Maps as Numbers

Getting Started with GIS

Chapter 3

Dursun Zafer ªeker

Chapter 3: Maps as Numbers

- 3.1 Representing Maps as Numbers
- 3.2 Structuring Attributes
- 3.3 Structuring Maps
- 3.4 Why Topology Matters
- 3.5 Formats for GIS Data
- 3.6 Exchanging Data

Representing Maps as Numbers

- Aim: The organisation of the map into digits how we capture, store, and use the map data in a GIS.
- The conversion of a visual or printed map to a set of digits can be done in many ways

Maps as Numbers

- GIS requires that both data and maps be represented as numbers.
- The GIS places data into the computer's memory in a physical data structure (i.e. files and directories).
- Files can be written in binary or as ASCII text.
- Binary is faster to read and smaller, ASCII can be read by humans and edited but uses more space.

the data are stored in a physical structure

- Ps: is not only how computer memory such as disk and RAM is used, but also how the files and directories store and access the map and attribute information.

the file encoded into *(1)*

- **binary digits or bits**

a number in base 10, decimal,
can be converted into base 2, binary,

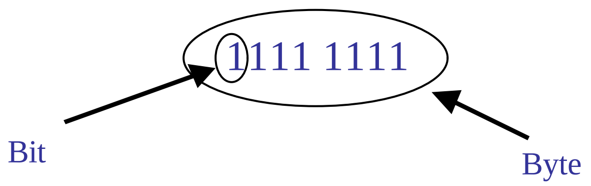
base 16, hexadecimal
1 byte goes from hexadecimal 00 to FF
exp: A=41, a=61 **(byte as hexadecimal)**

8bits = 1 byte

0 - 255
0000 0000-1111 1111

on/off 1/0

0,1,2,3,4,5,6,7,8,9,A,B,C,D, E,F.



Byte = 8 Bit

Example

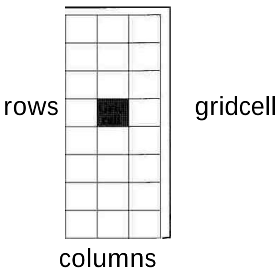
1010 1010 = $(1x2^7)+(0x2^6)+(1x2^5)+(0x2^4)+(1x2^3)+(0x2^2)+(1x2^1)+(0x2^0)$

= 128+32+8+2 = 170

The Data Model

- A logical data model is how data are organized for use by the GIS.
- GISs have traditionally used either raster or vector for maps.
- data model for attribute data: flat files

Raster data disadvantages



- cell size determines resolution of data
- rows and colums extend as often rectangular
- Imperfect fit for features shape
- hold one feature
- big enough cell for largest value of attribute or index

Raster data advantages

- easy to understand
- capability retrieval and analysis
- easy to draw on the screen

ITU Photogrammetry Division

the vector

- a set of points
- a list of coordinates
- line direction
- area enclosed by a surrounding ring of lines
    - Using vectors, we can draw an outline map with only a few thousand points, far fewer than the number of grid cells that would be required.

ITU Photogrammetry Division

Vector data advantages

- Accuracy follow features very closely
- efficient at storing features
- very suitable for plotting devices
- to store information about connectivity to other features (topology)

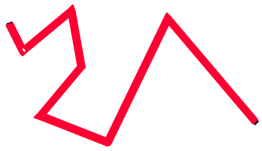
Vector data disadvantages

- not very good at representing continuous field variables
- not very good filling areas with shades or colour.

Rasters and vectors can be flat files

... if they are simple

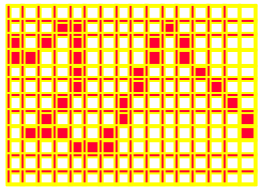
Vector-based line



Flat File

```
4753456 623412
4753436 623424
4753462 623478
4753432 623482
4753405 623429
4753401 623508
4753462 623555
4753398 623634
```

Raster-based line



Flat File

```
0000000000000000
0001100000100000
1010100001010000
1100100001010000
0000100010001000
0000100010000100
0001000100000010
0010000100000001
0111001000000001
0000111000000000
0000000000000000
```


Features and Maps

- A GIS map is a scaled-down digital representation of point, line, area, and volume features.
- While most GIS systems can handle raster and vector, only one is used for the internal organization of spatial data.

Attribute data

- Attribute data are stored logically in flat files.
- A flat file is a matrix of numbers and values stored in rows and columns, like a spreadsheet.
- Both logical and physical data models have evolved over time.
- DBMSs use many different methods to store and manage flat files in physical files.

A flat file

- numbers are stored in tables or in a spreadsheet
- rows for records and columns for attributes
- Both vectors and rasters can be stored in flat files, if they are simple

Flat file for the raster grids

- to store values in the cells of the table
- link the data in the flat file to the data in the map
- store index numbers in the grid and any number of attributes for the index numbers in the flat file

exp: land-use map 1:forest, 2:farmland, 3:urban

Flat file for the vector data

- Point data as coordinates
- Lines and areas variable numbers of points

polygon or area file

polygon attribute table file and a file of arcs by polygon

STRUCTURING ATTRIBUTES

- A flat file is a table (like matrix with rows and columns)

  Columns store attributes, and rows store records.
- information is stored in each attribute as text or numbers
- for each attribute values (records) can written by ASCII codes (as database terms: field)
- each record start from a new line like matrix

STRUCTURING ATTRIBUTES

sort of the data

search line by line until we found the correct one

encode the numbers in binary or sort them in the file so that the most commonly referenced records were first in the file.

DBMSs: **database management systems**

STRUCTURING ATTRIBUTES

- the *database dictionary:*attributes with all their characteristics must be written into the separate file

At the most basic it consists of a file.

At the most complex, it might be several files in a directory.

So the attribute database part of a GIS is fairly simple.

- Maps have at least two dimensions (x,y, lat., lon.)
- features can be points, lines, area, or even volumes

Point features: *x* and *y* can be stored just as regular attributes in a standard database

Line and area features: different shapes and sizes, it is not fit easily into the attribute database

Two basic models for map data, the vector and the raster

A raster data model uses a grid.

- One grid cell is one unit or holds one attribute.
- Every cell has a value, even if it is “missing.”
- A cell can hold a number or an index value standing for an attribute.
- A cell has a resolution, given as the cell size in ground units.

Raster Data Structures

- The data form an array or matrix of rows and columns.
- Each pixel or grid cell contains either a data value for an attribute, or an index number that points to a reference in the attribute database.
- An operation such as comparing a grid cell with its neighbours can be performed by looking at the values in the next and preceding row and column of the grid cells in question.
- the TIN in vector data structures fit easily into the raster. (RS & scanning data)

Generic structure for a grid

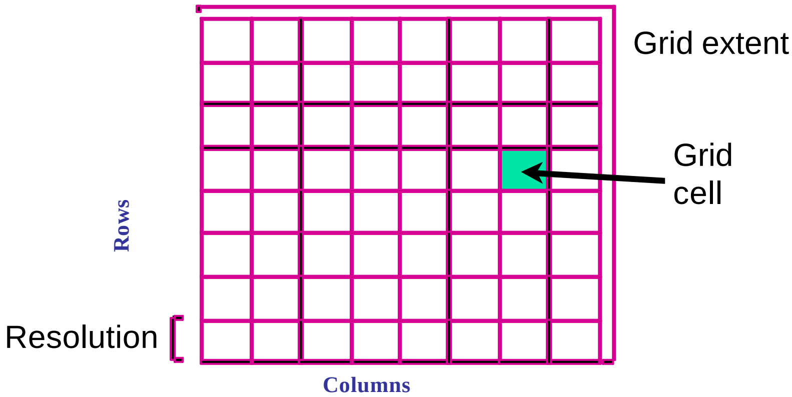


Figure 3.1 Generic structure for a grid.

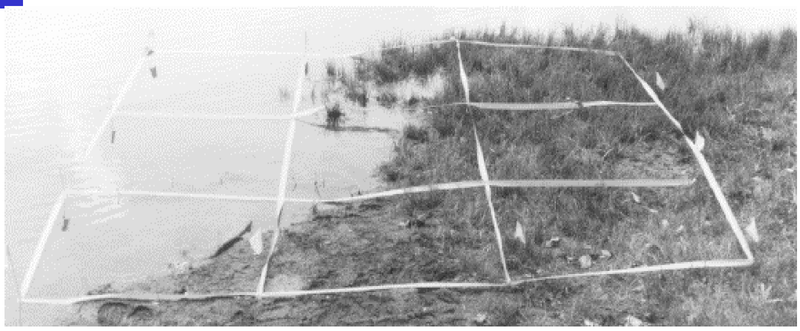
ITU Photogrammetry Division

Raster Data Structures problems

- the raster is not very good at representing lines or points, since each becomes a set of cells in the grid. Lines can become disconnected or "fat" if they cross the grid at too shallow an angle.
- the mixed pixel problem Even when the boundary is absolutely clear, a vector data representation may be better than the raster.

ITU Photogrammetry Division

The mixed pixel problem



Water dominates

W	W	G
W	W	G
W	W	G

Winner takes all

W	G	G
W	W	G
W	G	G

Edges separate

W	E	G
W	E	G
E	E	G

Grids and missing data

Figure 3.8 GIS data layer as a grid with a large section of "missing data," in this case, the zeros in the ocean off of New York and New Jersey.

Rasters are faster...

- Points and lines in raster format have to move to a cell center.
- Lines can become fat. Areas may need separately coded edges.
- Each cell can be owned by only one feature.
- As data, all cells must be able to hold the maximum cell value.
- Rasters are easy to understand, easy to read and write, and easy to draw on the screen.

RASTER

- A grid or raster maps directly onto a programming computer memory structure called an array.
- Grids are poor at representing points, lines and areas, but good at surfaces.
- Grids are good only at very localized topology, and weak otherwise.
- Grids are a natural for scanned or remotely sensed data.
- Grids suffer from the mixed pixel problem.
- Grids must often include redundant or missing data.
- Grid compression techniques used in GIS are run-length encoding and quad trees.

The quad-tree structure

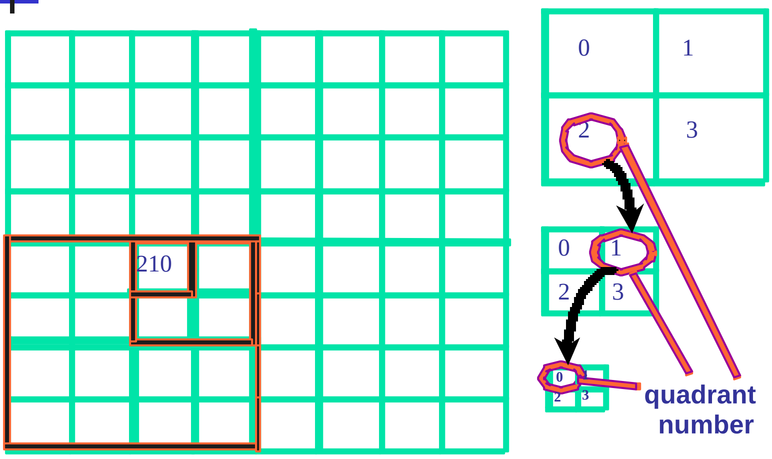


Figure 3.9 The quad-tree structure. Reference to code 210.

- Many GIS packages and many industry standard image formats use run-length encoding.
- Method takes a map unit row by row (or column by column). The pixel having same values on a row is coded the starting and ending column numbers. (Row 5 3,3) ...

Vector Data Structure

the first to be used for computer cartography and GIS because they were simply derived from digitising tablets, because they are more exact in representing complex features such as land parcels, and because they are easily drawn on pen-type output devices such as plotters.

- *cartographic spaghetti*

\- *vector data*

\- *Defence Mapping Agency's standard linear format*

\- *possible to convert to topological data*

The Vector Model

- A vector data model uses points stored by their real (earth) coordinates.
- Lines and areas are built from sequences of points in order.
- Lines have a direction to the ordering of the points.
- Polygons can be built from points or lines.
- Vectors can store information about topology.

VECTOR

- At first, GISs used vector data and cartographic spaghetti structures.
- Vector data evolved the arc/node model in the 1960s.
- In the arc/node model, an area consist of lines and a line consists of points.
- Points, lines, and areas can each be stored in their own files, with links between them.
- The topological vector model uses the line (arc) as a basic unit. Areas (polygons) are built up from arcs.
- The endpoint of a line (arc) is called a node. Arc junctions are only at nodes.
- Stored with the arc is the topology (i.e. the connecting arcs and left and right polygons).

Vectors just seemed more correcter.

- TIN must be used to represent volumes.
- Vector can represent point, line, and area features very accurately.
- Vectors are far more efficient than grids.
- Vectors work well with pen and light-plotting devices and tablet digitizers.
- Vectors are not good at continuous coverages or plotters that fill areas.

WHY TOPOLOGY MATTERS

- today the vector arc/node data structure with topology probably is the most widespread for GIS data
- a GIS maintains the arc as the basic unit, each line is stored only once and that the only duplication is the endpoints
- disadvantage is that whenever areas or polygons are to be used, some recomputing is necessary

Topology Matters

- The tolerances controlling snapping, elimination, and merging must be considered carefully, because they can move features.
- Complete topology makes map overlay feasible.
- Topology allows many GIS operations to be done without accessing the point files.

- allowed GIS for the first time to do error detection. If a set of polygons is fully connected, and there are no gaps at nodes or breaks in the lines defining the areas, the set of areas is called *topologically clean.*
- When maps are first digitised, the user can use the topology to check the polygons

TOPOLOGY

- Topological data structures dominate GIS software.
- Topology allows automated error detection and elimination.
- Rarely are maps topologically clean when digitized or imported.
- A GIS has to be able to build topology from unconnected arcs.
- Nodes that are close together are snapped.
- Slivers due to double digitizing and overlay are eliminated.

Basic arc topology

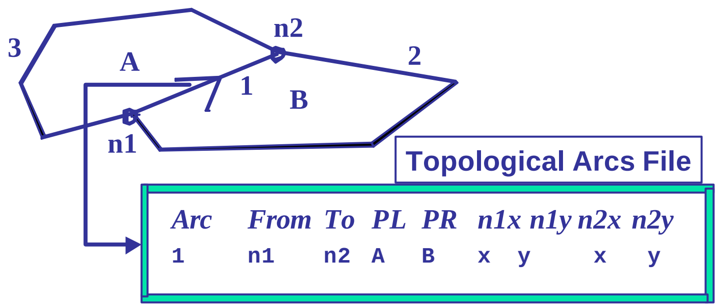


Figure 3.5 A topological structure for the arcs.

Arc/node map data structure with files

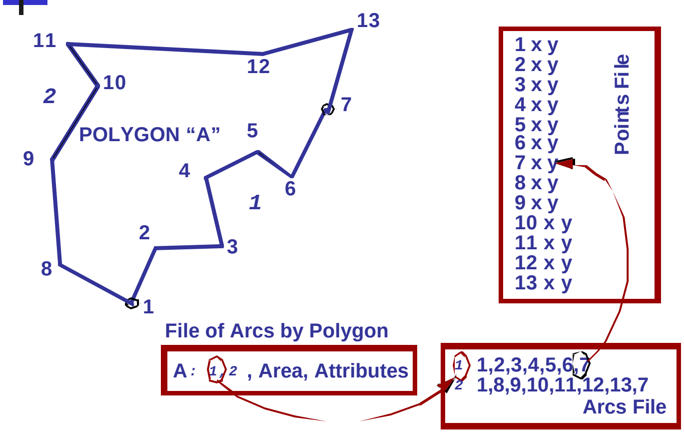
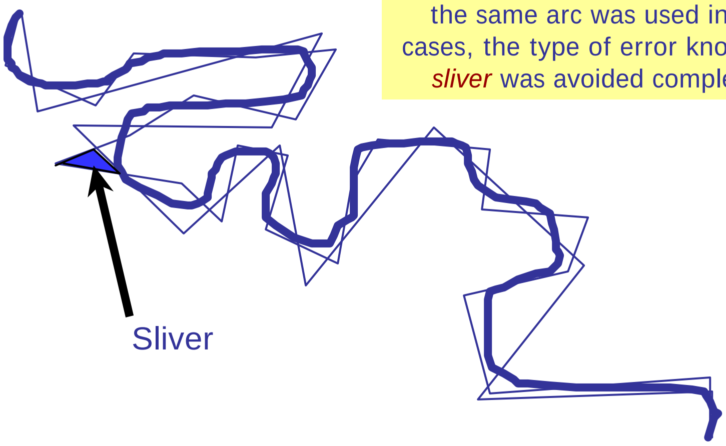


Figure 3.4 Arc/Node Map Data Structure with Files.

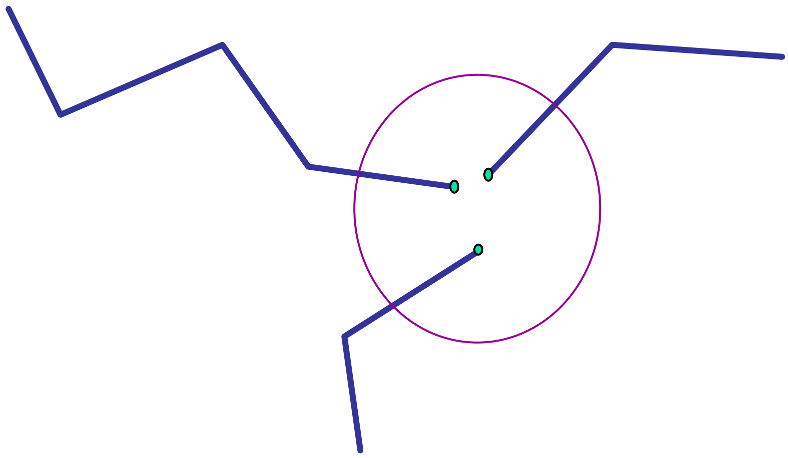
- A GIS will have the ability to build the topology from the unconnected arcs
- slivers (unmatched nodes along two lines) double digitising unsnapped nodes(end points of two lines that should be the same point)
- an error automatically detected can become an error automatically eliminated

Slivers

when a map was digitised, the user only needed trace each arc once, instead of twice if each area was traced around the edge. As exactly the same arc was used in both cases, the type of error known as a *sliver* was avoided completely.

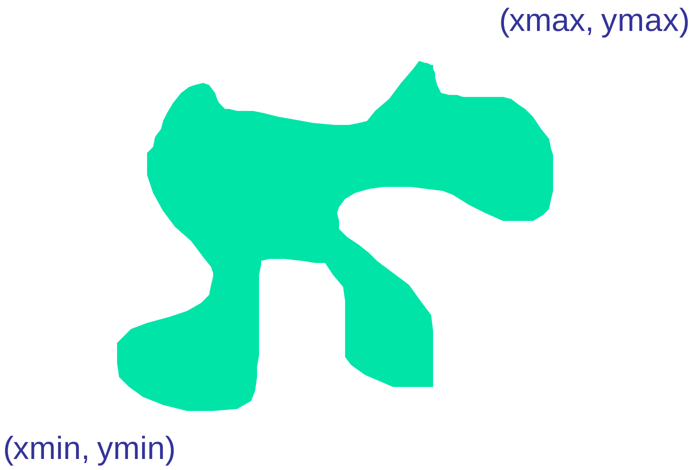


Unsnapped node



ITU Photogrammetry Division

The bounding rectangle



ITU Photogrammetry Division

topology

- *fuzzy tolerance:* Short lines, small polygons, or precisely measured point locations have critical significance and should not be deleted by automatic testing for topological completeness
- The primary advantage of having a topologically consistent map is that when two or more maps must be overlain
- Bounding rectangles: rectangles that contain a polygon completely. If a test point lies wholly outside the bounding rectangle, it must be outside the enclosed polygon.

Vectors and 3D

- Volumes (surfaces) are structured with the TIN model, including edge or triangle topology.
- TINs use an optimal Delaunay triangulation of a set of irregularly distributed points.
- TINs are popular in CAD and surveying packages.

triangulated irregular network *TIN*

- TIN is really just a list of points with their coordinates; stored with the points is a file containing information about the topology of a network.
- The network is a set of triangles, constructed by connecting the points in a network of triangles called a Delaunay triangulation.
- The **TIN** became popular as a way of storing topography or land elevation data for visualisation and engineering.

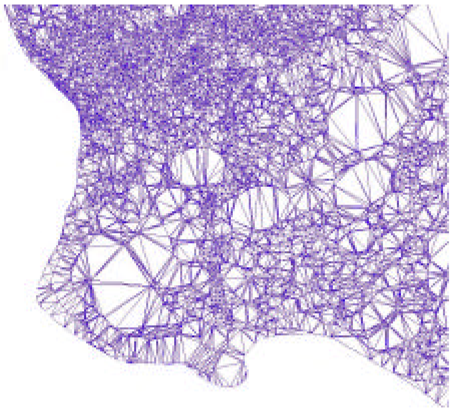
TIN

Easy to draw contours, make a three-dimensional view of an area, estimate how water would run downhill over a digital landscape, or calculate how much material would have to be moved in a construction project.

Many GIS programs that work with computer-aided drafting (CAD) systems or with surveying software use TIN as their data structure. TIN has proven to be both efficient in storing data and versatile in finding new uses within GIS.

TIN: Triangulated Irregular Network

- Way to handle field data with the vector data structure.
- Common in some GISs and most AM/FM packages.
- More efficient than a grid.



FORMATS

- Most GIS systems can import different data formats, or use utility programs to convert them.
- Data formats can be industry standard, commonly accepted or standard.

FORMATS FOR GIS DATA

The data structures used are often invisible as far as the GIS user is concerned. We might even not need to understand exactly what is happening when two maps are overlain. However, if we are to be objective, scientific GIS users, at the very least we must have a full understanding of the errors and transformations involved.

industry standard format:

- Hewlett-Packard Graphics Language (HPGL) page description language for plotters and printers.
- the PostScript page definition language for desktop and professional publishing products. In GIS, PostScript is usually used to export or print a finished map rather than data as such.
- Autocad digital exchange format (DXF) for drawing data. DXF does not support topology, only maintain information in separate layers, a familiar GIS concept.

others

- The digital line graph (DLG) format: USGS National Mapping Division (fig. 3.13)
- TIGER format: U. S. Census Bureau, it is vector files and does contain topology
- ...

- to store digital images and pictures
- TIF, GIF, JPEG, encapsulated PostScript

convert from raster to vector, and vice versa, such as CorelDraw

- a raster-to-vector or a vector-to-raster conversion

  from vector to raster, filling in grid cells as lines cross them or as polygons include them, is relatively simple. The opposite is quite complex.

- different computer systems

  different local authorities may develop their GIS operations around different software.

Exchange problems

- From a GIS context, we finish with data of unknown accuracy, source, projection, and with unspecified or imperfectly matched attributes. On one county map perhaps, all streams and rivers are shown, on the other only the major ones. One might think that the climate was wetter, but in fact the difference is one of interpretation and lack of standardisation.

the spatial data transfer standard 1980-1992

- Not only did the standard have to produce a bibliography, a terminology, and a complete list of geographic and map features, it also had to address the problems of data accuracy and the broader metadata issues for data description

the spatial data transfer standard

- Within the members of NATO, an exchange standard called DIGEST was developed, with a vector data profile called the vector product format (VPF).
- Digital Chart of the World data exchange in Germany, Australia, South Africa, the European Union

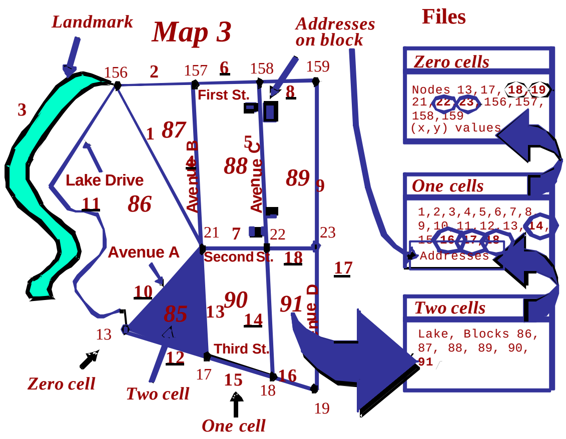
the spatial data transfer standard

- Data exchange by translation (export and import) can lead to significant errors in attributes and in geometry
- Efficient data exchange is important for the future of GIS.

Vector Data Formats

- Vector formats are either page definition languages or preserve ground coordinates.
- Page languages are HPGL, PostScript, and Autocad DXF.
- True vector GIS data formats are DLG and TIGER, which has topology.

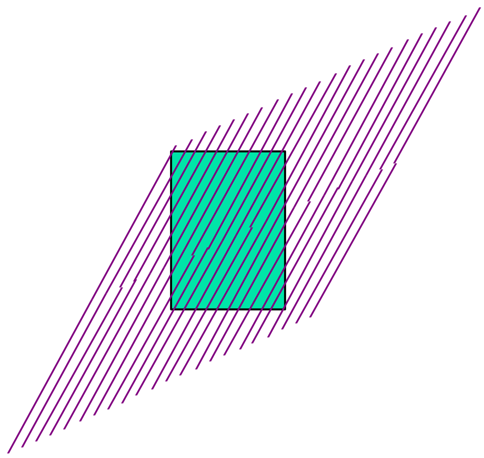
The TIGER data structure



Raster Data Formats

- Most raster formats are digital image formats.
- Most GISs accept TIF, GIF, JPEG or encapsulated PostScript, which are not georeferenced.
- DEMs are true raster data formats.

DEMs and UTM (7.5 minute 30m)



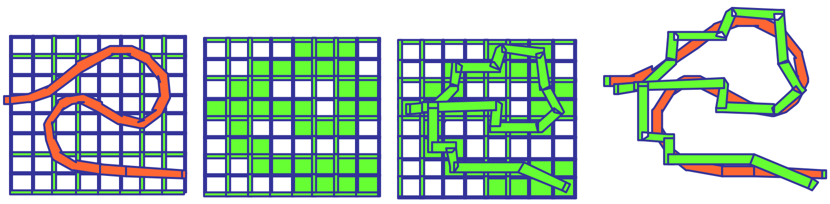
ITU Photogrammetry Division

EXCHANGE

- Most GISs use many formats and one data structure.
- If a GIS supports many data structures, changing structures becomes the user’s responsibility.
- Changing vector to raster is easy; raster to vector is hard.
- Data also are often exchanged or transferred between different GIS packages and computer systems.
- The history of GIS data exchange is chaotic and has been wasteful.

ITU Photogrammetry Division

Vector to raster exchange errors

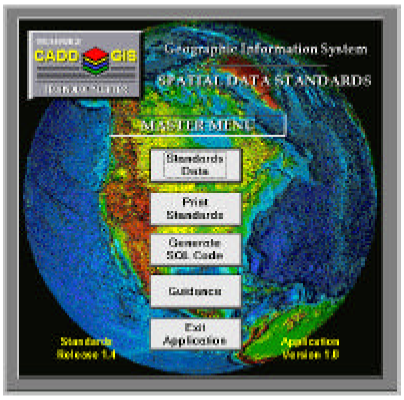


Exchange problems:

- none of the industry standards exchanges topology with the data, transferring instead only the graphic information.
- with many different formats, each package has to include a large number of format translators

Transfer Standards

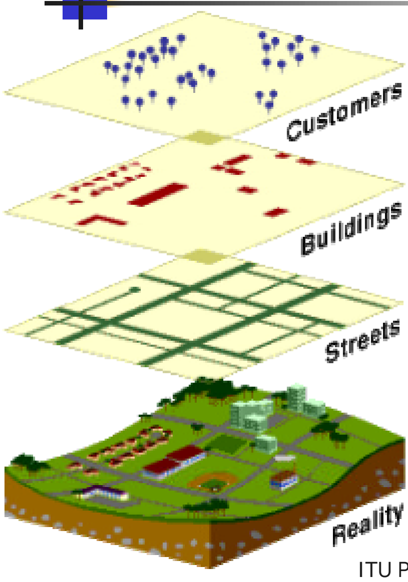
SDTS
Spatial Data Transfer Standard

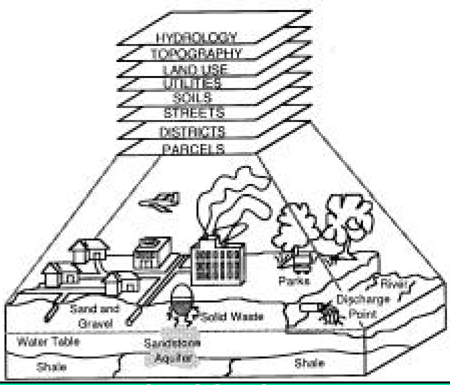


GIS Data Exchange

- Data exchange by translation (export and import) can lead to significant errors in attributes and in geometry.
- In the United States, the SDTS was evolved to facilitate data transfer.
- SDTS became a federal standard (FIPS 173) in 1992.
- SDTS contains a terminology, a set of references, a list of features, a transfer mechanism, and an accuracy standard.
- Both DLG and TIGER data are available in SDTS format.
- Other standards efforts are DIGEST, DX-90, the Tri-Service Spatial Data Standards, and many other international standards.
- Efficient data exchange is important for the future of GIS.

How GIS works?





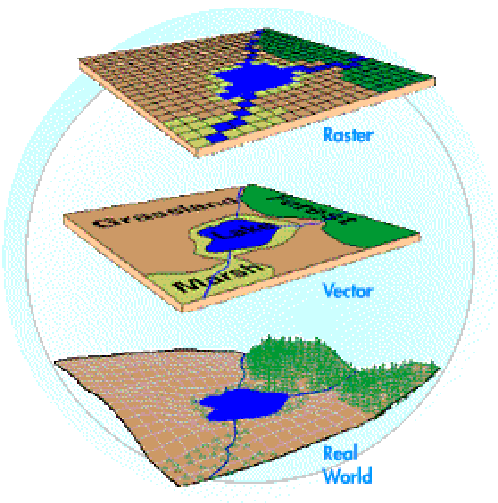
Related data layers

Geographic References

Geographic information contains either an explicit geographic reference, such as a latitude and longitude or national grid coordinate, or an implicit reference such as an address, postal code, census tract name, forest stand identifier, or road name. An automated process called geocoding is used to create explicit geographic references (multiple locations) from implicit references (descriptions such as addresses). These geographic references allow you to locate features, such as a business or forest stand, and events, such as an earthquake, on the earth's surface for analysis.

Vector and Raster Models

Geographic information systems work with two fundamentally different types of geographic models--the "vector" model and the "raster" model. In the vector model, information about points, lines, and polygons is encoded and stored as a collection of x,y coordinates.



Vector and Raster Models

The location of a point feature, such as a bore hole, can be described by a single x,y coordinate. Linear features, such as roads and rivers, can be stored as a collection of point coordinates. Polygonal features, such as sales territories and river catchments, can be stored as a closed loop of coordinates.

The vector model is extremely useful for describing discrete features, but less useful for describing continuously varying features such as soil type or accessibility costs for hospitals. The raster model has evolved to model such continuous features. A raster image comprises a collection of grid cells rather like a scanned map or picture. Both the vector and raster models for storing geographic data have unique advantages and disadvantages. Modern GISs are able to handle both models.

Coming next…

Getting the Map into the Computer