Geography 360: GIS & Mapping

Geographic Data Modeling

Data Quality

Vaishnavi Thakar

UNIVERSITY *of* WASHINGTON

Review

- **Representation is a fundamental issue in GI**
- **Discrete Objects and Continuous Fields**
  - Two fundamental ways of representing geography
- **Raster and Vector**
  - two methods of representing geographic data in digital computers
- **Topology : Mathematics and science of geometrical relationships.**
- **Surfaces**

Learning Objectives

- Understand the concept of uncertainty, and the ways in which it arises from imperfect representation of geographic phenomena.
- Be aware of the uncertainties introduced in the three stages (conception, measurement and representation, and analysis) of database creation and use.
- Understand the concepts associated with data quality.
- Understand how and why scale of geographic measurement and analysis can both create and propagate uncertainty .

Data Quality

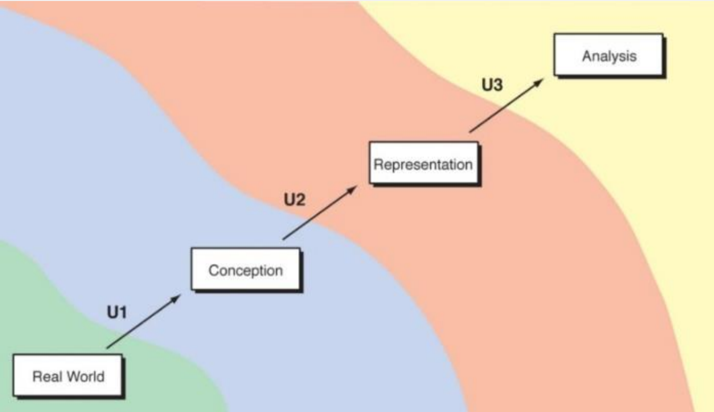
Effective spatial analysis requires an intelligent user, not just a powerful computer!

- GiGo: garbage in, garbage out.
- GIS can be thought of as a garbage magnifier.
- Typical users take digital data for granted, assuming their quality is high and fits the intended usage. However, most failed GIS projects are due to poor data quality.
- An increasing number of accidents result from the poor data quality.

Garbage in garbage out !

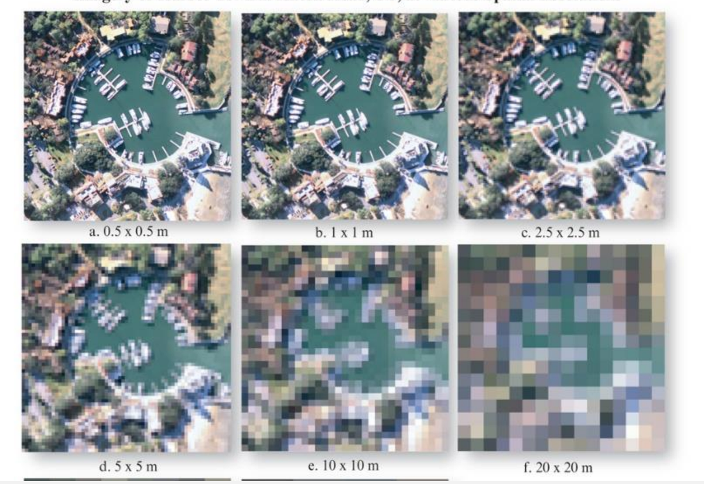
Data Quality

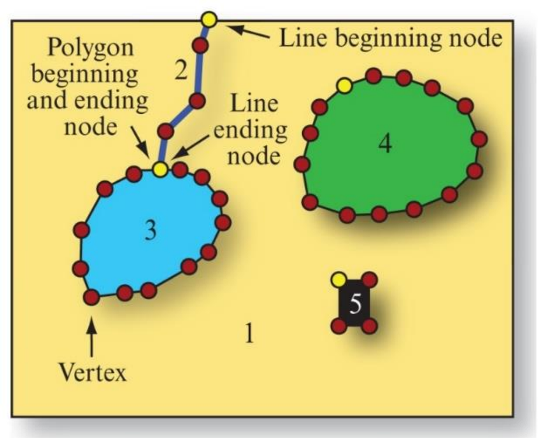
- **Spatial data** is just an abstraction of what is really there. Because of this abstraction, we can expect error due to
  - How we conceptualize the data in the first place
  - How we collect and represent the data
  - How we analyze the data



- It is **impossible** to make a **perfect representation** of the world, so **uncertainty** about it is inevitable.

Imagery of Harbor Town in Hilton Head, SC, at Various Spatial Resolutions





b. Vector data model.

© 2013 Pearson Education, Inc.

Sources of Error

- **Whenever you work with spatial data, you will deal with some sort of error due to the many steps involved in creating spatial data**
  - data collection
  - data input
  - data storage
  - data manipulation
  - data output
  - use of results

GIS definition: A set of tools for collecting, storing, retrieving, transforming, and displaying spatial data from the real world.

Sources of Error

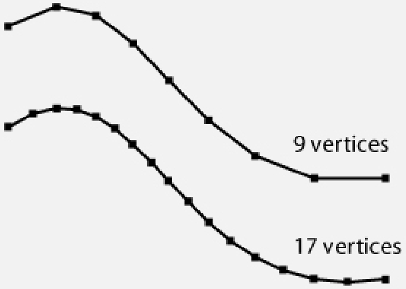
- **Data Collection**
  - Errors in field data collection
    - due to mistakes made by people operating equipment
    - due to technical problems with equipment.
  - Errors in existing maps used for digital data creation
    - Hard-copy maps may be folded, stretched, or shrunk

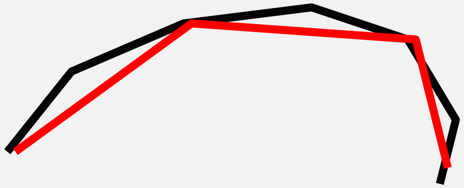
Sources of Error

- **Data Input**
  - Inaccuracies in digitizing (operator and equipment)
    - When digitizing a curve, a user can place many vertices to approximate the curve, or only a few vertices.

Digitizing: The process of converting the geographic features on an analog map into digital format. Features on a paper map are traced with a device such as a mouse, and the x,y coordinates of these features are automatically recorded and stored as spatial data.

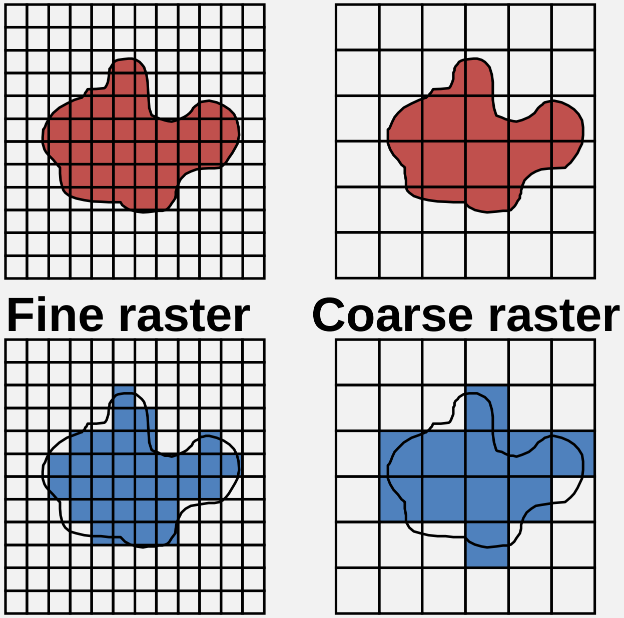
  - Surveyors may make mistakes or data may be entered into the database incorrectly.





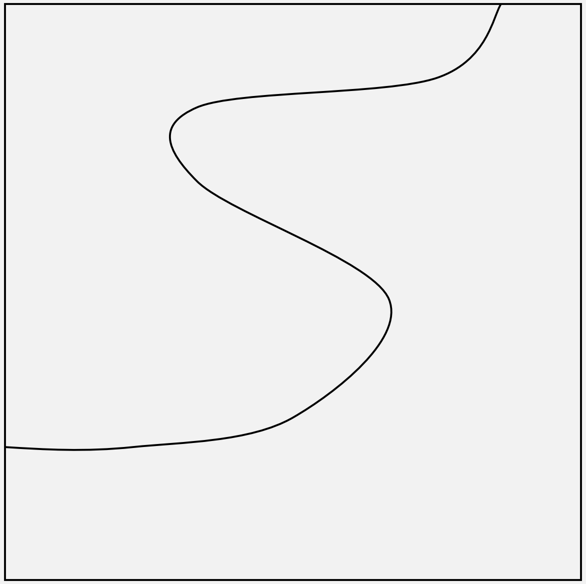
Sources of Error

- **Data Storage**
  - Numerical Precision (e.g., excel).
- **Data Processing**
  - Computational errors: numerical precision.
    - E.g., to what decimal point the data is represented?
  - Raster and vector conversion.
  - Mistakes in classification and would create attribute errors.

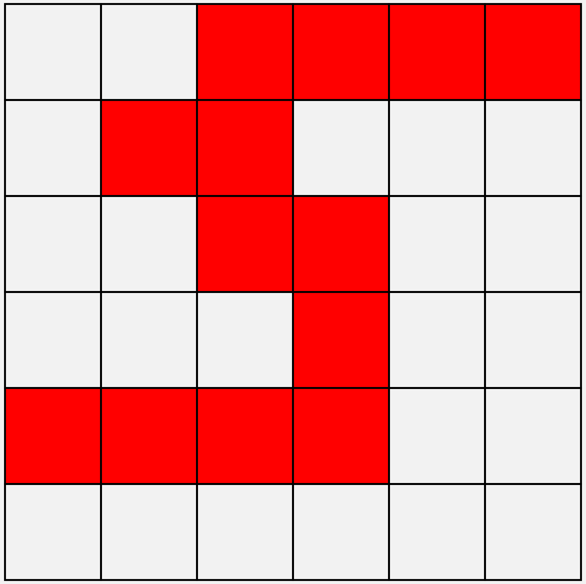


Sources of Error

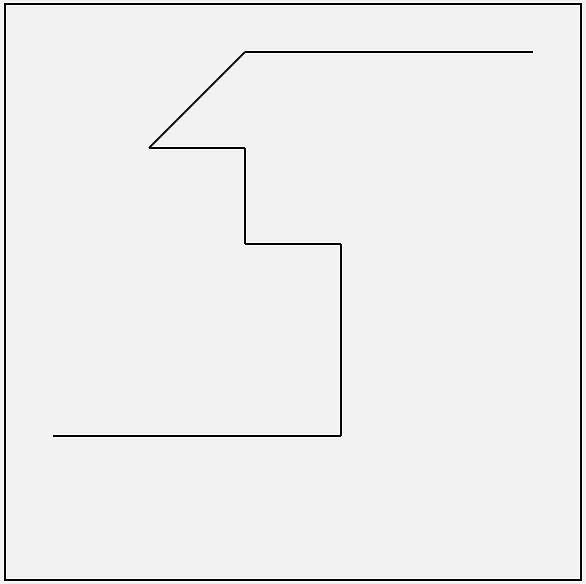
Original



Vector to Raster



Raster to Vector



Sources of Error

- **Data Output**
  - Scaling inaccuracies (printer dpi)
- **Use of Results**
  - Misinterpretation of data

Data Quality

- **Data Quality**
  - A measure of how well the GIS data represents the target domain.
  - data quality refers to the relative accuracy and precision of a particular GIS database.

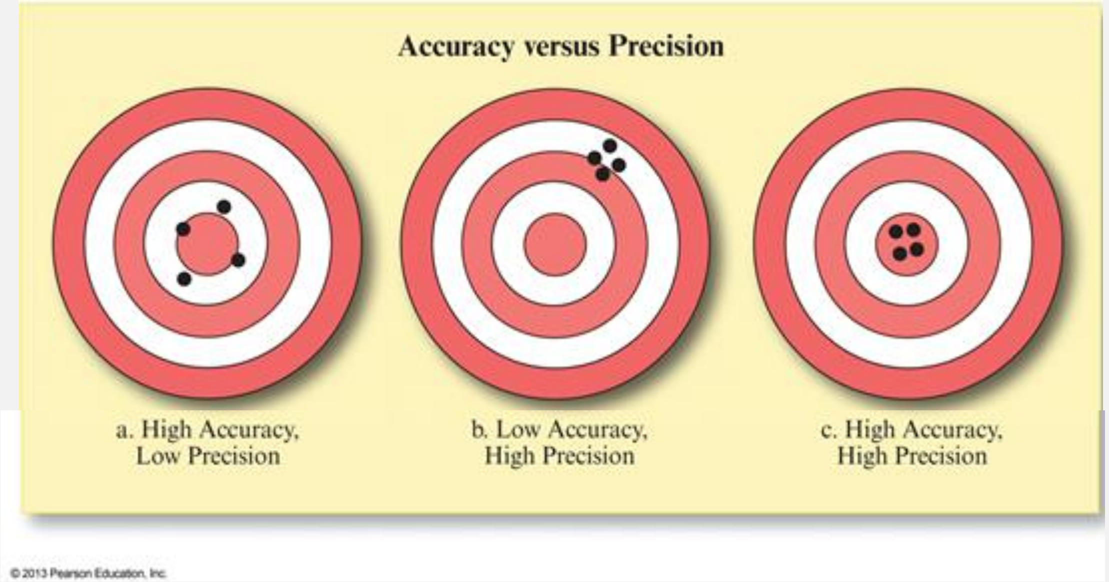
Data Quality

- Accuracy
  - Refers to the extent that both attribute and positional data correspond to their real-world counterparts.
- Precision
  - Refers to the “exactness” or the ability to “repeat” the measurements
  - Sometimes people think of precision as the number of decimal places that a device is capable of measuring
    - E.g., One thermometer measures temperature every other degree (e.g., 94, 96, and 98F) while the other one measures every half degree (94.5, 95, 95.5F) . Which thermometer is more precise?
- Highly precise and accurate data can be very difficult and costly to collect. Carefully surveyed locations needed by utility companies to record the locations of pumps, wires, and pipes cost $5-20 per point to collect.

For mapping, accuracy is associated with position of an object to its true position. Precision is then the ability to repeat a measurement, or how likely you are to return to the same location time and time again.

Accuracy vs. Precision

- Good accuracy does not necessarily guarantee good precision, and good precision does not necessarily guarantee good accuracy.



Accuracy versus Precision

a. High Accuracy, Low Precision

b. Low Accuracy, High Precision

c. High Accuracy, High Precision

© 2013 Pearson Education, Inc.

- There is a systematic **error of 3 degree** associated with the thermometer that measures temperature to **every half degree**.
- There is **no systematic error** in the other thermometer that measures **every other degree**.

Which thermometer is more accurate?

In this case , the less precise thermometer is actually more accurate than the more precise thermometer.

Types of Error in Geospatial Data

- Attribute error
- Positional error (x, y, z)
- Topological (geometric) error
- Temporal error
- Interpretation error due to ecological fallacy
- Error due to the modifiable areal unit problem

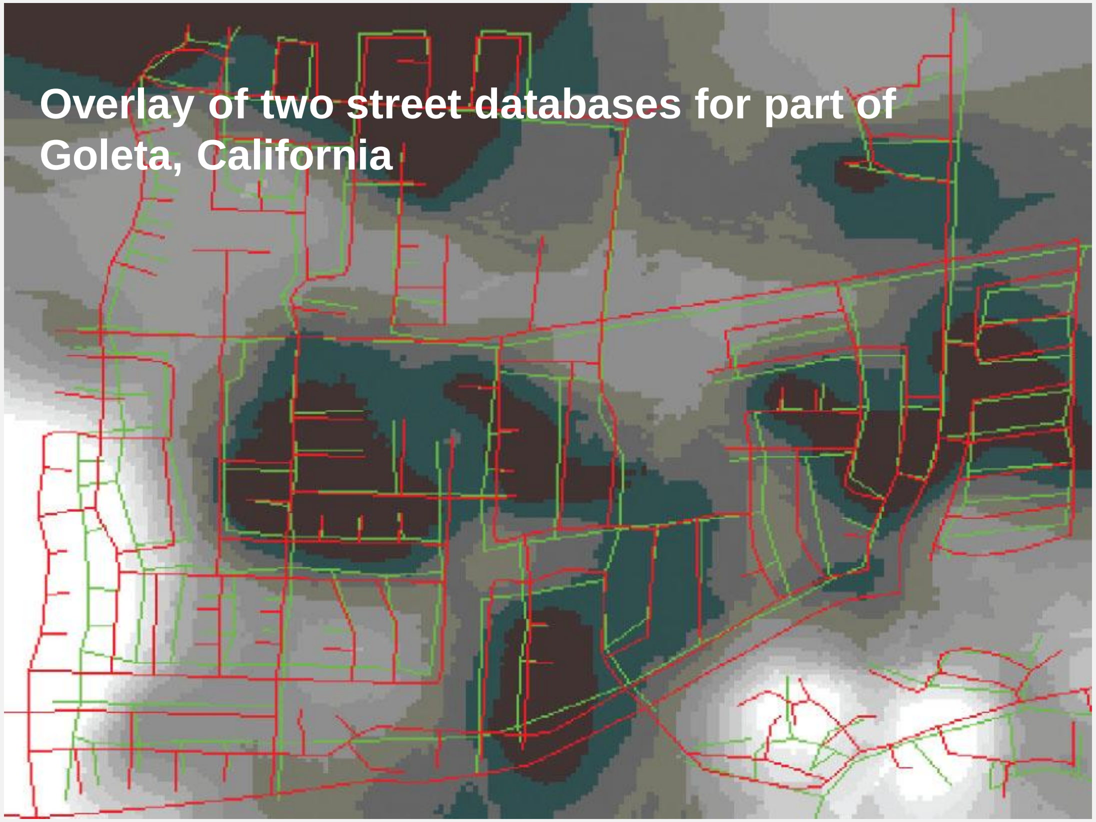
Attribute Error

- **Attribute accuracy and precision refer to quality of non-spatial, attribute data**
  - Wrong
  - Missing
- Determining attribute accuracy can be difficult when the database contains tremendous number of records.
- Consequently, it is sometimes useful to conduct a spatial sampling to determine attribute accuracy.

Positional Error

- **Positional accuracy measures how close the geographic coordinates of features in a spatial data layer are to their real-world geographical coordinates (both horizontal and vertical)**

Positional accuracy can be determined by examining the x, y, z position of features in the spatial database and comparing those positions with actual real-world location measurements based on the use of a more accurate measuring device (e.g., a survey-grade GPS unit)



Overlay of two street databases for part of Goleta, California

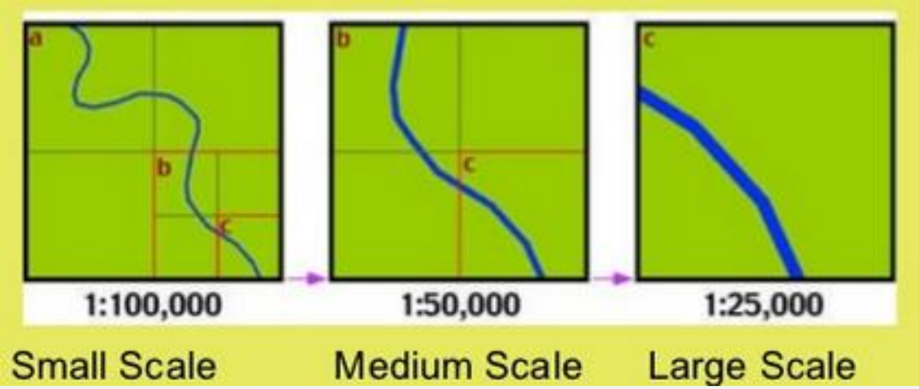
Positional Error

- **Positional accuracy** is **dependent** on the **scale** at which the data were acquired.
- Generally, larger-scale spatial database (e.g., scales>1:20,000) have better positional accuracy than smaller-scale spatial database (e.g., 1:50,000 or 1:100,000)

Scale:

– ratio of distance on a map to the equivalent distance on the earth's surface.

  - **Large scale** -->large detail, small area covered (1:2,400)
  - **Small scale** -->small detail, large area (1:250,000)



Large-scale
Small-scale

Source: 1) Image arcade 2) GIS Commons

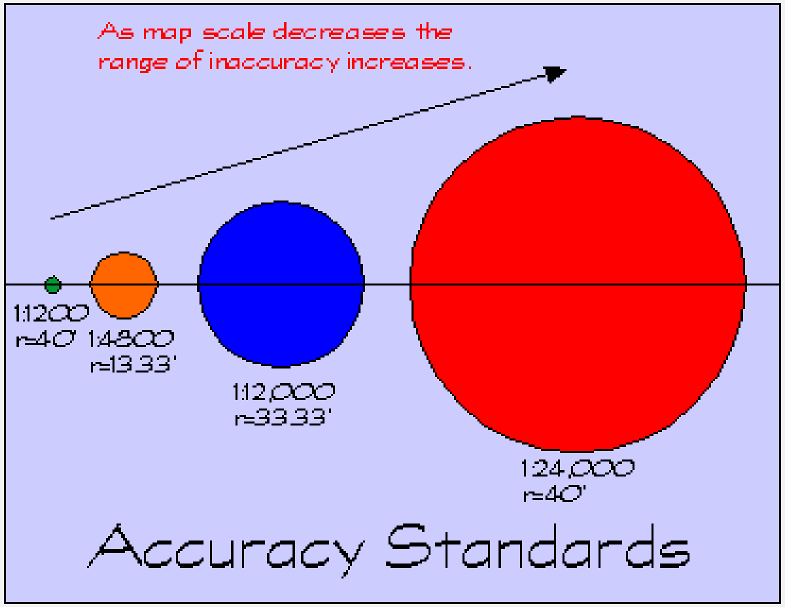
Positional Error

- **Positional Accuracy**

  standards specify that acceptable positional error varies with scale.

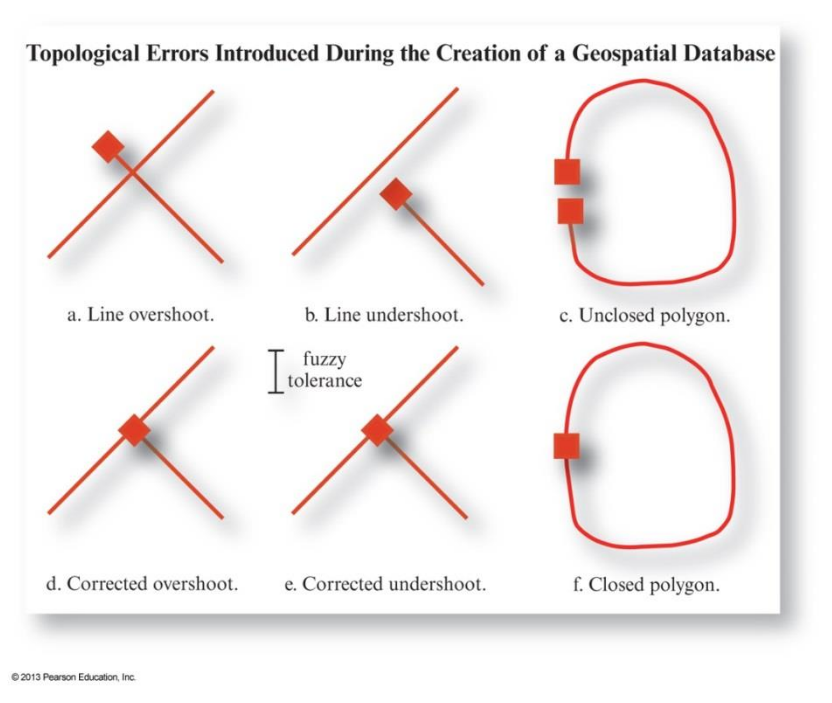
- **Map Accuracy Standards**
  - United States National Map Accuracy Standards (NMAS)
  - The American Society for Photogrammetry and Remote Sensing (ASPRS) Map Accuracy Standard for Large-Scale Maps
  - The Federal Geographic Data Committee (FGDC) National Standard for Spatial Data Accuracy

As the map scale decreases, the range of inaccuracy increases.



Topological Error

- **These type of errors are often introduced when creating new point, line, and polygon data using GIS editing tools.**


Topological Errors Introduced During the Creation of a Geospatial Database

a. Line overshoot.

b. Line undershoot.

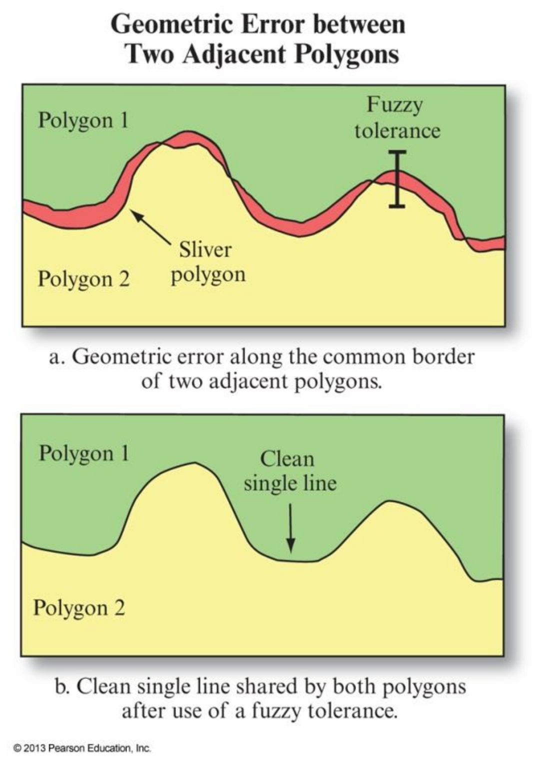
c. Unclosed polygon.

d. Corrected overshoot.

e. Corrected undershoot.

f. Closed polygon.

© 2013 Pearson Education, Inc.


Geometric Error between Two Adjacent Polygons

a. Geometric error along the common border of two adjacent polygons.

b. Clean single line shared by both polygons after use of a fuzzy tolerance.

© 2013 Pearson Education, Inc.

- **To correct thee errors, GIS users can**
  - Apply an appropriate fuzzy tolerance value
    - Fuzzy tolerance is the distance within which individual discrete points are snapped to form a single point during editing.
  - Delete the point, line, and area data and recreate the features

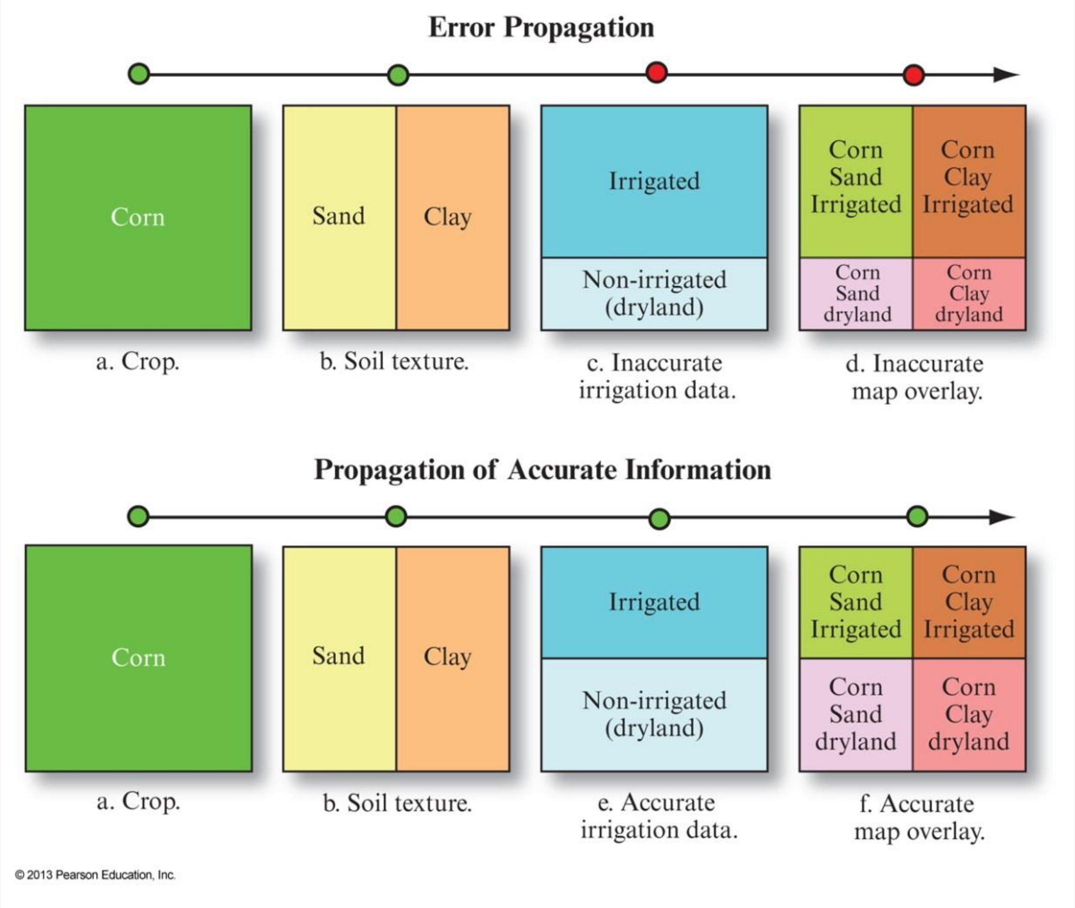
Temporal Accuracy

- **Temporal Accuracy refers to how up-to-date a geospatial database is.**
  - Some geospatial data need to be updated every half-hour or at even shorter time increments. **EXAMPLE ?**
    - E.g., traffic data
  - Some geospatial data are updated every several years or several decades. **EXAMPLE ?** (e.g, demographic data and land use).
  - Some data do not change significantly over the time data usage. **EXAMPLE ?**

    (e.g. elevation data).

Error Propagation

- If attribute or positional errors are not known, corrected, or accounted for at the beginning of a GIS project, they will propagate throughout the study and accumulate in interim or final products.

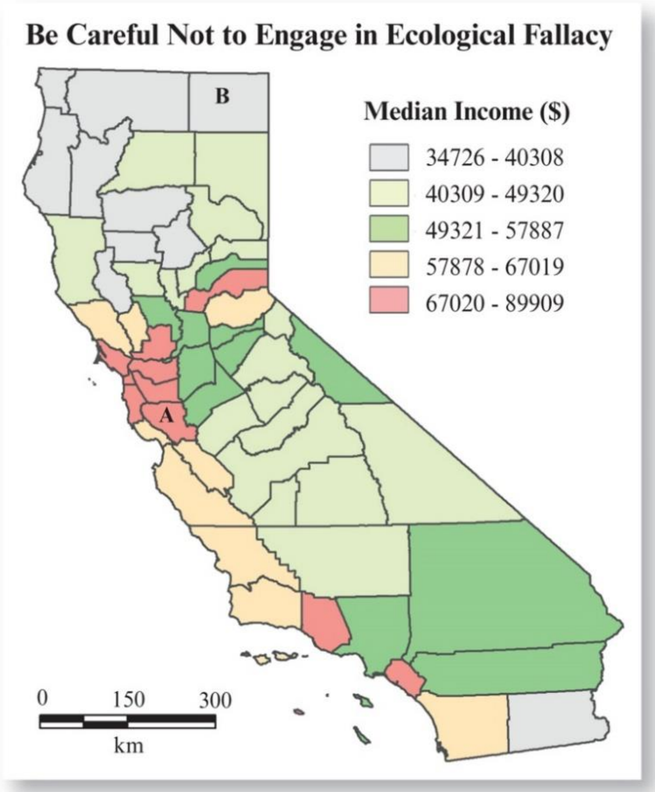


© 2013 Pearson Education, Inc.

Ecological Fallacy

- **Ecological fallacy refers to inappropriate inference from aggregate data about the characteristics of individuals.**
  - Ecological fallacy is the belief that all observations within an area will exhibit the same or similar values.
  - In other words, it is the belief that the characteristics or relationships of a group are the same for each individual within the group.

- **The likelihood of committing ecological fallacy in GIS depends on the nature of the aggregation being used**
  - **The members within a zone are all perfectly uniform and homogenous i.e. variation between zones**
  - **Nearly all zones are internally heterogeneity to some degree i.e. increases likelihood and severity of the ecological fallacy problem**

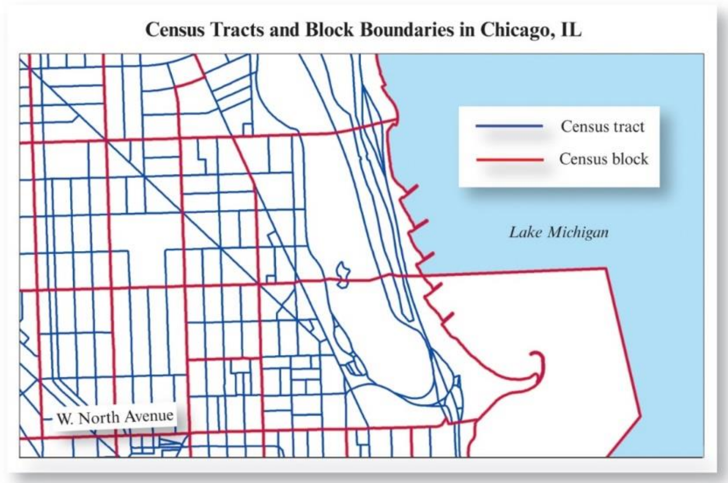

Be Careful Not to Engage in Ecological Fallacy

Modifiable Areal Unit Problem

- **Spatial data may be reported, mapped, and analyzed using districts of various sizes.**
  - Including countries, states, counties, cities, census block groups, census block, etc.
  - National census: collected at the household level but reported for practical and privacy reasons at various levels of aggregation (block, block group, tract, county, state, etc.)
- **When smaller units are combined into fewer, larger units, the variation present in the smaller units may decrease.**
- **The aggregation units used will affect statistics determined on the basis of data reported in this way. Generally, the correlation between variables increases as the size of the areal unit increases**

- Consequently, it is possible for researchers to “modify” the areal unites that are being mapped, introducing bias into the project. This is typically referred to as the modifiable areal unit problem.

- E.g. A researcher may find a better or higher correlation between variables when analyzed at census block level rather than census tract level.



Metadata

- **Metadata is data documentation, or “data about data”**
- **To avoid many of these errors, good documentation of source data is needed.**
  - Spatial data structure (e.g., raster or vector)
  - Projection (e.g., datum, coordinate system, and projection)
  - Scale of the original data
  - When the data were created
  - How the data were created
  - Database field names and properties (e.g., data types and formats)
  - Data quality
  - Accuracy and precision of the instruments used to collect the data

Documentation and Metadata

- The federal geographic data committee (FGDC) is a federal entity that developed a "Content Standard for Digital Geospatial Metadata" in 1998, which is a model for all spatial data users to follow.
- https://www.fgdc.gov/metadata/geospatial-metadata-standards
- Purpose is: "to provide a common set of terminology and definitions for the documentation of digital geospatial data."
- All federal agencies are required to use these standards.

Consolidation

- **Uncertainty is inevitable in GI**
- **Data obtained from others should never be taken as truth**
  - efforts should be made to determine quality
- **Effects on GI system outputs are often much greater than expected**
  - there is an automatic tendency to regard outputs from a computer as the truth
- **Use as many sources of data as possible**
  - and cross-check them for accuracy
- **Be honest and informative in reporting results**
  - add plenty of caveats and cautions

Conclusions

- **Uncertainty is much more than error**
- **Sources of error**
- **Uncertainties in three stages**
  - Conception
  - measurement and representation
  - analysis
- **Scale**

Questions ?

https://www.google.com/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwihuqvgkpjcAhU3IDQIHZbqB8QQjRx6BAgBEAU&url=http%3A%2F%2Fwww.cityofrockhill.com%2Fdepartments%2Finformation-technology-services%2Fmore%2Finformation-technology-services%2Fgeographic-information-systems-gis-%2Fgis-frequently-asked-questions&psig=AOvVaw2fEUXAIjbJY26W-bnjS0wY&ust=1531436220322311

Upcoming

- Monday (Lecture) : Data Collection
- **Lab 02 due** (Check Syllabus)
- Readings updated on canvas.
- DRS Accommodations for Exam
- Exam 1 In Week 6 (Check Syllabus)