Issues of Scale and Accuracy

Presented by:

Tim Haithcoat
University of Missouri Columbia

MU Department of Geography
Geographic Resources Center

12

Introduction

- The world is infinitely complex
- The contents of a spatial database represent a particular view of the world
- The user sees the real world through the medium of the database
  - The measurements & samples contained in the database must present as complete and accurate a view of the world as possible
  - The contents of the database must be relevant in terms of:
    - Themes and characteristics
    - The time period covered
    - The study area

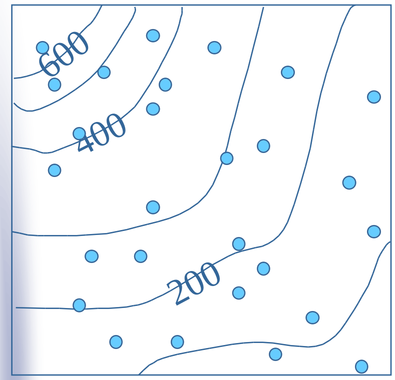
Representing Reality

- A database consists of digital representations of discrete objects
- The features shown on a map (lakes, benchmarks, contours) can be thought of as discrete objects
  - Thus the contents of a map can be captured in a database by turning map features into database objects

Representing Reality

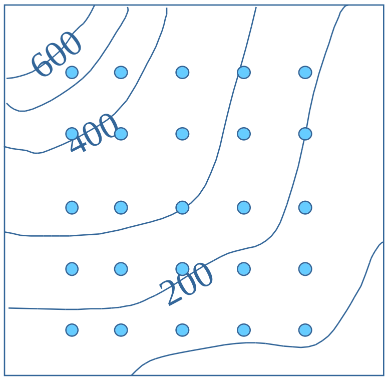
- Many of the features shown on a map are fictitious and do NOT exist in the real world
  - Contours do not really exist, but houses & lakes are real objects
- The contents of a spatial database include:
  - Digital versions of real objects (e.g. houses)
  - Digital versions of artificial map features (e.g. contours)
  - Artificial objects created for the purposes of the database (e.g. pixels)

Sampling Strategies



Random

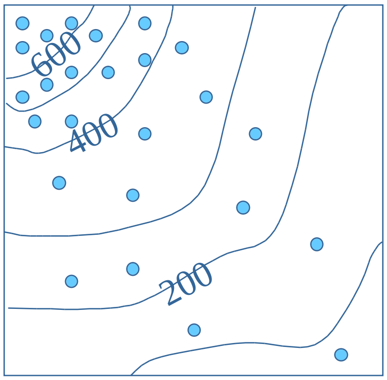
Every place or time is equally likely to be chosen.



Systematic

Samples are chosen according to a rule.

Example: every 1 km, but the rule is expected to create no bias in the results of analysis



Stratified

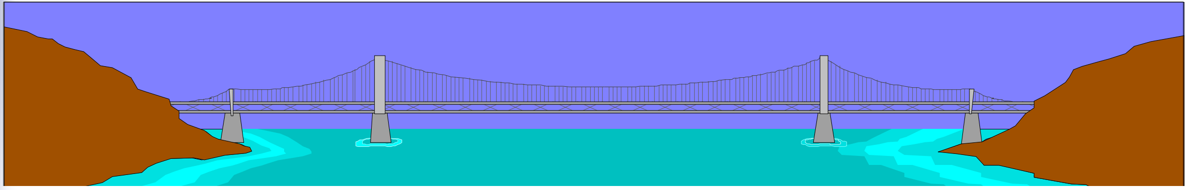
Research knows for some reason that the universe contains significantly different sub-populations, & samples within each sub-population in order to achieve adequate representation of each.

Errors and Accuracy

- There's nearly a universal tendency to lose sight of errors once the data are in digital form
- Errors:
  - Are implanted in original databases due to errors in the original sources
  - Are added during data capture and storage (processing errors)
  - Occur when data are extracted
  - Arise when the various layers of data are combined in an analytical exercise

Original Sin: Source Errors

- **Are extremely common in non-mapped source data, such as locations of wells, or lot descriptions**
- **Can be caused by misinterpreting aerial photography and images**
- **Often occur because base maps are relied on too heavily**

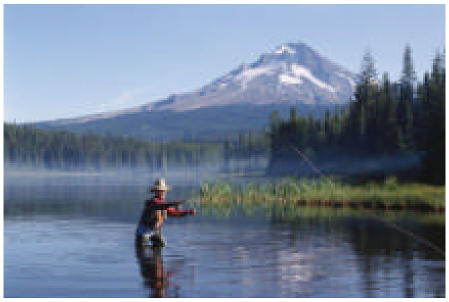


•An attempt to overlay bridge locations on transportation data resulted in bridges lying neither beneath roads, nor over water, and roads lying apparently under rivers. Until they were compared in this way, it was assumed that each data set was locationally acceptable. The ability of GIS to overlay may expose previously unsuspected errors.

Boundaries

- Boundaries of soil types are actually transition zones, but are mapped by lines less than 0.5 mm wide

- Lakes fluctuate widely in area, yet have permanently recorded shorelines

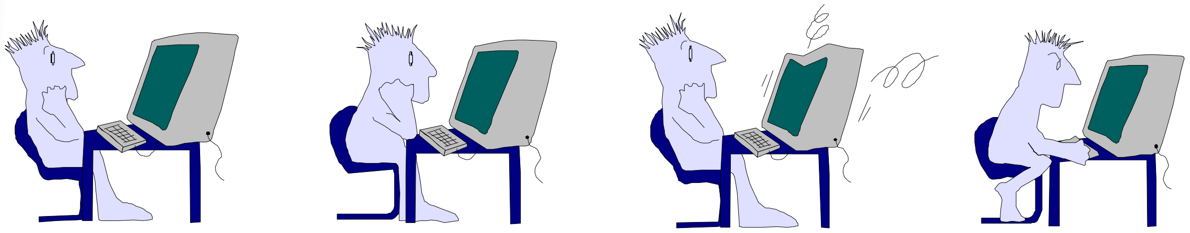


Classification Errors

- Common when tabular data are rendered in map form
- Simple typing errors may be invisible until presented graphically
  - Floodplain soils may appear on hilltops
  - Pastureland may appear to be misinterpreted marsh
- More complex classification errors may be due to sampling strategies that produced the original data
  - Information may exist that documents the error of the sampling technique
  - However, such information is seldom included in the GIS database

Data Capture Errors

- Manual data input induces another set of errors
- Eye-hand coordination varies from operator to operator and from time to time
  - Data input is a tedious task, it is difficult to maintain quality over long periods of time

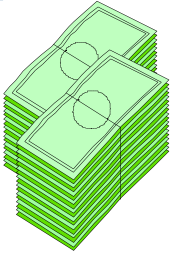


Accuracy Standards

- Many agencies have accuracy standards for geographical data
  - These are more often concerned with accuracy of locations of objects than with accuracy of attributes
- Location accuracy standards are commonly decided from the scale of source materials
  - For natural resource data 1:24,000 scale accuracy is a common target
  - At this scale, 0.5 mm line width = 12 m on the ground

Accuracy Standards (continued)

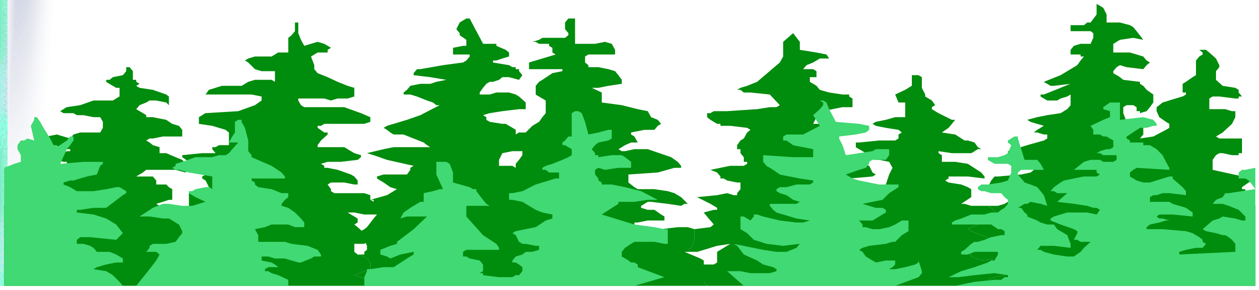
- USGS topographic information is currently available in digital form at 1:100,000
  - 0.5 mm line width = 50 m on the ground
- Higher accuracy requires better source materials
  - Is the added cost justified by the objectives of the study?



- Accuracy standards should be determined by considering both the value of information and cost of collection

Data Quality Properties

- Positional Accuracy
- Attribute Accuracy
- Topological Correctness
- Completeness



Quality Control

Cover Two Topics:

- **General Principles of Quality Assurance**
- **Quality Management**

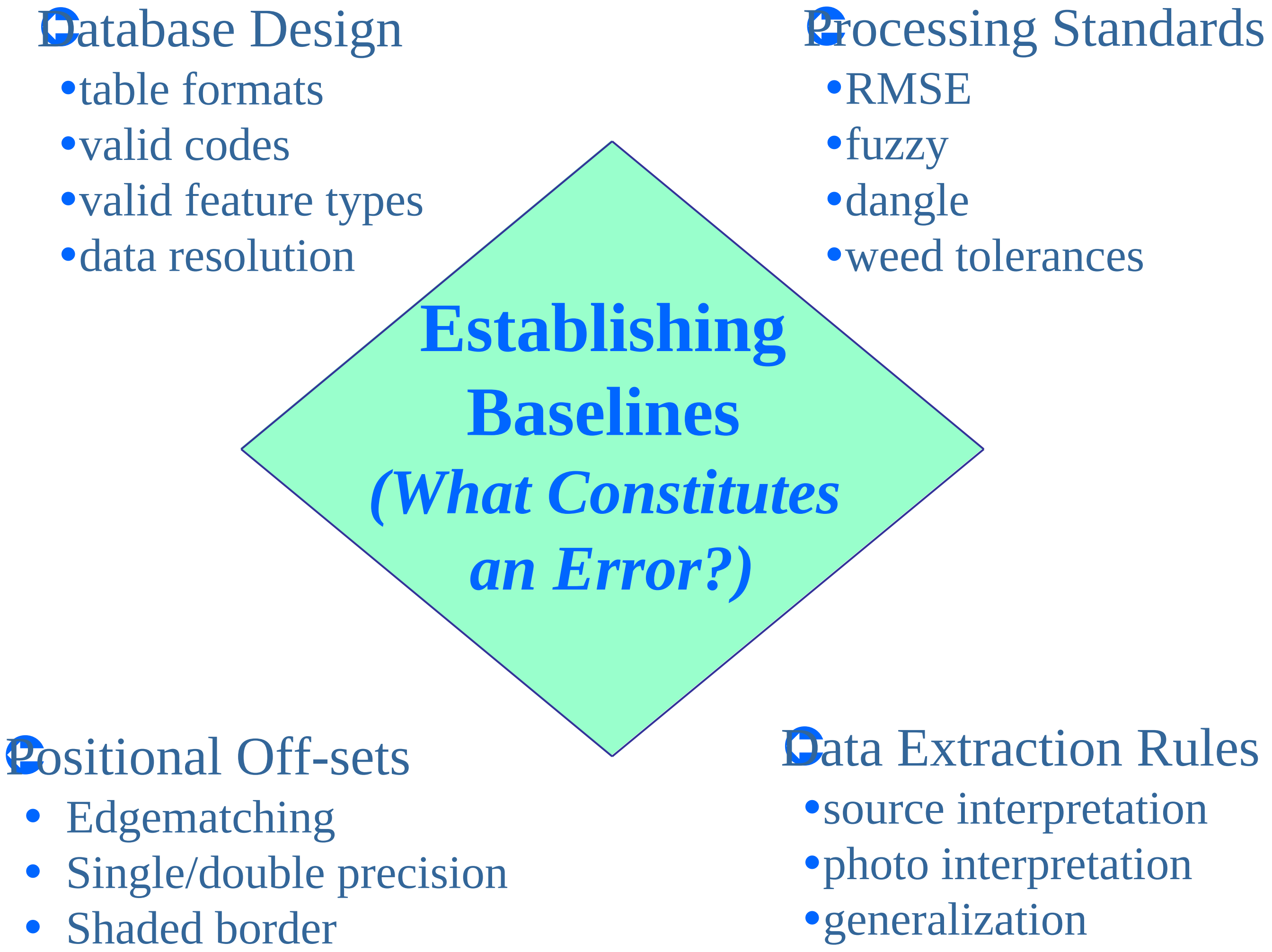
Definitions

Quality Assurance (QA)

Pertaining to a comprehensive approach or system for ensuring product quality.

Quality Control (QC)

Pertaining to an action or step designed to test product quality for any number of properties.



Database Design

- table formats
- valid codes
- valid feature types
- data resolution

Processing Standards

- RMSE
- fuzzy
- dangle
- weed tolerances

Establishing Baselines

(What Constitutes an Error?)

Positional Off-sets

- Edgematching
- Single/double precision
- Shaded border

Data Extraction Rules

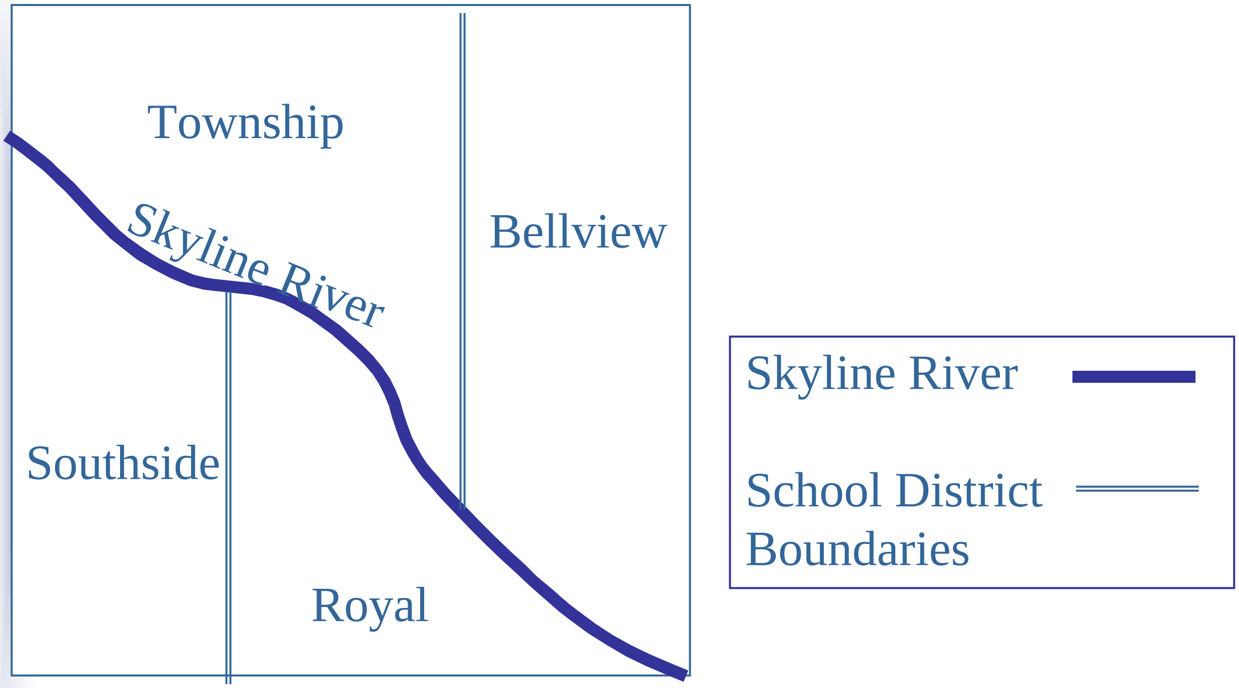
- source interpretation
- photo interpretation
- generalization

Analyzing QC Requirements

- Develop a process-oriented flow
  - (i.e., edge matching standards, digitizing standards)
- Identify key points in the conversion process where data properties could or are change(d)
  - (i.e., digitization, transformation, re-projection)
- Adjust process flow to concentrate QC property checking into as few steps as possible

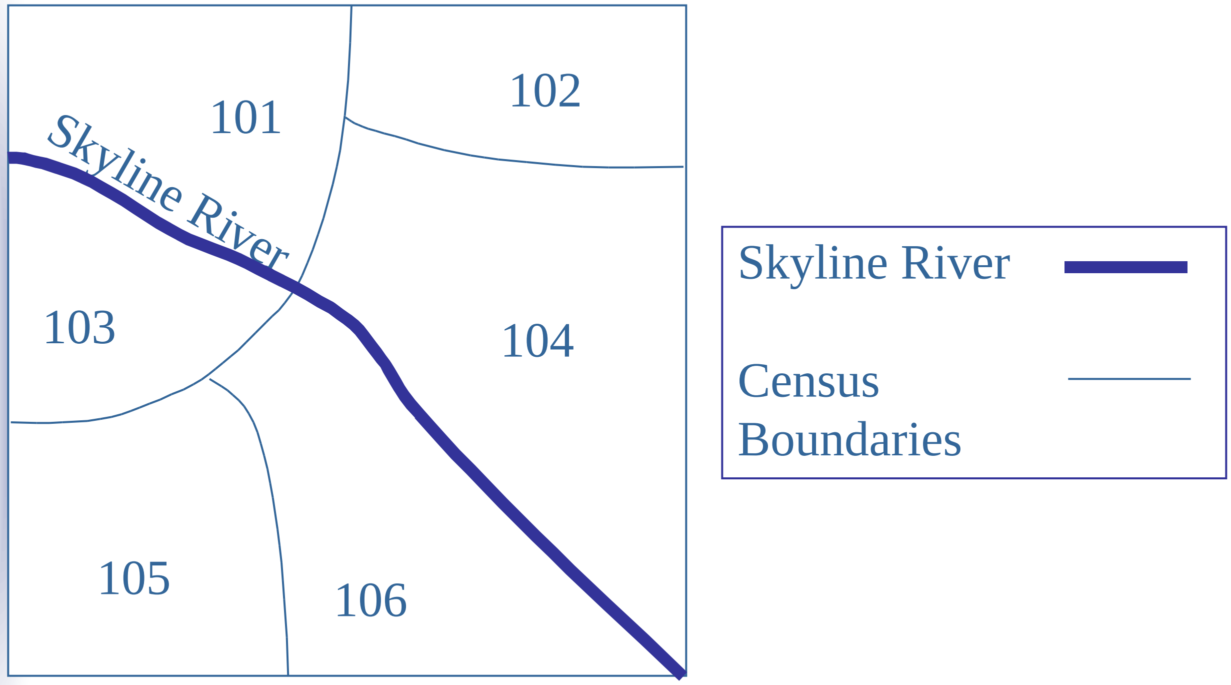
Sliver or Spurious Polygons

(page 1 of 3)



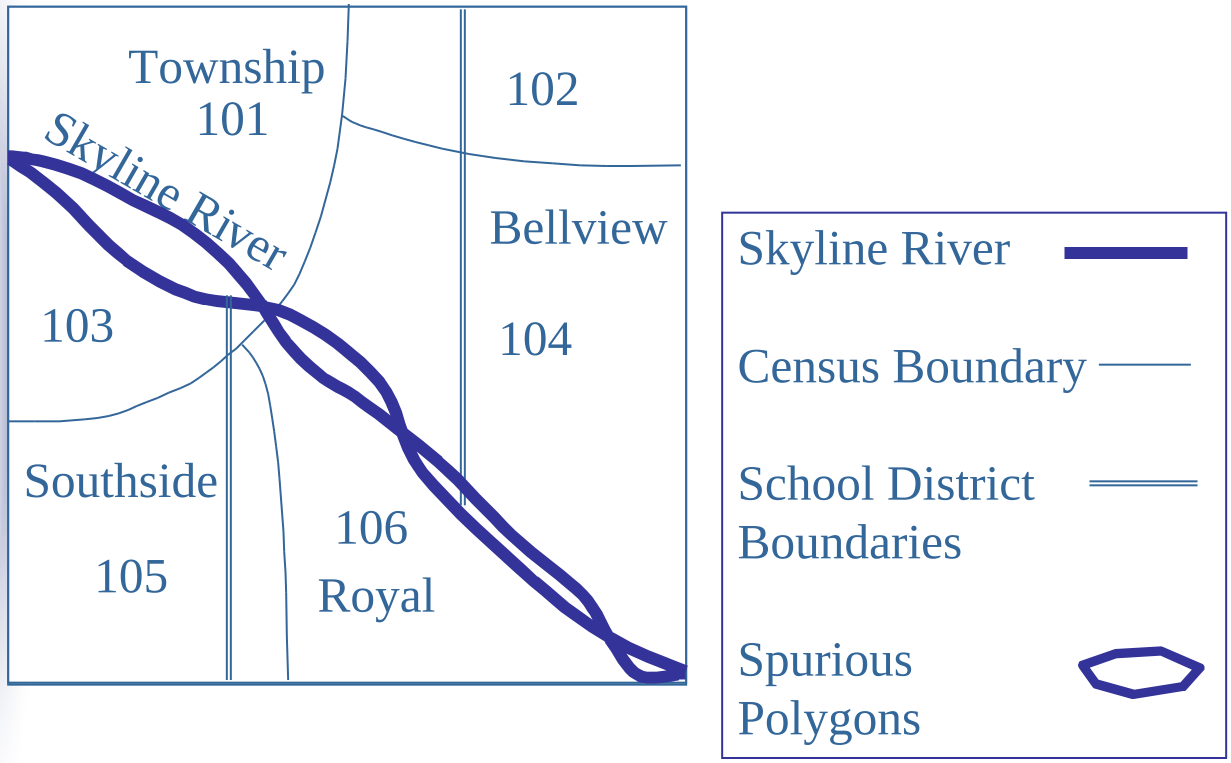
Sliver or Spurious Polygons

(page 2 of 3)

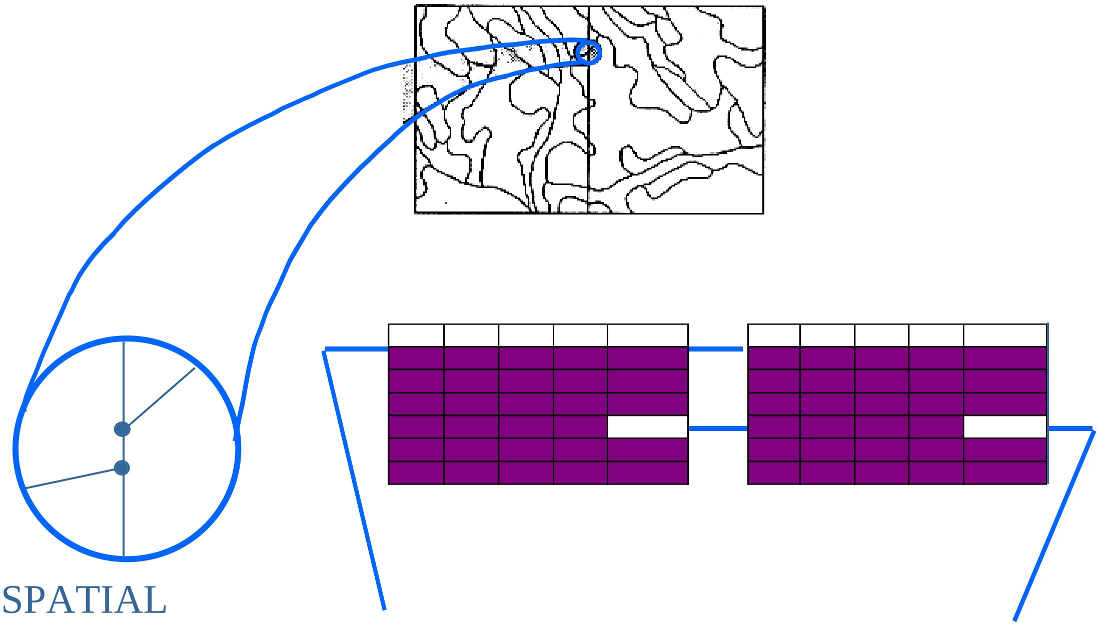


Sliver or Spurious Polygons

(page 3 of 3)



Criteria Used When Matching Data

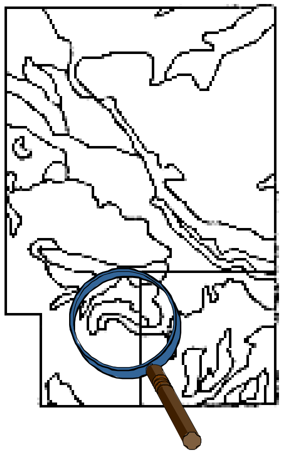


SPATIAL
COORDINATES

ITEM DEFINITIONS

ATTRIBUTE
VALUES

Visually Compare Spatial Data

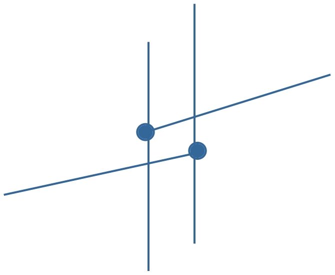


Small-scale display

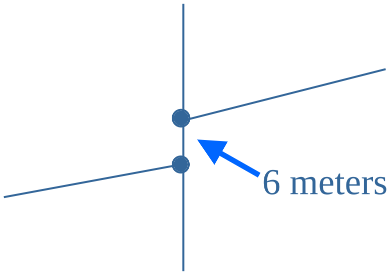
(map extent set to both coverages)

Large-scale display

(zoomed-in map extent)



Do the boundaries Overlap?



How far apart are
Corresponding nodes?

Managing the Database

To be functional, a project database should contain coverages that have the following characteristics:

- ☑ **Each coverage contains clean topology**
- ☑ **The accuracy of all feature locations has been verified**
- ☑ **Attribute tables are present**
- ☑ **The accuracy of the attribute values has been verified**
- ☑ **A system of ground control points exist**
- ☑ **All geographic features are recorded using real-world coordinates**
- ☑ **All related coverages are in one common coordinate system and datum**
- ☑ **The features of each coverage are spatially referenced to features in associated coverages**

How to Test Positional Accuracy

- Use an independent source of higher accuracy
- Use internal evidence
- Compute accuracy from knowledge of the errors introduced by different sources

Sensitivity also refers to…

Spatial Resolution

- Would increasing resolution give a better result?
- Would cost of additional data collection at higher resolution be justified?
- Can we put a value on spatial resolution required?

This information can be used in assessing the level of input accuracy that is needed.

For example, if the additional accuracy will not change the results, it may be unnecessary to carry out costly detailed surveys.

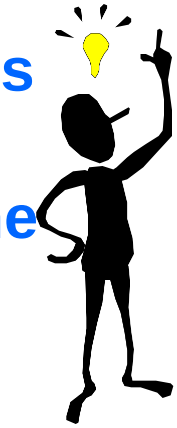
Can also use sensitivity analysis to assess the effects of uncertainty in the data - “confidence interval” measure for the results.

Consider...

- Use full observed range to test sensitivity
  - Response of the result to a change in one of the inputs from its minimum observed value to its maximum
- Layers which are important, but nevertheless do not show geographic variation over the study area will not have high sensitivity in this definition

Consider...

- Brings out the distinction between sensitivity *in principle* and *in practice*
  - A layer may be important in principle, but have no impact within the study area
- Examine both the decision rules & the value ranges to help determine which layers have the highest impact on the result



Sensitivity can be defined for:

Data Inputs

How much does the result change when the data input changes?

Data Weights

How much does the result change when the weight given to a factor changes?

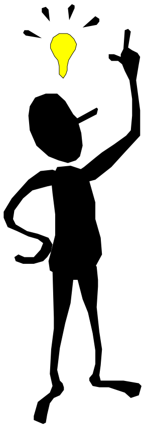
Errors in determining weights may be just as important as error in the database.

Sensitivity Analysis

- It is the response of the result (suitability) to a unit change in one of the inputs.
- Easy to see what a unit change means for temperature or precipitation data, but what does it mean for a vegetation class?

In some types of operations…

the accuracy of suitability is determined by the accuracy of the least accurate layer.



In other cases…

the accuracy of the result is significantly better than the accuracy of the least accurate layer.

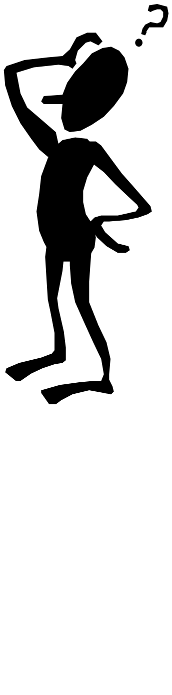
How then do we determine the impact of inaccuracy on the result?

Effects of Cascading on an Error will be Complex

✓Do errors get worse?

✓Do errors cancel out?

✓Are errors independent or related?



Suppose two maps, each with percent correctly classified of .90 are overlaid…

- **Studies have shown that the accuracy of the resulting map is little better than .9 x .9 = .81**
- **When many maps are overlaid the accuracy of the resulting composite can be very poor.**
- **However, we are more interested in the accuracy of the composite suitability index than in the overlaid attributes themselves.**

Inaccuracy arises primarily from...

- Randomness
  - May occur when an observation can assume a range of values
- Vagueness
  - May result from imprecision in taxonomic definitions
- Incompleteness of Evidence
  - May occur when sampling has been applied, there are missing values, or surrogate variables have been employed.

- **The interdependence between location and value in spatial databases gives rise to spatial dependence and heterogeneity.**
- **In GAP analysis, we combine data from different sources with different levels of accuracy.**
- **What impact does error in each data layer have on the final result?**
- **Reliability is a function of both cartographic & ecological factors.**
- **Inaccuracy is often inadvertent but may also be intentional since generalization methods are frequently applied to enhance cartographic ease.**

Useful Resolution Groups for Engineering & Planning

From: “Selection of Maps for Engineering & Planning”, Committee on Cartographic Surveying, Journal of the Surveying and Mapping Division, Proceedings of the American Society of Civil Engineers, July, 1972. Table 1, p. 112

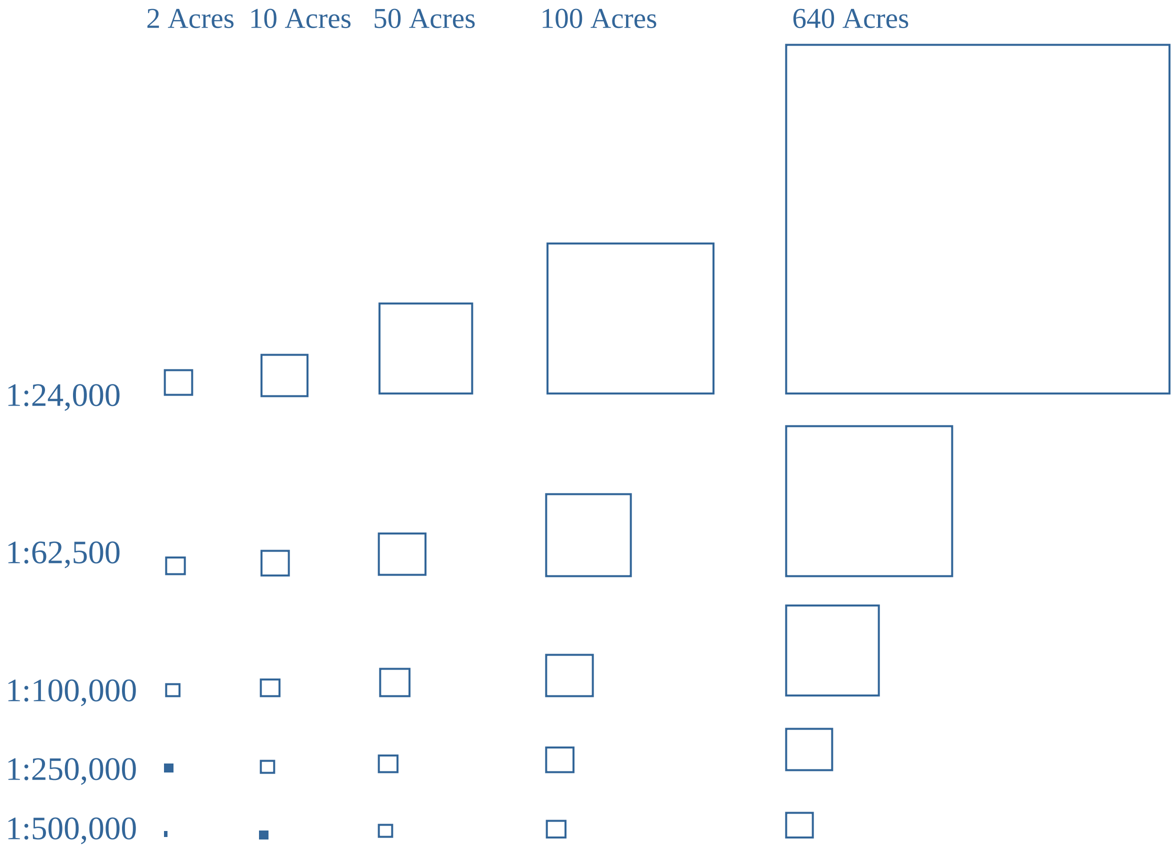
Type of Map	**Scale**			
	Feet per Inch	**Representative Fraction**	**Feet**	**Meters**
Design				
Critical	10 to 50	1:100 to 1:500	.2 to 5	.1 to 1
General	40 to 200	1:500 to 1:2,000	.05 to 10	.1 to 2
Planning				
Micro	100 to 1,000	1:1,000 to 1: 10,000	1 to 20	.2 to 5
Local	400 to 2,000	1:5,000 to 1: 25,000	2 to 50	.5 to 10
Regional	1,000 to 10,000	1:10,000 to 1:100,000	5 to 100	1 to 20
National	10,000 to 100,000 (2 miles) (20 miles)	1:100,000 to 1:1,000,000	10 to 1,000	2 to 200

Scale - Data Resolution

	Polygon	Lines	
	Acres	Mile	Feet
1:24,000	2-3	.05	250
1:62,500	5-10	.12	650
1:100,000	25-50	.2	1050
1:250,000	250-500	.5	2600
1:500,000	500-1000	1.0	5280

Data below these resolutions are generally merged into surrounding data, converted to a point or deleted.

Data Resolution



2 Acres 10 Acres 50 Acres 100 Acres 640 Acres

1:24,000

1:62,500

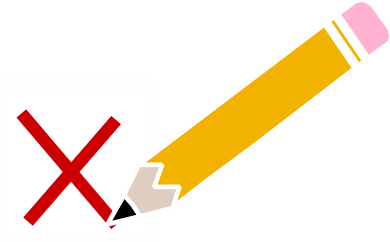
1:100,000

1:250,000

1:500,000

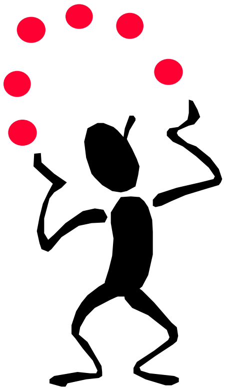
Scale ~ Map Resolution

- **Definition:** The accuracy with which the location & shape of map features can be depicted for a given scale.
- Decreasing map scale results in lower map resolution as selected features are:
  - Smoothed
  - Simplified
  - Aggregated
  - Eliminated
  - Reduced in Dimension
    - Area (2) to Line (1)
    - Area (2) to Point (0)
- Understand/document all GIS data source resolutions
- Categorize sources by resolution groups
- Make careful choices regarding upward & lower bounds of resolution groups.
- Begin to define scales of the GIS database



Cartographic Considerations

- GIS Database is NOT an ordinary database
- Location is explicit in design
- Designer has to be aware of cartographic base that describes/specifies location
- Considerations include:
  - Scale
  - Coordinate Systems
  - Map Projections
  - Datums
  - Geodetic Control - GPS



Points to Remember...

✗ The precision of GIS processing is effectively infinite.

✗ All spatial data are of limited accuracy.

✗ The precision of GIS processing exceeds the accuracy of the data.

✗ In conventional map analysis, precision is usually adapted to accuracy.

✗ The ability to change scale and combine data from various sources and scales in a GIS means that precision is usually not adapted to accuracy.

✗ We have no adequate means to describe the accuracy of complex spatial objects.

✗ The objective should be a measure of uncertainty on every GIS product.

There is a nearly universal tendency to lose sight of errors once the data are in digital form.

Errors...

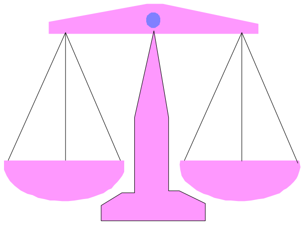
…are implanted in databases because of errors in the original source

…are added during data capture and storage

…occur when data are extracted from the computer

…arise when the various layers of data are combined in an analytical exercise.

Accuracy & Scale



Accuracy The closeness of results,
computations,
or estimates to true values.

Precision **Computer-based:** The number of decimal places or significant digits in a measurement

Application based: The regularity or consistency of a result, computation, or estimate.