., elevation
Analysis on attribute data - categorical / texts

- Word frequency summary (word cloud)
- Geospatial semantic analysis ("mountains" vs "hills")
- Natural language processing
- ....

Semantic analysis on geographic information is an active research field!
Analysis on attribute data - numbers

- For attributes that are in numeric values, mature statistical metrics can be applied

- Statistical metrics:
  - Max, min, range
  - Mean, median, quantiles
  - Variance and standard deviation
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

● Quartiles: the three points that divide the data set into four equal groups
  ○ Lower (25%), median (50%) and upper (75%) quartiles
  ○ Order all values from the minimum to the maximum; divide the data into two halves and get the median; the median of the lower half is the lower quartile; the median of the upper half is the upper quartile

● Examples:
  ○ A set of elevation values: {7, 15, 36, 39, 40, 41}, what is the lower, median and upper quartiles?
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?
Analysis on attribute data - numbers

- Data visualization: explore the data intuitively
- Typical visualizations for numeric values:
  - Histogram
  - Bar chart, pie chart
  - Line chart
  - Box plot
  - Scatter plot
  - Radar plot
  - ...
Histogram

- Divide the data into several groups
- Summarize the frequency of each group
Bar chart and pie chart

- Summarize the frequency of data in each category
Use bar chart to visualize a classified raster data
Line chart

- Often used for temporal sequence data

Travel flows in different hours
Box plot

- Based on min, lower quartile, median, upper quartile and max
Scatter plot

- Visualize the potential relations between two attributes
Radar plot

- Visualize multiple attributes
Data distributions

Attribute data can be distributed in different ways:
Normal distribution

- A very common distribution
- Also called Gaussian distribution
- Widely used in various statistical methods/assumptions
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?