UNIT 30: VALIDATING DATABASES

Written by Chris Cogan, University of California, Santa Cruz

Context

What is Validation?

Database validation is the process of determining if database values are reasonably accurate, complete, and logically consistent. "Reasonable accuracy" will depend on the intended use of the data, so values relative to stated and required standards are often more appropriate than absolute data accuracies.

Validation will often consist of several steps, including logical checks, accuracy assessments, and error analysis. In this context, accuracy assessment is the determination of spatial and thematic accuracy relative to a known standard (e.g. Goodchild 1995) and error analysis involves the evaluation of data with regard to measurement uncertainty (Taylor 1982), and includes source errors, use errors, and process errors (Beard 1989).

How is Validation done?

The simplest form of validation is not a direct analysis of the data itself; rather the potential user reads through the accompanying documentation, (see metadata, Unit 7) to determine if the data in question seem appropriate for the application. Once the metadata have been examined, it is often necessary to validate the actual data. This can be thought of as a two level process, with primary validation looking at directly verifiable properties of the data, and secondary validation involving various levels of user judgement during the assessment. In some cases the primary validation will be unnecessary, for example when metadata inform the user that certain forms of logical consistency have been machine validated during initial data construction. Examples of primary and secondary validation are listed in Table 1.

The major kinds of validation listed in Table 1 show that a single type of data can be validated in several different ways. For example, suppose you have a point map with elevation data as the attribute. You could have automated validation performed in the GIS software that ensures that each point is drawn with the same symbol, and that the elevation for each point is an integer value. These are examples of primary validation which should be (but

are not always) performed by the GIS software. Secondary validation of the same point data could include statistical tests to look for outliers, queries of the data to determine if the elevation values fall into a reasonable range, or checks to see if all the points occur within a reasonable area.

Primary Validation:

- logical cartographic consistency
    - closed polygons
    - one label for each polygon
    - no duplicate arcs
    - no overshoot arcs (dangles)
    - similar features use similar symbols
- logical attribute consistency
    - values within logical range (look for illegal values)
        - dates (e.g. month less than or equal to 12)
        - time of day less than 24:00 hours
        - nominal data illegally resampled into ratio data
        - precipitation values equal to or greater than zero
    - linkage of features with attribute fields
        - does a polygon map of lakes include depth and salinity data for each lake?

Secondary Validation:

- logical query and statistical tests of the spatial and attribute data (look for unlikely values)
    - points placed in distant locations on the map
    - elevations with reasonable values
- ground truth or comparison to known standards
    - sample ground areas and compare to database
        - evaluate spatial accuracy
        - evaluate attribute accuracy
    - completeness of data ("model type" completeness - relative to user needs)
        - in a map of roads, are all the roads important to the us er included?
- sensitivity analysis
    - change the data, and see if those changes affect the results for your application.

Table 1. Primary and secondary levels of database validation

Who performs database validation?

Ideally the producer will perform some basic validation and document it in the metadata. However, every data user should conduct some level of validation before using the data to ensure its suitability for their application. This may only involve a quick overview of the database statistical summaries, or a review of the metadata documentation. In many cases however, more extensive validation efforts are needed, requiring more in-depth assessment of the data.

Why is validation increasingly performed?

Digital data:

Traditional data products such as government topographic maps are based on familiar analog formats with published accuracy standards. Data users are often experienced with the spatial and thematic accuracy of these products, and are aware of the cartographic necessities of increasing line thickness, object representation, and varying resolution with changing map scales. As we move away from analog products towards the digital domain, data quality becomes less intuitive and less understood, placing more demand on users to assess the data before using them in an application.

Digital data are also more complex than many analog products, often maintaining extensive links to ancillary thematic information, and this complexity increases the need for improved understanding of the data. It is generally understood that all maps are generalizations, but when multiple spatial and aspatial data sets are combined into a GIS map layer, an understanding of data quality becomes an essential element in the value of a dataset.

Increased data sharing and decentralized data production:

In addition to the increased demands of digital data, improved portability and accessibility of data is allowing more people to publish their own data, compared to the more traditional centralized distribution of analog products. With more users creating application specific data sets, there is increasing potential for misunderstanding data quality, and inappropriate application of the data. In many cases, users can be guided by the metadata accompanying the dataset; however at other times the user will need to validate the data for themselves to ensure they are being appropriately applied.

More custom data production:

Even data users who have generated their own information are often unsure of data accuracy and consistency. With increasing in-house production of custom generated data designed for specific applications, users are responsible for their own validation. The results of the validation are often reported as part of the metadata, a required data component for many projects.

Example Application

A biologist and GIS analyst are working with a database on the endangered California condor. Part of their research is to determine if condor foraging and feeding activity is correlated with land cover vegetation (e.g. Stoms et al 1992).

Given a vegetation polygon map of the condor range, and a point map of condor sightings, the researchers must determine if the points representing condor locations are appropriate for use with the vegetation map. The vegetation data appear to be well suited for condor studies, having an appropriately detailed classification scheme and level of spatial resolution.

The condor researchers begin with logical consistency checks of the condor sighting information, looking for valid dates and times. If a problem is found such as the 32nd day of the month or 25:00 hours as time of day, the entire record is reviewed. Secondary validation steps require user judgement and include logical queries of the attribute information for reasonable values of date, time, and bird activity. Suspicious records such as those reporting nighttime flight observations or records for distant locations are marked for further validation. As a final measure, the researchers employ a sensitivity analysis to see if minor displacement of the condor sighting locations results in significant changes in reported species-habitat correlations. The study shows that even if the condor sighting points are moved 250 meters in random directions, several patterns of behavior are still strongly correlated with specific land cover types. Thus, despite some uncertainty about the exact locations of the observations, the data are valid for this application.

Learning Outcomes

The following list describes the expected skills which students should master for each level of training, i.e. Awareness/Competency/Mastery.

Awareness:

Students should be able to discuss the issues of database validation as they relate to GIS; be familiar with the vocabulary terms used with validation issues; and demonstrate a working knowledge of some common types of validation in use.

Competency:

Students should be able to evaluate datasets for appropriateness; identify potential problems with the data which could affect dataset use, and recommend more complete measures for validation.

Mastery:

Students should be able to conduct each of the primary and secondary validation procedures, including statistical testing, and sensitivity analysis. Be able to discuss issues of automated primary validation, including the future potential for all GIS data to be machine validated for logical consistency.

Preparatory Units

Recommended:

1. UNIT 1 - Data acquisition
2. UNIT 7 - Using and interpreting metadata
3. UNIT 24 - Collecting GPS data

Complementary:

1. UNIT 8 - Error checking
2. UNIT 11 - Registration and conflation
3. UNIT 26 - Editing point data
4. UNIT 27 - Editing linear data
5. UNIT 28 - Editing polygon data
6. UNIT 29 - Editing raster data
7. UNIT 36 - Using distance and connectivity operators
8. UNIT 39 - Performing statistical analyses

Awareness

Learning Objectives for Awareness level skills:

1. Be able to define "validation" and give examples of several methods used.
2. Become fluent with the vocabulary of validation, including the terms listed in this section.

Vocabulary* - (Suggestions)

- accuracy - (relative to a more accurate standard)
- attribute accuracy - (correctly identified data about points, lines, and areas)
- buffer - (extension of area about points, lines, and areas)
- completeness - (data completeness=data quality, model completeness=data fitness)
- coordinate pair - (longitude latitude; x,y)
- data dictionary - (reference for codes used in the database)
- error analysis - (error budget, analysis of measurement error)
- global positioning system - (GPS- field mapping tools, independent locations)
- line data - (arcs, roads, boundaries, small rivers)
- lineage - (ancestral source data, how the data was derived)
- logical consistency - (logical rules for structure and attribute rules for spatial data)
- map extent - (map domain, area covered by the map)
- map scale - (representative fraction, large scale, small scale, inconsistent term)
- metadata - (data about the data, a description of the data)
- minimum map unit - (grain, dimension of the smallest area allowed in the data)
- point data - (intersections, centroids, areas below the minimum map unit)
- positional accuracy - (position relative to "true" position)
- polygon data - (area data, at some map scales becomes point or line data)
- query language - (SQL, language to retrieve, modify, add, or delete data)
- raster data - (remote sensing, imagery, pixels, matrix of cells with values)
- spatial registration - (alignment for coincident positioning)
- spatial resolution - (smallest area identifiable, one raster cell)
- thematic resolution - (number of classes)
- truth table - (error matrix, producers and consumers accuracy)

*Many of these terms are defined in GIS text books, though not all definitions are consistent. For additional information, see Guptill and Morrison, 1995) and the following web sites:

Glossary for Exploring Geographic Information Systems *[outdated link removed]*

The GIS Glossary from ESRI *[outdated link removed]*

Glossary for the Geographer's Craft - University of Texas *[outdated link removed]*

Tasks

1. Using Table 1 as a guide, explain how validation can operate at several different levels of user interaction.
2. Briefly describe a database and a potential application of the data, then give at least four examples of database validation procedures appropriate to the situation.

Competency

Learning Objectives:

1. Be able to recognize potential problems with a dataset.
2. Be prepared to implement several types of validation procedures.

Task:

1. You are given a polygon database of the boundaries of all the U.S. Geological Survey (USGS) 1:24,000 scale topographic maps which cover the States of Washington, Oregon, and Idaho. This kind of database is used to index maps, see the USGS web page for an example: http://edcwww.cr.usgs.gov/doc/edchome/ndcdb/ndcdb.html

Each polygon represents the area covered by the USGS map, and includes a label with the name of the topographic map. Looking through the metadata, you see the map was generated by a computer program finding the corners of the 1:24,000 quadrangle maps at intervals of 7.5 degrees latitude and longitude.

Your task is to validate the GIS data layer, for cartographic as well as attribute quality.

Steps:

1. To begin this task, it is helpful to visualize the data, using resources such as the USGS web page listed above, or other map sources. Notice that the three state area is approximately 7 degrees latitude by 14 degrees longitude so there are too many 7.5 minute maps to examine individually. (i.e. in 14 deg. longitude, 14 deg = 14 deg x 60

min/deg x 1 quad/7.5 min. = 112 quads)

2. Beginning with logical cartographic consistency, check to verify that each of the polygons is closed and properly labeled with a single label. This step will catch polygons without a label, and polygons with more than one label. Multiple labels can occur in this type of data when a polygon is not quite closed, so two adjacent quadrangles will have the correct labels, but are stored in the database as one contiguous polygon.

   If you are using Arc/Info, check the Arc command labelerrors. Other useful Arc commands are nodeerrors (to find nodes connecting less than three arcs), and log to see if the coverage has been built and cleaned.

   

   Arc/Info Example

3. This data can also be checked for attribute consistency and values within a reasonable range. Is there a text string (the map name) for each polygon? Are the map coordinates within the range of values expected? Are the quadrangle polygons of reasonable size?

4. Another step in the validation of the data could be a comparison to known standards. Begin by looking at several areas of the map in question, and compare this to a reference map such as the USGS web site listed above, or a paper map indexing the USGS quadrangles in the areas of interest. Does the map being validated have correct polygon placement along the coastline? Are there any map outlines that appear to be completely offshore? When are the corners of the map areas not at even 7 1/2 minute increments of latitude and longitude?

Mastery

Learning Objectives:

1. Given a complex dataset of landcover, outline the appropriate steps to conduct a validation.
2. Given a new application of a previously validated dataset, be able to critique the existing validation, and explain what additional measures will be required.

Task:

You are given a one hectare minimum map unit (mmu) polygon land cover map of Idaho USA. Over 200 people have worked on this monumental effort over a period of four years. Despite the large number of people and long duration of the mapping project, some of the metadata is incomplete; however this new map represents the most detailed data of its kind in the country. In addition to several categories of urbanization and agriculture, the map includes

200 categories of native vegetation.

Your research team wants to use the map to evaluate the spatial relationship of native vegetation to urbanized areas. Area, patch size, and distance to roads are critical to the analysis. At your disposal are 100 person/hours of field work, 50 digital orthographic photo quadrangles (1 meter resolution air photos terrain corrected and registered to 1:24,000 scale USGS topographic maps), and three months of laboratory access to GIS hardware and software.

Your job is to outline an appropriate validation strategy for the land cover map, describing the methods you have selected based on the proposed use of the data.

Suggestions:

Each of the steps listed in Table 1 are worth consideration.

Incorporate ancillary data into your validation. For example, suppose several of the vegetation types in the map are known to exist only above certain elevations. Use a digital elevation model to validate these assumptions.

Part of the proposed research will involve measuring the distance of particular vegetation types to roads and urban areas. This measurement will require a certain level of cartographic accuracy. In order to avoid the added complexity of fuzzy natural vegetation boundaries, select a landcover type which is clearly distinguishable (lakes, permanent agricultural boundaries etc.) and determine if these polygons are registered with reasonable accuracy. GPS and Digital Ortho Photos could be useful for this task.

Since this type of validation is always dependant on the requirements of the data user, you will need to assume some specific applications for your validation to be useful. If you are working with a dataset that has been validated for another use (e.g. under a different mmu), how might the previous validation be inappropriate for your needs?

Follow-up Units

1. Unit 31 - Managing database files
2. Unit 33 - Using buffering operators
3. Unit 34 - Using overlay operators; Pre/Post overlay tasks
4. Unit 35 - Point in polygon operations; Line in polygon operations
5. Unit 41 - Using boolean search techniques

Resources

- Aronoff, S. 1985. The minimum accuracy value as an index of classification accuracy. Photogrammetric Engineering and Remote Sensing, 51(1):99-111.

- Beard, M.K. 1989. Use error: the neglected error component. Proceedings, AUTO-CARTO 9, Ninth International Symposium on Computer-Assisted Cartography, Baltimore, April 2-7, pp. 808-817.

- Blakemore, M. 1984. Generalization and error in spatial data bases. Cartographica, 21:131-139.

- Chrisman, N.R. 1991. The error component in spatial data, p. 165-174. In D. J. Maguire, M. F. Goodchild, and D. W. Rhind (ed.), Geographical Information Systems: Principles and Applications, vol. 1. Longman Scientific & Technical.

- Congalton, R.G. 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment, 37:35-46.

- Congalton, R.G., and K. Green. 1993. A practical look at the sources of confusion in error matrix generation. Photogrammetric Engineering & Remote Sensing. 59(5):641-644.

- Edwards, T.C., G.G. Moisen, and D.R. Cutler. 1988. Assessing map accuracy in a remotely sensed, ecoregion-scale cover map. Remote sensing of environment, 63(1):73-83.

- Fenstermaker, L.K. 1994. Remote sensing thematic accuracy assessment: a compendium. American Society for Photogrammetry and Remote Sensing, Bethesda, Md.

- Ginevan, M.E. 1979. Testing land-use map accuracy: another look. Photogrammetric Engineering and Remote Sensing, 45(10): 1371-1377.

- Goodchild, M.F. 1995. "Attribute accuracy", In Elements of spatial data quality, edited by S.C. Guptill and S. Gopal Elsevier Science. Oxford, U.K., pp. 59-79.

- Goodchild, M.F., and S. Gopal. 1989. Accuracy of Spatial Databases. London: Taylor and Francis.

- Guptill, S.C., and J.L. Morrison (eds.). 1995. Elements of spatial data quality. Elsevier Science. Oxford, U.K.

- Mark, D.M., and F. Csillag. 1989. The nature of boundaries on 'area-class' maps. Cartographica 26(1): 65-78.

- McDonnell, R. and K. Kemp. 1995. International GIS Dictionary. Wiley New York. 111pp.

- Merchant, J., et. al. 1993. Validation of continental-scale land cover databases developed from AVHRR data. Presented at Pecora 12, Sioux Falls, South Dakota, August 24 - 26.

- Slonecker, E.T., and N. Tosta. 1991. National map accuracy standards: out of sync, out

of time. Geo Info Systems. 2(1):20, 24-26.

- Stoms, D.M., F.W. Davis, and C.B. Cogan. 1992. Sensitivity of wildlife habitat models to uncertainties in GIS data. Photogrammetric Engineering and Remote Sensing, 58: 843-850.

- Story, M., and R.G. Congalton. 1986. Accuracy assessment: a user's perspective. Photogrammetric Engineering and Remote Sensing, 52(3): 397-399.

- Taylor, J.R. 1982. An introduction to error analysis: the study of uncertainties in physical measurements. Oxford University Press.

Created: May 14, 1997. Last updated: October 5, 1998.

Arc/Info example:

The following commands illustrate database queries of info tables associated with arc/info map coverages. Similar queries could also be done from arcplot, or in arcview. The INFO example used here was selected to show the fundamental steps in logical queries, though this method lacks the graphics available in arcplot or arcview.

```
Arc: listcoverages                                        list the available
coverages in the current workspace

Arc: list usgs24k.pat                                     look at the database
files for each polygon

             1

   AREA                      = ************

   PERIMETER                 =  4825542.500

   USGS24K#                  =     1

   USGS24K-ID                =     0

   USGS#                     =     0

   QUADNAME                  =

             2

   AREA                      = ************

   PERIMETER                 =    48446.836

   USGS24K#                  =     2

   USGS24K-ID                =  7699

   USGS#                     =     1

   QUADNAME                  = O'BRIEN (OR)


Arc: INFO                                                 start the info
database program

ENTER USER NAME> ARC                              use caps for all info
commands

ENTER COMMAND> SELECT  USGS24K.PAT      select the coverage polygon attribute
 table

    2850 RECORD(S) SELECTED

ENTER COMMAND >LIST                               look at the database files
for each polygon

               1                                          note that record #1
is the outside (neg.) area

 AREA            =-4.38082E+11

 PERIMETER       = 4825543.000

 USGS24K#        =    1

 USGS24K-ID      =    0
```

```
 USGS#           =    0
 QUADNAME        =
                2
 AREA            =1.436519E+08
 PERIMETER       =  48,446.840
 USGS24K#        =    2
 USGS24K-ID      =7,699
 USGS#           =    1
 QUADNAME        =O'BRIEN (OR)

ENTER COMMAND >RESELECT QUADNAME  =  ' '        look for polygons with no
quadname
       1 RECORD(S) SELECTED                     note that this is normal for
record one


ENTER COMMAND >ASELECT                          select all the records in the
 dataset
    2850 RECORD(S) SELECTED

ENTER COMMAND >RESELECT AREA > 1.63E+08 look for polygons with large area
       0 RECORD(S) SELECTED
```