Geographic Information System

What is a Geographic Information System?

An understanding of what Geographic Information Systems may be helped by considering the component parts of the term separately.

Geographic

This term is used because GIS tend to deal primarily with 'geographic' or 'spatial' features.

These are objects which can be referenced or related to a specific location in space

(may be physical, or economic in nature).

Information

This represents the large volumes of data.

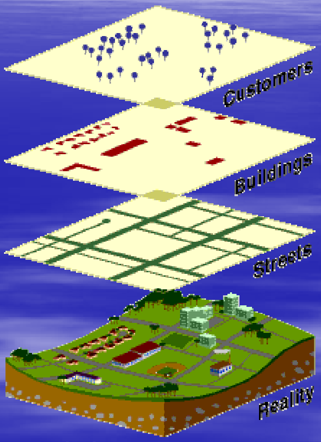
Systems

This term is used to represent the systems approach taken by GIS.

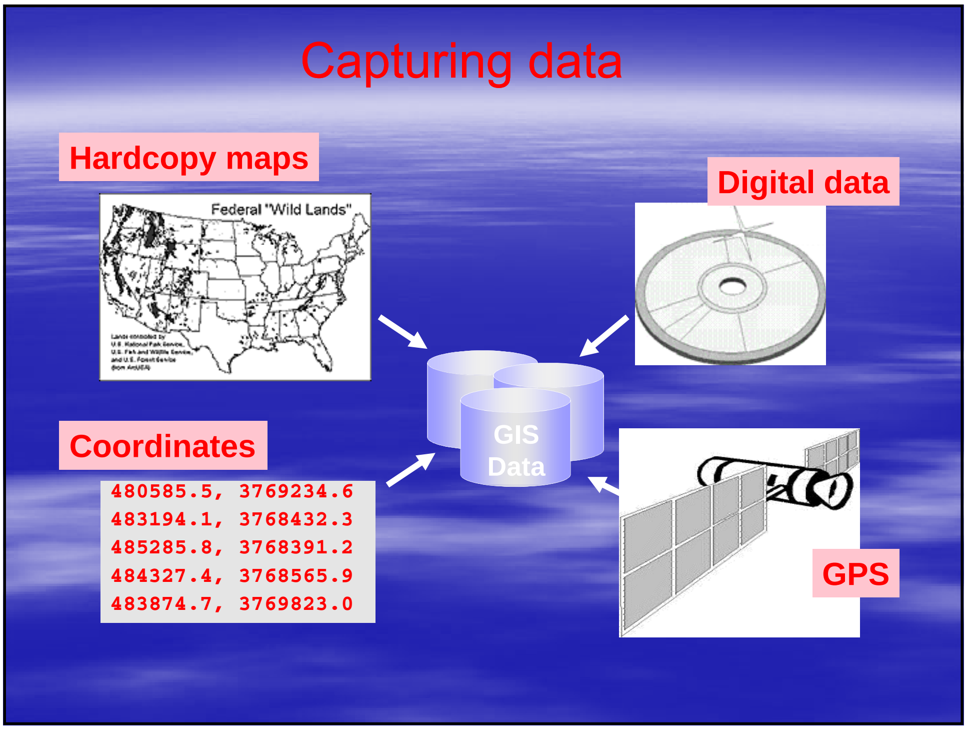
Computer systems are becoming vital for the storage and manipulation of the increasing volumes of data, the handling of complex spatial algorithms and me integration of data of different scales, projections and formats.

A geographic information system (GIS)
is a computer system for

Capture
Store
Query
Analyze
Display
Output



geographically referenced data
also called geospatial data.



Capturing data

Hardcopy maps

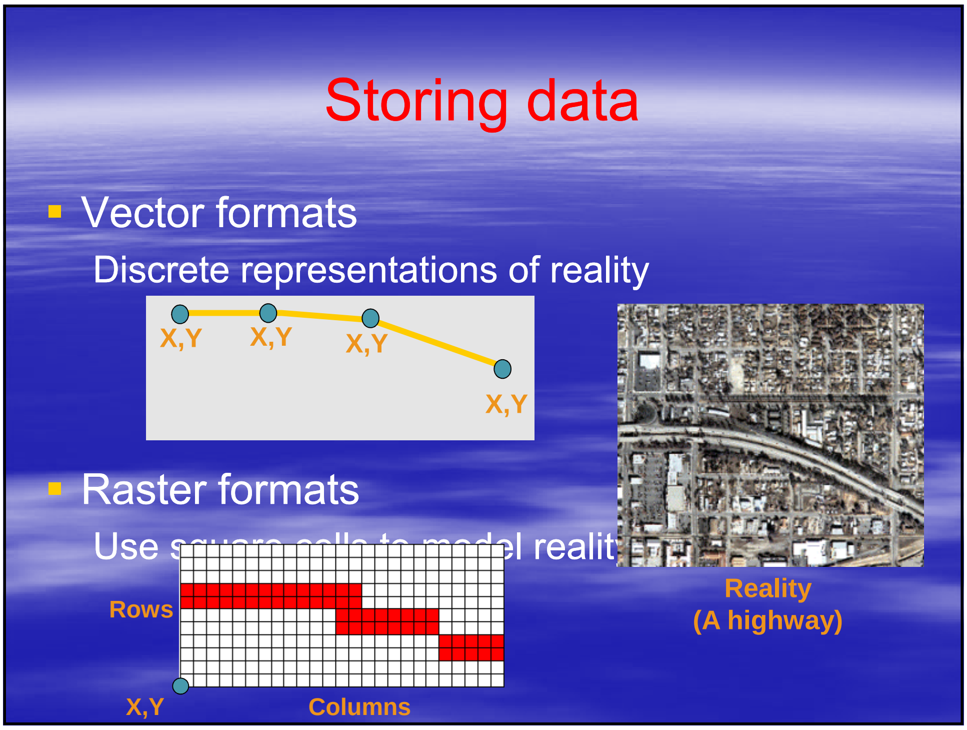
Digital data

GIS Data

Coordinates

480585.5,	3769234.6
483194.1,	3768432.3
485285.8,	3768391.2
484327.4,	3768565.9
483874.7,	3769823.0

GPS



Storing data

- Vector formats

  Discrete representations of reality

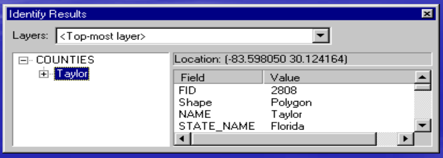
- Raster formats

  Use square cells to model reality

Reality
(A highway)

Query

Identifying specific features


Identify Results

Layers: <Top-most layer>

COUNTIES
Taylor

Location: (-83.598050 30.124164)

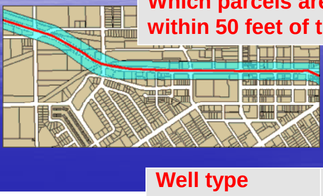
Field	Value
FID	2808
Shape	Polygon
NAME	Taylor
STATE_NAME	Florida

Identifying features based on conditions

Florida counties with a population greater than 300,000

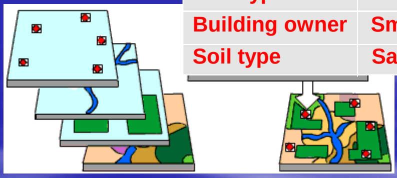
Analysis

Proximity



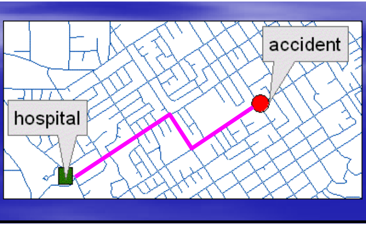
Which parcels are within 50 feet of the road?

Overlay

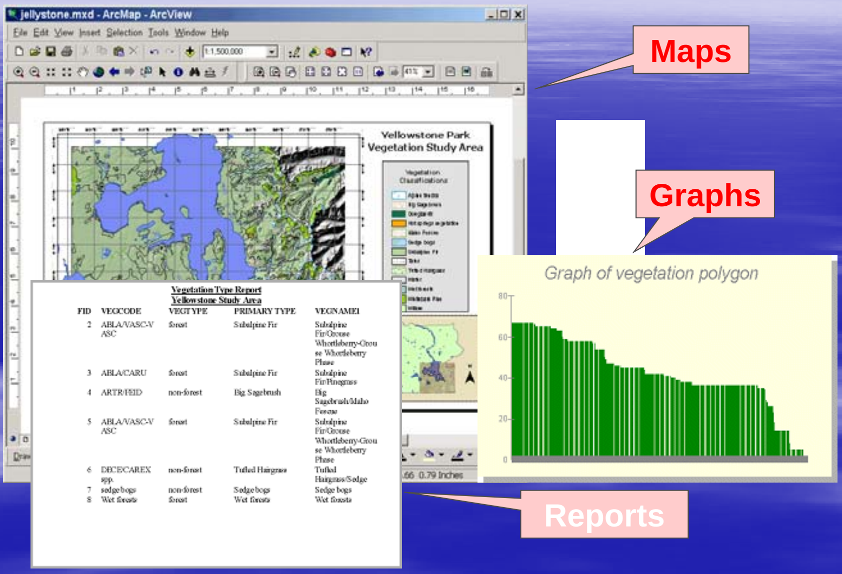


Well type	**Drilled**
Building owner	**Smith**
Soil type	**Sandy**

Network



Display



Vegetation Type Report
Yellowstone Study Area

FID	VEGCODE	VEGTYPE	PRIMARY TYPE	VEGNAME1
2	ABLA/VASC-VASC	forest	Subalpine Fir	Subalpine Fir/Grouse Whortleberry-Grouse Whortleberry Phase
3	ABLA/CARU	forest	Subalpine Fir	Subalpine Fir/Pinegrass
4	ARTR/FEID	non-forest	Big Sagebrush	Big Sagebrush/Idaho Fescue
5	ABLA/VASC-VASC	forest	Subalpine Fir	Subalpine Fir/Grouse Whortleberry-Grouse Whortleberry Phase
6	DECE/CAREX spp.	non-forest	Tufted Hairgrass	Tufted Hairgrass/Sedge
7	sedge bogs	non-forest	Sedge bogs	Sedge bogs
8	Wet forests	forest	Wet forests	Wet forests

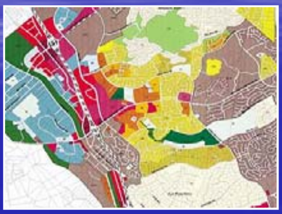
Maps

Graphs

Reports

Output

Paper map



Internet

GIS Data

Image

Florida.jpg

Document

Florida.mxd

Geographically referenced data are data that describe both the location and characteristics of spatial features such as roads, land parcels, and vegetation stands on the Earth's surface.

GIS Applications.

Since the beginning, GIS has been important in natural resource management

including land-use planning, natural hazard assessment, wildlife habitat analysis, and riparian zone monitoring.

In more recent years GIS has been used for crime analysis, emergency planning, land records management, market analysis, and transportation applications.

Integration of GIS with global positioning system (GPS) and the Internet has also introduced new and exciting applications such as precision farming, interactive mapping, and location-based services.

Components of a GIS.

Like any other information technology, GIS requires the following four components:

Computer System.

The computer system includes the computer and the operating system to run GIS.

Additional equipment may include

monitors for display, digitizers and scanners for spatial data input, and printers and plotters for hardcopy data display.

GIS Software.

GIS software includes the program and the user interface for driving the hardware.

Brainware.

Brainware refers to the purpose and objectives, and provides the reason and justification for using GIS.

Infrastructure.

Infrastructure refers to the necessary physical, organizational, administrative, and cultural environments that support GIS operations.

The infra-structure includes requisite skills, data standards, data clearinghouses, and general organizational patterns.

A Brief history of GIS.

GIS is not new.

Since the late 1960s computers have been used to store and process geographically referenced data.

Early examples of GIS work from the late 1960s and 1970s

For many years, GIS has been considered to be too difficult, expensive, and proprietary. the graphical user interface (GUI), powerful and affordable hardware and software, and public digital data has broadened the range of GIS applications and brought GIS to mainstream use in the 1990s.

According to a published survey, ESRI Inc. and Intergraph Corp. These two companies also led the GIS industry in software revenues in 2000.

The main software product from ESRI Inc. is ArcGIS, a scalable system with Arc View, ArcEditor, and ArcInfo.

Intergraph Corp. has two main products: GeoMedia and MGE.

Geographically Referenced Data.

Geographically referenced data separate GIS from other information systems.

We must understand the nature of geographically referenced data.

Take the example of roads.

- location
    - where it is
- characteristics
    - length
        - name
            - speed limit
                - direction

The location, also called geometry or shape, represents spatial data.

The characteristics are attribute data.

Thus a road, like any geographically referenced data.

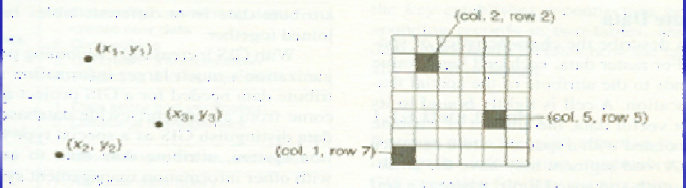
Two components of spatial data and attributes data.

Spatial Data.

Spatial data describe the locations of spatial features, which may be discrete or continuous.

(Discrete or Continuous) features include points (wells), lines (roads), and areas (land-use types).

The vector data model uses points and their x-, y-coordinates to construct spatial features of points, lines, and ar-eas.

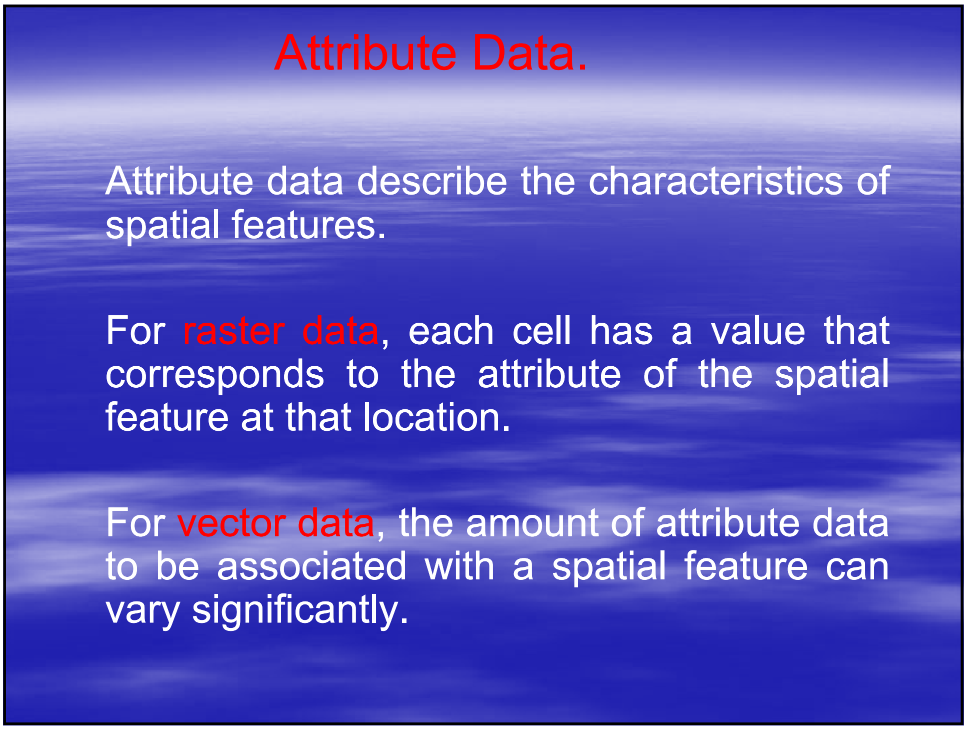


The raster data model uses a grid and grid cells to represent the spatial variation of a feature.

The two data models differ in concept and data structure :

Vector data are ideal for representing discrete features and may be topological or nontopological, and simple or higher-level ;

Raster data are better suited for representing continuous features and have a simple data structure with rows and columns and fixed cell locations.



Attribute Data.

Attribute data describe the characteristics of spatial features.

For raster data, each cell has a value that corresponds to the attribute of the spatial feature at that location.

For vector data, the amount of attribute data to be associated with a spatial feature can vary significantly.

GIS Operations.

Although GIS activities no longer follow a set sequence, to explain what GIS users do, we can group GIS activities (Operation) into

- Spatial data input.
- Attribute data management.
- Data display.
- Data exploration.
- Data analysis.
- GIS modeling.

GIS Operations.

1. Spatial data input.

1.1. Data entry:
(use existing data, create new data)
1.2. Data editing
1.3. Geometric transformation
1.4. Projection and reprojection

GIS Operations.

2. **Attribute data management**

- **2.1. Data entry and verification**
- **2.2. Database management**

GIS Operations.

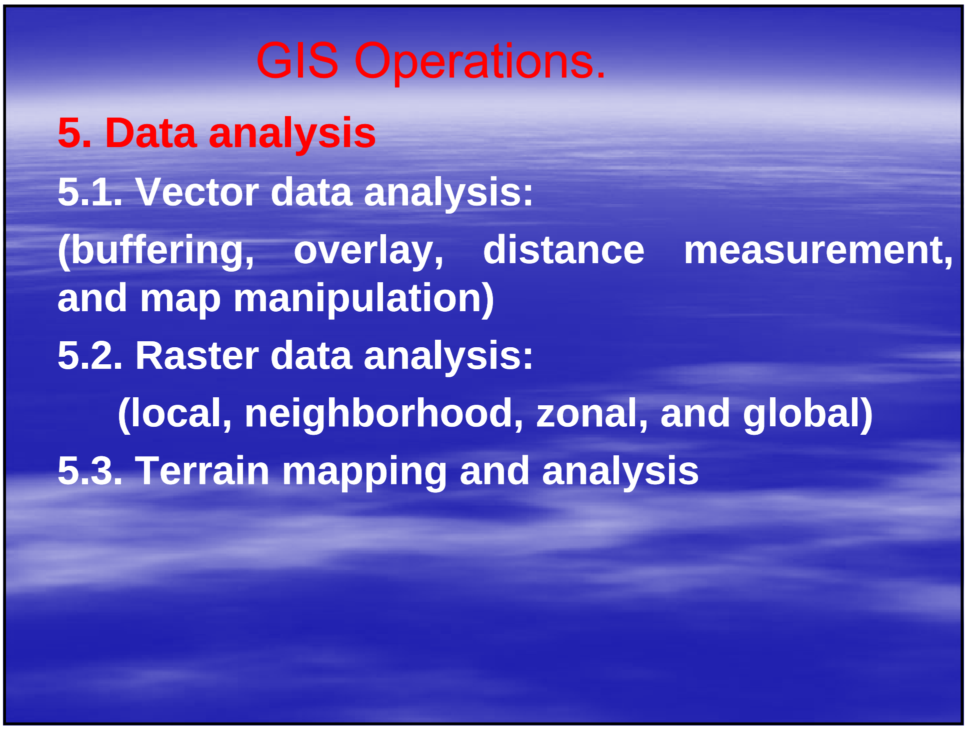
3. Data display.

3.1. Use of
- maps
- charts
- tables

GIS Operations.

4. Data exploration

- **4.1. Attribute data query**
- **4.2. Spatial data query**
- **4.3. Geographic visualization**



GIS Operations.

5. Data analysis

5.1. Vector data analysis:

(buffering, overlay, distance measurement, and map manipulation)

5.2. Raster data analysis:

(local, neighborhood, zonal, and global)

5.3. Terrain mapping and analysis

GIS Operations.

6. GIS modeling

6.1. Binary models
6.2. Index models
6.3. Regression models
6.4. Process models

Geographic Coordinate System.

The geographic coordinate system is the location reference system for spatial features on the Earth's surface.

The geographic coordinate system consists of meridians and parallels.

The geographic coordinate system is the location reference system for spatial features on the Earth's surface.

The geographic coordinate system

consists of meridians and parallels.

Meridians are lines of longitude.

Running north and south,

meridians are used for

measuring location

in the E-W direction.

Using the meridian passing through Greenwich, England, as the prime meridian or 0°,

one can measure the longitude value of a point on the Earth's surface as 0° to 180" east or west of the prime meridian.

one can measure the latitude value of a point as 0° to 90° north or south of the equator.

The origin of the geographic coordinate system is where the prime meridian meets the equator.

Thus, longitude values are similar to x values in a coordinate system and latitude values are similar to y values.

In GIS, to enter longitude and latitude values with positive or negative signs.

Latitude values are positive if north of the equator, and negative if south of the equator.

Longitude values are positive in the eastern hemisphere and negative in the western hemisphere.

Longitude and latitude values may be measured in the
Degrees-Minutes-Seconds (DMS) system or
Decimal Degree (DD) system.

In which 1 degree equals 60 minutes and 1 minute, equals 60 seconds.

This allows conversion between the two systems.

For example, a latitude value of 45°52'30" would be equal to 45.875° (45 + 52/60 + 30/3600).

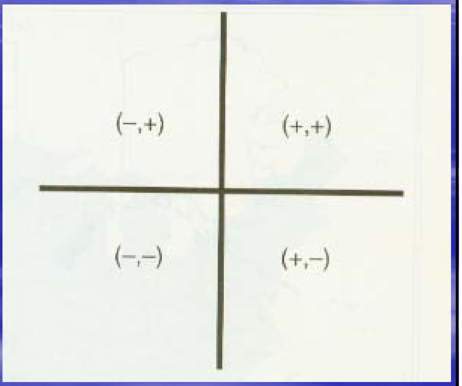
The central parallel and the central meridian divide a map projection into four quadrants. Points within

the NE quadrant have positive *x-* and y-coordinates,

the NW quadrant have negative x-coordinates and positive y-coordinates,

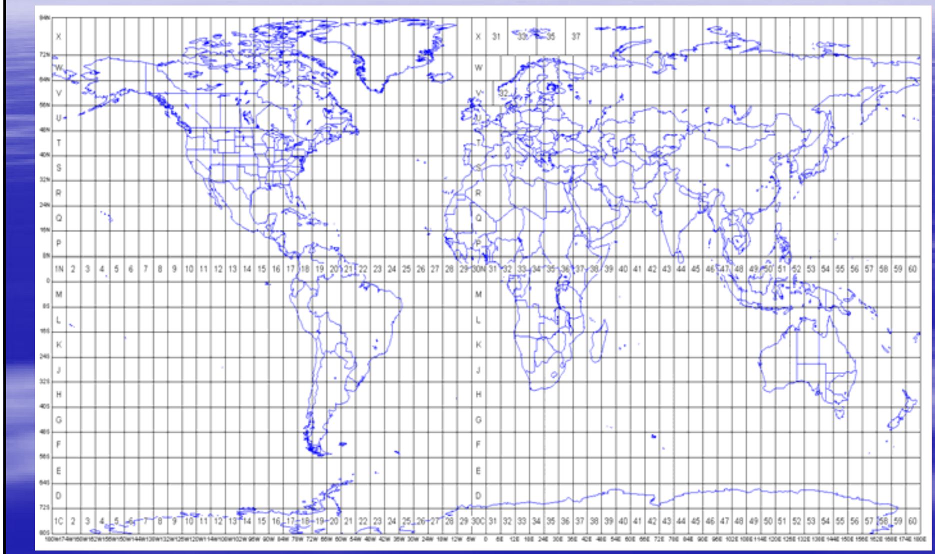
the SE quadrant have positive x-coordinates and negative y-coordinates,

the SW quadrant have negative *x-* and y-coordinates.



Commonly Used Map Projections.

1. **Transverse Mercator.**
2. **Lambert Conformed Conic.**
3. **Albers Equal-Area Conic.**
4. **Equidistant Conic.**

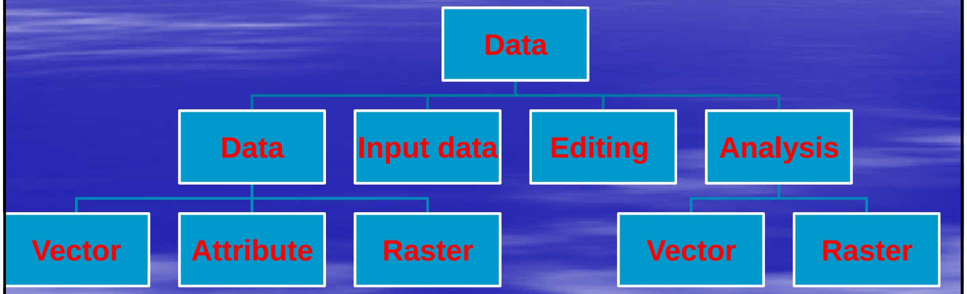


Working with Coordinate Systems in GIS.

Most GIS packages provide a wide variety of coordinate systems, datums, and spheroids.

	Predefined	**Custom**
Geographic	WGS84	Datum transformation
Plane	UTM	IDTM

(Data)



- Data
  - Data
    - Vector
    - Attribute
    - Raster
  - Input data
  - Editing
  - Analysis
    - Vector
    - Raster

Vector Data

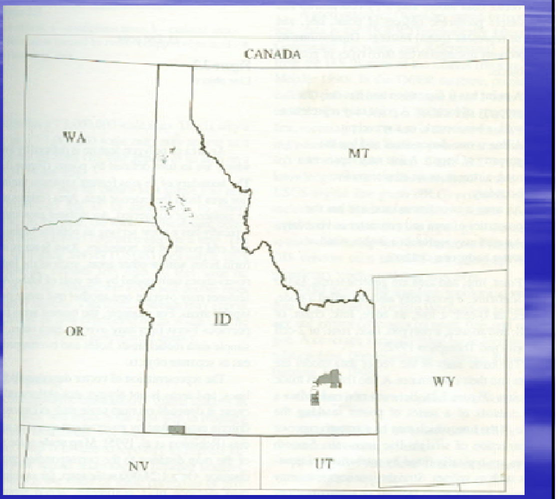
Introduction.

Looking at a paper map, we can see

what map features are like and

We can easily see the map features and their spatial relationships one another.

but how can the computer see these features and relationships?



Geometric Objects.

The vector data uses *x-,* y-coordinates and the simple geometric objects.

Three types of geometric objects.

A point

A line

An area

A point has 0 dimension and has only the property of location.

A point may represent a well, a benchmark, or a pit.

In the GIS literature, a point may also be called a
node,
vertex,
or 0-cell.

A **line** is one-dimensional and has the property of length.

A line may represent a road, a stream, or an administrative boundary.

In the GIS literature, a line may also be called an

edge,
link,
chain,
or 1-cell.

An **area** is two-dimensional and has the properties of area and perimeter or boundary.

An area may represent, a water body, or a sinkhole.

In the GIS literature, an area may also be called a

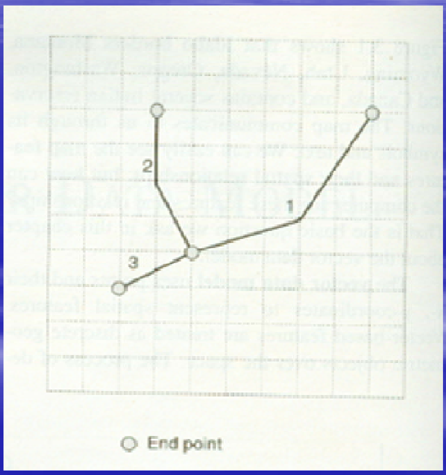
polygon,
face,
zone,
or 2-cell.

The basic units of the vector data are points and their coordinates.

A line feature is made of points.

Between two end points a line consists of a series of points marking the shape of the line, which may be a smooth curve or a connection of straight-line segments.

Smooth curves are typically fitted by mathematical equations.



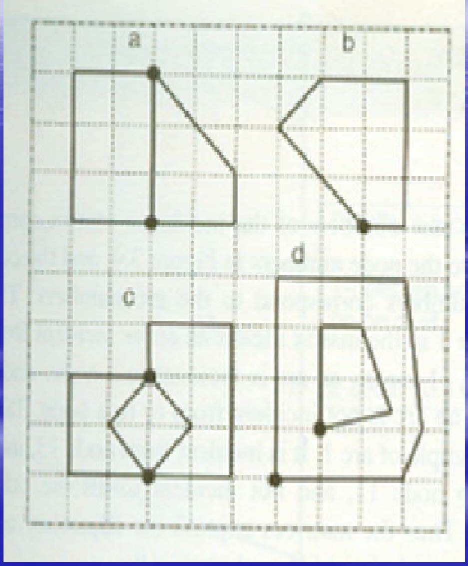
An area or polygon feature is defined by lines, which are in turn defined by points.

a-contiguous areas;
b-isolated area;
c-three areas formed by two overlapped objects
d-a hole within an area.

The boundary of an area feature separates the interior area from the exterior area.

Area features may be isolated or connected.

Area features may overlap one another and create overlapped areas.

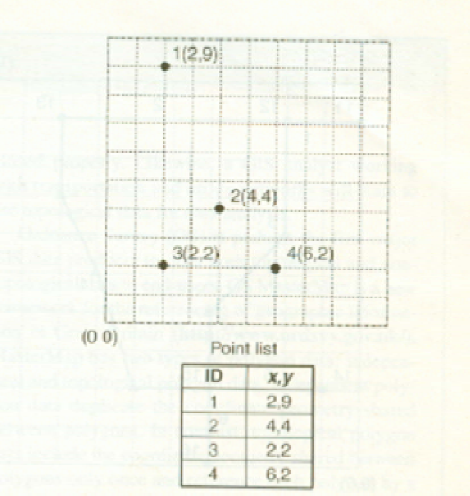


Vector Data Structure.

Data structure of a point features:

they can be coded with their identification numbers (IDs) and pairs of *x*- and *y*-coordinates.

Topology does not apply to points because points are separate from one another.



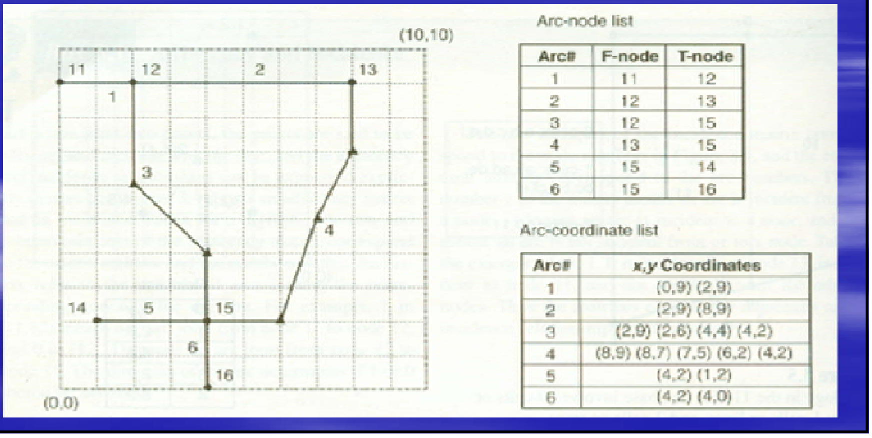
Point list

ID	x,y
1	2,9
2	4,4
3	2,2
4	6,2

Data structure of a line feature.

An arc is a line segment, which is connected to two end points called nodes.

The starting point is the from-node and the ending point the to-node.



Arc-node list

Arc#	F-node	T-node
1	11	12
2	12	13
3	12	15
4	13	15
5	15	14
6	15	16

Arc-coordinate list

Arc#	x,y Coordinates
1	(0,9) (2,9)
2	(2,9) (8,9)
3	(2,9) (2,6) (4,4) (4,2)
4	(8,9) (8,7) (7,5) (6,2) (4,2)
5	(4,2) (1,2)
6	(4,2) (4,0)

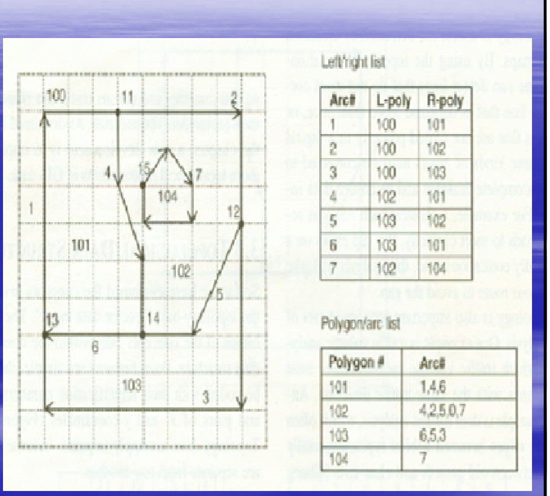
Data structure of an area or polygon feature.

The polygon/arc list shows the relationship between polygons and arcs.

arcs 1, 4 and 6 connect to define polygon 101.

Polygon 104 differs from the other polygons in being surrounded by polygon 102.

To show that polygon 104 is a hole within polygon 102, the arc list for polygon 102 contains a zero to separate the external and internal boundaries.



Left/right list

Arc#	L-poly	R-poly
1	100	101
2	100	102
3	100	103
4	102	101
5	103	102
6	103	101
7	102	104

Polygon/arc list

Polygon #	Arc#
101	1,4,6
102	4,2,5,0,7
103	6,5,3
104	7

Raster Data

The vector data uses the geometric objects of point, line, and area to represent spatial features.

Although ideal for discrete features with well-defined locations and shapes.

the vector data does not work well with spatial phenomena that vary continuously over the space such as precipitation, elevation, and soil erosion.

A better option for representing continuous phenomena is the raster data model.

The raster data uses a regular grid to cover the space and the value in each grid cell to correspond to the characteristic of a spatial phenomenon at the cell location.

Conceptually, the variation of the spatial phenomenon is reflected by the changes in the cell value.

Raster data have been described as field-based, as opposed to object-based vector data.

A wide variety of data used in GIS are encoded in raster format.

They include digital elevation data, satellite images, digital orthophotos, scanned maps, and graphic files.

Commercial GIS packages can display raster and vector data simultaneously, and can convert from raster to vector data or from vector to raster data.

Raster data also introduce a large set of data analysis functions and applications to GIS.

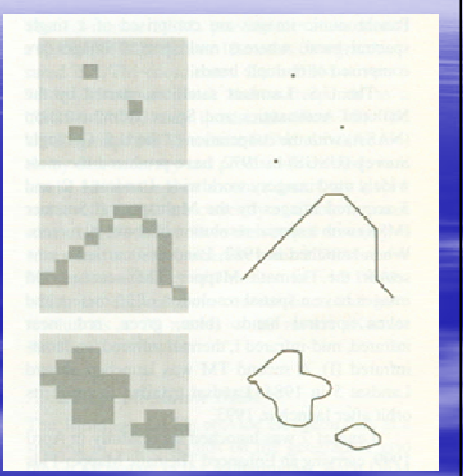
Integration of both types of data has therefore become a common and desirable feature in a GIS project.

Raster data represent

points by single cells,

lines by sequences of neighboring cells,

areas by collections of contiguous cells.



The cell size determines the resolution of the raster data model.

A cell size of 30 meters means that each cell measures 900 square meters (30 X 30 meters).

A grid is normally projected onto a plane coordinate system such as the UTM (Universal Transverse Mercator) coordinate system.

There-fore, raster data can be displayed with vector data if they are based on the same coordinate system.

Type of Raster Data.

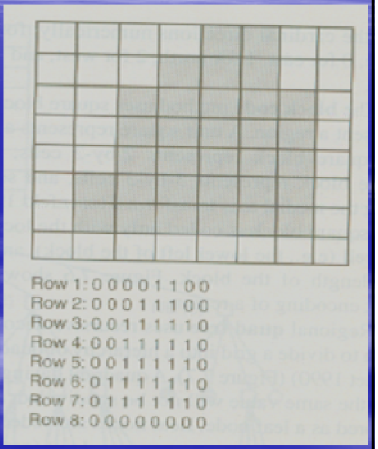
1. Satellite Imagery.
2. Digital Elevation Models (DEM).
3 Digital Orthophotos.
4. Scanned Files.
5. Digital Raster Graphics.
6. Graphic Files.
7. GIS Software-Specific Raster Data.

Data Structure.

Data structure refers to the storage of raster data so that they can be used and processed by the computer.

The simplest data structure is called the cell-by-cell encoding method

a raster model is stored as a matrix, and its cell values are written into a file by row and column.



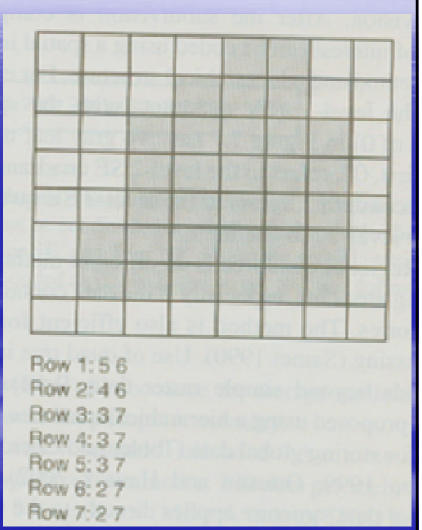
Row 1: 0 0 0 0 1 1 0 0
Row 2: 0 0 0 1 1 1 0 0
Row 3: 0 0 1 1 1 1 1 0
Row 4: 0 0 1 1 1 1 1 0
Row 5: 0 0 1 1 1 1 1 0
Row 6: 0 1 1 1 1 1 1 0
Row 7: 0 1 1 1 1 1 1 0
Row 8: 0 0 0 0 0 0 0 0

The cell-by-cell encoding method becomes inefficient if a raster model contains many redundant cell values.

Raster models such as binary scanned files can be more efficiently stored using the run length encoding (RLE) method, which records the cell values by row and by group.

Each group includes a cell value and the number of cells with that value.

If all cells in a row contain the same value, only one group is recorded, thus saving the computer memory.



Header File.

To import raster data for use, a GIS package must have information about the data structure and, if applicable, the compression method.

This information is contained in the header file, often denoted by the extension .hdr, which is like metadata in function.

A header file also contains information on

numbers of rows and columns,

number of spectral bands,

number of bits per pixel,

value for no data,

x- and y-coordinates of the origin,

pixel size.

Some raster data use other files besides the header file or organize the data information into different files.

Satellite images, for example, may have two optional files.

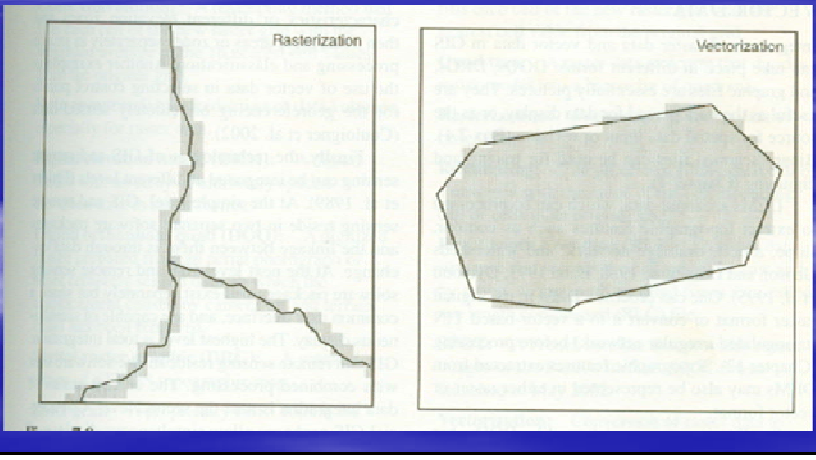
The statistics file, often with the extension .stx, describes statistics such as minimum, maximum, mean, and standard deviation for each spectral band in an image.

The color file, often with the extension .clr, associates colors with different pixel values in an image.

Data Conversion.

The conversion of vector data to raster data is called Rasterization, and the conversion of raster data to vector data is called Vectorization.

These two types of conversion use different computer algorithms.



The simpler of the two conversion methods, Rasterization involves three basic steps.

The first step is to set up a grid with a specified cell size to cover the area extent of the coverage and to assign initially all cell values as zeros.

The second step is to change the values of those cells that correspond to points, lines, or polygon boundaries. The cell value is set to 1 for a point, the line's value for a line, and the polygon's value for a polygon boundary.

The third step is to fill the interior of the polygon outline with the polygon value.

Errors from rasterization depend on the design of the computer algorithm and the size of raster cell.

Vectorization turns raster lines into vector lines in a process called tracing.

Tracing involves three basic elements:

line thinning

line extraction

topological reconstruction.

Lines in the vector data model have length but no width. Lines in a scanned file (raster lines), however, usually occupy several pixels in width.

Raster lines must be thinned, ideally to a 1-cell width, for vectorization.

Line extraction is the process of determining where individual lines begin and end.

Finally, topology is built between lines extracted from the raster model.

Results of raster-to-vector conversion often show steplike features along diagonal lines.

A line smoothing operation can eliminate those artifacts from raster data.

Attribute Data

Introduction

GIS involves both spatial and attribute data:

Spatial data relate to the geometry of map features,

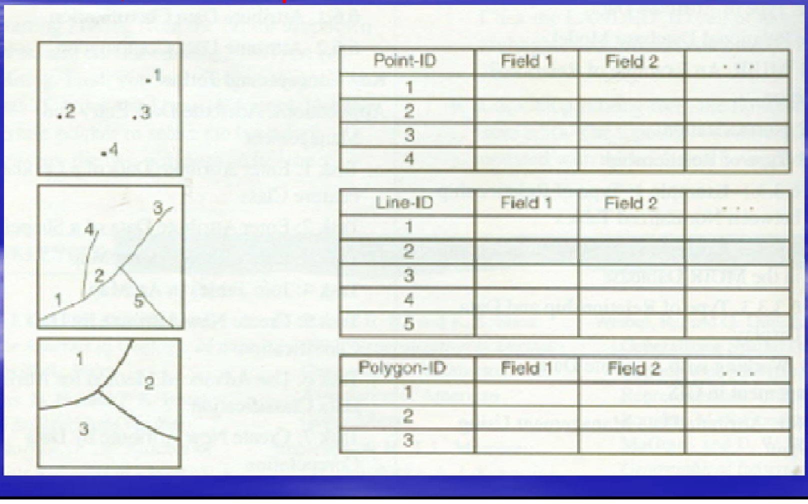
Attribute data describe the characteristics of the map features. Attribute data are stored in tables.

Each row of a table represents a map feature, and each column represents a characteristic.

The intersection of a column and a row shows the value of a particular characteristic for a particular map feature.

The difference between spatial and attribute data is well defined with vector-based map features.

In GIS, stores spatial data and attribute data



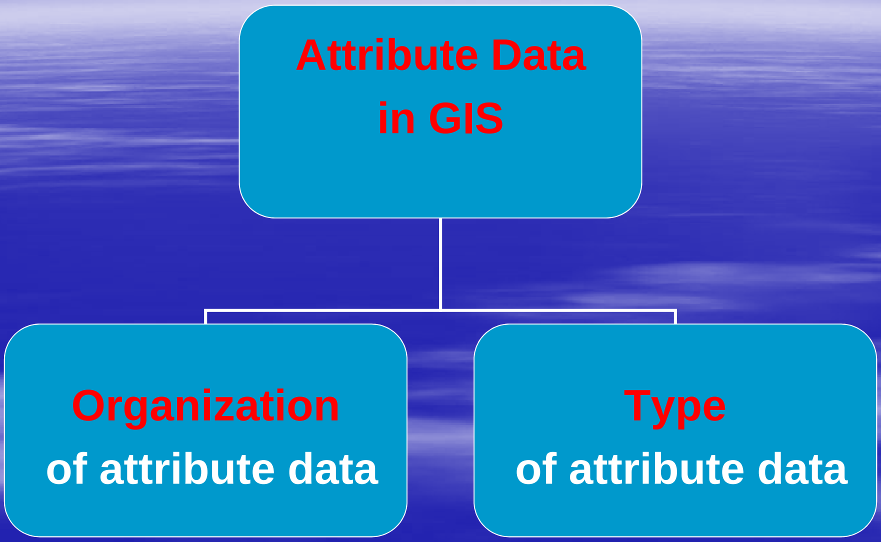
Point-ID	Field 1	Field 2	. . .
1			
2			
3			
4			

Line-ID	Field 1	Field 2	. . .
1			
2			
3			
4			
5			

Polygon-ID	Field 1	Field 2	. . .
1			
2			
3			

The two data sets must be synchronized so that the spatial and attribute data of map features can be queried, analyzed, and displayed.

Each map feature has a unique object ID and an attribute to store its geometry.



Attribute Data in GIS

Organization of attribute data

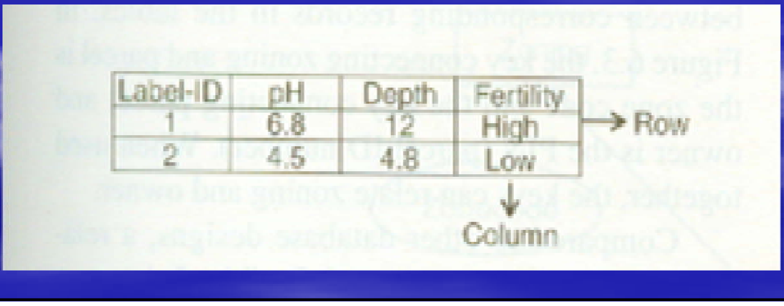
Type of attribute data

Organization of Attribute Data

Attribute data are stored in feature attribute tables. An attribute table is organized by row and column.

Each row represents a map feature, which has a unique label ID or object ID also called a record or a tuple.

Each column represents a property or characteristic of the map feature and also called a field or an item.



Label-ID	pH	Depth	Fertility
1	6.8	12	High
2	4.5	4.8	Low

Line features have the default attribute of length, and area features have the default attributes of area and perimeter.

A feature attribute table may be the only table needed if a map has only several attributes to be associated.

But this is not the case with most GIS projects.

Type of Attribute Data

One method for classifying attribute data is the data type.

A data type determines the kind of data a GIS package can store. Common data types include character, integer, floating, and date.

Each common type may have subtypes.

The integer type can be short or long, and the floating type can be single or double.

Each field in an attribute table must be defined with a data type.

Another method is to define attribute data by measurement scale.

The measurement scale concept groups attribute data into

Nominal data

describe different kinds or different categories of data such as land-use types or soil types.

Ordinal data

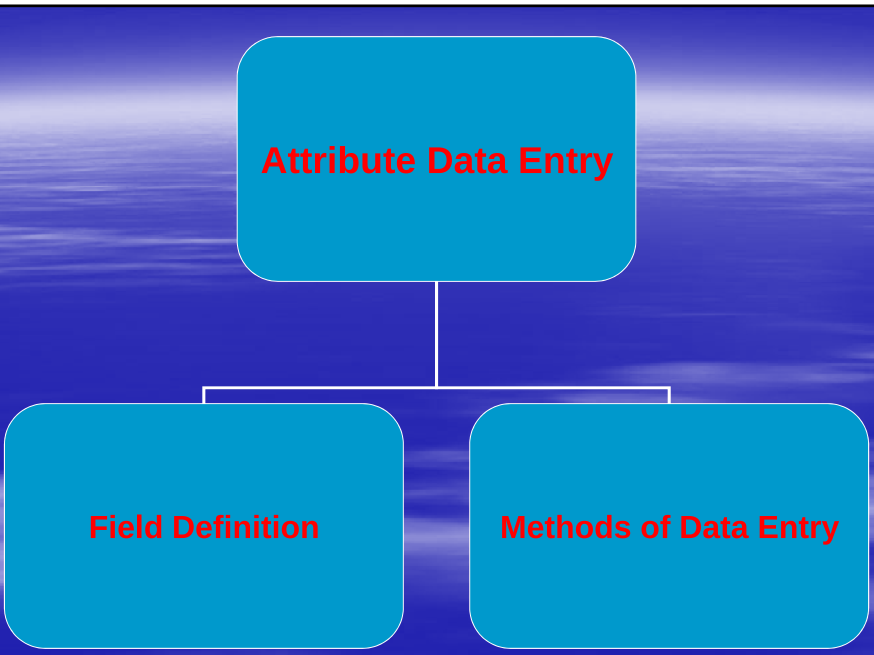
differentiate data by a ranking relationship, such as , soil erosion may be ordered from severe, moderate, to light)

Interval data

have known intervals between values, such as a temperature reading of 70°F is warmer than 60°F by 10°F

Ratio data

are the same as interval data except that ratio data are based on a meaningful, or absolute, zero value, such as population densities, because a density of 0 is an absolute zero).



Attribute Data Entry

Field Definition

Methods of Data Entry

Field Definition

The first step in attribute data entry is to define each field in the table.

A field definition usually includes

data width

The width refers to the number of spaces to be reserved for a field. The width should be large enough for the largest number or the longest string in the data. Spaces for the negative sign and the decimal point should also be included in the width.

data type

The data type must follow data types allowed in the GIS package.

number of decimal places.

The number of decimal places is part of' the definition for the real numeric data type.

Methods of Data Entry

Attribute data entry is like digitizing for spatial data entry.

One must enter attribute data by typing.

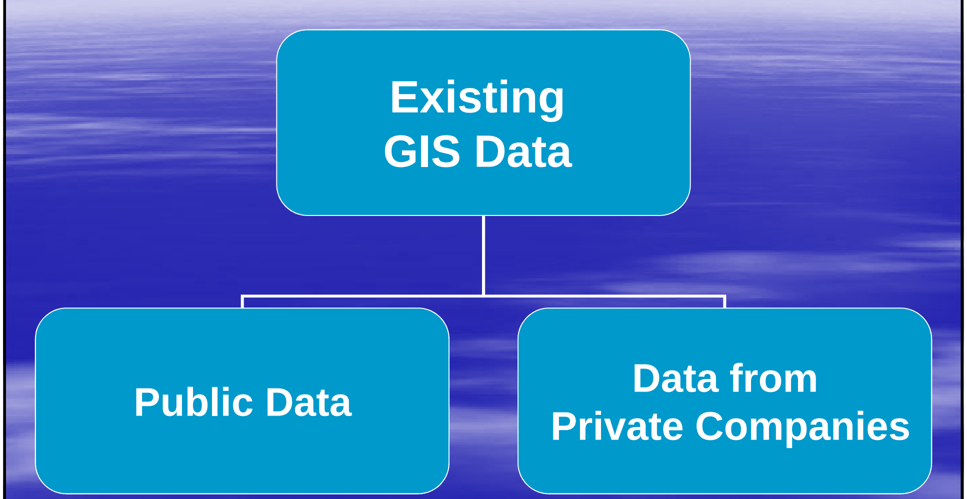
A map with 4000 polygons, each with 50 fields of attribute data, could require entering 200,000 values.

How to reduce time and effort in attribute data entry is of interest to any GIS user.

Input Data

The most expensive part of a GIS project is database construction.

Converting from paper maps to digital maps used to be the first step in constructing a database.



Existing GIS Data

Public Data

Data from Private Companies

Existing GIS Data.

- To find existing GIS data for a project is often a matter of knowledge, experience, and luck.
- Government agencies at different levels have set up websites for sharing public data and for directing users to the source of the desired information.

Public Data.

- Public data are often free and downloadable from the Internet.
- All levels of government let GIS users access their public data.
- To GIS users looking for data, the website maintained by the Federal Geographic Data Committee (FGDC) is a good start.
- One can download from the USGS website digital line graphs (DLGs) and land-use/land-cover data.

- Most GIS data on the Internet are data that many organizations regularly use for GIS activities.
- These are called framework data, which typically include seven basic layers:
  Geodetic control (accurate positional framework for surveying and mapping),
  Orthoimagery (rectified imagery such as orthophotos),
  Elevation,
  Transportation,
  Hydrography,
  Governmental units,
  Cadastral information.

Metadata.

- Metadata provide information about spatial data.
- They are particularly important to GIS users who want to use public data for their projects.
- First, metadata let GIS users know if the data meet their specific needs for area coverage, data quality, and data currency.
- Second, metadata show GIS users how to transfer, process, and interpret spatial data.
- Third, metadata include the contact information for GIS users who need more information.

- Metadata standards describe a data set based on the following:

1. Identification information (basic information about the data set, including title, geographic data covered, and currency).

2. Data quality information (information about the quality of the data set, including positional and attribute accuracy, completeness, consistency, sources of information, and methods used to produce the data).

3. Spatial data organization information (information about the data representation in the data set such as method for data representation (e.g., raster or vector) and number of spatial objects).

---

4. Spatial reference information (description of the reference frame for and means of encoding coordinates in the data set such as the parameters for map projections or coordinate systems, horizontal and vertical datums, and the coordinate system resolution).

5. Entity and attribute information (information about the content of the data set, such as the entity types and their attributes and the domains from which attribute values may be assigned).

6. Distribution information (information about obtaining the data set).

7. Metadata reference information (information on the currency of the metadata information and the responsible party).

Creating New Data.

A variety of data sources that can be used to create new digital spatial data.

1. Remotely Sensed Data.
2. Global Positioning System (GPS) Data.
3. Street Addresses.
4. Text Files with xy-Coordinates.
5. Survey Data.
6. Digitizing Using a Digitizing Table.

7 Scanning.

8. On-Screen Digitizing.

---

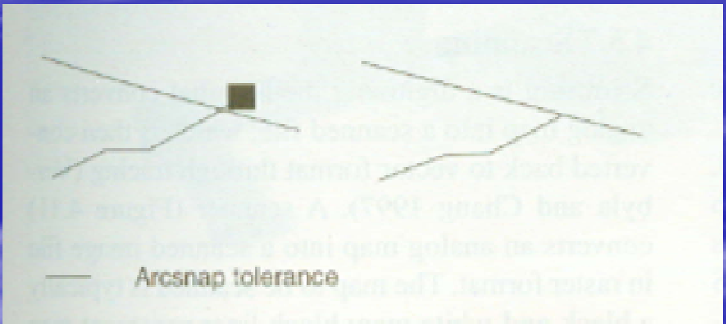
1. Remotely Sensed Data.

- Remotely sensed data such as digital orthophoto quads (DOQs) and satellite images are data acquired by a sensor from a distance.
- Although they are raster data, DOQs and satellite images are useful for vector data input. DOQs are digitized aerial photographs that have been differentially rectified or corrected to remove image displacements by camera tilt and terrain relief.
- Black-and-white DOQs have a 1-meter ground resolution and have pixel values representing 256 gray levels.
- GIS users can process satellite images and extract data for a variety of maps in vector format such as land use and land cover, hydrography, water quality, and areas of eroded soils.

6. Digitizing Using a Digitizing Table.

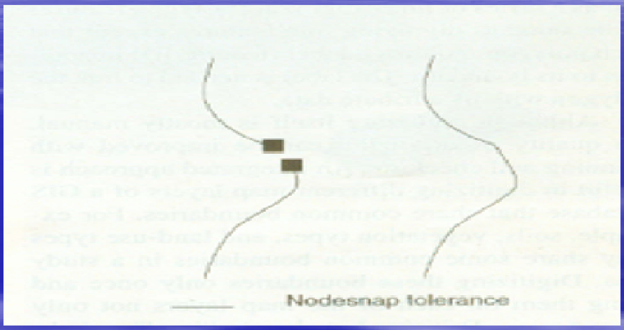
- Digitizing is the process of converting data from analog to digital format. Manual digitizing uses a digitizing table (Figure 6.2).
- A digitizing table has a built-in electronic mesh, which can sense the position of the cursor. To transmit the *x-,* y-coordinates of a point to the connected computer, the operator simply clicks on a button on the cursor after lining up the cursor's cross hair with the point. Large-size digitizing tables typically have an absolute accuracy of 0.001 inch (0.003 centimeter).

- Many GIS packages have a built-in digitizing module for manual digitizing.
- A topology-based GIS package has functionalities to ensure the topology of the digitized map.
- ARCSNAP allows the user to specify a distance within which an arc will be snapped to an existing arc.



The arcsnap tolerance allows the end of an arc to be snapped to an existing arc. The left diagram shows the overextension of a digitized arc. Because the overextension is smaller than the arcsnap tolerance, the end of the digitized arc is snapped to an existing arc, as shown in the right diagram.

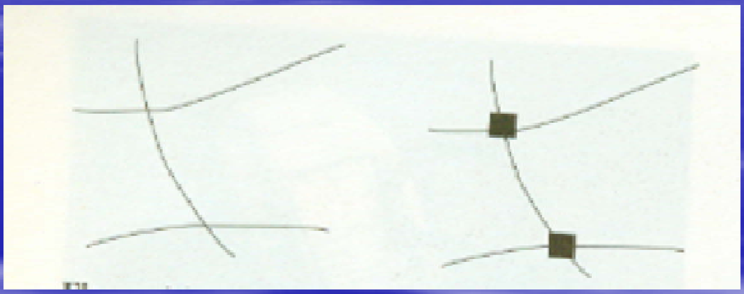
- NODESNAP works the same way as ARCSNAP except it snaps a node to an existing node.



The nodesnap tolerance allows the snapping of nodes.

The left diagram shows the digitized arc not reaching its intended end point. Because the gap between the nodes is smaller than the nodesnap tolerance, the end node of the arc is snapped to the other node.

- INTERSECT ARCS calculates arc intersections and adds nodes at inter-sections.

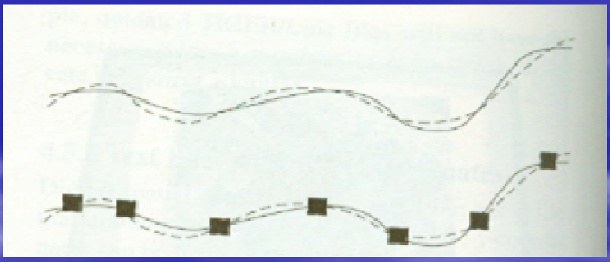


The intersect arcs command calculates arc intersections and adds nodes at the intersections.

The left diagram shows the arcs were digitized without adding nodes at the intersections. The right diagram shows the creation of nodes with the intersect arcs command.

- Although digitizing itself is mostly manual, the quality of digitizing can be improved with planning and checking.
- An integrated approach is useful in digitizing different map layers of a GIS database that share common boundaries. For example, soils, vegetation types, and land-use types may share some common boundaries in a study area.
- Digitizing these boundaries only once and using them on each of the map layers not only saves time in digitizing but also ensures the matching of the layers.

- A rule of thumb in digitizing line or polygon features is to digitize each line once and only once to avoid duplicate lines.
- Duplicate lines are seldom on top of one another because of the high accuracy of a digitizing table. In fact, duplicate lines form a series of polygons between them and are difficult to detect and correct in editing.



- One way to reduce the number of duplicate lines is to put a transparent sheet on top of the source map and to mark off each line on the transparent sheet after the line is digitized. This method can also re-duce the number of missing lines.

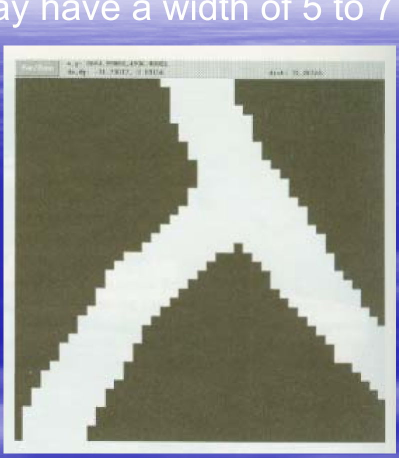
7 Scanning.

- Scanning is a digitizing method that converts an analog map into a scanned file, which is then converted back to vector format through tracing.
- A scanner converts an analog map into a scanned image file in raster format. The map to be scanned is typically a black-and-white map: black lines represent map features, and white areas represent the background. The map may be a paper map or a Mylar map, and inked or penciled. Scanning converts the map into a binary scanned file in raster format; each pixel has a value of either 1 (map feature) or 0 (back-ground).

- Map features are shown as raster lines, a series of connected pixels on the scanned file.
- The pixel size depends on the scanning resolution, which is often set at 300 dots per inch (dpi) or 400 dpi for digitizing.

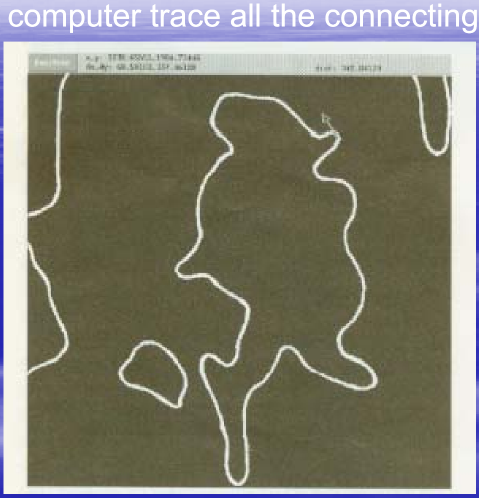
A binary scanned file: the lines are soil lines, and the black areas are the background.

- A raster line representing a thin inked line on the source map may have a width of 5 to 7 pixels.



A raster line in a scanned file has a width of several pixels.

- Tracing can be semiautomatic or manual. In semiautomatic mode, the user selects a starting point on the image map and lets the computer trace all the connecting raster lines.



- In manual mode, the user determines the raster line to be traced and the direction of tracing.

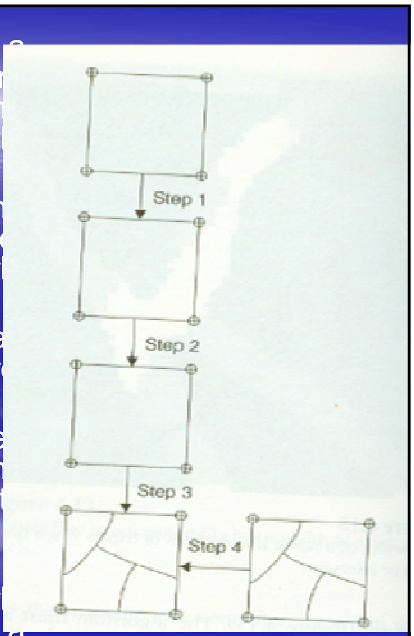
8. On-Screen Digitizing.

- On-screen digitizing is manual digitizing on the computer monitor using a data source such as a DOQ as the background.
- On-screen digitizing is an efficient method for editing or updating an existing map such as adding new trails or roads that are not on an existing map but are on a new DOQ. Likewise, the method can be used for editing a vegetation map based on new information from a new DOQ that shows recent clear-cuts or burned areas.

Geometric Transformation.

- A newly digitized map is measured in the same measurement unit as the source map used for digitizing or scanning.
- Before the digitized map can be used in a GIS project, it must be converted to real-world coordinates such as UTM coordinates.
- Geometric transformation is the process of converting a digital map from one coordinate system to another by using a set of control points and transformation equations.

- The process begins with digitizing set of control points, which ar points with known longitude an latitude values or real-wor coordinates.
- The next step is to convert th longitude and latitude readings the control points to real-wor coordinates.
- This step can be saved if the rea world coordinates of the contr points are already known.
- A map without map features is the copied from the digitized map, an its control points are updated t their real-world coordinates.
- Finally, the digitized map converted into real-wor coordinates through the use of a transformation method and the control points.



Root Mean Square (RMS) Error.

- Control points play a crucial role in geometric transformation.
- The quality of the control points with respect to their estimated and actual locations determines the accuracy of geometric transformation and the positional accuracy of digitized map features.
- One measure of the goodness of the control points is the root mean square (RMS) error, which measures the deviation between the actual location of the control points, as projected from their longitude and latitude readings into the output map, and the estimated location, as digitized on the input map.
- Typically, RMS errors are listed for each control point, and for the average from all control points.
- If the RMS error exceeds the specified tolerance, either the control points need to be redigitized or a new set of control points has to be entered.
- If the RMS is within the acceptable range, then the assumption is that this same level of accuracy can also be applied to map features, which are usually bounded within the control points.

Editing Data

- Spatial data editing refers to the process of adding, deleting, and modifying features in digital maps.
- A major part of spatial data editing is to remove digitizing errors.
- Digital maps from any source may contain errors from initial digitizing or from outdated data sources.
- Therefore, roads, land parcels, forest inventory, and other data require regular revision and updating.
- The updating process is basically the same as correcting errors on a newly digitized map.

- There are two types of digitizing errors.

1. Location errors such as missing polygons or distorted lines relate to inaccuracies of map features,
2. Topological errors such as dangling arcs and unclosed polygons relate to logical inconsis-tencies among map features.

To correct location errors, one often has to reshape individual arcs and digitize new arcs.

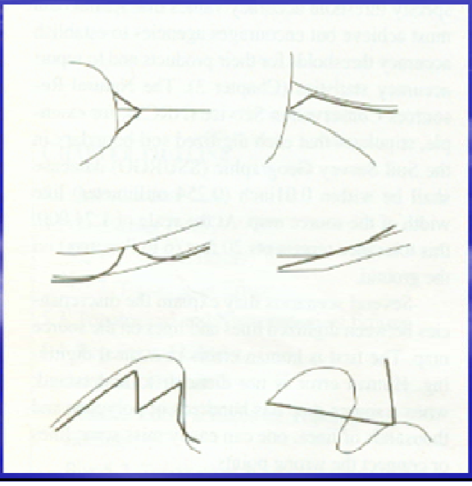
To correct topological errors, one must be knowledgeable about the topological relationships and use a topology-based GIS package to help make corrections.

- Location errors relate to the location of map features on a digitized map.
- Several scenarios may explain the discrepancies between digitized lines and lines on the source map.
- The first is human errors in manual digitizing.

Human error is not difficult to understand: when a source map has hundreds of polygons and thousands of lines, one can easily miss some lines or connect the wrong points.

- The second scenario consists of errors in scanning and tracing.

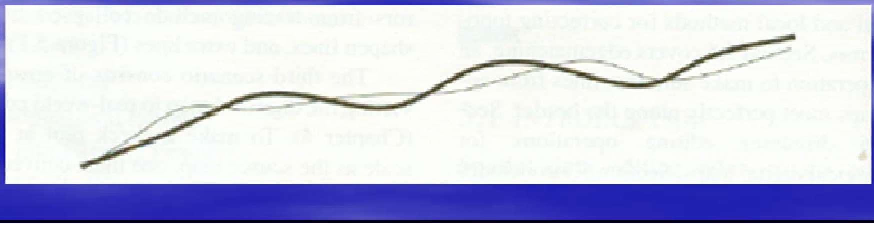
  A tracing algorithm usually has problems when raster lines meet or intersect, are too close together, are too wide, or are too thin and broken. Digitizing errors from tracing include collapsed lines, misshapen lines, and extra lines.



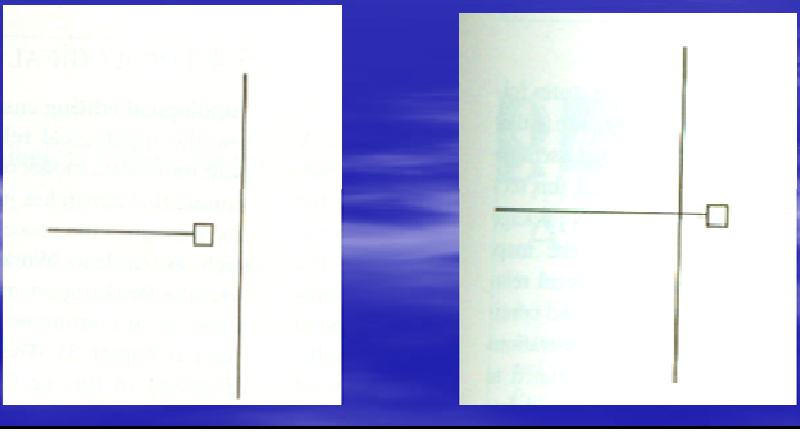
- The third scenario consists of errors in converting the digitized map to real-world coordinates.

  To make a check plot at the same scale as the source map, one must convert the digitized map to real-world coordinates by using a set of control points.

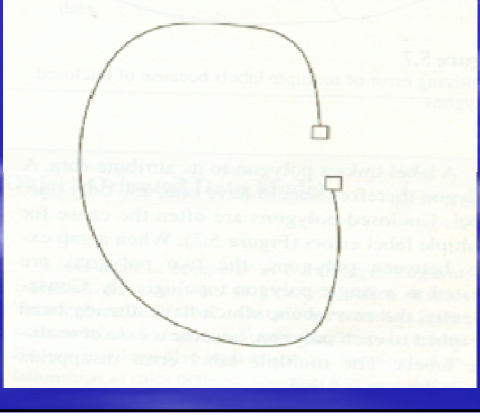
- Location errors also include duplicate lines from manual digitizing or tracing.
- Each arc on a line or polygon map is supposed to be digitized only once.
- Duplicate lines also occur frequently in tracing because semiautomatic tracing follows continuous lines even if some of the lines have already been traced.
- Duplicate lines are sometimes difficult to detect without zooming in.



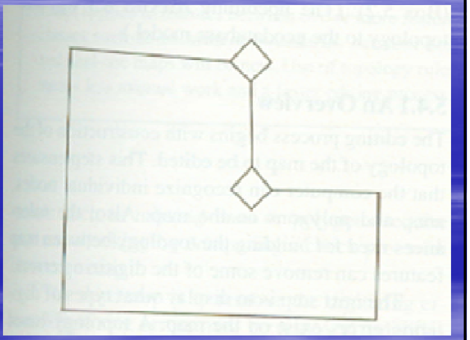
- Topological errors
- A common error occurs when arcs that are supposed to meet at a point (node) do not meet perfectly.
- This type of error becomes an undershoot if a gap exists between the arcs, and an overshoot if one arc is overextended.



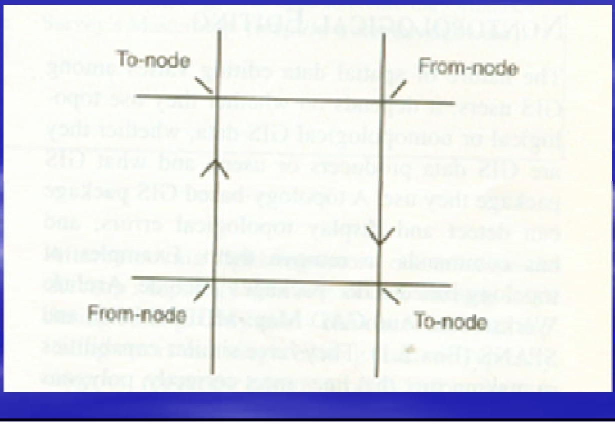
- The result of both cases is a dangling arc, which has the same polygon on its left and right sides, and a dangling node at the end of the arc.
- In general, it is easier to identify and remove an overshoot with its dangling node than an undershoot.
- Dangling nodes also occur when a polygon is not closed perfectly.



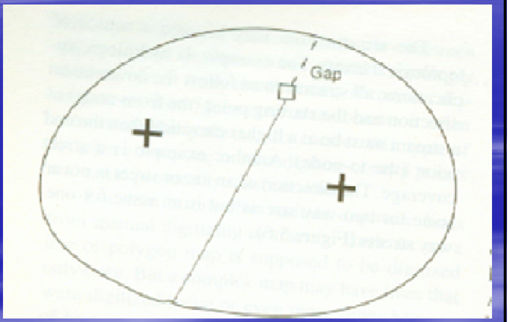
- A pseudo node appears along a continuous arc and divides the arc unnecessarily into separate arcs.
- Some pseudo nodes are, how-ever, acceptable or even necessary.
- For example, to show different segments of a road having different attributes such as speed limit, pseudo nodes may be placed at points where attribute values change.
- Another example applies to isolated polygons not connected to other polygons.
- An arbitrary starting point (and end point) must be chosen to digitize the polygon.
- That point then becomes a pseudo node.



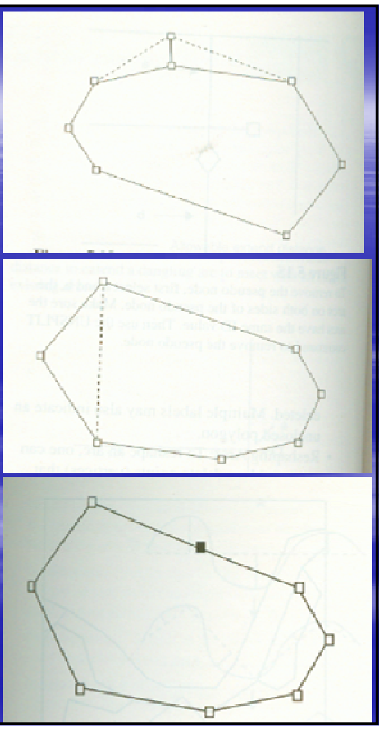
- The arc direction may become a source of topological errors.
- For example, in hydrologic applications, all streams must follow the downstream direction and the starting point (the from-node) of a stream must be at a higher elevation than the end point (the to-node).



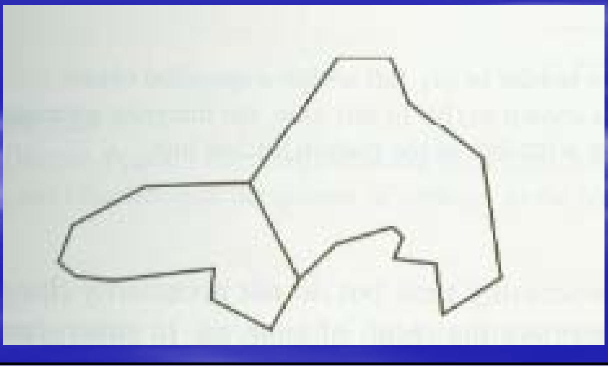
- A label links a polygon to its attribute data.
- A polygon therefore should have one, and only one, label.
- Unclosed polygons are often the cause for multiple label errors.
- When a gap exists between polygons, the two polygons are treated as a single polygon topologically.
- Consequently, the two labels, which have already been assigned to each polygon, become a case of multi-ple labels.
- The multiple label error disappears when the gap is filled.



- **Reshaping Features:** One can alter the shape of a line by moving, deleting, or adding points (also called vertices) on the line.
- The same method can be used to reshape a polygon.
- If the reshaping is intended for a polygon and its connected polygons, one must use a different tool (called the shared edit tool in ArcMap) so that, when a boundary is moved, all polygons that share the same boundaries are reshaped simultaneously.

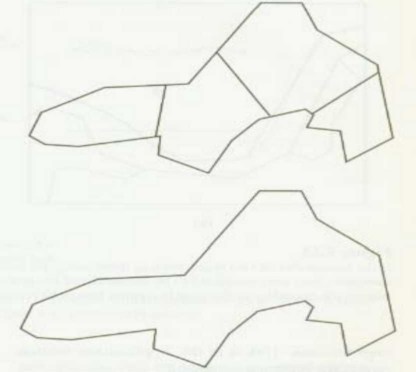


- Split Lines and Polygons: One can sketch a new line that crosses an existing line to split the line, or sketch a split line through a polygon to split the polygon.
- A polygon can be split into two by drawing a split line across the polygon boundary.



The following summarizes nontopological operations that create new features from existing features.

- **Merge Features:** One can group selected line or polygon features into one feature.
- If the merged features are not spatially adjacent, they form a feature with multiple parts.



- **Buffer Features:** One can create a buffer around a line or polygon feature at a specified distance.
- **Union Features:** One **can** combine features from different maps into one feature. This operation differs from the merge operation because it works with different maps.
- **Intersect Features:** One can create a new feature from the intersection of overlapped features from different maps.

Vector Data Analysis

The vector data uses points and their *X*-, *y*-coordinates to construct spatial features of points, lines, and polygons.

Vector data analysis is based on

the geometric objects of point, line, and polygon

the accuracy of analysis results depends on the accuracy of these objects in terms of location and shape.

Because vector data may be topology-based.

Topology can also be a factor for some vector data analysis.

1. Buffering.

Based on the concept of proximity

Buffering separates a map into two areas:

One area that is within a specified distance of selected map features (is called the buffer zone)

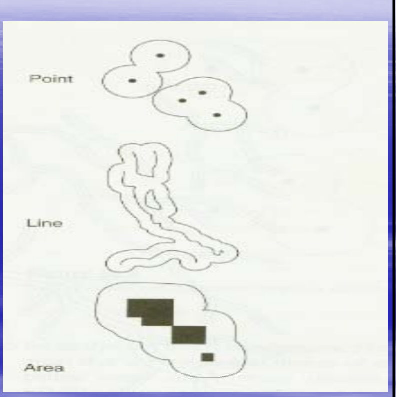
The other area that is beyond.

Selected map features for buffering may be points, lines, or areas.

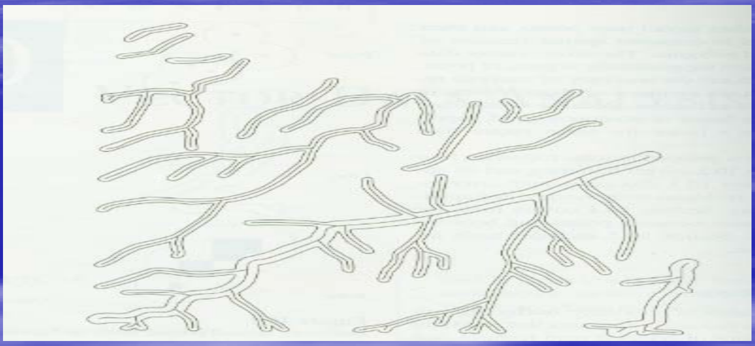
Buffering around points creates circular buffer zones.

Buffering around lines creates a series of elongated buffer zones.

Buffering around polygons creates buffer zones, extending outward from the polygon boundaries.



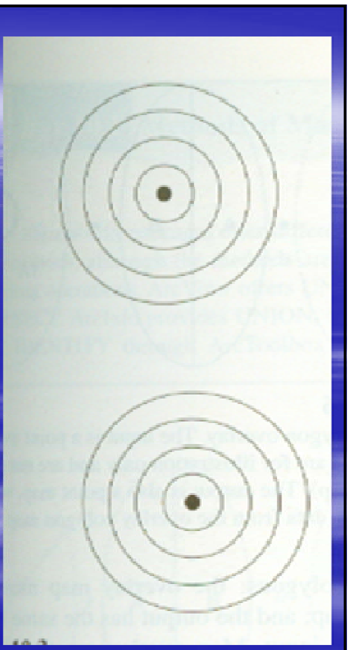
The buffer distance or buffer size does not have to be constant; it can vary according to the values of a given field.

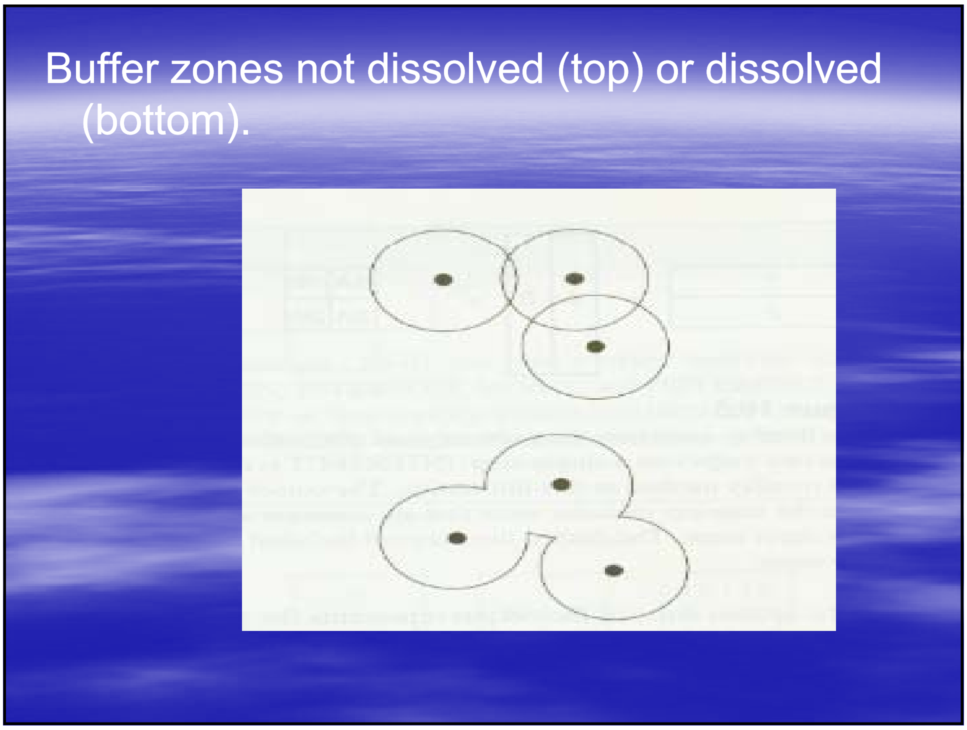


For example, stream buffer sizes may vary depending on the intensity of adjacent land use.

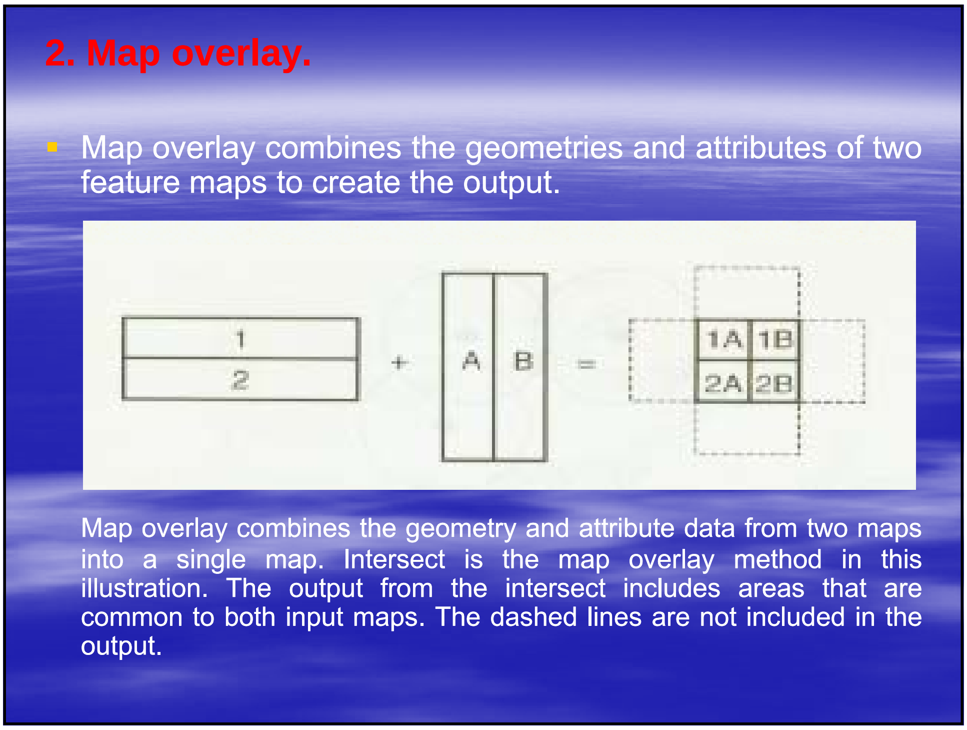
A map feature may have more than one buffer zone.

For example, a nuclear power plant may be buffered with distances of 5, 10, 15, and 20 miles, thus forming multiple rings around the plant.





Buffer zones not dissolved (top) or dissolved (bottom).



2. Map overlay.

- Map overlay combines the geometries and attributes of two feature maps to create the output.

Map overlay combines the geometry and attribute data from two maps into a single map. Intersect is the map overlay method in this illustration. The output from the intersect includes areas that are common to both input maps. The dashed lines are not included in the output.

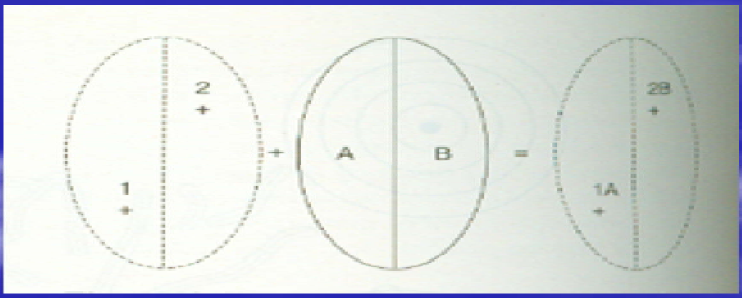
- One of the two maps is called the input map and the other the overlay map.
- The geometry, or the spatial data, of the output represents the geometric intersection of map features from the input and overlay maps.
- Therefore, the number of map features on the output map is not the sum of map features on the input and overlay maps but is usually much larger than the sum.
- Each map feature on the output contains a combination of attributes from the input and overlay maps, and this combi-nation differs from its neighbors.

- Map overlay can work with only two polygon maps at a time.
- For instance, if three polygon maps are to be overlaid, the operation begins with the first two and then uses the output from them with the third.
- The process, especially the tracking of the intermediate outputs, can be tedious.
- The problem of using only two maps in an overlay operation disappears with the regions data model, which allows overlaying more than two layers in a single operation.

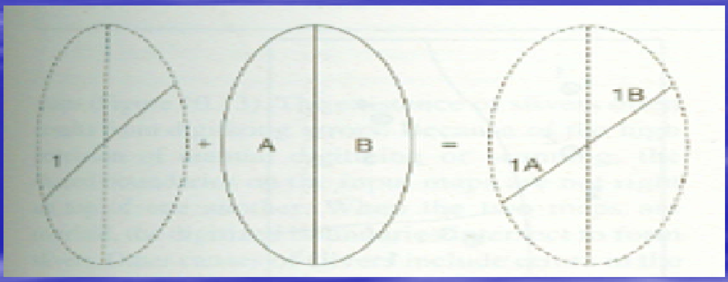
2.1. Feature Type and Map Overlay.

- The first consideration of map overlay is feature type.
- The input map may contain points, lines, or polygons; the overlay map must be a polygon map; and the output has the same feature type as the input.
- Map overlay can therefore be grouped by feature type into point-in-polygon, line-in-polygon, and polygon-on-polygon.

- In a point-in-polygon operation, the same point features in the input are included in the output but each point is assigned with attributes of the polygon within which it falls.

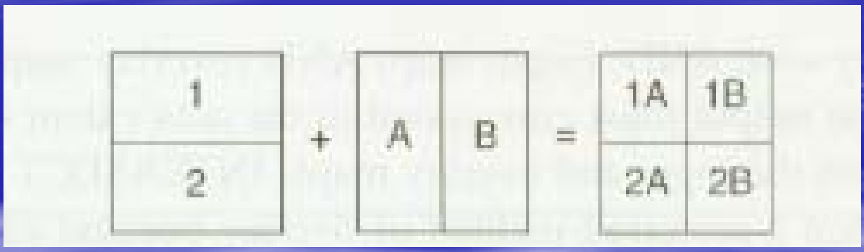


- In a line-in-polygon operation, the output contains the same line features as in the input but each of them is dissected by the polygon bound-aries on the overlay map.



- Thus the output has more line segments than does the input, each segment on the output combines attributes from the line map and the polygon within which it falls.

- The most common overlay operation is polygon-on-polygon, involving two polygon maps. The output combines the polygon bound-aries from the input and overlay maps to create a new set of polygons.



- Each new polygon carries attributes from both maps, and these attributes differ from those of adjacent poly-gons

2.2. Map Overlay Methods.

- If the input map has the same area extent as the overlay map, then that area extent also applies to the output.

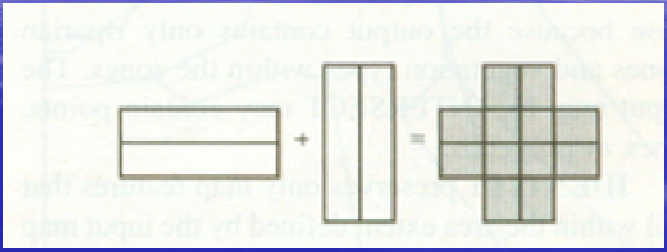
- But, if the input map has a different area extent than the overlay map, then the area ex-tent of the output may vary depending on the overlay method used.

- The three common map overlay methods are called
  1. UNION.

     2. INTERSECT

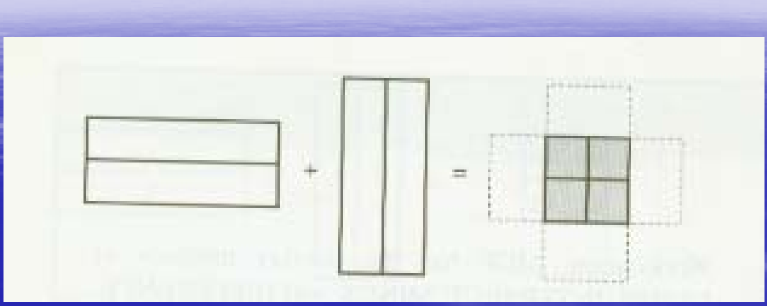
     3. IDENTITY.

- Union preserves all map features from the input and overlay maps by combining the area extents from both maps.



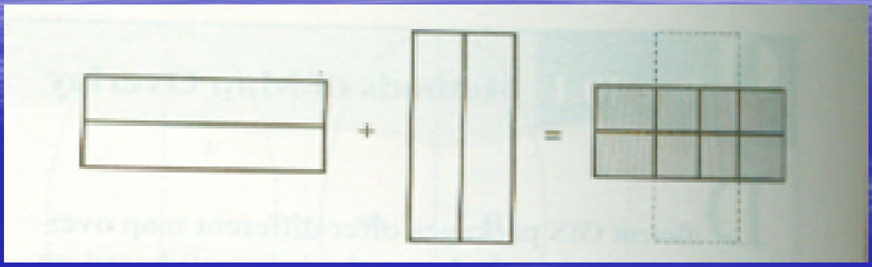
- It is the Boolean operation that uses the keyword OR: (in-put map) OR (overlay map).
- The output corresponds to the area extent of the input map, or the overlay map, or both union requires that both the input and overlay maps be polygon maps.

- Intersect preserves only those features that fall within the area extent common to both the input and overlay maps.



- Intersect is the Boolean operation that uses the keyword AND: (input map) AND (overlay map).
- The output must correspond to the area extent of both the input and overlay maps. Intersect is often a preferred method of overlay because any map feature on its output has attribute data from both of its inputs.

- Identity preserves only map features that fall within the area extent defined by the input map.



- Expressed in a Boolean expression, IDENTITY represents the operation: [(input map) AND (overlay map)] OR (input map). Features of the overlay map that fall outside the area extent of the input map are left out of the output. The input map for identity may contain points, lines, or polygons.

3. Distance Measurement.

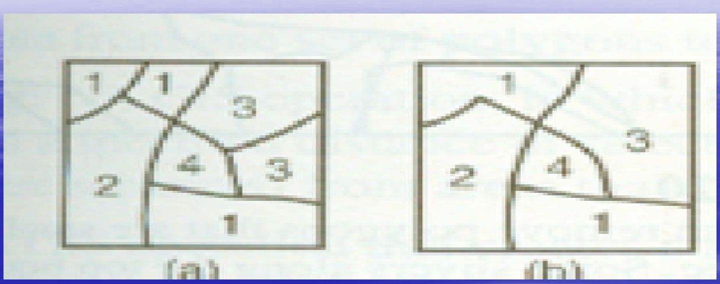
- Distance measurement refers to measuring straight-line (Euclidean) distances between points, or between points and their corresponding nearest points or lines.
- Euclidean distance is one type of distance measure used in GIS.
- Distance measures can be used directly in data analysis.

---

4. Map Manipulation.

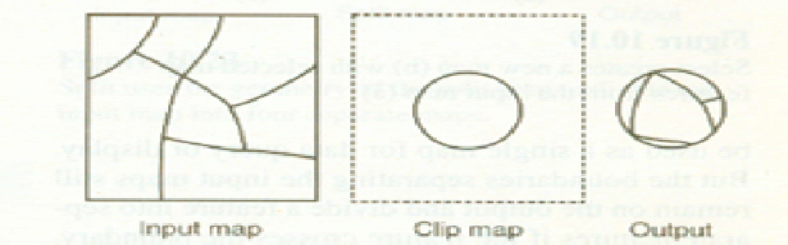
- Many GIS packages provide tools for manipulating and managing maps in a database. Like buffering and map overlay, these tools are considered basic GIS tools often needed for data preprocessing and data analysis.
- Map manipulation is easy to follow graphically, even though terms describing the various tools may differ between GIS packages.
- Most tools covered in this section involve two maps and are sometimes classified as map overlay methods.
- however, limits map overlay to only those methods that can combine spatial and attribute data from two maps into a single map.

- **Dissolve** aggregates map features that have the same value of a selected attribute.

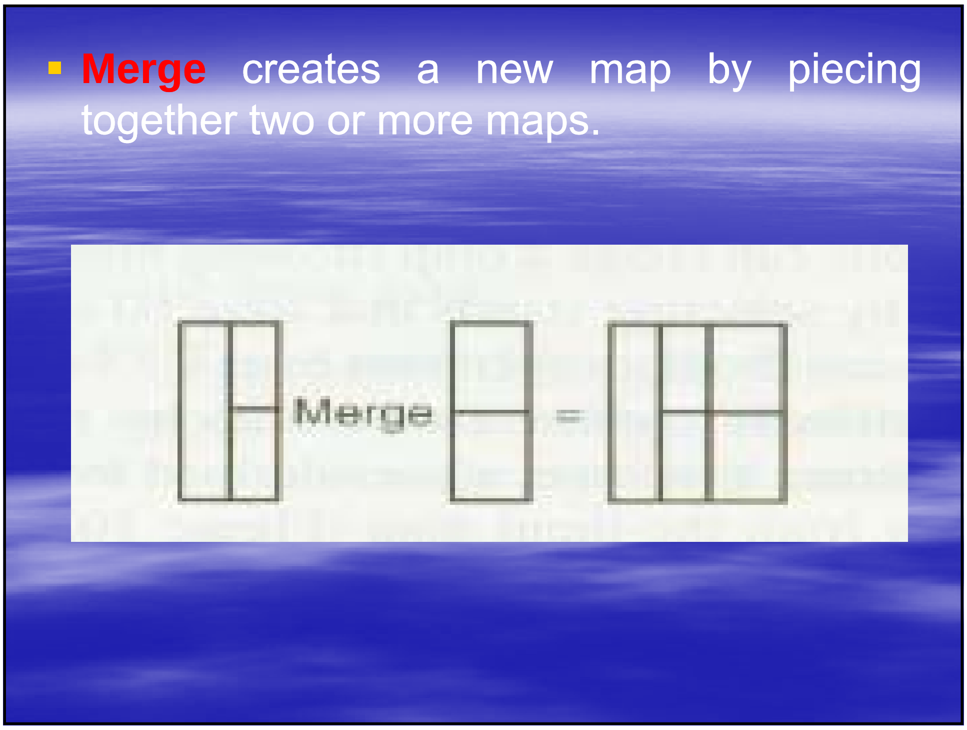


- An important application of dissolve is to simplify a classified polygon map. Classification groups values of a selected attribute into classes and makes obsolete boundaries of adjacent polygons, which have dif-ferent values initially but are now grouped into the same class. Dissolve can remove these unnecessary boundaries and creates a new, simpler map with the classification results as its attribute values. Another application of dissolve is to aggregate both spatial and attribute data of the input map.

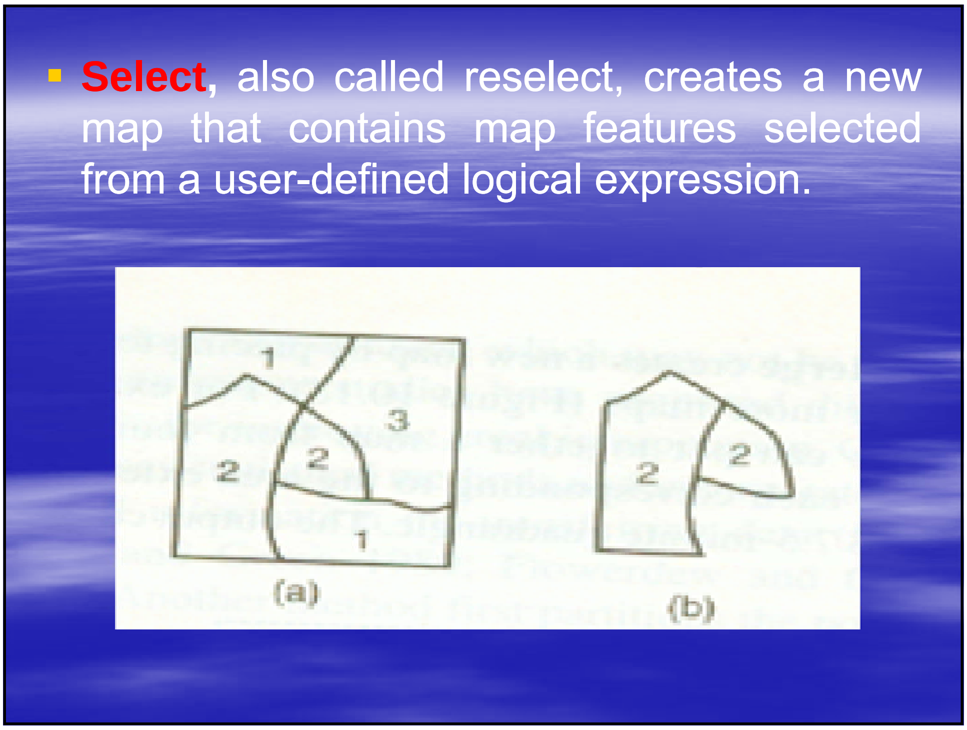
- **Clip** creates a new map that includes only those features of the input map that falls within the area extent of the clip map.



- Clip is a useful tool, for example, for cutting a map acquired elsewhere to fit a study area. The input may be a point, line, or polygon map, but the clip map must be a polygon map.



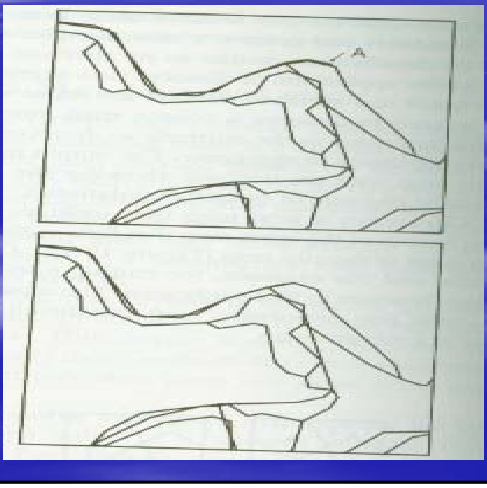
- **Merge** creates a new map by piecing together two or more maps.



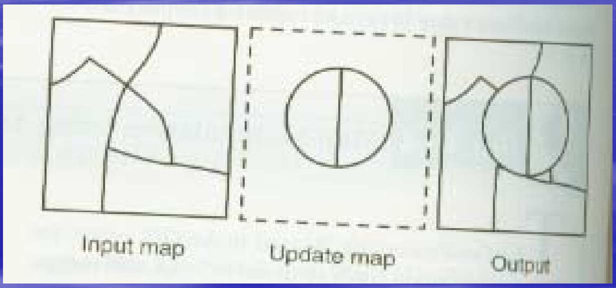
- **Select**, also called reselect, creates a new map that contains map features selected from a user-defined logical expression.

(a) (b)

- **Eliminate** creates a new map by removing map features that meet a user-defined logical expression from the input map.

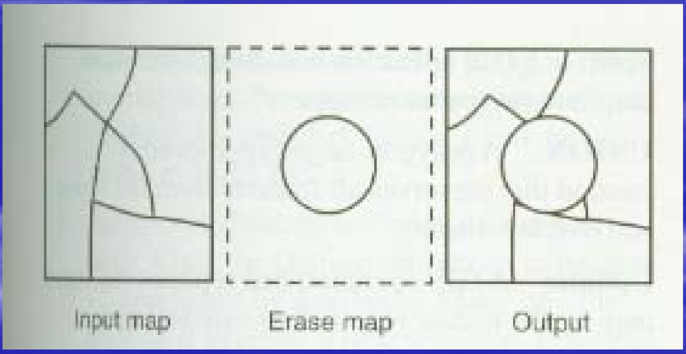


- **Update** uses a "cut and paste" operation to replace the input map by the update map and its map features.

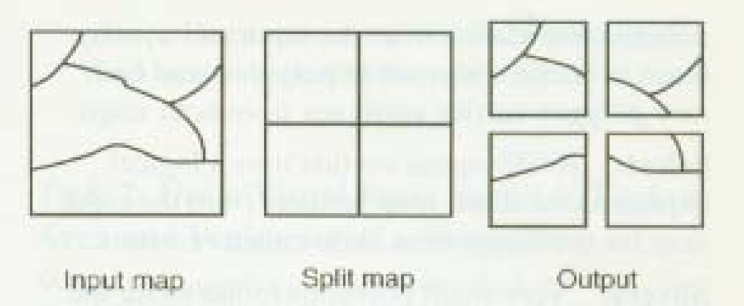


- As the name suggests, up-date is useful for updating an existing map with new map features in limited areas of the map. It saves the GIS user from redigitizing the map.

- **Erase** removes from the input map those map features that fall within the area extent of a map named as the erase map.



- **Split** divides the input map into two or more maps.



- A split map, which shows area subunits, is used as the template for dividing the input map.