Data Input

presented by:

Tim Haithcoat

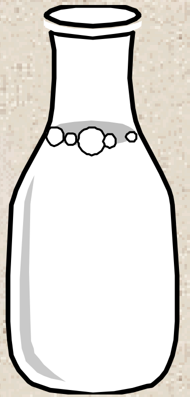
University of Missouri
Columbia

compiled with materials from:

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

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 cost
  - 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



Introduction ~ Continued

- 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 cots 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

Introduction ~ Continued

- 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 & occasionally locational data

Manual locating devices: user directly manipulates a device whose location is recognized by the computer (i.e., digitizing)

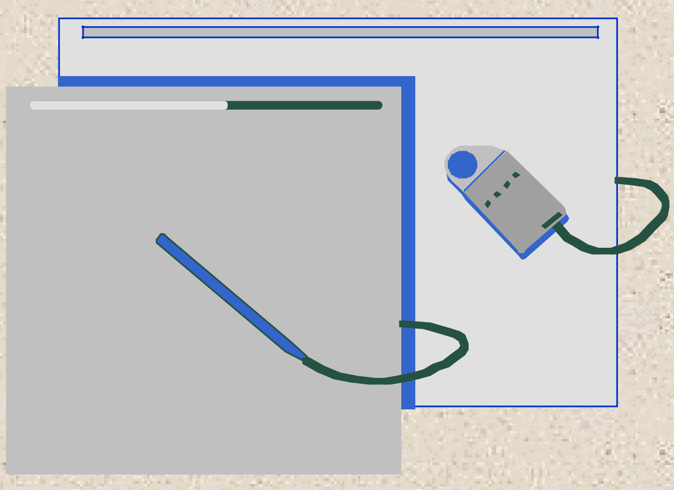
Conversion: directly from other digital sources

Automated devices: automatically extract spatial data from maps and photography (i.e., scanning)

Voice Input: has been tried, particularly for controlling digitizer operations but not very successful - machine has to be recalibrated for each user, after coffee breaks, etc.

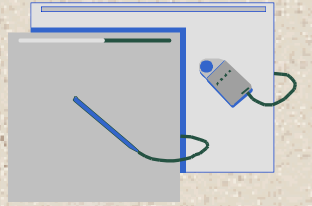
Digitizers

- 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

Hardware ~ Continued

- Digitizing tablets can be purchased in sizes from 25x25 cm to 200x150 cm, at approximate costs from $500 to $5,000
  - http://www.calcomp.com/p_tablets.htm
- 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 set to a host processor
  - Some early low-cost systems had mechanically linked cursors - the free-cursor digitizer was initially much more expensive

Hardware ~ Continued

- 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

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 (i.e., 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 Operation ~ Continued

- Digitizing the map contents can be done in 2 different modes:

Point mode:
operator identifies the points to be captured explicitly by pressing a button

Stream mode:
points are captured at set time intervals (typically 10 per second) or on movement of the cursor by a fixed amount

Digitizing Operations ~ Continued

- Advantages and disadvantages
  - In point mode the operator selects points subjectively (2 point 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 judgment 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 form 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

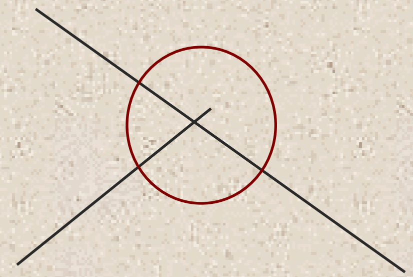
Problems with Digitizing Maps ~ Continued

- 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

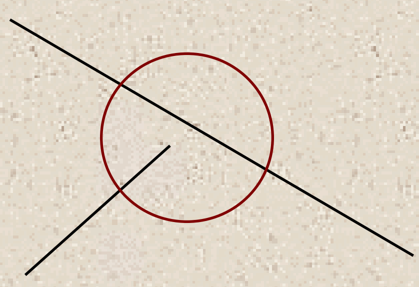
- Edgematching: discrepancies across map sheet boundaries can cause discrepancies in the total GIS database
  - For example, roads or streams that do not meet exactly when two map sheets are placed next to each other

Problems with Digitizing Maps ~ Continued

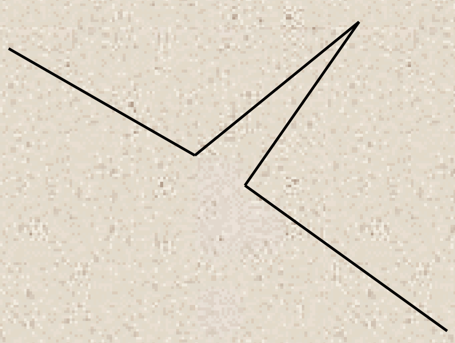
- User fatigue and boredom
- User error causes overshoots, undershoots (gaps) and spikes at intersection of lines



Overshoots



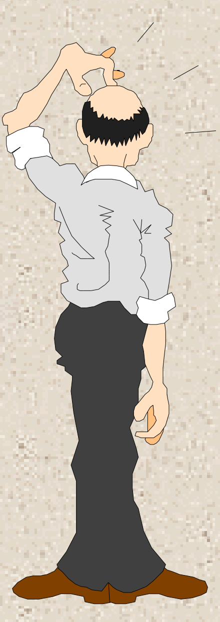
Undershoots



Spikes

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



Digitizing Costs

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

Video Scanner

- Essentially television cameras, with appropriate interface electronics to create a computer-readable dataset
  - Available in either b/w 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 & 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

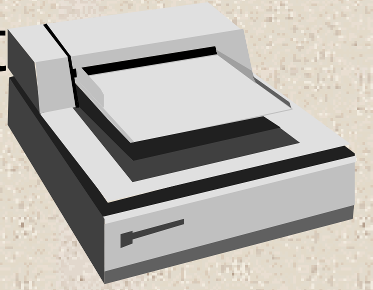
- Unlike video scanning systems, electromechanical systems are typically more expensive ($10,000 to $100,000) and slower, but can crate better quality products
- One common class of scanners involves attaching the graphic to a drum
  - As the drum rotates about its axis, a scanner head containing a light source and photo-detector 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

Electromechanical Scanner ~ Continued

- This controls distortion by brining 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 a s 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 photo-detectors 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

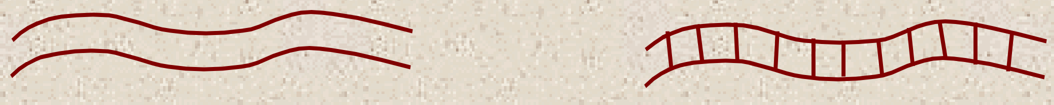
Requirements for Scanning

- 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 accidentally scanned as line features
- Contour lines cannot be broken with text

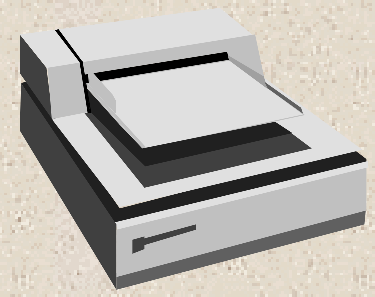


Requirements for Scanning ~ Continued

- Automatic feature recognition is not easy (two contour lines vs. road symbols)



- Special symbols (I.e., 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



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(Digital Exchange File))
  - 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 $1,000, tape drive $14,000

Automated Surveying

- Allows you to create a DIG File while on survey and makes a coordinate file
- Directly determines the actual horizontal and vertical positions of objects
- Two kinds of measurements are made: distance & direction
  - Traditionally, distance measuring involved pacing, chains and tapes of various materials
  - Direction measurements were made with transits & theodolites
- Modern surveyors have a number of automated tools to make distance & direction measurements easier

Automated Surveying ~ Continued

- Electronic systems measure distance using the time of travel of beams of light or radio waves
  - By measuring the round-trip time to travel, from the observing instrument to the object in question and back, we can use the relationship (d= v x 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
- A 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 of GIS & other programs

Global Positioning System (GPS)

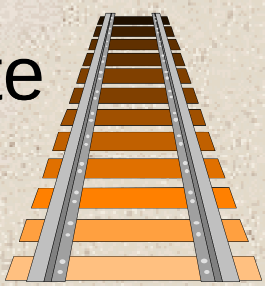
- A tool for determining accurate positions on the surface of the earth
- Computes positions from signals received from a series of satellites (NAVSTAR)
  - As of ____ there are ___ in orbit,
  - Are currently ____ active, but eventually will be __
- Depends on precise information about the orbits of the satellites
- GPS accuracy is already as good as the largest scale base mapping available for the continental US

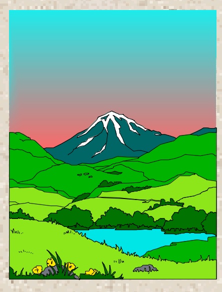
GPS ~ Continued

- A radio receiver with appropriate electronics is connected to a small antenna and depending on the method used, in one hours or less than 1 second, the system is able to determine its location in 3D 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 $__________

GPS ~ Continued

- Railroad companies are using GPS to create the first accurate survey of the US rail network and to track train positions



-  The use of GPS has resulted in corrections to the elevations of many of the world's peaks, including Mount Blanc and K2
- Current GPS positional accuracies are order 5-10 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 & experts fine tune the software & hardware

Criteria for Choosing Modes of Input

Type of data source:

~ images favor scanning

~ maps can be scanned or digitized

Database model of the GIS

~ scanning easier for raster,

~ digitizing for vector

Density of data:

~ dense linework makes for difficult digitizing

~ *example:* Mt. Everest 20 ft elevation contours, too dense to scan

Expected applications of the GIS implementation

Rasterization of Digitized Data

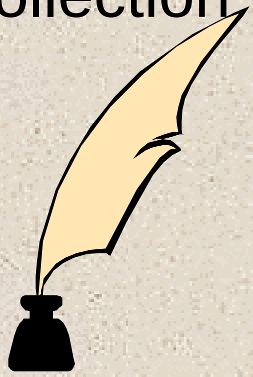
- For some data, entry in vector form is more efficient, followed by conversion to raster
- We might digitize the county boundary in vector form by:
  - Mounting a map on a digitizing table
  - Capturing the locations of points along the boundary
  - Assuming that the points are connected by straight 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 digitizers may go directly into the GIS

Rasterization of Digitized Data ~ Continued

- 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
  - i.e. 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 and 20 pixels across
  - The vectorization 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
- Since the scanner can be color sensitive, vectorizing may be aided by the use of special inks for certain features



Vectorization of Scanned Images (continued)

- 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
    - Ex: the Canada Geographic Information System redrafted each of its approximately 10,000 input documents
- 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:
    - Edit the raster image prior to vectorization
    - Edit the vectorized features

Integrating Different Data Sources
FORMATS

- Many different format standards exist for geographical data
- Some of these have been established by public agencies
  - ex: the USGS in cooperation with other federal agencies has developed an SDTS (Standard Data Transfer Standard) for geographical data, it became a national standard in _____
  - ex: The Defense Mapping Agency (DMA) has developed the DIGEST data transfer standard
- Some have been defined by vendors
  - ex:SIF (Standard Interchange Format) is an Intergraph standard for data transfer
- A good GIS can accept & generate datasets in a wide range of standard formats

- There are many ways of representing the curved surface of the earth of a flat map
  - Some of these map projections are very common (Mercator, Universal Transverse Mercator (UTM), Lambert Conformal Conic)
  - Each state has a standard SPC (State Plane Coordinate system) based on one or more 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, & must use conversion routines if data are to be integrated or compared

Integrating Different Data Sources
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
  - Features are generalized at smaller scales, enhanced in detail at larger scales
- Variation in scales can be a major problem in integrating data
  - The scale of most input maps for a GIS project is 1:250,000 (topography, soils, landcover) 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

Integrating Different Data Sources

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