UNIT 46 - MANAGING ERROR

UNIT 46 - MANAGING ERROR

UNIT 46 - MANAGING ERROR

A. ERROR PROPAGATION

- in GIS applications we combine data from different sources, with different levels of accuracy
  - What impact does error in each data layer have on the final result?

Example application

Problem: find the best route for a power transmission corridor from a given origin point to a given destination point about 150 km away, across an area of the Midwest with comparatively high densities of agriculture and settlement

- the study area has been divided up into about 30,000 raster cells, each 500 m on a side
- have identified about 100 factors likely to influence the choice of route, including:

    - agricultural productivity (dollars per hectare)
    - settlement (presence or absence)
    - existing rights of way for power lines (presence or absence)

- the 100 factors have been combined, or cascaded, to a single measure of suitability on a scale of 0 through 6
    - the cascading rules group factors into composites such as "social impact", "agricultural impact" and then weight each group against the others

- the rules used in cascading include weighted addition:

  suitability = w1x1 + w2x2

  as well as simple conditions:

  suitability = 0 if settlement = "present"

  and reclassifications:

  suitability = 3 if x1 = A and x2 = d suitability = 4 if x1 = B and x2 = d

Error analysis

- the effects of cascading on error will be complex
    - do errors get worse, i.e. multiply?
    - do errors cancel out?
    - are errors in each layer independent or are they related?

- suppose two maps, each with percent correctly classified of 0.90 are overlaid
    - studies have shown that the accuracy of the resulting map (percent of points having both of the overlaid classes) is little better than 0.90x0.90=0.81
    - when many maps are overlaid the accuracy of the resulting composite can be very poor

- however we're more interested in the accuracy of the composite suitability index than in the overlaid attributes themselves

- for some types of operations the accuracy of suitability is determined by the accuracy of the least accurate layer
    - this is true if reclassification and the and operator are used extensively, or if simple conditions are used based on inaccurate layers

- in other cases the accuracy of the result is significantly better than the accuracy of the least accurate layer
    - this is true if weighted addition is used, or if reclassification uses the or operator
        - e.g. suitability = 4 if x1 = A or x2 = d

Sensitivity analysis

how to determine the impact of inaccuracy on the results?

- two types of answers are needed:
  - the impact of error on the suitability map
  - the impact of error on the best route
  - the answers will likely be very different
- it will also be useful to ask the question the other way:
  - what accuracy is needed in each layer in order to produce a required level of accuracy in the result?
- sensitivity is the response of the result (suitability, or the route location) to a unit change in one of the inputs
  - easy to see what a unit change means for agricultural productivity in dollars per acre, but what does it mean for vegetation class?
- sensitivity can be defined for: 1. the data inputs:
    - how much does result change when data input changes? 2. the weights
    - how much does result change when the weight given to a factor changes?
    - error in determining weights may be just as important as error in the database
- may be better to use full observed range to test sensitivity
  - i.e. response of the result to a change in one of the inputs from its minimum observed value to its maximum
  - e.g. suppose one layer is settlement (present/absent)
    - set the entire layer to settlement=present and recompute suitability and the best route
    - then set the entire layer to settlement=absent and recompute
    - the difference will be a measure of the sensitivity of the analysis to the settlement layer
  - layers which are important but nevertheless do not show geographical variation over the study area will not have high sensitivity in this definition
- this serves to point up the distinction between sensitivity in principle and in practice
  - a layer may be important in principle, but have no impact in this study area
    - e.g. in principle the agricultural productivity layer may be very important in the decision framework, but if all the land is equally productive, then it will not be important in practice
- in practice, only a few layers (out of our original 100) will have much impact on the final route
  - it is critical to know which these are in order to defend the methodology effectively (or to attack it!)
  - must examine both the decision rules and the value ranges to determine which layers have the highest impact in the suitability product
  - this information can be used in assessing the level of input accuracy that is needed

        - e.g. 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
    - compute the impact of values at each end of the uncertainty range and compare the results
    - provides a measure of the "confidence interval" of the results

- sensitivity may also refer 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?

B. ARTIFACTS OF ERROR

- artifacts are unwanted effects which result from using a high- precision GIS to process low-accuracy spatial data
    - usually result from positional errors, not attribute errors

Raster data

- since raster data has finite resolution, determined by pixel size
    - as long as pixel size is greater than the positional accuracy of the data, we have no risk of unwanted effects or artifacts

Vector data

- often have precision different than accuracy

- significant problems occur in two areas:
    - digitizing
    - polygon overlay

Digitizing artifacts

- a digitizer operator will not be able to close polygons exactly, or to form precise junctions between lines
    - a tolerance distance must be established, so that gaps and overshoots can be corrected (lines snapped together) as long as they fall within the tolerance distance

- most digitizer operators can work easily to a tolerance of 0.02 inches or 0.5 mm

- problems arise whenever the map has real detail at this resolution or finer
    - e.g. polygon with a narrow isthmus:

  diagram

    - e.g. two lines close together - which one to snap to?

diagram

- e.g. removing overshoot - must look back along line to form correct topology:

diagram

Strategies used to avoid problems:

- essentially, we try to find a balance between:

  1. asking the operator to resolve problems, which slows down the digitizing, and 2. having the system resolve problems, which requires good software and lots of CPU usage

- each system establishes its own ways of avoiding or reducing these problems
  - some are more successful than others

  1. require the user to enlarge the map photographically
     - increases the scale of the map while holding tolerance constant, so problem detail is now bigger than the tolerance
     - difficult or impossible to get error-free enlargement cheaply and easily

  2. require the user to digitize each arc separately
     - e.g. if the following is digitized as one arc then it there is no intersection

  diagram

     - program then only needs to check for snaps and overshoots at ends of arcs
     - tedious for the digitizer operator

  3. require the user to identify snap points
     - press a different digitizer button when a point needs to be snapped
     - wait for system response indicating successful snap

  diagram

  4. have the system check for snaps continuously during digitizing

     - requires fast, dedicated processor
     - computing load gets higher as database accumulates
     - requires continuous display of results
     - no good for imported datasets

  5. use rules to assist CPU in making decisions
     - e.g. two labels in a polygon indicates that it's really two polygons, not one with a narrow isthmus
     - might use expectations about polygon shape
     - puts heavy load on the processor

- the best current solutions use a combination of strategies 3 and 4

- it is almost always useful to keep track of digitizing by marking work done on a transparent overlay
    - a cursor in the form of a pen is a good practical solution

Polygon overlay artifacts

- covered algorithms for dealing with sliver polygons earlier

- another strategy for avoiding the sliver polygon problem is to allow objects to share primitives
    - this departs from the database model in which every set of polygons is thought of as a different layer
    - e.g. suppose a woodlot (polygon) shares part of its boundary with a road (line)
        - the shared part becomes a primitive object which is stored only once in the database, and shared by the two higher level features

- by using shared primitives, can avoid artifacts which might result when comparing or overlaying the two versions of the woodlot/road line, one belonging to the road object and one to the woodlot object

- to identify shared primitives during digitizing they must be on the same document
    - need an operation which allows two separate primitives to be identified as shared and replaced by one
    - need a converse operation to unshare a primitive if one version of the line must be moved and not the other

diagram

C. STORING ACCURACY INFORMATION

- how to store information on accuracy in a database?

Raster data

- uncertainty in each cell's attributes might be stored by giving each cell a set of probability attributes, one for each of the possible classes
    - in classified remote sensing images this information can come directly from the classification procedures

- uncertainty in elevation in a DEM is more likely constant over the raster and can be stored as part of the descriptive or metadata for the raster as a whole

- positional uncertainty is also likely constant for the raster
    - can be stored once for the whole map

Vector data

there are five potential levels for storage of uncertainty information in a vector database:

- map
- class of objects
- polygon
- arc
- point

Positional uncertainty

- positional accuracy at one level may not imply similar accuracy at other levels
  - positional accuracy about a point says little about the positional accuracy of an arc

diagram

  - similarly, positional accuracy at the polygon level may cause confusion along shared arcs

diagram

- for lines and polygons, accuracy can be stored as an attribute of:
  - arc (e.g. width of transition zone between two polygons)
  - class of objects (e.g. error in position of railroads)
  - map as a whole (e.g. all boundaries and lines on the map have been digitized to specified accuracy)
- for points, can be stored as an attribute of point, class or map

Attribute uncertainty

- uncertainty in an object's attributes can be stored as:
  - an attribute of the object (e.g. polygon is 90% A)
  - an attribute of the entire class of objects (e.g. soil type A has been correctly identified 90% of the time)

REFERENCES

Burrough, P.A., 1986. Principles of Geographical Information Systems for Land Resources Assessment. Clarendon, Oxford. Chapter 6 on error in GIS.

Chrisman, N.R., 1983. "The role of quality information in the long-term functioning of a geographic information system," Cartographica 21:79.

Goodchild, M.F. and S. Gopal, editors, 1989. The Accuracy of Spatial Databases, Taylor and Francis, Basingstoke, UK. Edited papers from a conference on error in spatial databases.

EXAM AND DISCUSSION QUESTIONS

1. Define the difference between sensitivity to error in principle and in practice.

2. Imagine that you represent a community trying to fight the proposed route of the powerline discussed in this unit. What arguments would you use to attack the power utility company's methods?

3. Compare the methods available in any digitizing system to which you have access, to those discussed in this unit. Does your system offer any significant advantages?

4. Some GIS processes can be very sensitive to small errors in data. Give examples of such processes, and discuss ways in which the effects of errors can be managed.

Last Updated: August 30, 1997.