1. Data Quality for GIS

Data Quality for GIS

Barbara Parmenter, PhD
GIS for Urban and Environmental Analysis
Tufts University
Fall 2007

(c) 2007, Barbara Parmenter, PhD

2. GIS quality-related issues

GIS quality-related issues

- Cost and difficulty of database creation
- Higher accuracy, higher costs
- Dependency on data we have rather than data we need
- Integration of data from diverse sources/scales can result in poor overall accuracy if not well managed

(c) 2007, Barbara Parmenter, PhD
3. Data Quality

Data Quality

- Higher quality data costs more to develop or to purchase
- Think about what the users of the information need to do their work
- Obtain / create data that meets those needs
- Do not create data that is of higher accuracy or resolution than what the purposes require

(c) 2007, Barbara Parmenter, PhD

4. GIS Projects – Cost versus Accuracy: Hypothetical Exam...

GIS Projects – Cost versus Accuracy

Hypothetical Example

A increase in accuracy, e.g., from 90-99% results in geometrically higher costs, not just a 9% cost increase

(c) 2007, Barbara Parmenter, PhD
5. Request to develop land use layer for Texas town

Request to develop *land use* layer for Texas town

- For watershed / water quality / flood modeling
- My question: "What are the accuracy requirements?"
- Person making request: "Accuracy does not matter"
- My response: "OK, I will complete it in 3 minutes…"
- Accuracy always matters!

(c) 2007, Barbara Parmenter, PhD

6. Data Quality Concepts

**Data Quality Concepts**

- **Accuracy**
  - Closeness to true value
  - Expressed as a probability (e.g., 90% of all points are within +/- 10m of their true location)
- **Resolution**
  - Minimum size of features that are discernible
- **Precision**
  - Ability to distinguish small differences
- **Currency**
  - Is the data current (up to date)?

(c) 2007, Barbara Parmenter, PhD
Accuracy, precision, resolution

- High resolution or precision does not automatically equal high accuracy
- Precision: A watch can be precise to the 1/100th of a second, and be 10 minutes slow.
- Resolution: Satellite imagery of 10km cells can provide accurate information for climate modeling at a global or national scale even though they are lower resolution than 1m orthophotos
- Understanding purpose and accuracy needs if critical!

Types of accuracy requirements

- Geographical position of features – 97% +/- xx meters
  - Relative or absolute?
  - Specify as compared to a source of known accuracy (e.g., compared to a USGS topographic map or to high resolution orthophotos)
  - Cartographic placement requirements
- Attribute accuracy – 99%
- Connectivity (e.g., roads, electric, stream networks) – 100%
- Database relationships / integrity – 100%
9. Types of error

**Types of error**

- Positional
  - Absolute
  - Relative
- Attribute
- Logical consistency
- Completeness
- Currency (time period)

*We’ll take these one by one…*

(c) 2007, Barbara Parmenter, PhD

10. Positional error

**Positional error**

- Positional
  - Absolute
  - Relative
- **Expressed as a probability of a feature being within +/- x units of its:**
  - True location on the face of the earth (absolute)
  - Or, its location in relation to other features in the map/database (relative)
- **E.g., 98% of the parcel corner points in a parcel layer are +/- 5 feet of their true location; or 90% of highway centerlines are +/- 3m of the centerline visible in a 1m orthophoto**

(c) 2007, Barbara Parmenter, PhD
11. Source of the “true value” for estimating accuracy?

Source of the “true value” for estimating accuracy?

When we say we want a map or data layer to be +/- x feet from its true position, how do we know that true value? We must have some other source in mind as representing this true value. What is the source of the “true” position? In reality, we mean compared to a source of known and higher accuracy. Some possibilities:

- Surveyed control points
- GPS points of known accuracy
- Maps of known higher accuracy
- Digital orthophotos of known accuracy

(c) 2007, Barbara Parmenter, PhD

12. US National Map Accuracy Standards (Paper Maps)

US National Map Accuracy Standards (Paper Maps)

- The National Map Accuracy Standard (NMAS) defines the requirements for meeting horizontal accuracy as 90% of all measurable points must be within 1/30th of an inch for maps at a scale of 1:20,000 or larger, and 1/50th of an inch for maps at scales smaller than 1:20,000.

(c) 2007, Barbara Parmenter, PhD
13. “Source scale” is one determinant of positional ... 

“Source scale” is one determinant of positional accuracy

- GIS data comes from some source (e.g., paper map, aerial orthophoto, satellite image, GPS)
- If from map, orthophoto, or satellite image, this source has a scale or resolution
- GIS layers derived from one of these sources are no MORE accurate than the source, and almost always LESS accurate

(c) 2007, Barbara Parmenter, PhD

14. About Scale

About Scale

- A map scale tells you what a certain distance on the paper map represents in the real world

<table>
<thead>
<tr>
<th>Scale expressed in words</th>
<th>One unit of something on the map = this many of those same units in the real world</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 inch = 1 mile</td>
<td>1:63,630 (5,280 feet x 12 inches)</td>
</tr>
<tr>
<td>1 inch = 2000 feet</td>
<td>1:24,000 (2000 feet x 12 inches)</td>
</tr>
<tr>
<td>1 inch = 100 feet</td>
<td>1:1200 (100 feet x 12 inches)</td>
</tr>
</tbody>
</table>

(c) 2007, Barbara Parmenter, PhD
15. Large and Small Scale Maps

Large and Small Scale Maps

1:5,000 is large scale
1:50,000,000 is small scale

The larger the denominator, the smaller the scale
1/4 of a pie is smaller than 1/2 of a pie

(c) 2007, Barbara Parmenter, PhD

16. Scale 1:100,000 (1 inch approx 1.58 miles)

(c) 2007, Barbara Parmenter, PhD
17. Scale 1:24,000

(c) 2007, Barbara Parmenter, PhD

18. Scale 1:10,000

(c) 2007, Barbara Parmenter, PhD
19. Scale 1:2,000

20. Relation between positional error and scale

Relation between positional error and scale

US National Map Accuracy Standard (for paper maps)
The values below are what are expected for accuracy on paper maps at
these different scales. All paper maps produced by the US government are
expected to meet these requirements (90% of measurable points must
meet the stated standard).

<table>
<thead>
<tr>
<th>Scale</th>
<th>Engineering Scale</th>
<th>National Map Accuracy Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1,200</td>
<td>1&quot;=100'</td>
<td>+/- 3.33 feet</td>
</tr>
<tr>
<td>1:2,400</td>
<td>1&quot;=200'</td>
<td>+/- 6.67 feet</td>
</tr>
<tr>
<td>1:4,800</td>
<td>1&quot;=400'</td>
<td>+/- 13.33 feet</td>
</tr>
<tr>
<td>1:9,600</td>
<td>1&quot;=800'</td>
<td>+/- 26.67 feet</td>
</tr>
<tr>
<td>1:10,000</td>
<td></td>
<td>+/- 27.78 feet</td>
</tr>
<tr>
<td>1:12,000</td>
<td>1&quot;=1000'</td>
<td>+/- 33.33 feet</td>
</tr>
<tr>
<td>1:24,000</td>
<td>1&quot;=2000'</td>
<td>+/- 40.00 feet</td>
</tr>
<tr>
<td>1:63,360</td>
<td>1&quot;=one mile</td>
<td>+/- 105.80 feet</td>
</tr>
<tr>
<td>1:100,000</td>
<td></td>
<td>+/- 166.67 feet</td>
</tr>
</tbody>
</table>

(c) 2007, Barbara Parmenter, PhD
21. NMAS Standards

This 1:24,000 USGS topographic map should meet NMAS standards - 90% of measurable points are within +/- 40 feet. What about all the other features?

22. Digitizing paper maps adds error

Digitizing paper maps adds error

- Errors in registering the map on the digitizer or onscreen
- Paper map can shrink, stretch, or tear
- Georeferencing is not perfect
- Errors by person digitizing
- Converting between coordinate systems and transforming between datums may add error

1:24,000 paper map with +/- 40 feet accuracy
Error when digitized is more than 40 feet, but difficult to know exactly how much more (RMS error value can be known).
23. Attribute accuracy

Attribute accuracy

- Wrong value in an attribute field
- For continuous values, like elevation, the error estimate could be expressed like positional error (e.g., +/- 15m)
- Categorical data (e.g., land use)
  - Classification error (example: value is "industrial" when it should be "residential multifamily")
  - Boundary error (example: a residential land use polygon extends into an industrial area)
- Missing attribute values
- Zero values that should be null?
- Misspelled word or other typographical error

(c) 2007, Barbara Parmenter, PhD

24. Logical Consistency

Logical Consistency

- Do the data fit the data model?
- Do water pipelines connect at a node?
- Does the river lie inside the floodplain?
- Is the soil type one of the possible categories for New England soils?
- Is there one parcel line (no gaps) between two adjacent parcels?
- Are there gaps or overlaps between census tracts?

(c) 2007, Barbara Parmenter, PhD
25. Completeness Is the data set “complete”?

Completeness
Is the data set “complete”?

- Are all parcels in the city represented in the parcel layer?
- Does data stop at a city or county boundary when you were expecting it to continue?
- Are all streets in the database? How about streets within apartment complexes or industrial parks or shopping malls? Highway ramps?
- Are all grocery stores in the database? Does that include convenience stores or not?
- Are all schools in a database? Only public? All that are licensed by state? Source for knowing what a complete set of schools might be?


Currency
Is the data set current? No!

- Data sets are never completely current
- But just how old is it?
- Distinguish between:
  - Publication date (when the data set was published)
  - Field date – when the data were actually collected (or the field dates of the sources from which the data was digitized)
  - Forecast period and date of forecast
  - Should be specified in metadata as “time period of content”
- Is there a regular update program? How often?

(c) 2007, Barbara Parmenter, PhD
Ensuring data quality

- During data conversion process the following methods are often used to prevent errors from occurring in the first place:
  - Use default values (e.g., if most water pipelines are 100 cm in diameter, set the default value at 100)
  - Restrict type of entry to domain or range of values
    - Water pipelines are not just any diameter – there may be 6 different standard diameters – list these as the only choices
    - Elevation values may range between 0 and 10,000 meters.
  - Required fields – must be filled before continuing – prevents missing data values

(c) 2007, Barbara Parmenter, PhD

Ensuring data quality

- After data development, other methods can be used, including:
  - 100% manual check against sources of higher accuracy (e.g., paper maps, printed information)
  - Manual or field check of a sample of the data
  - Automated logical consistency checks (typically these will put some kind of flag in the map or on a report to aid correction of errors)
    - Are network lines connected
    - Overshoots or undershoots in lines or polygons
    - Are all fields filled? No null fields?
    - Do attribute values match domain or range of possible values?

(c) 2007, Barbara Parmenter, PhD
Ensuring data quality

- Checking data takes a lot of time and costs a lot of money
- Often data is not checked fully or properly
- Look for automated checks whenever possible
- Example of unusual method: double data entry – less expensive to pay two typists to enter the data twice, and then do an automated comparison, rather than pay one person to type once and one person to check it manually once
- Always document data quality in metadata

(c) 2007, Barbara Parmenter, PhD

Documenting data quality

- Describe positional accuracy in quantitative terms if possible (e.g., +/- 2m) or refer to scale and lineage descriptions (see below)
- Describe attribute accuracy (also in quantitative terms if possible (e.g., 98% of attribute values are correct; elevation values are within +/- 5m of their correct elevation)
- Describe data source and source scale
- Describe lineage – the process used to create the data, including registration RMS error (to be discussed later)
- Document time period represented by the data (not only the date of digitizing)

(c) 2007, Barbara Parmenter, PhD
31. What positional accuracy do you need?

**What positional accuracy do you need?**

- Childhood asthma study in a city/state
- Water/wastewater pipeline maintenance system
- Urban land development planning (suitable areas for development)
- Agricultural suitability modeling
- Analyzing access to healthy food for urban populations
- Land use and environmental information to support development decision-making in rural regions

(c) 2007, Barbara Parmenter, PhD

32. What attribute accuracy do you need?

**What attribute accuracy do you need?**

- Childhood asthma study in a city/state
- Water/wastewater pipeline maintenance system
- Urban land development planning (suitable areas for development)
- Agricultural suitability modeling
- Analyzing access to healthy food for urban populations
- Land use and environmental information to support development decision-making in rural regions

(c) 2007, Barbara Parmenter, PhD
33. GIS quality-related issues

GIS quality-related issues

- Cost and difficulty of database creation
- Higher accuracy, higher costs
- Because of costs, we tend to depend on the data we have rather than data we need
- Integration of data from diverse sources can result in poor overall accuracy if not well managed
- Representation of features that don’t have sharp boundaries in reality

(c) 2007, Barbara Parmenter, PhD

34. Metadata is critical!

Metadata is critical!

- To know the quality of a data set and whether it is appropriate for your purpose, access to good metadata is extremely important
- Good metadata should document positional and attribute accuracy, time period, source scale, purpose, lineage of the data (how it was created), completeness, and logical consistency

(c) 2007, Barbara Parmenter, PhD
Web sites with more information

- From *The Geographers’ Craft Project*
  Kenneth E. Foote and Donald J. Huebner, Department of Geography, The University of Colorado at Boulder
  - Error, Accuracy and Precision
    http://www.colorado.edu/geography/gcraft/notes/error/error_f.html
  - Managing Error
    http://www.colorado.edu/geography/gcraft/notes/manerror/manerror_f.html
- From the US Geological Survey (USGS)
  - National Mapping Standards
    http://nationalmap.gov/gio/standards/
  - Map Accuracy Standards
- Other useful overviews:

(c) 2007, Barbara Parmenter, PhD

Learn more about metadata

- Federal Geographic Data Committee – Geospatial Metadata
  http://www.fgdc.gov/metadata
- Minnesota Land Management Information Center
  http://www.lmic.state.mn.us/chouse/meta_help.html
  (good guides for a variety of metadata and data quality assessment issues)
- National States Geographic Information Center – Metadata Resources
  http://www lick.wisc.edu/metadata/metaprim.htm

(c) 2007, Barbara Parmenter, PhD
37. Metadata examples

**Metadata examples**

- National Hydrography Data for Nevada - [http://www.epa.gov/esd/land-sci/nv_geospatial/pages/nvgeo_gis10_nhd_md.htm](http://www.epa.gov/esd/land-sci/nv_geospatial/pages/nvgeo_gis10_nhd_md.htm)
- MassGIS – Community Health Centers - [http://www.mass.gov/mgis/chcs.htm](http://www.mass.gov/mgis/chcs.htm)

(c) 2007, Barbara Parmenter, PhD