UNIT 7 - DATA INPUT

UNIT 7 - DATA INPUT

Compiled with assistance from Jeffrey L. Star, University of California at Santa Barbara, and Holly Dickinson, SUNY Buffalo

- NOTES

This unit examines the common methods of data input. This may be a good time to take a field trip to a local GIS shop to show students the operation of these various devices. If you can't find local examples, the slide set contains some examples of the hardware items described.

UNIT 7 - DATA INPUT

Compiled with assistance from Jeffrey L. Star, University of California at Santa Barbara, and Holly Dickinson, SUNY Buffalo

A. INTRODUCTION

- need to have tools to transform spatial data of various types into digital format
- data input is a major bottleneck in application of GIS technology
  - costs of input often consume 80% or more of project costs
  - data input is labor intensive, tedious, error-prone
  - there is a danger that construction of the database may become an end in itself and the project may not move on to analysis of the data collected
  - essential to find ways to reduce costs, maximize accuracy
- need to automate the input process as much as possible, but:
  - automated input often creates bigger editing problems later
  - source documents (maps) may often have to be redrafted to meet rigid quality requirements of automated input
- because of the costs involved, much research has gone into devising better input methods - however, few reductions in cost have been realized
- sharing of digital data is one way around the input bottleneck
  - more and more spatial data is becoming available in digital form
- data input to a GIS involves encoding both the locational and attribute data
- the locational data is encoded as coordinates on a particular cartesian coordinate system
  - source maps may have different projections, scales
  - several stages of data transformation may be needed to bring all data to a common coordinate system
- attribute data is often obtained and stored in tables

Modes of data input

- keyboard entry for non-spatial attributes and occasionally locational data
- manual locating devices
  - user directly manipulates a device whose location is recognized by the computer

e.g. digitizing

- automated devices
  - automatically extract spatial data from maps and photography
  - e.g. scanning

- conversion directly from other digital sources

- voice input has been tried, particularly for controlling digitizer operations
  - not very successful - machine needs to be recalibrated for each operator, after coffee breaks, etc.

B. DIGITIZERS

- digitizers are the most common device for extracting spatial information from maps and photographs
  - the map, photo, or other document is placed on the flat surface of the digitizing tablet

Hardware

- the position of an indicator as it is moved over the surface of the digitizing tablet is detected by the computer and interpreted as pairs of x,y coordinates
  - the indicator may be a pen-like stylus or a cursor (a small flat plate the size of a hockey puck with a cross-hair)

- frequently, there are control buttons on the cursor which permit control of the system without having to turn attention from the digitizing tablet to a computer terminal

- digitizing tablets can be purchased in sizes from 25x25 cm to 200x150 cm, at approximate costs from $500 to $5,000

- early digitizers (ca. 1965) were backlit glass tables
  - a magnetic field generated by the cursor was tracked mechanically by an arm located behind the table
  - the arm's motion was encoded, coordinates computed and sent to a host processor
  - some early low-cost systems had mechanically linked cursors - the free-cursor digitizer was initially much more expensive

- the first solid-state systems used a spark generated by the cursor and detected by linear microphones
  - problems with errors generated by ambient noise

- contemporary tablets use a grid of wires embedded in the tablet to generate a magnetic field which is detected by the cursor
  - accuracies are typically better than 0.1 mm
  - this is better than the accuracy with which the average operator can position the cursor
  - functions for transforming coordinates are sometimes built into the tablet and

used to process data before it is sent to the host

The digitizing operation

- the map is affixed to a digitizing table
- three or more control points ("reference points", "tics", etc.) are digitized for each map sheet
  - these will be easily identified points (intersections of major streets, major peaks, points on coastline)
  - the coordinates of these points will be known in the coordinate system to be used in the final database, e.g. lat/long, State Plane Coordinates, military grid
  - the control points are used by the system to calculate the necessary mathematical transformations to convert all coordinates to the final system
  - the more control points, the better
- digitizing the map contents can be done in two different modes:
  - in point mode, the operator identifies the points to be captured explicitly by pressing a button
  - in stream mode points are captured at set time intervals (typically 10 per second) or on movement of the cursor by a fixed amount
- advantages and disadvantages:
  - in point mode the operator selects points subjectively
    - two point mode operators will not code a line in the same way
  - stream mode generates large numbers of points, many of which may be redundant
  - stream mode is more demanding on the user while point mode requires some judgement about how to represent the line
- most digitizing is currently done in point mode

Problems with digitizing maps

- arise since most maps were not drafted for the purpose of digitizing
  - paper maps are unstable: each time the map is removed from the digitizing table, the reference points must be re-entered when the map is affixed to the table again
  - if the map has stretched or shrunk in the interim, the newly digitized points will be slightly off in their location when compared to previously digitized points
  - errors occur on these maps, and these errors are entered into the GIS database as well
  - the level of error in the GIS database is directly related to the error level of the source maps
- maps are meant to display information, and do not always accurately record locational information
  - for example, when a railroad, stream and road all go through a narrow mountain pass, the pass may actually be depicted wider than its actual size to allow for the three symbols to be drafted in the pass

- discrepancies across map sheet boundaries can cause discrepancies in the total GIS database
  - e.g. roads or streams that do not meet exactly when two map sheets are placed next to each other

- user error causes overshoots, undershoots (gaps) and spikes at intersection of lines

  diagram

- user fatigue and boredom

- for a complete discussion on the manual digitizing process, see Marble et al, 1984

Editing errors from digitizing

- some errors can be corrected automatically
  - small gaps at line junctions
  - overshoots and sudden spikes in lines

- error rates depend on the complexity of the map, are high for small scale, complex maps

- these topics are explored in greater detail in later Units
  - Unit 13 looks at the process of editing digitized data
  - Units 45 and 46 discuss digitizing error

Digitizing costs

- a common rule of thumb in the industry is one digitized boundary per minute
  - e.g. it would take 99/60 = 1.65 hours to digitize the boundaries of the 99 counties of Iowa

C. SCANNERS

Video scanner

- essentially television cameras, with appropriate interface electronics to create a computer-readable dataset
  - available in either black and white or color
  - extremely fast (scan times of under 1 second)
  - relatively inexpensive ($500 - $10,000)

- produce a raster array of brightness (or color) values, which are then processed much like any other raster array
  - typical data arrays from video scanners are of the order of 250 to 1000 pixels on a side

- typically have poor geometrical and radiometrical characteristics, including various kinds of spatial distortions and uneven sensitivity to brightness across the scanned field
  - video scanners are difficult to use for map input because of problems with

distortion and interpretation of features

Electromechanical scanner

- unlike the video scanning systems, electromechanical systems are typically more expensive ($10,000 to 100,000) and slower, but can create better quality products
- one common class of scanners involves attaching the graphic to a drum
  - as the drum rotates about its axis, a scanner head containing a light source and photodetector reads the reflectivity of the target graphic, and digitizing this signal, creates a single column of pixels from the graphic
  - the scanner head moves along the axis of the drum to create the next column of pixels, and so on through the entire scan
  - compare the action of a lathe in a machine shop
- this controls distortion by bringing the single light source and detector to position on a regular grid of locations on the graphic
- systems may have a scan spot size of as little as 25 micrometers, and be able to scan graphics of the order of 1 meter on a side
- an alternative mechanism involves an array of photodetectors which extract data from several rows of the raster simultaneously
  - the detector moves across the document in a swath
  - when all the columns have been scanned, the detector moves to a new swath of rows
- for an in-depth discussion scanning techniques, see Peuquet and Boyle (1984)

Requirements for scanning

- documents must be clean (no smudges or extra markings)
- lines should be at least 0.1 mm wide
- complex line work provides greater chance of error in scanning
- text may be accidently scanned as line features
- contour lines cannot be broken with text
- automatic feature recognition is not easy (two contour lines vs. road symbols)

  diagram
- special symbols (e.g. marsh symbols) must be recognized and dealt with
- if good source documents are available, scanning can be an efficient time saving mode of data input

D. CONVERSION FROM OTHER DIGITAL SOURCES

- involves transferring data from one system to another by means of a conversion program

- more and more data is becoming available in magnetic media
  - USGS digital cartographic data (DLGs - Digital Line Graphs)
  - digital elevation models (DEMs)
  - TIGER and other census related data
  - data from CAD/CAM systems (AutoCAD, DXF)
  - data from other GIS

- these data generally are supplied on digital tapes that must be read into the computer
  - however, CD-ROM is becoming increasingly popular for this purpose
    - provides better standards
    - CD-ROM hardware is much less expensive - CD-ROM drive $1000, tape drive $14,000

Automated Surveying

- directly determines the actual horizontal and vertical positions of objects

- two kinds of measurements are made: distance and direction
  - traditionally, distance measuring involved pacing, chains and tapes of various materials
  - direction measurements were made with transits and theodolites

- modern surveyors have a number of automated tools to make distance and direction measurements easier

- electronic systems measure distance using the time of travel of beams of light or radio waves
  - by measuring the round-trip time of travel, from the observing instrument to the object in question and back, we can use the relationship ($d = v \times t$) to determine the distance
  - an instrument based on timing the travel of a pulse of infrared light can measure distances on the order of 10 km with a standard deviation of +/- 15 mm

- the total station (cost about $30,000) captures distance and direction data in digital form
  - the data is downloaded to a host computer at the end of each session for direct input to GIS and other programs

Global Positioning System (GPS)

- a new tool for determining accurate positions on the surface of the earth

- computes positions from signals received from a series of satellites (NAVSTAR)
  - as of April, 1990 there are 20 in orbit, by 1991 there should be the full set of 24
  - are currently 7 active but eventually will be 21

- depends on precise information about the orbits of the satellites
- a radio receiver with appropriate electronics is connected to a small antenna, and depending on the method used, in one hour to less than 1 second, the system is able to determine its location in 3-D space
- developed and operated by the US armed forces, but access is generally available and civilian interest is high
- particularly valuable for establishing accurate positional control in remote areas
- current GPS receivers cost about $5,000 to $15,000 (mid 1990) but costs will decline rapidly
- railroad companies are using GPS to create the first accurate survey of the US rail network and to track train positions
- recently, the use of GPS has resulted in corrections to the elevations of many of the world's peaks, including Mont Blanc and K2
- current GPS positional accuracies are order 5 to 10 m with standard equipment and as small as 1 cm with "survey grade" receivers
    - accuracy will continue to improve as more satellites are placed in orbit and experts fine tune the software and hardware
- GPS accuracy is already as good as the largest scale base mapping available for the continental US

E. CRITERIA FOR CHOOSING MODES OF INPUT

- the type of data source
    - images favor scanning
    - maps can be scanned or digitized
- the database model of the GIS
    - scanning easier for raster, digitizing for vector
- the density of data
    - dense linework makes for difficult digitizing
- expected applications of the GIS implementation

F. RASTERIZATION AND VECTORIZATION

Rasterization of digitized data

- for some data, entry in vector form is more efficient, followed by conversion to raster
- we might digitize the county boundary in vector form by

    - mounting a map on a digitizing table
    - capturing the locations of points along the boundary
    - assuming that the points are connected by straight line segments

- this may produce an ASCII file of pairs of xy coordinates which must then be processed by the GIS, or the output of the digitizer may go directly into the GIS

- the vector representation of the boundary as points is then converted to a raster by an operation known as vector-raster conversion
    - the computer calculates which county each cell is in using the vector representation of the boundary and outputs a raster

- digitizing the boundary is much less work than cell by cell entry

- most raster GIS have functions such as vector-raster conversion to support vector entry
    - many support digitizing and editing of vector data

Vectorization of scanned images

- for many purposes it is necessary to extract features and objects from a scanned image
    - e.g. a road on the input document will have produced characteristic values in each of a band of pixels
    - if the scanner has pixels of 25 microns = 0.025 mm, a line of width 0.5 mm will create a band 20 pixels across
    - the vectorized version of the line will be a series of coordinate points joined by straight lines, representing the road as an object or feature instead of a collection of contiguous pixels

- successful vectorization requires a clean line scanned from media free of cluttering labels, coffee stains, dust etc.
    - to create a sufficiently clean line, it is often necessary to redraft input documents
        - e.g. the Canada Geographic Information System redrafted each of its approximately 10,000 input documents

- since the scanner can be color sensitive, vectorizing may be aided by the use of special inks for certain features

- although scanning is much less labor intensive, problems with vectorization lead to costs which are often as high as manual digitizing
    - two stages of error correction may be necessary: 1. edit the raster image prior to vectorization 2. edit the vectorized features

G. INTEGRATING DIFFERENT DATA SOURCES

Formats

- many different format standards exist for geographical data

- some of these have been established by public agencies

    - e.g. the USGS in cooperation with other federal agencies is developing SDTS (Standard Data Transfer Standard) for geographical data, will propose it as a national standard in 1990
    - e.g. the Defense Mapping Agency (DMA) has developed the DIGEST data transfer standard

- some have been defined by vendors
    - e.g. SIF (Standard Interchange Format) is an Intergraph standard for data transfer

- see Unit 69 for more on GIS standards

- a good GIS can accept and generate datasets in a wide range of standard formats

Projections

- there are many ways of representing the curved surface of the earth on a flat map
    - some of these map projections are very common, e.g. Mercator, Universal Transverse Mercator (UTM), Lambert Conformal Conic
    - each state has a standard SPC (State Plane Coordinate system) based on one or more projections
    - see Unit 27 for more on map projections

- a good GIS can convert data from one projection to another, or to latitude/longitude

- input derived from maps by scanning or digitizing retains the map's projection

- with data from different sources, a GIS database often contains information in more than one projection, and must use conversion routines if data are to be integrated or compared

Scale

- data may be input at a variety of scales

- although a GIS likely will not store the scale of the input document as an attribute of a dataset, scale is an important indicator of accuracy

- maps of the same area at different scales will often show the same features
    - e.g. features are generalized at smaller scales, enhanced in detail at larger scales

- variation in scales can be a major problem in integrating data
    - e.g. the scale of most input maps for a GIS project is 1:250,000 (topography, soils, land cover) but the only geological mapping available is 1:7,000,000
    - if integrated with the other layers, the user may believe the geological layer is equally accurate
    - in fact, it is so generalized as to be virtually useless

Resampling rasters

raster data from different sources may use different pixel sizes, orientations, positions, projections

- resampling is the process of interpolating information from one set of pixels to another
- resampling to larger pixels is comparatively safe, resampling to smaller pixels is very dangerous

REFERENCES

Burrough, P.A., 1986. Principles of Geographical Information Systems for Land Resources Assessment, Clarendon, Oxford. Chapter 4 reviews alternative methods of data input and editing for GIS.

Chrisman, N.R., 1978. "Efficient digitizing through the combination of appropriate hardware and software for error detection and editing," International Journal of Geographical Information Systems 1:265-77. Discusses ways of reducing the data input bottleneck.

Drummond, J., and M. Bosman, 1989. "A review of low-cost scanners," International Journal of Geographical Information Systems 3:83-97. A good review of current scanning technology.

Ehlers, M., G. Edwards and Y. Bedard, 1989. "Integration of remote sensing with GIS: a necessary evolution," Photogrammetric Engineering and Remote Sensing 55(11):1619-27. A recent review of the relationship between the two technologies.

Goodchild, M.F. and B.R. Rizzo, 1987. "Performance evaluation and work-load estimation for geographic information systems," International Journal of Geographical Information Systems 1:67-76. Statistical analysis of costs of scanning.

Lai, Poh-Chin, 1988. "Resource use in manual digitizing. A case study of the Patuxent basin geographical information system database," International Journal of Geographical Information Systems 2(4):329-46. A detailed analysis of the costs of building a practical database.

Marble, D.F., J.P. Lauzon, and M. McGranaghan, 1984. "Development of a Conceptual Model of the Manual Digitizing Process," Proceedings of the International Symposium on Spatial Data Handling, Volume 1, August 20- 24, 1984, Zurich Switzerland, Symposium Secretariat, Department of Geography, University of Zurich-Irchel, 8057 Zurich, Switzerland. Conceptual discussion of the digitizing process.

Peuquet, D. J., 1981. "Cartographic data, part I: the raster-to-vector process," Cartographica 18:34-48.

Peuquet, D. J., 1981. "An examination of techniques for reformatting digital cartographic data, part II: the vector-to-raster process," Cartographica 18:21-33.

Peuquet, D. J., and A. R. Boyle, 1984. Raster Scanning, Processing and Plotting of Cartographic Documents, SPAD Systems, Ltd., P.O. Box 571, Williamsville, New York,

14221, U.S.A. A comprehensive discussion of scanning technology.

Tomlinson, R.F., H.W. Calkins and D.F. Marble, 1976. Computer Handling of Geographical Data, UNESCO Press, Paris. Comparison of input methods and costs of 5 GISs.

DISCUSSION AND EXAM QUESTIONS

1. In his book Computers and the Representation of Geographical Data (Wiley, New York, 1987), E.E. Shiryaev argues that maps must be redesigned to be equally readable by humans and computer scanners, and that this would ultimately make scanning much more cost-effective than digitizing. How might this be done, and what advantages would it have?

2. The cost of digitizing has remained remarkably constant over the past 20 years despite dramatic reductions in computer hardware and software cost. Why is this, and what impact has it had on GIS? Do you predict any change in this situation in the future?

3. "Digitizing is a suitable activity for convicted criminals." Discuss.

4. As manager of a GIS operation, you have the task of laying out rules which your staff must follow in digitizing

complex geographical lines. What instructions would you give them to ensure a reasonable level of accuracy? Assume they will be using point mode digitizing, and that points will be connected by straight lines for analysis and output.

5. What type of documents are best suited for automatic scanning?

6. After reading the article by Marble, Lauzon and McGranaghan on the conceptual model of digitizing, describe and explain the importance of map pre-processing.

Last Updated: August 30, 1997.