Title

Unit 45 - Accuracy of Spatial Databases

Authors

Unit 45, CC in GIS
Chrisman, Nicholas R.
McGranaghan, Matt

Publication Date

1990

Peer reviewed

UNIT 45 - ACCURACY OF SPATIAL DATABASES

UNIT 45 - ACCURACY OF SPATIAL DATABASES

Compiled with assistance from Nicholas R. Chrisman, University

of Washington and Matt McGranaghan, University of Hawaii

UNIT 45 - ACCURACY OF SPATIAL DATABASES

Compiled with assistance from Nicholas R. Chrisman, University

of Washington and Matt McGranaghan, University of Hawaii

A. INTRODUCTION

- the course thus far has looked at technical issues in:
    - georeferencing, i.e. describing locations
    - data structures - how to create digital representations of spatial data
    - algorithms - how to process these digital representations to generate useful results

- among other technical issues in GIS, accuracy is perhaps the most important - it covers concerns for data quality, error, uncertainty, scale, resolution and precision in spatial data and affects the ways in which it can be used and interpreted

- all spatial data is inaccurate to some degree but it is generally represented in the computer to high precision

- need to consider:
    - how well do these digital structures represent the real world?
    - how well do algorithms compute the true values of products?

B. DEFINITIONS

Accuracy

- defined as the closeness of results, computations or estimates to true values (or values accepted to be true)
    - since spatial data is usually a generalization of the real world, it is often difficult to identify a true value, and we work instead with values which are accepted to be true
        - e.g., in measuring the accuracy of a contour in a digital database, we compare to the contour as drawn on the source map, since the contour does not exist as a real line on the surface of the earth

- the accuracy of the database may have little relationship to the accuracy of products computed from the database
    - e.g. the accuracy of a slope, aspect or watershed computed from a DEM is not easily related to the accuracy of the elevations in the DEM itself

Precision

- defined as the number of decimal places or significant digits in a measurement
    - precision is not the same as accuracy - a large number of significant digits doesn't necessarily indicate that the measurement is accurate
    - a GIS works at high precision, mostly much higher than the accuracy of the data itself

- since all spatial data are of limited accuracy, inaccurate to some degree, the important questions are:
    - how to measure accuracy
    - how to track the way errors are propagated through GIS operations
    - how to ensure that users don't ascribe greater accuracy to data than it deserves

Components of data quality

- recently a National Standard for Digital Cartographic Data (see reference) was developed by a coordinated national effort in the US
    - this is a standard model to be used for describing digital data accuracy
    - similar standards are being adopted in other countries

- this standard identifies several components of data quality:
    - positional accuracy
    - attribute accuracy
    - logical consistency
    - completeness
    - lineage

- each of these will now be examined

B. POSITIONAL ACCURACY

- defined as the closeness of locational information (usually coordinates) to the true position

- conventionally, maps are accurate to roughly one line width or 0.5 mm
    - equivalent to 12 m on 1:24,000, or 125 m on 1:250,000 maps

- within a database, a typical UTM coordinate pair might be:

  Easting 579124.349m Northing 5194732.247m

    - if the database was digitized from a 1:24,000 sheet, the last four digits in each coordinate (units, tenths, hundredths and thousandths) would be spurious

How to test positional accuracy?

- use an independent source of higher accuracy
    - find a larger scale map
    - use the Global Positioning System (GPS)
    - use raw survey data

- use internal evidence
    - unclosed polygons, lines which overshoot or undershoot junctions, are indications of inaccuracy - the sizes of gaps, overshoots and undershoots may be used as a measure of positional accuracy

- compute accuracy from knowledge of the errors introduced by different sources, e.g
  - 1 mm in source document
  - 0.5 mm in map registration for digitizing
  - 0.2 mm in digitizing
  - if sources combine independently, we can get an estimate of overall accuracy by summing the squares of each component and taking the square root of the sum:

    (12 + 0.52 + 0.22)0.5 = 1.14 mm

C. ATTRIBUTE ACCURACY

- defined as the closeness of attribute values to their true value

- note that while location does not change with time, attributes often do

- attribute accuracy must be analyzed in different ways depending on the nature of the data

- for continuous attributes (surfaces) such as on a DEM or TIN:
  - accuracy is expressed as measurement error
    - e.g. elevation accurate to 1 m

- for categorical attributes such as classified polygons:
  - are the categories appropriate, sufficiently detailed and defined?
  - gross errors, such as a polygon classified as A when it should have been B, are simple but unlikely

    - e.g. land use is shopping center instead of golf course
  - more likely the polygon will be heterogeneous:
    - e.g. vegetation zones where the area may be 70% A and 30% B
  - worse, A and B may not be well-defined, may not be able to identify the class clearly as A or B
    - e.g. soils classifications are typically fuzzy
  - at the center of the polygon, may be confident that the class is A, but more like B at the edges

How to test attribute accuracy?

- prepare a misclassification matrix as follows:
  - take a number of randomly chosen points
  - determine the class according to the database
  - then determine the class in the field by ground check
  - complete the matrix:

    Class in Class on ground database A B C D A . . . . B . . . . C . . . . D . . . .

- ideally, want all points to lie on the diagonal of the matrix - this indicates that the same class was observed on the ground as is recorded in the database

- an error of omission occurs when a point's class on the ground is incorrectly recorded in the database
    - the number of class B points incorrectly recorded is the sum of column B row A, column B row C and column B row D, i.e. the number of points that are B on the ground but something else in the database
    - that is, the column sum less the diagonal cell

- an error of comission occurs when the class recorded in the database does not exist on the ground
    - e.g. the number of errors of comission for class A is the sum of row A column B, row A column C, row A column D, i.e. the points falsely recorded as A in the database
    - that is, the row sum less the diagonal cell

How to summarize the matrix?

- the percent of cases correctly classified is often used
    - this is the percent of cases located in the diagonal cells of the matrix
    - however, even in the worst case we would expect some cases in the diagonal cells by chance

- an index kappa (Cohen's kappa) adjusts for this by subtracting the number expected by chance
    - the number expected by chance in each diagonal cell is found by multiplying the appropriate row and column totals and dividing by the total number of cases

  overhead - Calculating Kappa

- then:

  $k = (d-q)/(N-q)$

  where d is the number of cases in diagonal cells q is the number of cases expected in diagonal cells by chance N is the total number of cases

- kappa is 1 for perfectly accurate data (all N cases on the diagonal), zero for accuracy no better than chance

- compare a map with a few large polygons to one with a large number of smaller polygons
    - is it easier to get a high kappa in the first case?
    - if so, is there a way of adjusting kappa to account for this difference?

- we expect attribute accuracy to vary over the map, so it would be useful to have an indication of the spatial variation in misclassification probability, not just a summary statistic

- the remaining aspects of data quality apply to the database as a whole, rather than to the objects, attributes or coordinates within it

E. LOGICAL CONSISTENCY

- refers to the internal consistency of the data structure, particularly applies to topological consistency
    - is the database consistent with its definitions?
        - if there are polygons, do they close?
        - is there exactly one label within each polygon?
        - are there nodes wherever arcs cross, or do arcs sometimes cross without forming nodes?

F. COMPLETENESS

- concerns the degree to which the data exhausts the universe of possible items
    - are all possible objects included within the database?
    - affected by rules of selection, generalization and scale

G. LINEAGE

- a record of the data sources and of the operations which created the database
    - how was it digitized, from what documents?
    - when was the data collected?
    - what agency collected the data?
    - what steps were used to process the data?
        - precision of computational results
- is often a useful indicator of accuracy

H. ERROR IN DATABASE CREATION

- error is introduced at almost every step of database creation
    - what are these steps, and what kinds of error are introduced?

Positional measurement error

Geodetic Control and GPS

- the most accurate basis of absolute positional data is the geodetic control network, a series of points whose positions are known with high precision
    - however, it is often difficult to tie a dataset to one of these high quality monuments
- global positioning systems is a powerful way of augmenting the geodetic network

Aerial Photography and Satellite Imagery

- most positional data is derived from air photos
    - here accuracy depends on the establishment of good control points
- data from remote sensing is more difficult to position accurately because of the size of each pixel

Text Descriptions

- some positional data comes from text descriptions
  - old surveys tied in to marks on trees
  - boundary follows watershed, or midline of river
- this type of source is often of very poor positional accuracy

Digitizing

- digitizers encode manuscript lines as sets of x-y coordinate pairs
  - see Units 7 and 13 for introductions to digitizing

- resolution of coordinate data is dependent on mode of digitizing:
- point-mode
  - digitizing operator specifically selects and encodes those points deemed "critical" to represent the geomorphology of the line or politically-significant coordinate pairs
  - requires intelligence, knowledge about the line representation that will be needed
- stream-mode
  - digitizing device automatically selects points on a distance or time parameter
  - generally, an unnecessary high density of coordinate pairs is selected.

- two types of errors normally occur in stream-mode digitizing:
- physiological errors are caused by involuntary muscular spasms that tend to parallel the longitudinal axis of the centerline
  - these errors are caused by agitations as the operator's hand twitches and jerks when digitizing
  - three specific types may be identified: spikes, switchbacks and polygonal knots (loops)

diagram

  - these are fairly simple to remove automatically
  - software has been developed to clean the initial digital data of duplicate coordinate pairs and simple physiological errors
  - a related problem in point mode digitizing is duplicate coordinate pairs which occur when the button is hit twice
- psychological errors are caused by psychomotor problems in line-following
  - the digitizing operator either cannot see the line or cannot properly move the crosshairs along the line
  - results in the diagonal line being displaced laterally from the intended position

  - may also involve misinterpretation, too much generalization
  - these are not easy to remove automatically

- in spite of the above, digitizing itself is not a major source of positional error

        it is not difficult for a digitizer operator to follow a line to an accuracy equal to the line's width
        - typical error in 0.5 mm range

    - a common test of digitizing accuracy is to compare the original line with its digitized and plotted version, and to see if daylight can be seen between the two
    - errors in registration and control points affect the entire dataset
    - errors are also introduced because of poor stability of base material
        - paper can shrink and stretch significantly (as much as 3%) with change in humidity

Coordinate transformation

- coordinate transformation introduces error, particularly if the projection of the original document is unknown, or if the source map has poor horizontal control

Attribute errors

- attributes are usually obtained through a combination of field collection and interpretation

- categories used in interpretation may not be easy to check in the field
    - concepts of "diversity" and "old growth" used in current forest management practice are highly subjective

- attributes obtained from air photo interpretation or classified satellite images may have high error rates

- for social data, the major source of inaccuracy is undercounting
    - e.g. in the Census undercount rates can be very high (>10%) in some areas and in some social groups

Compilation errors

- common practices in map compilation introduce further inaccuracies:
    - generalization
    - aggregation
    - line smoothing

    - separation of features
        - e.g. railroad moved on map so as not to overlap adjacent road

- however, many of these may also be seen as improving the usefulness and meaning of the data

Processing errors

- processing of data produces error
    - misuse of logic
    - generalization and problems of interpretation

  - mathematical errors
  - accuracy lost due to low precision computations
  - rasterization of vector data
    - e.g., true line position is somewhere in the cell
    - boundary cells may actually contain parts of all adjacent cells

I. DATA QUALITY REPORTS

- because there are so many, diverse sources of error it is probably not possible to measure the error introduced at each step independently - the strategy of combining errors arithmetically probably won't work

USGS

- require that no more the 10% of the points tested be in error by more than 1/30 inch, measured at publication scale (scale >1:20,000)

- question are "How far out are the 10%?" "Where are the 10%?"
  - e.g. in a particularly bad case, all of the 10% might be accounted for by one boundary line which is out by several inches

British Ordnance Survey

- carry out an ongoing accuracy assessment and re-survey

- to verify a survey, a large number of points (typically n = 150 to 500 of a single type) are used to calculate:
  - root mean total square displacement:

    e = / (S (xi2) / n) where xi is displacement at each point i

  - systematic error:

    s = S (xi)/n

  - standard error:

    se = / (e2 - s2)

  - if the error is "excessive", then the survey is carefully reviewed
  - see Merchant (1987) for an implementation

US National standards

- National Map Accuracy Standards from the Bureau of the Budget, 1947
  - not completed

- current standards developed by the National Committee for Digital Cartographic Data Standards chaired by Hal Moellering

- purpose: to set standards for compatibility of:
  - definitions of cartographic objects
  - interchange formats
  - DATA QUALITY documentation

- dates: 1982 Jan NCDCDS formed 1985 Jan Interim Proposed Standard 1988 Jan Proposed Standard 1988 Testing in the field

  handout - Interim Proposed Standard for Digital Cartographic Data Quality (2 pages)

REFERENCES

Bureau of the Budget, 1947. National Map Accuracy Standards, Washington DC, GPO, reprinted in M.M.Thompson, 1979, Maps for America, USGS, Reston VA, p 104.

Burrough, P.A., 1986. Principles of Geographical Information Systems for Land Resources Assessment, Clarendon Press, Oxford. See pp. 103-135.

DCDSTF, 1988. "The Proposed Standard for Digital Cartographic Data," The American Cartographer 15(1):entire issue.

Federal Geodetic Control Committee, 1974. Classification, Standards of Accuracy, and General Specifications of Geodetic Control Surveys, Washington DC, GPO, 1980-0- 333-276 (also NOAA--S/T 81-29).

Harley, J. B., 1975. Ordnance Survey Maps: A Descriptive Manual, Ordnance Survey, Southampton, England.

Merchant, D.C., 1987. "Spatial accuracy specification for large scale topographic maps," Photogrammetric Engineering and Remote Sensing 53:958-61. Reports a recent effort by ASPRS to revise the US National Map Accuracy Standard.

National Committee for Digital Cartographic Data Standards, Moellering, H., ed, 1985. Digital Cartographic Data Standards: An Interim Proposed Standard, Report #6.

DISCUSSION AND EXAM QUESTIONS

1. Explain the difference between accuracy and precision, and show how these ideas apply to GIS.

2. "In manual map analysis, precision and accuracy are similar, but in GIS processing, precision frequently exceeds the accuracy of the data". Discuss

3. Design an experiment to measure the accuracy achieved by an agency in its digitizing operations. How would you measure the accuracy with which lines are being digitized?

4. What is meant by data lineage, and why is it important in understanding the accuracy of

spatial databases?

Last Updated: August 30, 1997.