ERROR SOURCE IDENTIFICATION AND MANAGEMENT

Linda Kirkland, OEI - *Presenter*

MODULE OBJECTIVES

- Relate error identification and management to QAPP systematic planning
- Identify requirements and authorities
- Identify components of error and their propagation related to QA planning
- Emphasize importance of product metadata

GIS AND QA: AUTHORITIES

- Executive Order 12906
- OMB Circular A-16
- EPA IRM Policy Manual 2100 Locational Data
- EPA Order 5360.1 CHG1
- EPA Quality Manual 5300

SYSTEMATIC PLANNING AREAS PERTAINING TO USE OF GIS

- Identification of data and purpose
- Determination of data quality and specification of performance criteria
- Determination of data source (lineage) and constraints on data collection
- Specification of QA/QC assessments needed
- Description of analysis, evaluation, and assessment techniques

COMPONENTS OF DATA QUALITY (FIPS 173)

- **Accuracy**
  - positional
  - attribute
- **Logical consistency**
- **Completeness**
- **Lineage**

Note: Consider hardware/software configuration controls (e.g. ARC/INFO).

ACCURACY

- Defined as the closeness of results to "true" values

Error (r) = estimated value minus true value.

All spatial data are inaccurate to some degree.

- **POSITIONAL ACCURACY**

  - Closeness of locational information to true position

- **ATTRIBUTE ACCURACY**

  - Closeness of attribute values to true value

  Continuous attributes → measurement error

  Categorical → misclassification

DETERMINING ACCURACY

- **Reporting requirements:**

  FGC NSSDA

  EPA LRS

- **Compare to source map or data of higher accuracy**

  Root mean square error

  95% confidence level

  20 check point minimum

- POSITIONAL ACCURACY

Determine maximum error and see if it meets the project needs.

Get statistics of accuracy from producer or similar product.

Identify steps to determine source, transfers between coordinate systems, formatting.

Estimate the error of each and potential for propagation.

  - resolution, compare projection to known values and compute root mean square error.

- ACCURACY TEST EXAMPLE - HORIZONTAL ACCURACY

Evaluation Data Set: Envirofacts Address Matching Points

Higher Accuracy Source: Texas GPS border survey (20 points)

Projection: National Lambert Meters (NAD 1983)

Geographic Area: Brownsville, TX to Las Cruces, NM

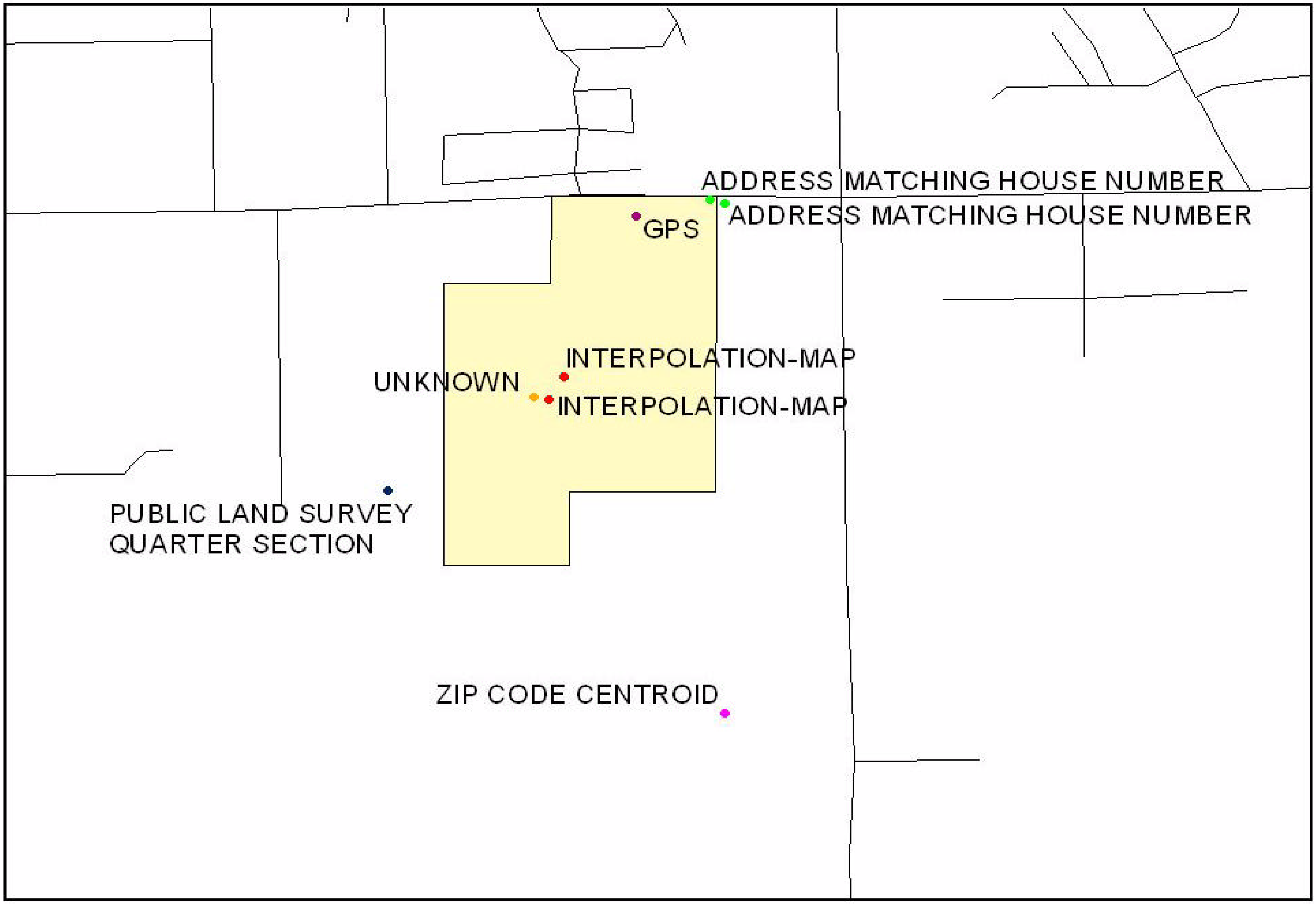
Absolute Difference in x range 8-669 m; y 8-1090 m

RMSE (x) = 187; RMSE (y) = 257

Accuracy = 2.4477*0.5*(RMSE(x) + RMSE(y)) = 544

Reporting: Tested 544 meters horizontal accuracy at 95% confidence level

Positional Accuracy Example



STANDARDS

- **EPA's Locational Data Policy (1991)**
  - principles for collecting and documenting (MAD codes) Lat/Long coordinates for facilities, sites, and monitoring points
- **FIPS 173 (NIST 1994)** data transfer standard with 5 quality report elements
- **NSDI/FGDC (1998)**
  - content standards for digital geospatial metadata (availability, access, transfer and data quality information on fitness for use)

- ATTRIBUTE ACCURACY

Attributes are facts tied to the earth's surface

- qualitative facts like soil classification
- quantitative facts like slope, population
- location of attribute by point, line, area

Error from direct observation, remote sensing interpretation, interpolation

Producers need to provide accuracy information as proof of product!

- ATTRIBUTE ACCURACY (continued)

Fact producer

- Census Bureau population surveys
- USDA soil surveys

Map producer

- USGS elevations

Accuracy determination for quantitative attributes:

- standard error (e.g. 7 m uncertainty in slope based upon 1 m S.D. in elevations)
- known instrument error (Landsat stripping)
- uncertainty models and Monte Carlo analysis

ATTRIBUTE ACCURACY (continued)

Accuracy determination for quantitative attributes:

- standard error
- known instrument error

Accuracy determination for qualitative attributes:

- can't calculate means or S.D.
- can report error matrix

comparisons to a higher accuracy source → uncertainty indices

- percent correctly classified
- Kappa Index

- ATTRIBUTE ACCURACY (continued)

Develop evaluation criteria as part of the systematic planning process!

Example 1:

- consumer's classification accuracy = 85%
- producer ground truth pixels = 80 correct/104 total = 76.9%

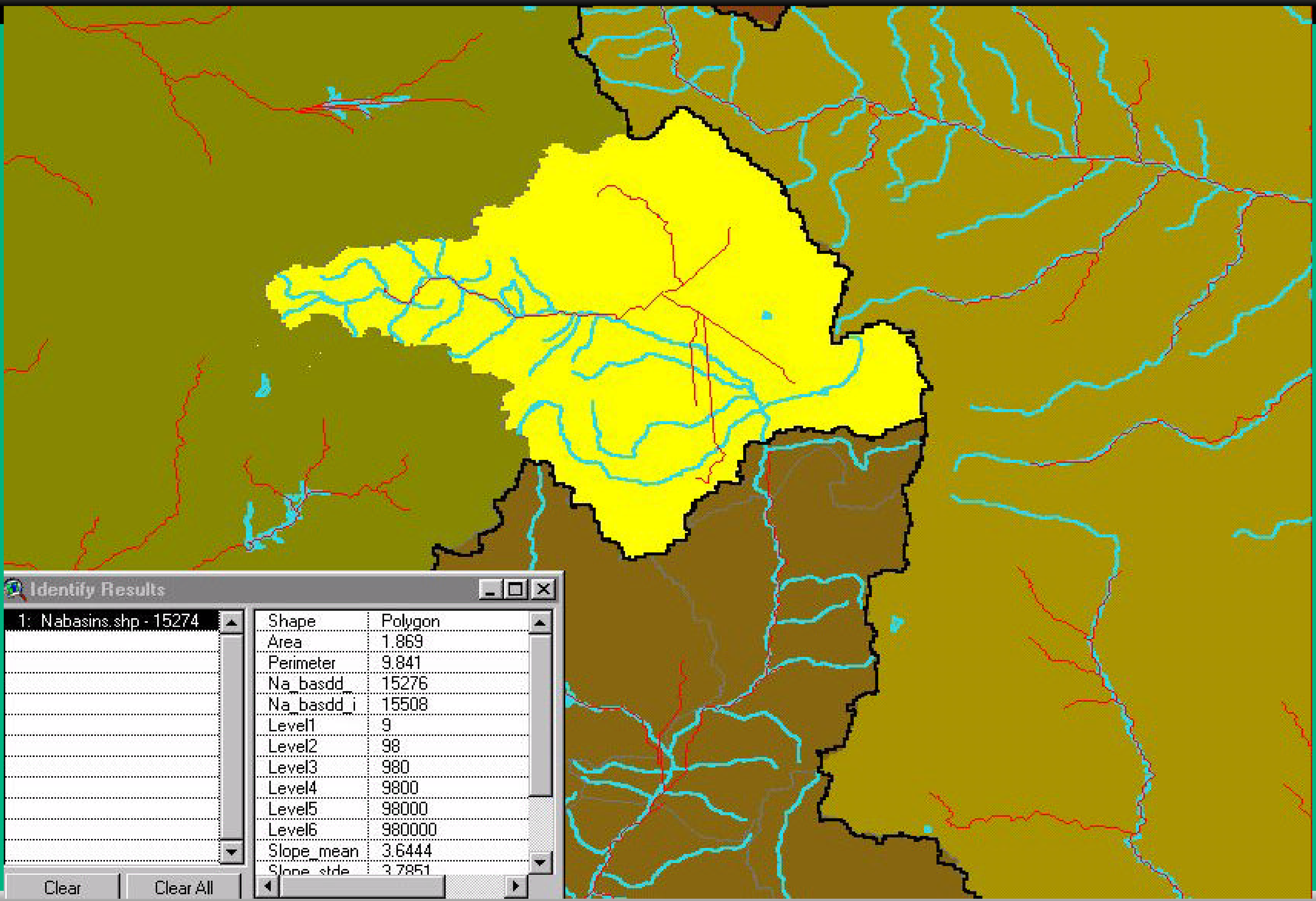
Example 2:

- consumer's slope criterion is 10% uncertainty
- elevation error correlations in digital elevation model = 33% uncertainty

OOPS!

ATTRIBUTE ACCURACY - EXAMPLE

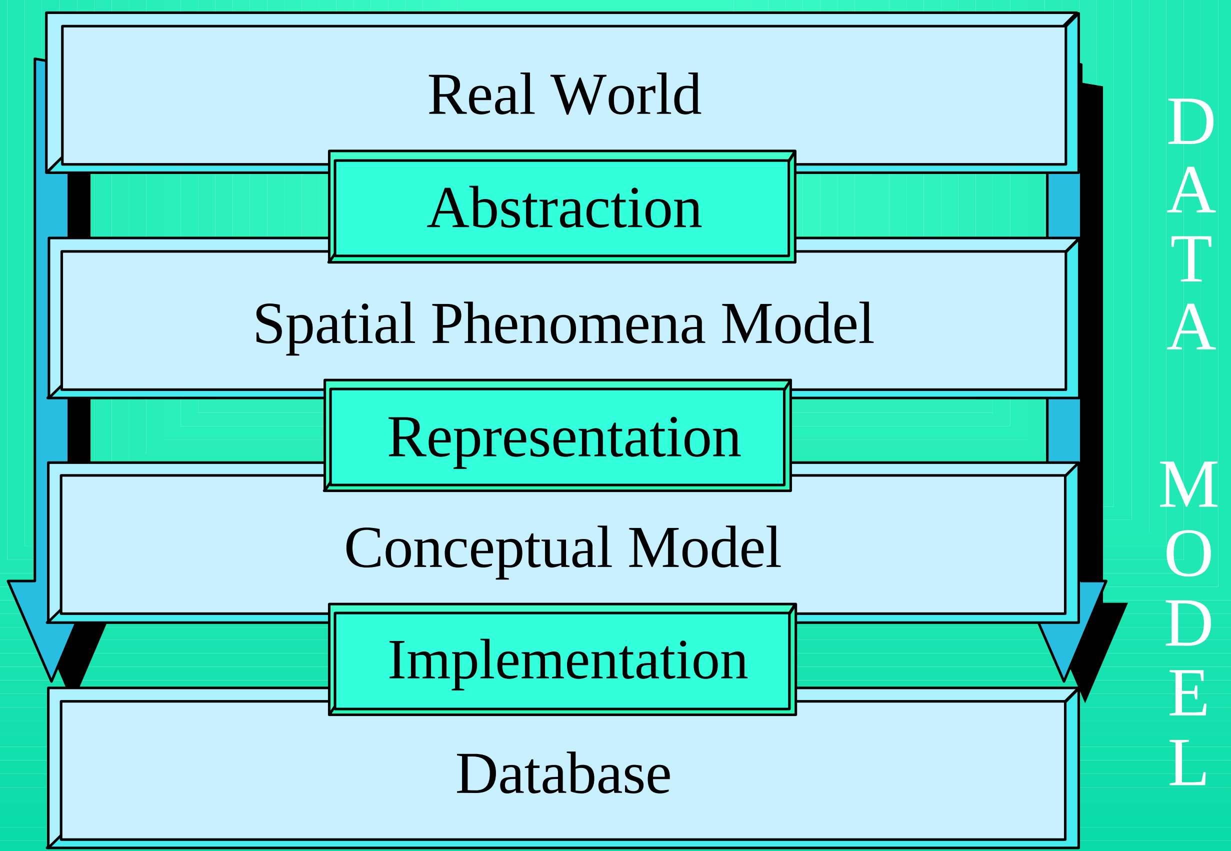
Upper Rio Grande Basin in southern Colorado



COMPLETENESS

Defined as the degree to which the entity objects (and attributes) represent the abstract universe

- Importance of metadata
  - a good definition of the abstract universe
  - defined selection criteria
- Missing data (incompleteness) can affect logical consistency



COMPLETENESS (continued)

Example:

- Need 95% coverage of lakes with surface area of > 1 km^2
- Discover data missing for some small lakes

Reject the data?
or
Relax our completeness criteria?
or
Change our universe definition to surface area > 5 km^2?

- If use is supported and logical consistency maintained

- PRECISION

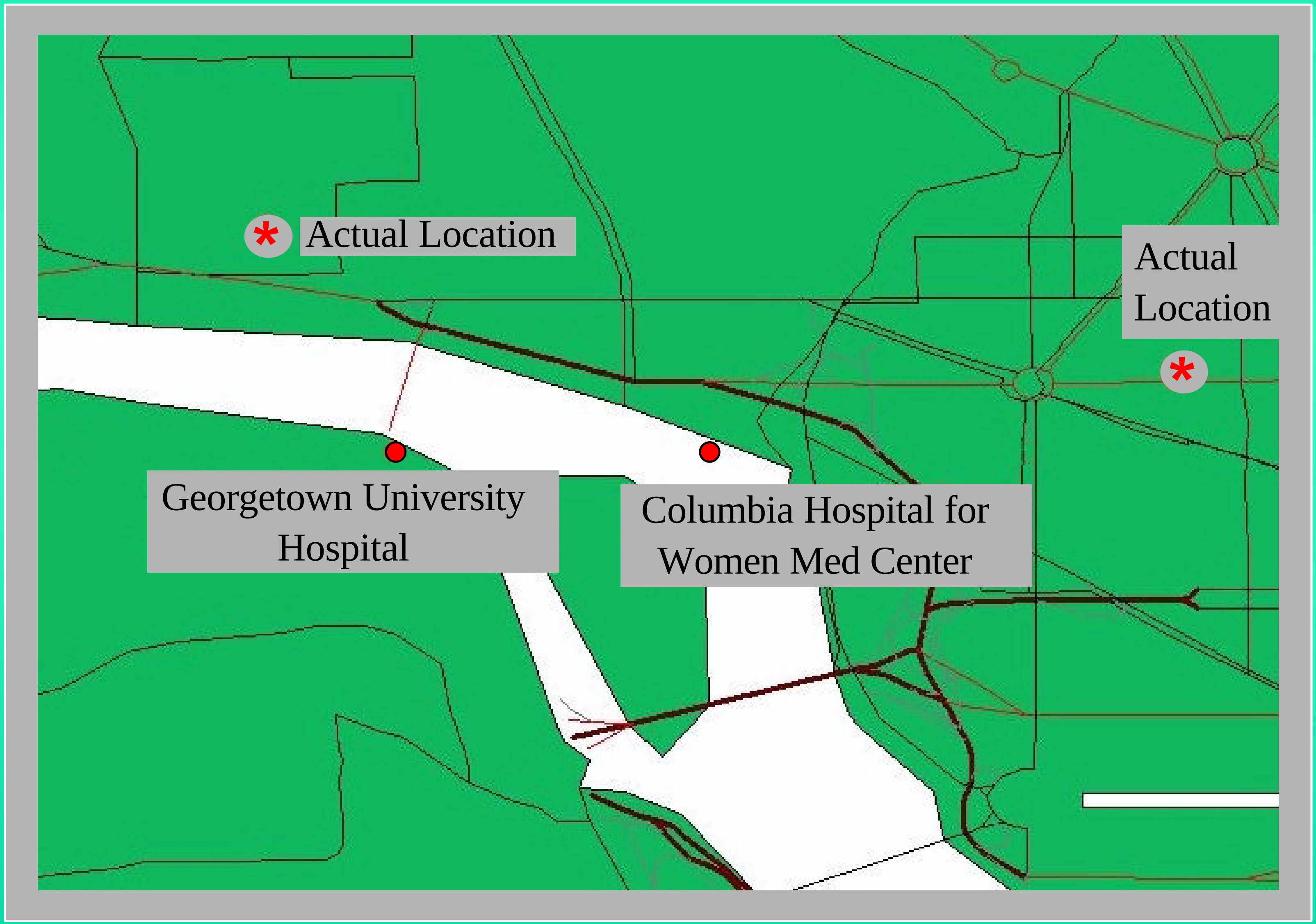
- Defined as the number of decimal places or significant digits in a measurement.

A GIS often works at higher precision than the accuracy of the data.

Precision in Software

An extreme example of data conversion problems:

- geographic coordinates converted from double to single precision



- LOGICAL CONSISTENCY
  - Refers to the internal consistency of the data structure.
  - A spatial data set is logically consistent when it complies with the structural characteristics of the data model and is compatible with attribute constraints.
  - Identify logical rules of structure needed for your use.
  - Identify rules for attribute consistency for your intended use.
  - Tests needed to check spatial data:
    - Compatible datum?
    - Valid attribute values?
    - Compatible with data model?

- LOGICAL CONSISTENCY (continued)
  - Inconsistencies violate rules/constraints.
    - attribute range
    - geometric and topological constraints
    - rules for spatial relationships and application
  - Consistency is needed for control of operational transactions.
  - Evaluations need to be reported.

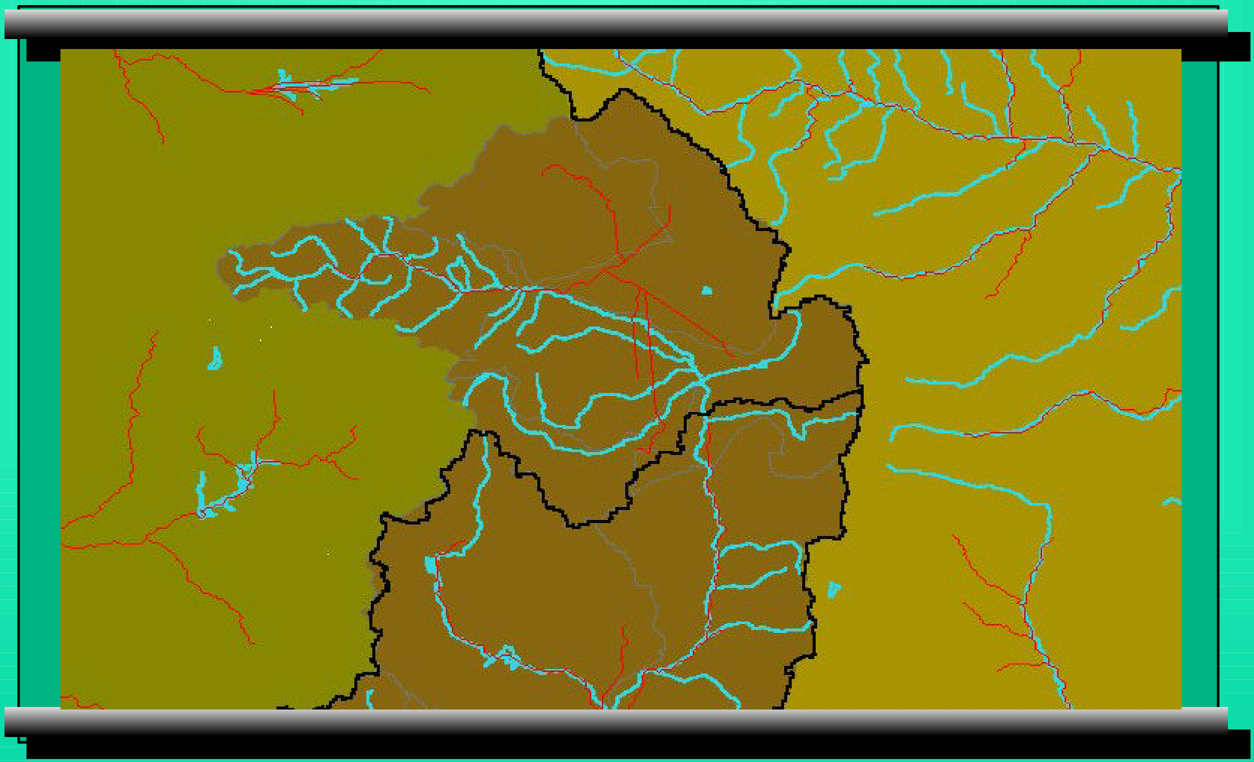
LOGICAL CONSISTENCY - EXAMPLE

Upper Rio Grande Basin in southern Colorado

Global Hydrography Data Base

- hydrography (red lines)
- basins (dark lines)

Secondary Hydrography EPA Reach 3 Files (blue lines)



- LINEAGE

- A record of data sources and database creation

Origin

Was there bias in the source?

In the methods?

References for data accuracy or corrections

Spatial data characteristics

Coordinate systems

- LINEAGE (continued)

Map Projections

- translation of spherical data to flat coordinates

Corrections or equipment calibrations

- number and type

Transformations and analyses

- 3-D transformation of a model's coordinates to terrain's coordinate system using control points (software version?)

Aerial Photography Source (ESD)

QAPP

- project objectives, responsibilities, criteria, reporting requirements

Method and Quality Control

- photogrammetric measurement, flying height, type of camera, type of digitizer, film, seasonal conditions, reference coordinates

Date/scale

- 1990/04/01; 1:30,000

Source Material Structure

- 60% stereo model to produce map

SOPs for verification, analysis and interpretation

PROPAGATION OF ERROR

Input error

- locational data
- maps
- digitized input
- environmental attributes from measurements

PROPAGATION OF ERROR

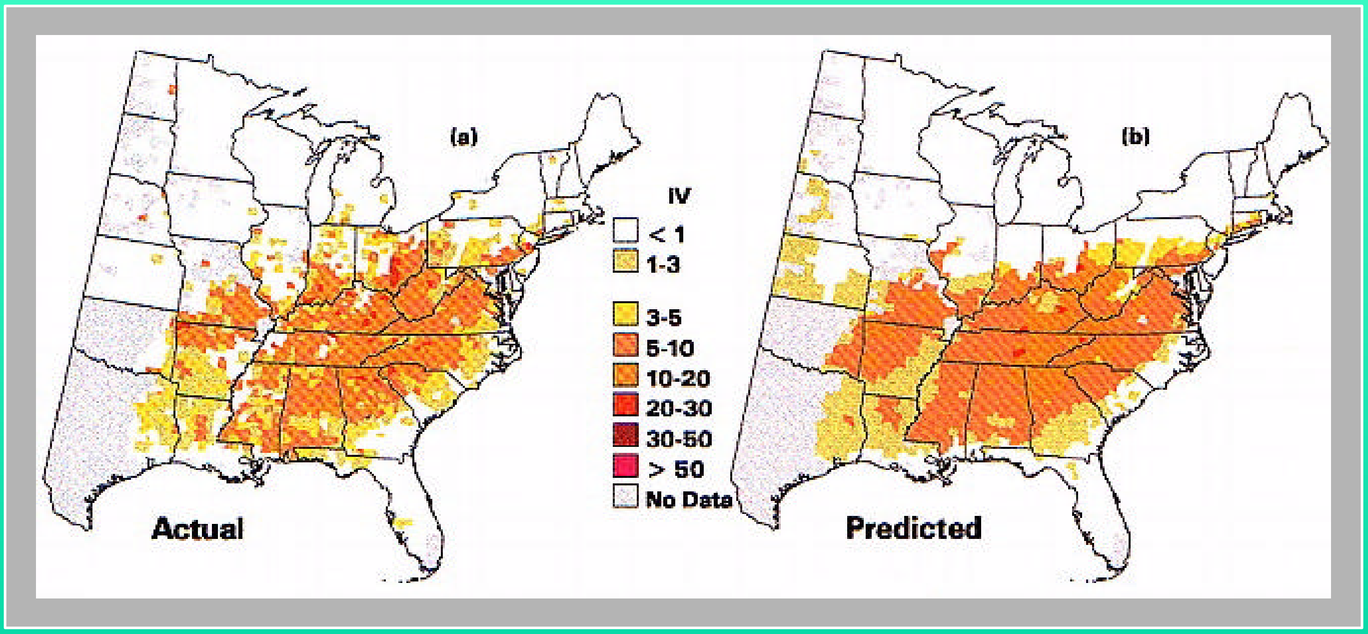
Processing Error

- GIS software
- Sequential processing steps
- Software predictive model

Projections into the Future

Future Modeling using GIS

- Urbanization Effects
  - San Francisco
  - Baltimore-Washington Corridor
- Vegetation Change
  - Global Warming: Predicting Abundance of Tree Species Following Climate Change in the Eastern United States



Limitations in Predictive Modeling

- overlaying GIS layers of differing sources and scales

  → Error Propagation
- uncertainty associated with variables
- inappropriate model assumptions
- extrapolation
- model verification and validation

PROPAGATION OF ERROR

Product Uncertainty?

- Source error documentation needed
- Process traceability needed
- Uncertainty estimates needed
- Problem: More research in error propagation needed

MODULE SUMMARY

- Error identification and management need to be addressed in QAPP systematic planning.
- The 5 major components of error (FIPS 173) and their propagation need to be evaluated in both primary and secondary data uses.
- Lineage (metadata) is needed for useful products.