UNIT - 2
SPATIAL DATA MODELS

2.1. INTRODUCTION

Spatial data are what drive a GIS. Every functionality that makes a GIS separate from another analytical environment is rooted in the spatially explicit nature of the data.

Spatial data are often referred to as layers, coverage's, or layers. We will use the term layers from this point on, since this is the recognized term used in Arc-GIS. Layers represent, in a special digital storage format, features on, above, or below the surface of the earth. Depending on the type of features they represent, and the purpose to which the data will be applied, layers will be one of two major types.

a) Vector data represent features as discrete points, lines, and polygons.

b) Raster data represent the landscape as a rectangular matrix of square cells.

Depending on the type of problem that needs to be solved, the type of maps that need to be made, and the data source, either raster or vector, or a combination of the two can be used. Each data model has strengths and weaknesses in terms of functionality and representation. As you get more experience with GIS, you will be able to determine which data type to use for a particular application.

2.2. DATABASE STRUCTURES

The two basic data structures in any fully-functional GIS are:

Vector, e.g,

- ArcInfo Coverages
- ArcGIS Shape Files
- CAD (AutoCAD DXF & DWG, or Micro Station DGN files)
- ASCII coordinate data

Raster, e.g,

- ArcInfo Grids
- Images

- Digital Elevation Models (DEMs)
- generic raster datasets

2.3. DATA STRUCTURE MODELS

Data models are the conceptual models that describe the structures of databases. The structure of a database is defined by the data types, the constraints and the relationships for the description or storage of data. Following are the most often used data models:

1) Hierarchical Data Structure Model
2) Network Data Structure Model
3) Relational Data Structure Model
4) Object Oriented Data Structure Model

2.3.1. Hierarchical Data Structure Model

It is the earliest database model that is evolved from file system where records are arranged in a hierarchy or as a tree structure shown in the **figure.2.1.** Records are connected through pointers that store the address of the related record.

Each pointer establishes a parent-child relationship where a parent can have more than one child but a child can only have one parent. There is no connection between the elements at the same level. To locate a particular record, you have to start at the top of the tree with a parent record and trace down the tree to the child.

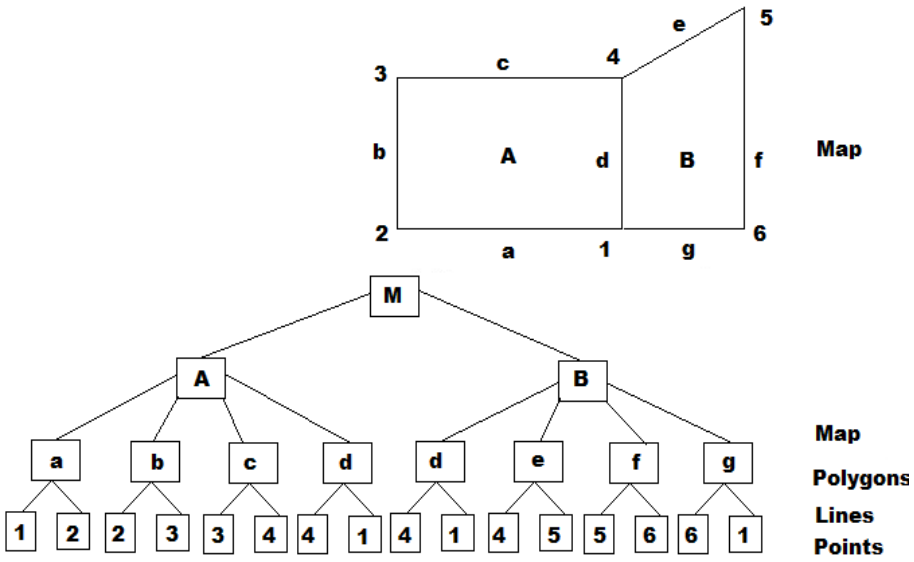


Fig.2.1 Hierarchical Database Structure Based on a Simple Map

Advantages

- Easy to understand: The organization of database parallels a family tree understanding which is quite easy.

- Accessing records or updating records are very fast since the relationships have been predefined.

Disadvantages

- Large index files are to be maintained and certain attribute values are repeated many times which lead to data redundancy and increased storage.
- The rigid structure of this model doesn't allow alteration of tables, therefore to add a new relationship entire database is to be redefined.

2.3.2. Network Data Structure Model

A network is a generalized graph that captures relationships between objects using connectivity shown figure.2.2. A network database consists of a collection of records that are connected to each other through links. A link is an association between two records. It allows each record to have many parents and many children thus allowing a natural model of relationships between entities.

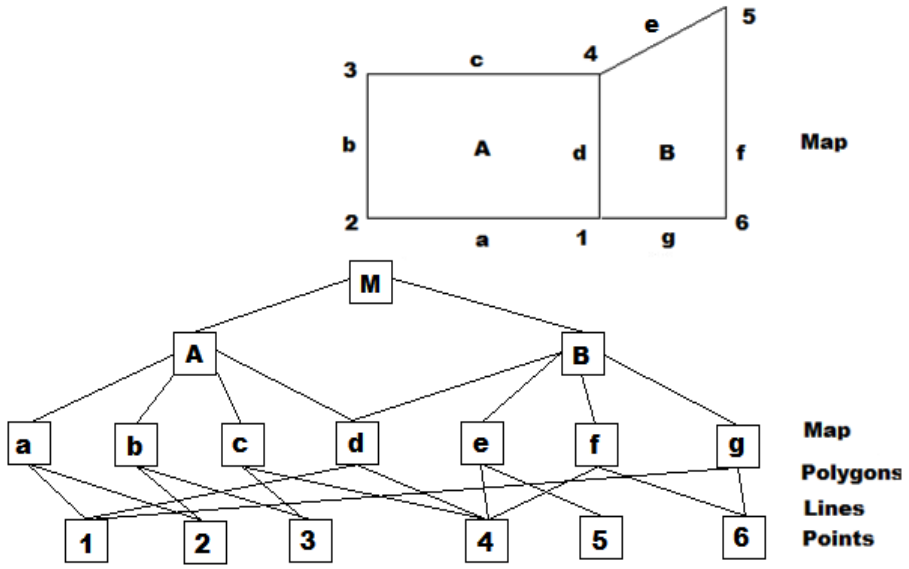


Fig.2.2. Network Data Structure Model

Advantages

- The many too many relationships are easily implemented in a network data model.
- Data access and flexibility in network model is better than that in hierarchical model. An application can access an owner record and the member records within a set.
- It enforces data integrity as a user must first define owner record and then the member records.
- The model eliminated redundancy but at the expense of more complicated relationships.

2.3.3. Relational Data Structure Model

- The relational data model was introduced by Codd in 1970. The relational database relates or connects data in different files through the use of a common field.
- A flat file structure is used with a relational database model. In this arrangement, data is stored in different tables made up of rows and columns as shown in **figure.2.3.**
- The columns of a table are named by attributes. Each row in the table is called a tuple and represents a basic fact.
- No two rows of the same table may have identical values in all columns.

Advantages

- The manager or administrator does not have to be aware of any data structure or data pointer. One can easily add, update, delete or create records using simple logic.

Disadvantages

- A few search commands in a relational database require more time to process compared with other database models.

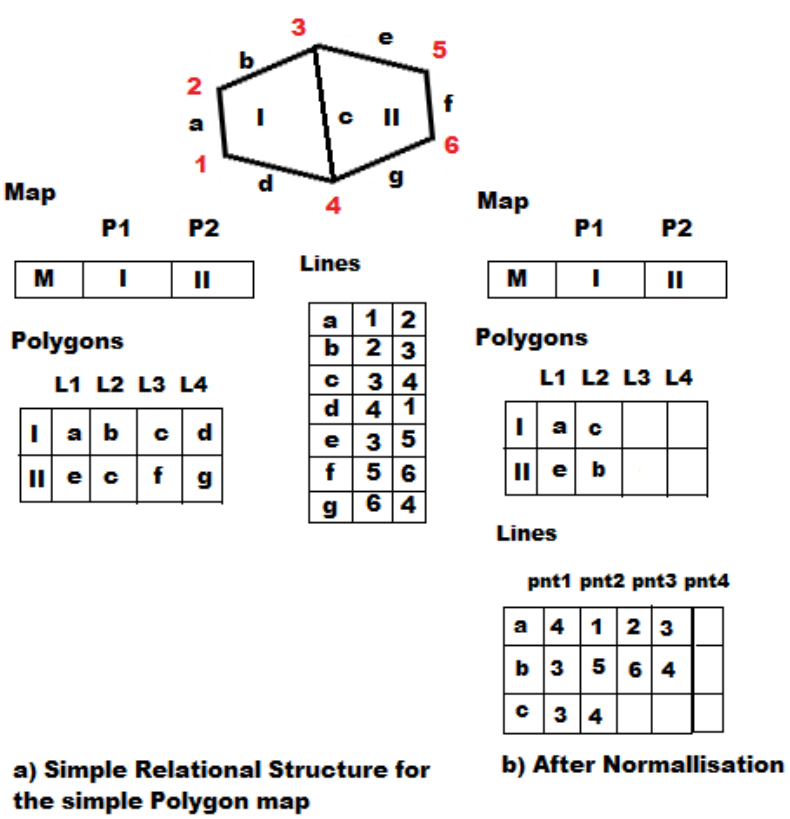


Fig.2.3. Relational Data Structure Model

2.3.4. Object Oriented Database Structure

- An Object Oriented model uses functions to model spatial and non-spatial relationships of geographic objects and the attributes.
- An object is an encapsulated unit which is characterized by attributes, a set of orientations and rules.

An object-oriented model has the following characteristics.

Generic Properties: there should be an inheritance relationship.

Abtraction: objects, classes and super classes are to be generated by classification, generalisation, association and aggregation.

Adhoc Queries: users can order spatial operations to obtain spatial relationships of geographic objects using a special language.

- **For example,** let us try to represent a thought: "Hawaii is an island that is a state of USA" in GIS. In this case, we don't mind the geographic location with latitude and longitude in the conventional GIS model. This is not appropriate to use the layers. In an object-oriented model, we are more careful with spatial relationships for example, "is a" (the island is a land) and "part of" (the state is a part of the country).

- In addition, Hawaii (state) has Honolulu City and also is in Pacific Region. Figure **2.4 (a)** shows "is an" inheritance for the super class of land, while Figure **2.4 (b)** shows the spatial relationships for the object of the state.

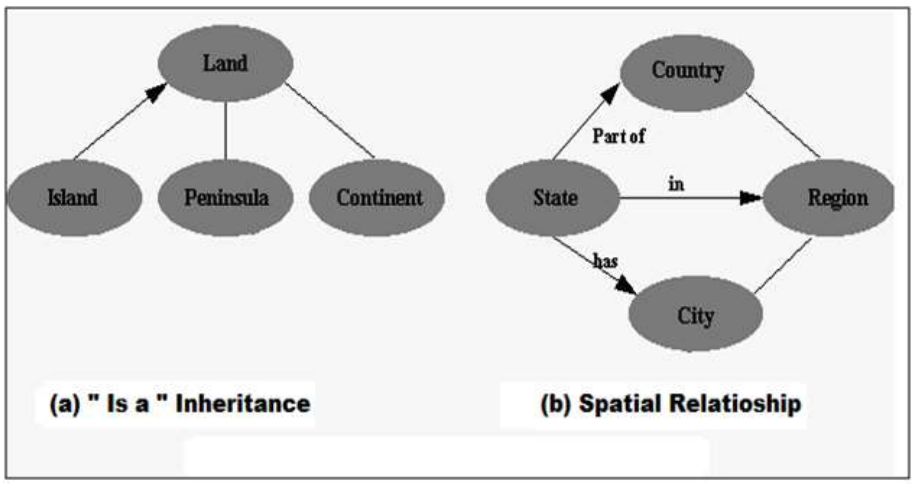


Fig.2.4. Object Oriented Database Structure

An object-oriented database is based on a semantic model as shown in **Figure 2.5.** Which is usually managed by a spatial language although the language has not yet been fully completed.

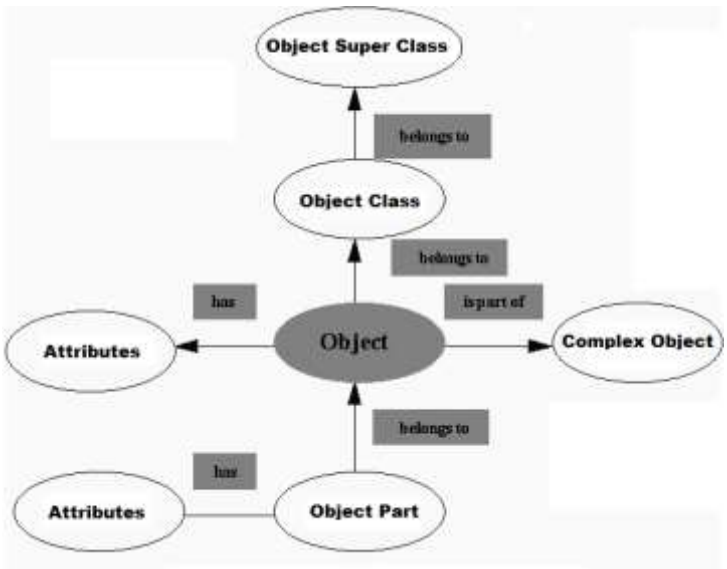


Fig.2.5. Object-Oriented Database is based on a Semantic Model

2.4. DEFINITION - WHAT DOES ENTITY-RELATIONSHIP DIAGRAM (ERD)

An entity-relationship diagram (ERD) is a data modeling technique that graphically illustrates an information system's entities and the relationships between those entities. An ERD is a conceptual and representational model of data used to represent the entity framework infrastructure.

The elements of an ERD are:

- Entities
- Relationships
- Attributes

Steps involved in creating an ERD include:

1) Identifying and defining the entities
2) Determining all interactions between the entities
3) Analyzing the nature of interactions/determining the cardinality of the relationships
4) Creating the ERD

2.4.1. Entity-Relationship Diagram (ERD)

An entity-relationship diagram (ERD) is crucial to creating a good database design. It is used as a high-level logical data model, which is useful in developing a conceptual design for databases.

An entity is a real-world item or concept that exists on its own. Entities are equivalent to database tables in a relational database, with each row of the table representing an instance of that entity.

An attribute of an entity is a particular property that describes the entity. A relationship is the association that describes the interaction between entities. Cardinality, in the context of ERD, is the number of instances of one entity that can, or must, be associated with each instance of another entity. In general, there may be one-to-one, one-to-many, or many-to-many relationships.

For example, let us consider two real-world entities, an employee and his department. An employee has attributes such as an employee number, name, department number, etc. Similarly, department number and name can be defined as attributes of a department. A department can interact with many employees, but an employee can belong to only one department, hence there can be a one-to-many relationship, defined between department and employee.

In the actual database, the employee table will have department number as a foreign key, referencing from department table, to enforce the relationship.

2.4.2. E-R DIAGRAM

E-R Diagram example from Database Management course

As mentioned before, the entity-relationship (E-R) diagram is one of the most commonly implemented conceptual data models used with GIS. Entities, attributes, and relationships are used to represent real-world features, what their properties are, and what the relationships are between these entities. Hardware and software issues are not explored in the E-R Diagram. These are addressed later in logical and physical data models. This first level of data abstraction is used by geospatial analysts as a starting point when analyzing and assessing the data available to them and how it fits together. The example below illustrates an E-R Diagram built during one of my Geospatial Data Structures course assignments.

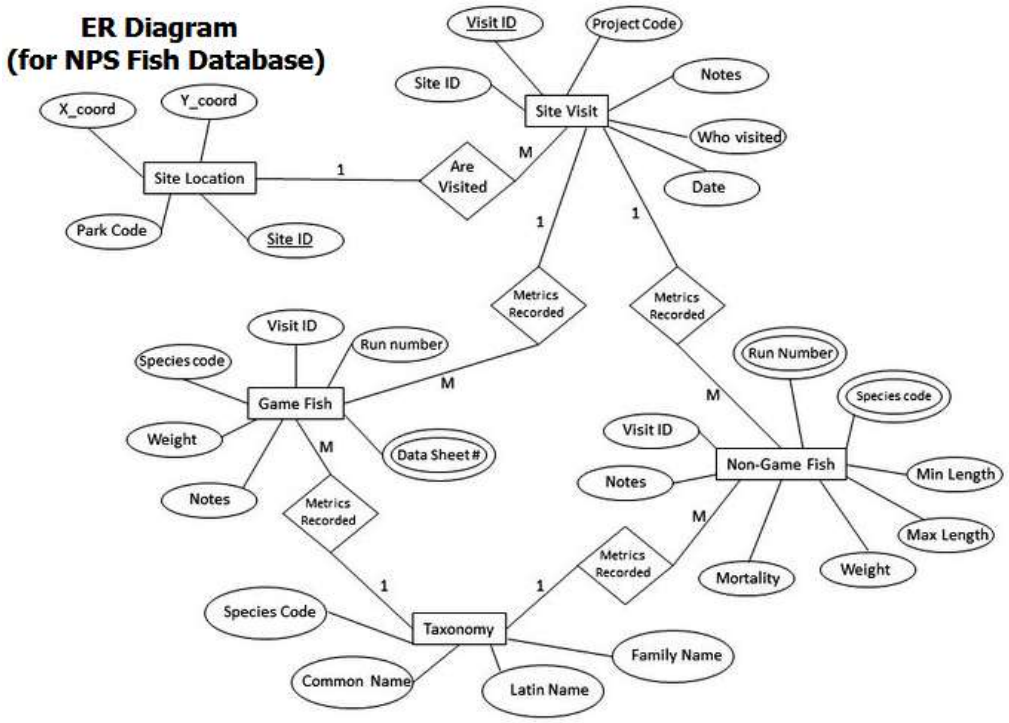
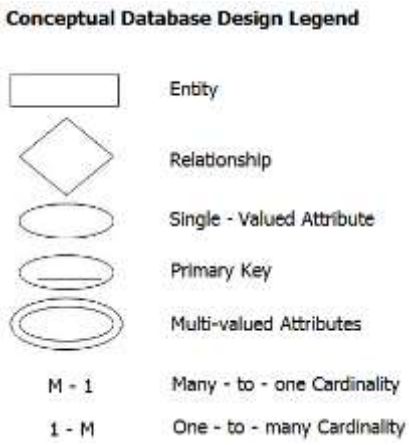


Fig.2.6. E-R Diagram



You could use this database to query how many of a particular species of game fish were examined at a specific park during a data range of interest. This would be non-spatial query because we are just counting an occurrence at one particular location. We are not using the coordinates to perform some type of buffer analysis or other spatial analysis to query the data.

2.5. SPATIAL DATA MODELS

Computers and GIS cannot directly be applied to the real world: a data gathering step comes first. Digital computers operate in numbers and characters held internally as binary digits. The real-world phenomenon of interest must be represented in symbolic form. The abstraction process of representing any property of the earth's surface in a computer accessible form involves the use of symbolic models.

Models are simplification of reality. A map is a symbolic model, because it is a simplified representation of part of the real world. The components of the model are spatial objects, approximating spatial entities of the real world; they are represented on the map by graphical symbols.

- The process of defining and organizing data about the real world into a consistent digital dataset that is useful and reveals information is called data modeling.
- The logical organization of data according to a scheme is known as data models
- **Data** can be defined as verifiable facts.
- **Information** is data organized to reveal patterns, and to facilitate search.
- **Spatial information** is difficult to extract from spatial data, unless the data are organized primarily by spatial attributes.
- **Spatial objects** are characterized by attributes that are both spatial and non-spatial, and the digital description of objects and their attributes comprise spatial datasets.
- **Spatial data** can be organized in different ways, depending on the way they are collected, how they are stored, and the purpose they are put.
- **A database** is a collection of inter-related data and everything that is needed to maintain and use it.

- A Database Management System is a collection of software for storing, editing and retrieving data in a database.

Traditionally spatial data has been stored and presented in the form of a map. Three basic types of spatial data models have evolved for storing geographic data digitally. These are referred to as:

- Vector;
- Raster;
- Image.

The following diagram reflects the two primary spatial data encoding techniques. These are vector and raster. Image data utilizes techniques very similar to raster data, however typically lacks the internal formats required for analysis and modeling of the data. Images reflect pictures or photographs of the landscape.

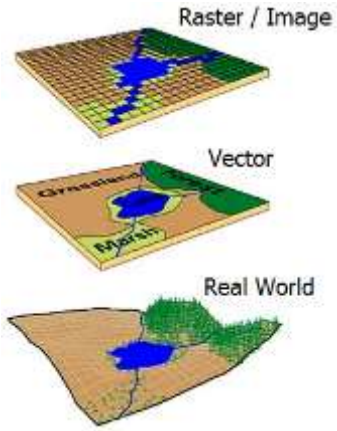


Fig.2.7. Two primary spatial data encoding techniques

2.5.1. Vector Data Formats

All spatial data models are approaches for storing the spatial location of geographic features in a database. Vector storage implies the use of vectors (directional lines) to represent a geographic feature. Vector data is characterized by the use of sequential points or vertices to define a linear segment. Each vertex consists of an X coordinate and a Y coordinate.

Vector lines are often referred to as arcs and consist of a string of vertices terminated by a node. A node is defined as a vertex that starts or ends an arc segment. Point features are defined by one coordinate pair, a vertex. Polygonal features are defined by a set of closed coordinate pairs. In vector representation, the storage of the vertices for each feature is important, as well as the connectivity between features, e.g. the sharing of common vertices where features connect.

Several different vector data models exist, however only two are commonly used in GIS data storage.

The most popular method of retaining spatial relationships among features is to explicitly record adjacency information in what is known as the topologic data model. Topology is a mathematical concept that has its basis in the principles of feature adjacency and connectivity.

The topologic data structure is often referred to as an intelligent data structure because spatial relationships between geographic features are easily derived when using them. Primarily for this reason the topologic model is the dominant vector data structure currently used in GIS technology. Many of the complex data analysis functions cannot effectively be undertaken without a topologic vector data structure. Topology is reviewed in greater detail later on in the book.

The secondary vector data structure that is common among GIS software is the computer-aided drafting (CAD) data structure. This structure consists of listing elements, not features, defined by strings of vertices, to define geographic features, e.g. points, lines, or areas. There is considerable redundancy with this data model since the boundary segment between two polygons can be stored twice, once for each feature. The CAD structure emerged from the development of computer graphics systems without specific considerations of processing geographic features. Accordingly, since features, e.g. polygons, are self-contained and independent, questions about the adjacency of features can be difficult to answer. The CAD vector model lacks the definition of spatial relationships between features that is defined by the topologic data model.

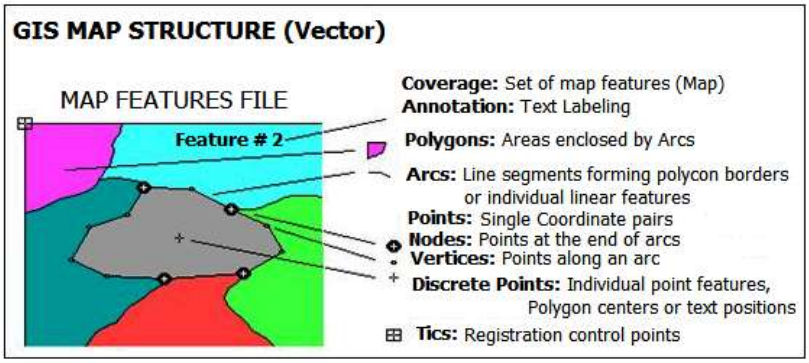


Fig.2.8. GIS MAP Structure - VECTOR systems (Adapted from Berry)

2.5.2. Raster Data Formats

Raster data models incorporate the use of a grid-cell data structure where the geographic area is divided into cells identified by row and column. This data structure is commonly called raster. While the term raster implies a regularly spaced grid other tessellated data structures do exist in grid based GIS systems. In particular, the quadtree data structure has found some acceptance as an alternative raster data model.

The size of cells in a tessellated data structure is selected on the basis of the data accuracy and the resolution needed by the user. There is no explicit coding of geographic coordinates required since that is implicit in the layout of the cells. A raster data structure is in fact a matrix where any coordinate can be quickly calculated if the origin point is known, and the size of the grid cells is known. Since grid-cells can be handled as two-dimensional arrays in computer encoding many analytical operations are easy to program. This makes tessellated data structures a popular choice for many GIS software. Topology is not a relevant concept with tessellated

structures since adjacency and connectivity are implicit in the location of a particular cell in the data matrix.

Several tessellated data structures exist, however only two are commonly used in GIS's. The most popular cell structure is the regularly spaced matrix or raster structure. This data structure involves a division of spatial data into regularly spaced cells. Each cell is of the same shape and size. Squares are most commonly utilized.

Since geographic data is rarely distinguished by regularly spaced shapes, cells must be classified as to the most common attribute for the cell. The problem of determining the proper resolution for a particular data layer can be a concern. If one selects too coarse a cell size then data may be overly generalized. If one selects too fine a cell size then too many cells may be created resulting in a large data volume, slower processing times, and a more cumbersome data set. As well, one can imply accuracy greater than that of the original data capture process and this may result in some erroneous results during analysis.

As well, since most data is captured in a vector format, e.g. digitizing, data must be converted to the raster data structure. This is called vector-raster conversion. Most GIS software allows the user to define the raster grid (cell) size for vector-raster conversion. It is imperative that the original scale, e.g. accuracy, of the data be known prior to conversion. The accuracy of the data, often referred to as the resolution, should determine the cell size of the output raster map during conversion.

Most raster based GIS software requires that the raster cell contain only a single discrete value. Accordingly, a data layer, e.g. forest inventory stands, may be broken down into a series of raster maps, each representing an attribute type, e.g. a species map, a height map, a density map, etc. These are often referred to as one attribute maps. This is in contrast to most conventional vector data models that maintain data as multiple attribute maps, e.g. forest inventory polygons linked to a database table containing all attributes as columns. This basic distinction of raster data storage provides the foundation for quantitative analysis techniques. This is often referred to as raster or map algebra. The use of raster data structures allow for sophisticated mathematical modelling processes while vector based systems are often constrained by the capabilities and language of a relational DBMS.

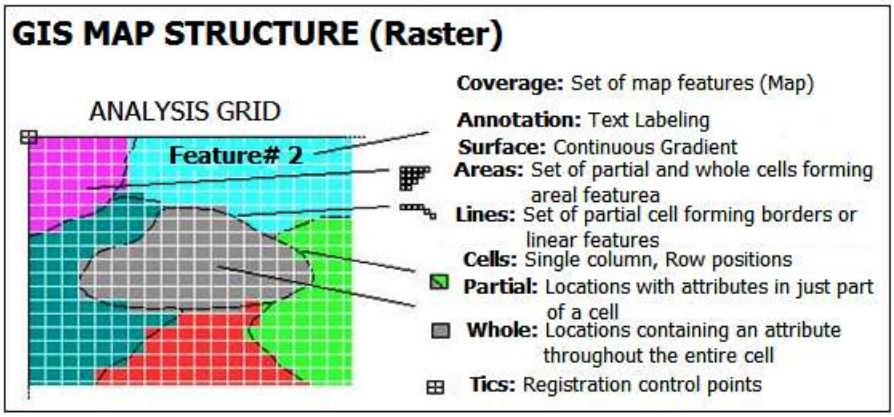


Fig.2.9. GIS MAP Structure - RASTER systems (Adapted from Berry)

This difference is the major distinguishing factor between vector and raster based GIS software. It is also important to understand that the selection of a particular data structure can provide advantages during the analysis stage. For example, the vector data model does not handle continuous data, e.g. elevation, very well while the raster data model is more ideally suited for this type of analysis. Accordingly, the raster structure does not handle linear data analysis, e.g. shortest path, very well while vector systems do. It is important for the user to understand that there are certain advantages and disadvantages to each data model.

The selection of a particular data model, vector or raster, is dependent on the source and type of data, as well as the intended use of the data. Certain analytical procedures require raster data while others are better suited to vector data.

2.5.3. Image Data

Image data is most often used to represent graphic or pictorial data. The term image inherently reflects a graphic representation, and in the GIS world, differs significantly from raster data. Most often, image data is used to store remotely sensed imagery, e.g. satellite scenes or orthophotos, or ancillary graphics such as photographs, scanned plan documents, etc. Image data is typically used in GIS systems as background display data (if the image has been rectified and georeferenced); or as a graphic attribute. Remote sensing software makes use of image data for image classification and processing. Typically, this data must be converted into a raster format (and perhaps vector) to be used analytically with the GIS.

Image data is typically stored in a variety of de facto industry standard proprietary formats. These often reflect the most popular image processing systems. Other graphic image formats, such as TIFF, GIF, PCX, etc., are used to store ancillary image data. Most GIS software will read such formats and allow you to display this data.

Fig.2.10. Image data is most often used for remotely sensed imagery such as satellite imagery or digital orthophotos.

2.5.4. Vector and Raster – Advantages and Disadvantages

There are several advantages and disadvantages for using either the vector or raster data model to store spatial data. These are summarized below.

Vector Data:

Advantages:

- Data can be represented at its original resolution and form without generalization.
- Graphic output is usually more aesthetically pleasing (traditional cartographic representation);
- Since most data, e.g. hard copy maps, is in vector form no data conversion is required.
- Accurate geographic location of data is maintained.
- Allows for efficient encoding of topology, and as a result more efficient operations that require topological information, e.g. proximity, network analysis.

Disadvantages:

- The location of each vertex needs to be stored explicitly. For effective analysis, vector data must be converted into a topological structure. This is often processing intensive and usually requires extensive data cleaning. As well, topology is static, and any updating or editing of the vector data requires re-building of the topology. Algorithms for manipulative and analysis functions are complex and may be processing intensive. Often, this inherently limits the functionality for large data sets, e.g. a large number of features.
- Continuous data, such as elevation data, is not effectively represented in vector form. Usually substantial data generalization or interpolation is required for these data layers.
- Spatial analysis and filtering within polygons is impossible

Raster Data

Advantages:

- The geographic location of each cell is implied by its position in the cell matrix. Accordingly, other than an origin point, e.g. bottom left corner, no geographic coordinates are stored.
- Due to the nature of the data storage technique data analysis is usually easy to program and quick to perform.
- The inherent nature of raster maps, e.g. one attribute maps, is ideally suited for mathematical modeling and quantitative analysis.
- Discrete data, e.g. forestry stands, is accommodated equally well as continuous data, e.g. elevation data, and facilitates the integrating of the two data types.
- Grid-cell systems are very compatible with raster-based output devices, e.g. electrostatic plotters, graphic terminals.

Disadvantages:

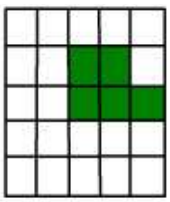
- The location of each vertex needs to be stored explicitly. For effective analysis, vector data must be converted into a topological structure. This is often processing intensive and usually requires extensive data cleaning. As well, topology is static, and any updating or editing of the vector data requires re-building of the topology.
- Algorithms for manipulative and analysis functions are complex and may be processing intensive. Often, this inherently limits the functionality for large data sets, e.g. a large number of features.
- Continuous data; such as elevation data, is not effectively represented in vector form.
- Usually substantial data generalization or interpolation is required for these data layers.
- Spatial analysis and filtering within polygons is impossible.

2.6. RASTER DATA STRUCTURE

In a simple raster data structure the geographical entities are stored in a matrix of rectangular cells. A code is given to each cell which informs users which entity is present in which cell. The simplest way of encoding a raster data into computers can be understood as follows:

(a) Entity model:

It represents the whole raster data. Let us assume that the raster data belongs to an area where land is surrounded by water. Here a particular entity (land) is shown in green color and the area where land is not present is shown by white.



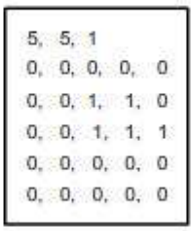
(b) Pixel values:

The pixel value for the full image is shown. Cells having a part of the land are encoded as 1 and others where land is not present are encoded as 0.

0	0	0	0	0
0	0	1	1	0
0	0	1	1	1
0	0	0	0	0
0	0	0	0	0

(c) File structure:

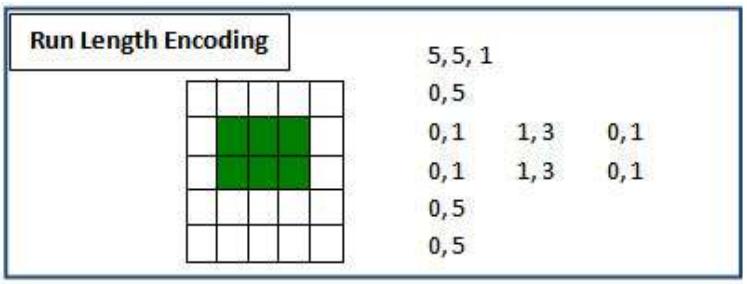
It demonstrates the method of coding raster data. The first row of the file structure data tells that there are 5 rows and 5 columns in the image, and 1 is the maximum pixel value. The subsequent rows have cells with value as either 0 or 1 (similar to pixel values).



The huge size of the data is a major problem with raster data. An image consisting of twenty different land-use classes takes the same storage space as a similar raster map showing the location of a single forest. To address this problem many data compaction methods have been developed which are discussed below:

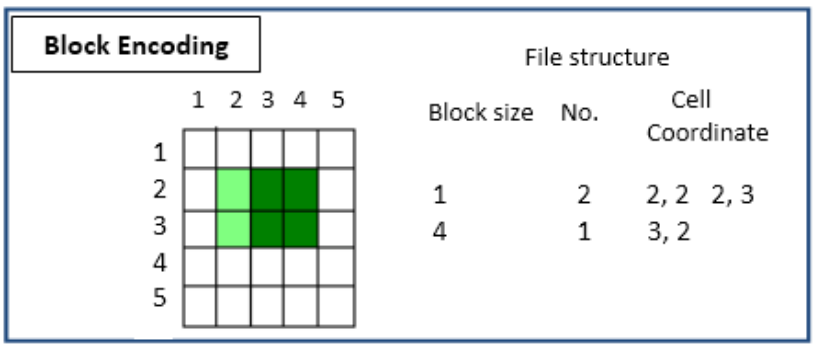
2.6.1. Run length encoding

- Reduction of data on a row by row basis
- Stores a single value for a group of cells rather than storing values for individual cells
- First line represents the dimension of the matrix (5×5) and the number of entities (1) present. In second and subsequent lines, the first number in the pair represents absence (0) or presence (1) of the entity and the second number indicates the number of cells referenced.



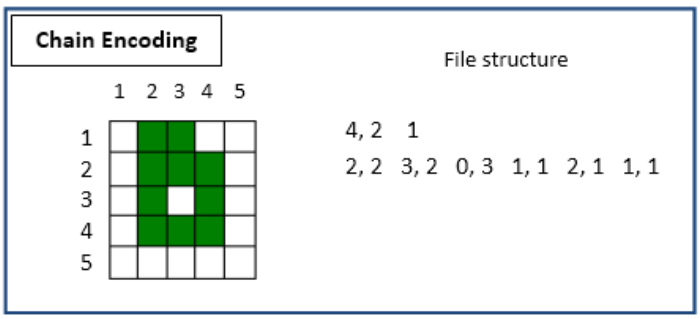
2.6.2. Block encoding

- Data is stored in blocks in the raster matrix.
- The entity is subdivided into hierarchical blocks and the blocks are located using coordinates.
- The first cell at top left hand is used as the origin for locating the blocks



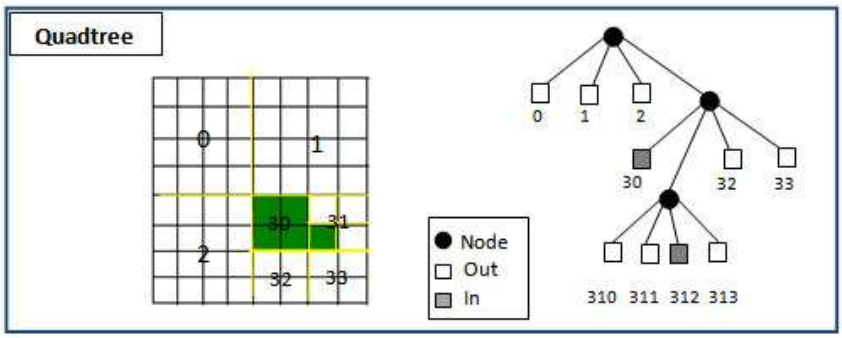
2.6.3. Chain encoding

- Works by defining boundary of the entity i.e. sequence of cells starting from and returning to the given origin
- Direction of travel is specified using numbers. (0 = North, 1 = East, 2 = South, 3 = West)
- The first line tells that the coding started at cell (4, 2) and there is only one chain. In the second line the first number in the pair tells the direction and the second number represents the number of cells lying in this direction.



2.6.4. Quadtree

- A raster is divided into a hierarchy of quadrants that are subdivided based on similar value pixels.
- The division of the raster stops when a quadrant is made entirely from cells of the same value.
- A quadrant that cannot be subdivided is called a leaf node.



A satellite or remote sensing image is a raster data where each cell has some value and together these values create a layer. A raster may have a single layer or multiple layers. In a multi-layer/ multi-band raster each layer is congruent with all other layers, have identical numbers of rows and columns, and have same locations in the plane. Digital elevation model (DEM) is an example of a single-band raster dataset each cell of which contains only one value representing surface elevation.

2.6.5. A single layer raster data can be represented using

(a) Two colors (binary):

The raster is represented as binary image with cell values as either 0 or 1 appearing black and white respectively.

Gray-scale:

Typical remote sensing images are recorded in an 8 bit digital system. A grayscale image is thus represented in 256 shades of gray which range from 0 (black) to 255 (white). However a human eye can't make distinction between the 255 different shades. It can only interpret 8 to 16 shades of gray.

A satellite image can have multiple bands, i.e. the scene/details are captured at different wavelengths (Ultraviolet- visible- infrared portions) of the electromagnetic spectrum. While creating a map we can choose to display a single band of data or form a color composite using multiple bands. A combination of any three of the available bands can be used to create RGB composites. These composites present a greater amount of information as compared to that provided by a single band raster.

2.6.6. Comparison between Vector and Raster Data Models

Data Model	Advantages	Disadvantages
Raster	Simple data structure	Cell size determines the resolution at which the data is represented
	Compatible with remote sensing or scanned data	Requires a lot of storage space
	Spatial analysis is easier	Projection transformations are time consuming
	Simulation is easy because each unit has the same size and shape	Network linkages are difficult to establish

Vector	Data is represented at its original resolution and form without generalization	The location of each vertex is to be stored explicitly
	Require less storage space	Overlay based on criteria is difficult
	Editing is faster and convenient	Spatial analysis is cumbersome
	Network analysis is fast	Simulation is difficult because each unit has a different topological form
	Projection transformations are easier	

2.7. DATA COMPRESSION

[Computing] The process of reducing the size of a file or database. Compression improves data handling, storage, and database performance. Examples of compression methods include quadtrees, run-length encoding, and wavelets.

Compression ratio:

- The compression ratio (that is , the size of the compressed file compared to that of the uncompressed file) of lossy video codec's is nearly always far superior to that of the audio and still-image equivalents Wavelet compression, used by raster formats such as MrSID,JPEG2000,andER Map per's ECW, takes time to decompress before drawing.
- Compression a series of techniques used for the reduction of space, bandwidth, cost, transmission, generating time, and the storage of data.
- It's a computer process using algorithms that reduces the size of electronic documents so they occupy less digital storage space.

2.7.1. Raster Data Compression

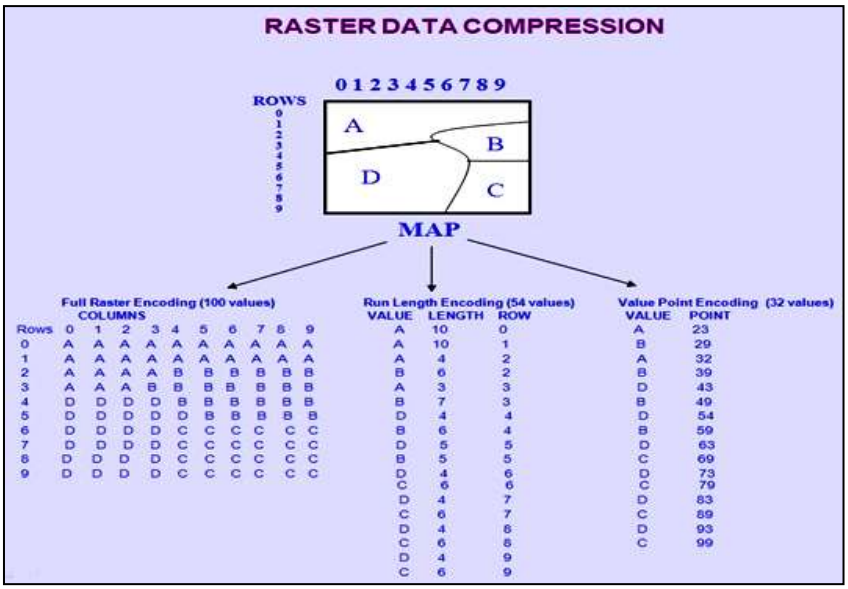


Fig.2.11. Raster Data Compression

Raster Data Compression

- Huge raster data has be to stored, retrieved, manipulated and analyzed.
- Large no. of thematic map layer is involved.
- Many repetitive characters are involved.
- Therefore, for better storage and to preserve highest possible degree of accuracy, we need to go for compact methods of storing.
- Common method is elimination of repetitive characters.

Original Data		**Compacted Data**	
Northing	**Easting**	**10,000**	**70,000**
10,234	70,565	234	565
10,245	70,599	245	599
10,167	70,423	167	423

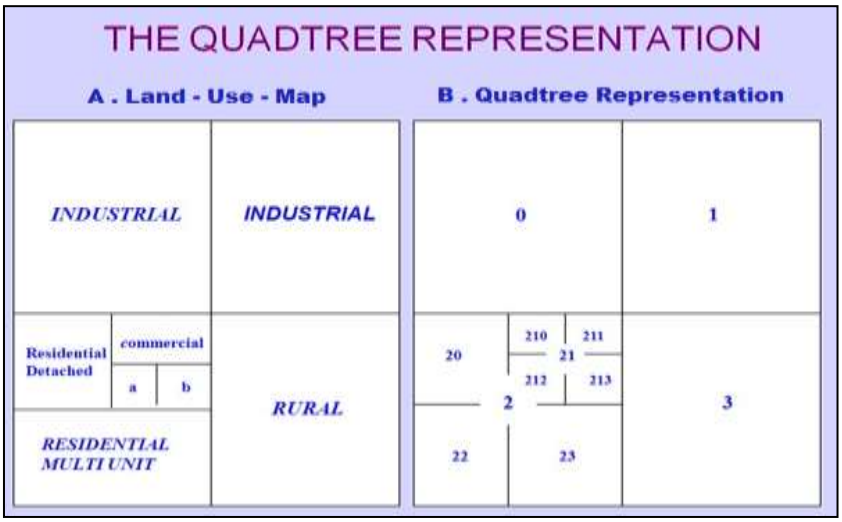


Fig.2.12. Raster Data Compression

2.7.2. Run length Encoding

- Value often occur in runs across several cells, i.e., cells of the same value are often neighbours, like same soil type, or similar parameters.
- spatial auto-correlation exists –a tendency for nearby things to be more similar than distant things
- In run length encoding, the cells of the same value in arrow may be compacted by stating the value and their total.
- Thematic maps storage sizes get reduced using runlength encoding.
- Some raster GIS packages have the capability to handle run length encoded files.

Value point encoding

- Cells are assigned position numbers starting in the upper left corner proceeding from left to right and from top to bottom.
- The position no. for end of each run is stored in the point columns. The value for each cell in the run is in the value column.

2.7.3. Quadtree

- Typical type of raster model is dividing area into equal-sized rectangular cells .
- However, many cases, variable sized grid cell size used for more compact raster representation as shown figure.2.13.
- Larger cells used to represent large homogenous areas and smaller cells for finely details.
- Process involves regularly subdividing a map into four equal sized quadrants. Quadrant that has more than one class is again subdivided. Then; it is further subdivided within each quadrant until a square is found to be so homogenous that it is no longer needed to be divided.
- Then a Quadtree is prepared, resembling an inverted tree with "Root", i.e., a point from which all branches expand; Leaf is a lower most point and all other points in the tree are nodes.

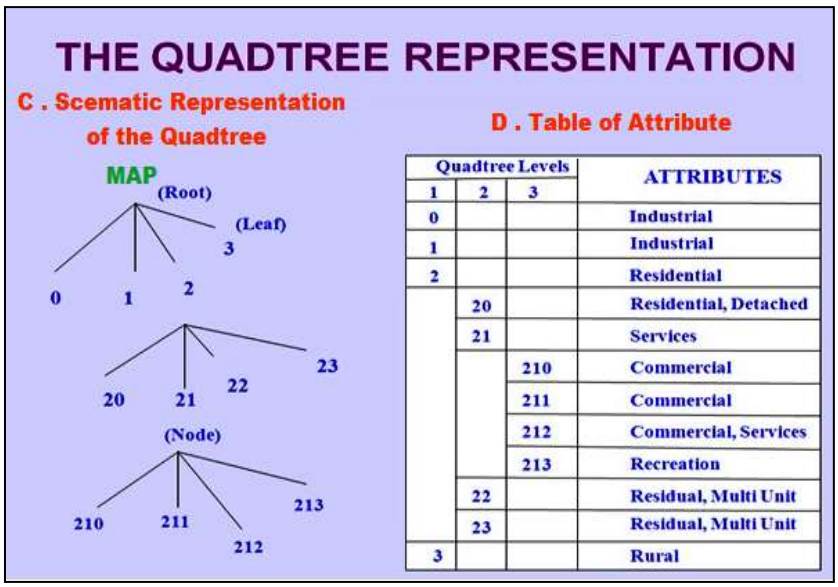


Fig.2.13. The Quadtree Representation

2.8. VECTOR DATA STRUCTURE

Geographic entities encoded using the vector data model, are often called features. The features can be divided into two classes:

a) Simple features

These are easy to create, store and are rendered on screen very quickly. They lack connectivity relationships and so are inefficient for modeling phenomena conceptualized as fields.

b) Topological features

A topology is a mathematical procedure that describes how features are spatially related and ensures data quality of the spatial relationships. Topological relationships include following three basic elements:

1) Connectivity: Information about linkages among spatial objects
2) Contiguity: Information about neighbouring spatial object
3) Containment: Information about inclusion of one spatial object within another spatial object

2.8.1. Connectivity

Arc node topology defines connectivity - arcs are connected to each other if they share a common node. This is the basis for many network tracing and path finding operations.

Arcs represent linear features and the borders of area features. Every arc has a from-node which is the first vertex in the arc and a to-node which is the last vertex. These two nodes define the direction of the arc. Nodes indicate the endpoints and intersections of arcs. They do not exist independently and therefore cannot be added or deleted except by adding and deleting arcs.

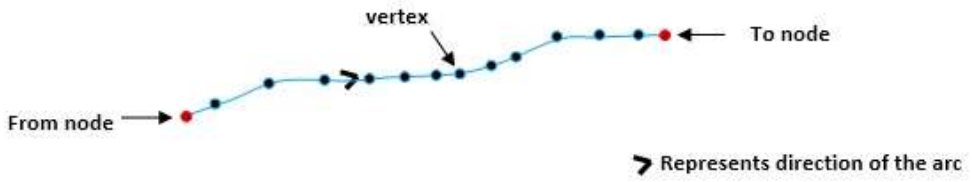


Fig.2.14. Arc-node Topology

Nodes can, however, be used to represent point features which connect segments of a linear feature (e.g., intersections connecting street segments, valves connecting pipe segments).

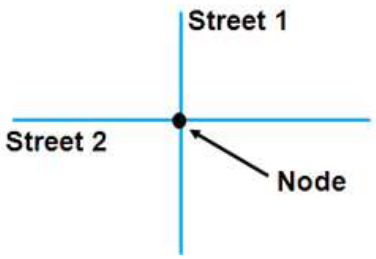
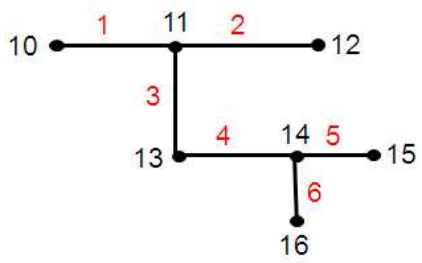


Fig.2.15. Node showing intersection

Arc-node topology is supported through an arc-node list. For each arc in the list there is a from node and a to node. Connected arcs are determined by common node numbers.

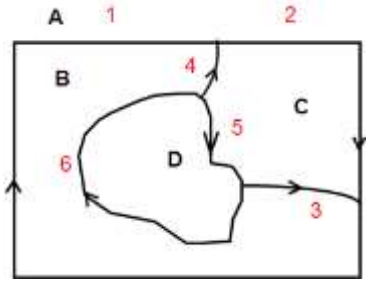


Arc-Node List		
Arc	*From node*	*To node*
1	10	11
2	11	12
3	11	13
4	13	14
5	14	15
6	14	16

Fig.2.16. Arc-Node Topology with list

2.8.2. Contiguity

Polygon topology defines contiguity. The polygons are said to be contiguous if they share a common arc. Contiguity allows the vector data model to determine adjacency.



Left-Right Topology		
Arc	Left Polygon	Right Polygon
1	A	B
2	A	C
3	C	B
4	B	C
5	C	D
6	B	D

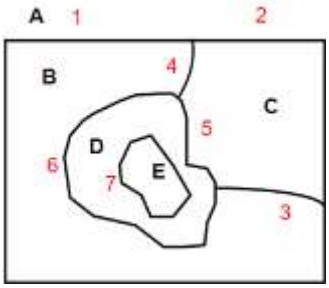
Fig2.17. Polygon Topology

The from node and *to* node of an arc indicate its direction, and it helps determining the polygons on its left and right side. Left-right topology refers to the polygons on the left and right sides of an arc. In the illustration above, polygon B is on the left and polygon C is on the right of the arc 4.

Polygon A is outside the boundary of the area covered by polygons B, C and D. It is called the external or universe polygon, and represents the world outside the study area. The universe polygon ensures that each arc always has a left and right side defined.

2.8.3. Containment

Geographic features cover distinguishable area on the surface of the earth. An area is represented by one or more boundaries defining a polygon.



Polygon arc topology	
Polygon	Arc List
B	1, 4, 6, 3
C	2, 3, 5, 4
D	5, 6, 0, 7
E	7

Fig2.18. Polygon arc topology

The polygons can be simple or they can be complex with a hole or island in the middle. In the illustration given below assume a lake with an island in the middle.

The lake actually has two boundaries, one which defines its outer edge and the other (island) which defines its inner edge. An island defines the inner boundary of a polygon. The polygon D is made up of arc 5, 6 and 7. The 0 before the 7 indicates that the arc 7 creates an island in the polygon.

Polygons are represented as an ordered list of arcs and not in terms of X, Y coordinates. This is called ***Polygon-Arc topology***. Since arcs define the boundary of polygon, arc coordinates are stored only once, thereby reducing the amount of data and ensuring no overlap of boundaries of the adjacent polygons.

2.8.4. Simple Features

Point entities:

These represent all geographical entities that are positioned by a single XY coordinate pair. Along with the XY coordinates the point must store other information such as what does the point represent etc.

Line entities:

Linear features made by tracing two or more XY coordinate pair.

- **Simple line:** It requires a start and an end point.
- **Arc:** A set of XY coordinate pairs describing a continuous complex line. The shorter the line segment and the higher the number of coordinate pairs, the closer the chain approximates a complex curve.

Simple Polygons:

Enclosed structures formed by joining set of XY coordinate pairs. The structure is simple but it carries few disadvantages which are mentioned below:

- Lines between adjacent polygons must be digitized and stored twice, improper digitization give rise to slivers and gaps
- Convey no information about neighbour
- Creating islands is not possible

2.8.5. Topologic Features

Networks:

A network is a topologic feature model which is defined as a line graph composed of links representing linear channels of flow and nodes representing their connections. The topologic relationship between the features is maintained in a connectivity table. By consulting connectivity table, it is possible to trace the information flowing in the network

Polygons with explicit topological structures:

Introducing explicit topological relationships takes care of islands as well as neighbours. The topological structures are built either by creating topological links during data input or using software. Dual Independent Map Encoding (DIME) system of US Bureau of the Census is one of the first attempts to create topology in geographic data.

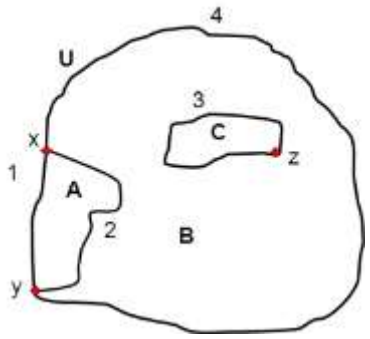


Fig.2.19. Polygon as a topological feature

- Polygons are formed using the lines and their nodes.
- Once formed, polygons are individually identified by a unique identification number.
- The topological information among the polygons is computed and stored using the adjacency information (the nodes of a line, and identifiers of the polygons to the left and right of the line) stored with the lines.

Poly ID	Arcs
A	1, 2
B	2, 3, 4
C	3

Arc ID	From	To
1	x	y
2	x	y
3	z	z
4	x	y

Arc ID	Left Poly	Right Poly
1	A	U
2	B	A
3	C	B
4	U	B

2.8.6. Fully topological polygon network structure

A fully topological polygon network structure is built using boundary chains that are digitized in any direction. It takes care of islands and lakes and allows automatic checks for improper polygons. Neighborhood searches are fully supported. These structures are edited by moving the coordinates of individual points and nodes, by changing polygon attributes and by cutting out or adding sections of lines or whole polygons. Changing coordinates require no modification to the topology but cutting out or adding lines and polygons requires recalculation of topology and rebuilding the database.

2.8.7. Triangular Irregular Network (TIN)

TIN represents surface as contiguous non-overlapping triangles created by performing Delaunay triangulation. These triangles have a unique property that the circum circle that passes through the vertices of a triangle contains no other point inside it. TIN is created from a set of mass points with x, y and z coordinate values. This topologic data structure manages information about the nodes that form each triangle and the neighbors of each triangle.

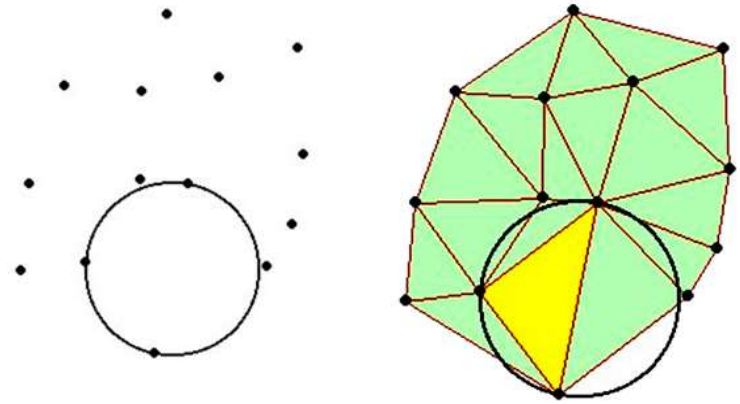
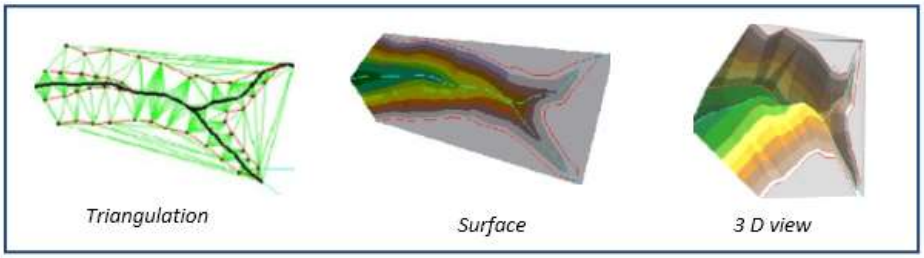


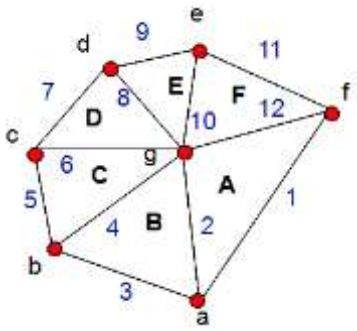
Fig.2.20. Delaunay Triangulation

Advantages of Delaunay triangulation

- The triangles are as equiangular as possible, thus reducing potential numerical precision problems created by long skinny triangles
- The triangulation is independent of the order the points are processed
- Ensures that any point on the surface is as close as possible to a node



Because points can be placed irregularly over a surface a TIN can have higher resolution in areas where surface is highly variable. The model incorporates original sample points providing a check on the accuracy of the model. The information related to TIN is stored in a file or a database table. Calculation of elevation, slope, and aspect is easy with TIN but these are less widely available than raster surface models and more time consuming in term of construction and processing.



Arc attribute table			
Edge ID	**Length**	**From node**	**To node**
1	160	f	a
2	140	a	g
3	130	a	b
4	140	b	g
...			

Arc attribute table			
Edge ID	Length	From node	To node
1	160	f	a
2	140	a	g
3	130	a	b
4	140	b	g
...			

Polygon attribute table					
Triangle ID	Area	Edge1	Edge2	Edge3	Neighbors
A	8200	1	2	12	B, F
B	7040	3	4	2	C, A
C	6000	5	6	4	D, B
D	5440	7	8	6	E, C
...					

The TIN model is a vector data model which is stored using the relational attribute tables. A TIN dataset contains three basic attribute tables: Arc attribute table that contains length, from node and to node of all the edges of all the triangles.

- Node attribute table that contains x, y coordinates and z (elevation) of the vertices
- Polygon attribute table that contains the areas of the triangles, the identification number of the edges and the identifier of the adjacent polygons.

Storing data in this manner eliminated redundancy as all the vertices and edges are stored only once even if they are used for more than one triangle. As TIN stores topological relationships, the datasets can be applied to vector based geo-processing such as automatic contouring, 3D landscape visualization, volumetric design, surface characterization etc.

2.9. RASTER VS VECTOR MODELS

The two primary types of spatial data are vector and raster data in GIS.

Data Model

The data model represents a set of guidelines to convert the real world (called entity) to the digitally and logically represented spatial objects consisting of the attributes and geometry. The attributes are managed by thematic or semantic structure while the geometry is represented by geometric-topological structure.

There are two major types of geometric data model;

1) Vector Data Model
2) Raster Data Model

Vector Data Model: [data models] A representation of the world using points, lines, and polygons (show in the figure 2.21). Vector models are useful for storing data that has discrete boundaries, such as country borders, land parcels, and streets.

Raster Data Model: [data models] A representation of the world as a surface divided into a regular grid of cells. Raster models are useful for storing data that varies continuously, as in an aerial photograph, a satellite image, a surface of chemical concentrations, or an elevation surface.

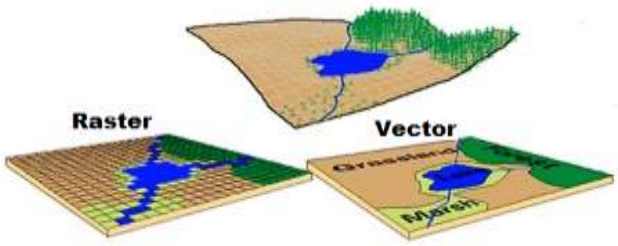


Fig.2.21. Example – Raster Data and Vector Data

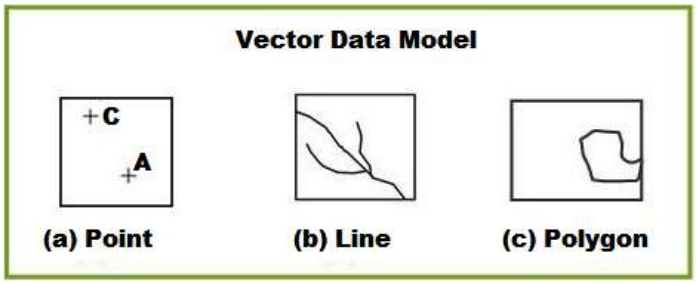


Fig .2.22. Example – Raster Data and Vector Data

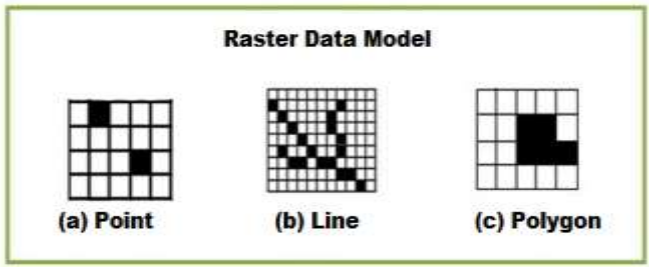


Figure.2.23. Example – Vector Data and Vector Data

2.9.1. Vector Data

Vector data (Show in Fig.2.22) is not made up of a grid of pixels. Instead, vector graphics are comprised of vertices and paths.

The three basic symbol types for vector data are

1) Points
2) Lines And
3) Polygons (areas).

Since the dawn of time, maps have been using symbols to represent real-world features. In GIS terminology, real-world features are called spatial entities.

The cartographer decides how much data needs to be generalized in a map. This depends on scale and how much detail will be displayed in the map. The decision to choose vector points, lines or polygons is governed by the cartographer and scale of the map.

(1) Points

Fig.2.24. Point Vector Data Type

Point Vector Data Type: Simple XY Coordinates

Vector points are simply XY coordinates. When features are too small to be represented as polygons, points are used indicate (fig.2.24.)

For Example: At a regional scale, city extents can be displayed as polygons because this amount of detail can be seen when zoomed in. But at a global scale, cities can be represented as points because the detail of city boundaries cannot be seen.

Vector data are stored as pairs of XY coordinates (latitude and longitude) represented as a point. Complementary information like street name or date of construction could accompany it in a table for its current use.

(2) Lines

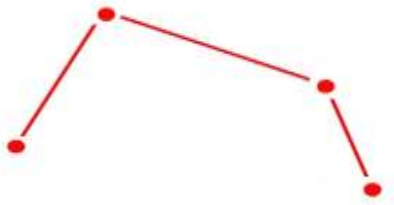


Fig.2.25. Vector Data Type Line

Vector Data Type Line:

Connect the dots and it becomes a line feature. Vector lines connect vertices with paths show in the fig (2.25). If you were to connect the dots in a particular order, you would end up with a vector line feature.

Lines usually represent features that are linear in nature. Cartographers can use a different thickness of line to show size of the feature. For Example, 500 meter Wide River may be thicker than a 50 meter wide river. They can exist in the real-world such as roads or rivers. Or they can also be artificial divisions such as regional borders or administrative boundaries.

Points are simply pairs of XY coordinates (latitude and longitude). When you connect each point or vertex with a line in a particular order, they become a vector line feature. Networks are line data sets but they are often considered to be different. This is because linear networks are topologically connected elements. They consist of junctions and turns with

connectivity. If you were to find an optimal route using a traffic line network, it would follow one-way streets and turn restrictions to solve an analysis. Networks are just that smart.

(3) Polygons

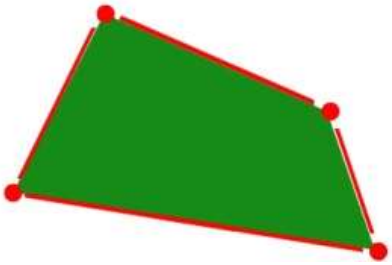


Fig.2.26. Vector Data Type Polygon

Vector Data Type Polygon: Connect the dots and enclose. It becomes a polygon feature when a set of vertices are joined in a particular order and closed; they become a vector Polygon feature shown the (fig.2.26). In order to create a polygon, the first and last coordinate pair is the same and all other pairs must be unique. Polygons represent features that have a two-dimensional area.

Examples of polygons are buildings, agricultural fields and discrete administrative areas. Cartographers use polygons when the map scale is large enough to be represented as polygons.

2.9.2. Raster Types: Discrete vs Continuous

Raster data is made up of pixels (also referred to as grid cells). They are usually regularly-spaced and square but they don't have to be. Rasters often look pixelated because each pixel has its own value or class.

For example:

Each pixel value in a satellite image has a red, green and blue value. Alternatively, each value in an elevation map represents a specific height. It could represent anything from rainfall to land cover.

Raster models are useful for storing data that varies continuously. For example, elevation surfaces, temperature and lead contamination.

Raster data models consist of 2 categories – discrete and continuous.

2.9.3. Discrete Raster's have Distinct Values

Discrete raster's have distinct themes or categories. For example, one grid cell represents a land cover class or a soil type.

In a discrete raster land cover/use map, you can distinguish each thematic class. Each class can be discretely defined where it begins and ends. In other words, each land cover cell is definable and it fills the entire area of the cell.

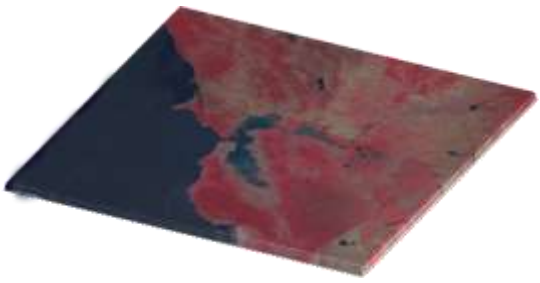
Discrete data usually consists of integers to represent classes. For example, the value 1 might represent urban areas; the value 2 represents forest and so on.

2.9.4. Continuous Rasters have Gradual Change

Continuous rasters (non-discrete) are grid cells with gradual changing data such as elevation, temperature or an aerial photograph.

A continuous raster surface can be derived from a fixed registration point. For example, digital elevation models use sea level as a registration point. Each cell represents a value above or below sea level. As another example, aspect cell values have fixed directions such as north, east, south or west.

Phenomena can gradually vary along a continuous raster from a specific source. In a raster depicting an oil spill, it can show how the fluid moves from high concentration to low concentration. At the source of the oil spill, concentration is higher and diffuses outwards with diminishing values as a function of distance.



2.10. VECTOR VS RASTER: SPATIAL DATA TYPES

It's not always straight-forward which spatial data type you should use for your maps.

In the end, it really comes down to the way in which the cartographer conceptualizes the feature in their map.

- **Do you want to work with pixels or coordinates?** Raster data works with pixels. Vector data consists of coordinates.
- **What is your map scale?** Vectors can scale objects up to the size of a billboard. But you don't get that type of flexibility with raster data
- **Do you have restrictions for file size?** Raster file size can result larger in comparison with vector data sets with the same phenomenon and area.

2.10.1. Vector and Raster – Advantages and Disadvantages

There are several advantages and disadvantages for using either the vector or raster data model to store spatial data. These are summarized below.

2.10.2. Vector Data:

Advantages:

- Data can be represented at its original resolution and form without generalization.
- Graphic output is usually more aesthetically pleasing (traditional cartographic representation);
- Since most data, e.g. hard copy maps, is in vector form no data conversion is required.
- Accurate geographic location of data is maintained.
- Allows for efficient encoding of topology, and as a result more efficient operations that require topological information, e.g. proximity, network analysis.

Disadvantages:

- The location of each vertex needs to be stored explicitly. For effective analysis, vector data must be converted into a topological structure. This is often processing intensive and usually requires extensive data cleaning. As well, topology is static, and any updating or editing of the vector data requires re-building of the topology. Algorithms for manipulative and analysis functions are complex and may be processing intensive. Often, this inherently limits the functionality for large data sets, e.g. a large number of features.
- Continuous data, such as elevation data, is not effectively represented in vector form. Usually substantial data generalization or interpolation is required for these data layers.
- Spatial analysis and filtering within polygons is impossible

2.10.3. Raster Data

Advantages :

- The geographic location of each cell is implied by its position in the cell matrix. Accordingly, other than an origin point, e.g. bottom left corner, no geographic coordinates are stored.
- Due to the nature of the data storage technique data analysis is usually easy to program and quick to perform.

- The inherent nature of raster maps, e.g. one attribute maps, is ideally suited for mathematical modeling and quantitative analysis.
- Discrete data, e.g. forestry stands, is accommodated equally well as continuous data, e.g. elevation data, and facilitates the integrating of the two data types.
- Grid-cell systems are very compatible with raster-based output devices, e.g. electrostatic plotters, graphic terminals.

Disadvantages:

- The location of each vertex needs to be stored explicitly. For effective analysis, vector data must be converted into a topological structure. This is often processing intensive and usually requires extensive data cleaning. As well, topology is static, and any updating or editing of the vector data requires re-building of the topology.
- Algorithms for manipulative and analysis functions are complex and may be processing intensive. Often, this inherently limits the functionality for large data sets, e.g. a large number of features.
- Continuous data; such as elevation data, is not effectively represented in vector form.
- Usually substantial data generalization or interpolation is required for these data layers.
- Spatial analysis and filtering within polygons is impossible.

2.11. TIN AND GRID DATA MODELS

TIN models

TIN stands for Triangular Irregular Network, which is a vector approach to handling a digital elevation model. TIN's are used to interpolate surfaces using multiple triangles. TIN's are able to interpolate surfaces by selecting representative points that are usually data points. TIN's connect these points to form a set of continuous and connected triangles. The data points consist of X, Y and Z values. The final result gives users a TIN surface.

Advantages of TIN models

TIN's give researchers the ability to view 2.5D and 3D at an area that was interpolated from minimal data collection.

- Users can describe a surface at different levels of resolution based on the points that were collected.
- TIN interpolation gives GIS users greater analytical capabilities. TIN models are easy to create and use.
- They provide users a simplified model that represents collected data points.
- Using a TIN surface in conjunction with Arc-Map extensions such as Spatial Analysis and 3D Analyst, TIN users can also derive slope, aspect, elevation, contour lines, hill shades, etc.

2.11.1. Different Types of TIN Methods and Processes

There are many different types of TIN interpolation methods. Some of the most popular TIN methods include:

- Natural Neighbour,
- Krigging,
- Spline,
- Nearest Neighbour and
- Inversed Distance Weighting.

These TIN interpolation methods use mathematical algorithms in order to generate interpolated surfaces. Each of these methods will produce different types of surfaces.

The TIN model (Triangulated Irregular Network):

A triangulated irregular network (TIN) is a digital data structure used in a geographic information system (GIS) for the representation of a surface.

A TIN is a vector based representation of the physical land surface or sea bottom, made up of irregularly distributed nodes and lines with three dimensional coordinates (x,y, and z) that are arranged in a network of non-overlapping triangles. TINs are often derived from the elevation data of a rasterized digital elevation model (DEM).

Structure of TIN Data Model

The TIN model represents a surface as a series of linked triangles, hence the adjective triangulated. Triangles are made from three points, which can occur at any location, giving the adjective, irregular. For each triangle, TIN records:

- The triangle number
- The numbers of each adjacent triangle
- The three nodes defining the triangle
- The x, y coordinates of each node
- The surface z value of each node
- The edge type of each triangle edge (hard or soft)

2.11.2. Components of TIN:

Nodes:

Nodes are the fundamental building blocks of the TIN. The nodes originate from the points and arc vertices contained in the input data sources. Every node is incorporated in the TIN triangulation. Every node in the TIN surface model must have a z value.

Edges:

Every node is joined with its nearest neighbors by edges to form triangles, which satisfy the Delaunay criterion. Each edge has two nodes, but a node may have two or more edges.

Because edges have a node with a z value at each end, it is possible to calculate a slope along the edge from one node to the other.

TIN:

Advantages **-** ability to describe the surface at different level of resolution, efficiency in storing data.

Disadvantages **-** in many cases require visual inspection and manual control of the network.

Automated hill shading:

The TIN model of terrain representation lends itself to development of an automated method of hill shading. Slope mapping is possible in TIN.

2.11.3. TIN Data Model

The Triangulated Irregular Network (TIN) data model is an alternative to the raster and vector data models for representing continuous surfaces. It allows surface models to be generated efficiently to analyze and display terrain and other types of surfaces. The TIN model creates a network of triangles by storing the topological relationships of the triangles. The fundamental building block of the TIN data is the node. Nodes are connected to their nearest neighbors by edges, according to a set of rules. Left-right topology is associated with the edges to identify adjacent triangles.

The TIN creates triangles from a set of points called mass points, which always become nodes. The user is not responsible for selecting; all the nodes are added according to a set of rules. Mass points can be located anywhere, the more carefully selected, the more accurate the model of the surface will be. Well-placed mass points occur when there is a major change in the shape of the surface, for example, at the peak of a mountain, the floor of a valley, or at the edge (top and bottom) of cliffs. By connecting points on a valley floor or along the edge of a cliff, a linear break in the surface can be defined. These are called break lines. Break lines can control the shape of the surface model.

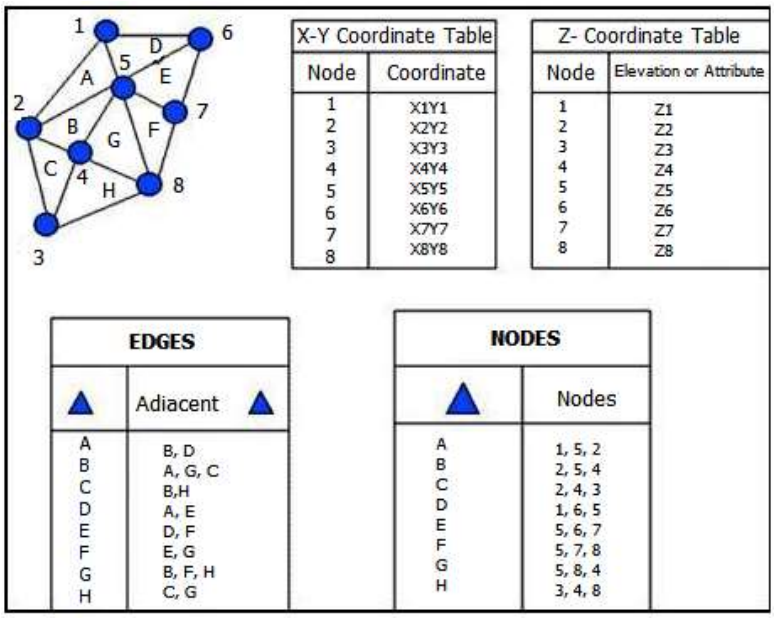
They always form edges of triangles and, generally, cannot be moved. A triangle always has three and only three straight sides, making their representation rather simple. A triangle is assigned a unique identifier that defines by its three nodes and its two or three neighboring triangles.

TIN is a vector-based topological data model that is used to represent terrain data. A TIN represents the terrain surface as a set of interconnected triangular facets. For each of the three vertices, the XY (geographic location) and the (elevation) Z values are encoded.

Four Tables for TIN Model

- Node Table it lists each triangle and the nodes which define it.
- Edge Table it lists three triangles adjacent to each facets. The triangles that border the boundary of the TIN show only two adjacent facets.

- XY Co-ordinate Table it lists the co-ordinate values of each node.
- Z Table it is the altitude value of each node.



2.12. GRID/LUNR/MAGI

In this model each grid cell is referenced or addressed individually and is associated with identically positioned grid cells in all other coverage's, rather than like a vertical column of grid cells, each dealing with a separate theme. Comparisons between coverage's are therefore performed on a single column at a time. Soil attributes in one coverage can be compared with vegetation attributes in a second coverage. Each soil grid cell in one coverage can be compared with a vegetation grid cell in the second coverage. The advantage of this data structure is that it facilitates the multiple coverage analysis for single cells. However, this limits the examination of spatial relationships between entire groups or themes in different coverage's.

2.12.1. Imgrid GIS

To represent a thematic map of land use that contains four categories: recreation, agriculture, industry and residence, each of these features have to be separated out as an individual layer. In the layer that represents agriculture 1 or 0 will represent the presence or absence of crops respectively. The rest of layer will be represented in the same way, with each variable referenced directly. The major advantage of IMGRID is its two-dimensional array of numbers resembling a map-like structure. The binary character of the information in each coverage simplifies long computations and eliminates the need for complex map legends. Since each coverage feature is uniquely identified, there is no limitation of assigning a single attribute value to a single grid cell. On the other side, the main problem related to information storage in an IMGRID structure is the excessive volume of data stored. Each grid cell will contain more than 1 or 0 values from more than one coverage and a large number of coverages are needed to store different types of information.

2.13. OPEN GEOSPATIAL CONSORTIUM (OGC)

The Open Geospatial Consortium (OGC), an international voluntary consensus standards organization, originated in 1994. In the OGC, more than 500 commercial, governmental, nonprofit and research organizations worldwide collaborate in a consensus process encouraging development and implementation of open standards for geospatial content and services, sensor web and Internet of Things, GIS data processing and data sharing.

2.13.1. Standards

Most of the OGC standards depend on a generalized architecture captured in a set of documents collectively called the Abstract Specification, which describes a basic data model for representing geographic features. Atop the Abstract Specification members have developed and continue to develop a growing number of specifications, or standards to serve specific needs for interoperable location and geospatial technology, including GIS.

The OGC standards baseline comprises more than thirty standards, including:

- CSW – Catalog Service for the Web: access to catalog information
- GML – Geography Mark-up Language: XML-format for geographical information
- Geo-XACML – Geospatial eXtensible Access Control Mark-up Language
- KML – Keyhole Mark-up Language: XML-based language schema for expressing geographic annotation and visualization on existing (or future) Web-based, two-dimensional maps and three-dimensional Earth browsers
- Observations and Measurements
- OGC Reference Model – a complete set of reference models
- OLS – Open Location Service (Open-LS)
- OGC Web Services Context Document defines the application state of an OGC Integrated Client
- OWS – OGC Web Service Common
- SOS – Sensor Observation Service
- SPS – Sensor Planning Service
- Sensor-ML – Sensor Model Language
- Sensor Things API - an open and unified framework to interconnect IoT devices, data, and applications over the Web. Currently a candidate standard waiting for votes.
- SFS – Simple Features – SQL

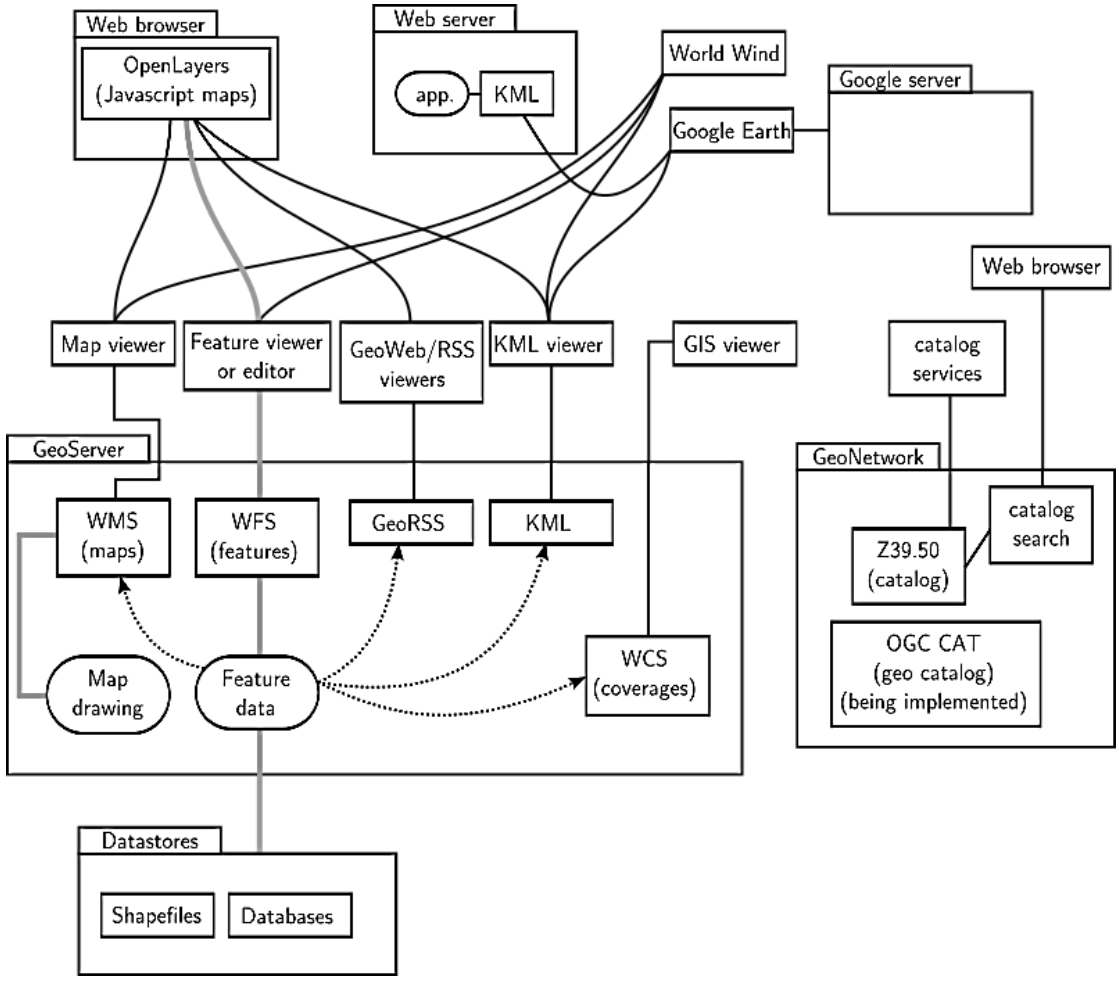


Fig.2.27. Relationship between clients/servers and OGC protocols

- SLD - Styled Layer Descriptor
- SRID, an identification for spatial coordinate systems
- Water-ML – Information model for the representation of hydrological observation data
- WCS – Web Coverage Service: provides access, sub setting, and processing on coverage objects
- WCPS – Web Coverage Processing Service: provides a raster query language for ad-hoc processing and filtering on raster coverage's
- WFS – Web Feature Service: for retrieving or altering feature descriptions
- WMS – Web Map Service: provides map images
- WMTS – Web Map Tile Service: provides map image tiles
- WPS – Web Processing Service: remote processing service
- Geo-SPARQL – Geographic SPARQL Protocol and RDF Query Language: representation and querying of geospatial data for the Semantic Web

- WTS – Web Terrain Service (WTS)

The design of standards were originally built on the HTTP web services paradigm for message-based interactions in web-based systems, but meanwhile has been extended with a common approach for SOAP protocol and WSDL bindings. Considerable progress has been made in defining Representational State Transfer (REST) web services, e.g., OGC Sensor Things API.

2.14. SPATIAL DATA QUALITY

Introduction

GIS developers and users have paid little attention to the problems caused by error, inaccuracy and imprecision in spatial data sets. There was awareness that all data suffers from inaccuracy and imprecision, but effects on GIS problems and solutions were not considered. It is now generally recognized that error, inaccuracy and imprecision can "make or break" GIS projects-making the results of a GIS analysis worthless. Spatial analyses done manually can easily align map boundaries to overlap and be registered. An automated GIS cannot do this, unless it is programmed to recognize the "undershoots, overshoots, and slivers" to connect lines. The level of the data quality must be made clear for the GIS to operate correctly. Assessing the quality of the data, however, may be costly. Data quality generally refers to the relative accuracy and precision of a particular GIS database. Error encompasses both the imprecision of data and its inaccuracies.

Although the term "garbage in, garbage out" certainly applies to GIS data, there are other important data quality issues besides the input data that need to be considered.

2.14.1. Components of Data Quality

There are three main components of data quality.

(i) Micro level components

(ii) Macro level components

(iii) Usage components.

Micro Level Components

Micro level components are data quality factors that pertain to the individual data elements. These components are usually evaluated by statistical testing of the data product against an independent source of higher quality information. They include positional accuracy, attribute accuracy and logical consistency given as follows:

a) Position Accuracy

b) Attribute Accuracy

c) Logical Consistency

Position Accuracy

Position accuracy is the expected deviance in the geographical location of an object in the data set (e.g. on a map) from its true ground position. Selecting a specified sample of points in a prescribed manner and comparing the position coordinates with an independent and more accurate

source of information usually test it. There are two components to position accuracy: the bias and the precision.

Attribute Accuracy

Attributes may be discrete or continuous variables. A discrete variable can take on only a finite number of values whereas a continuous variable can take on any number of values. Categories like land use class, vegetation type, or administrative area are discrete variables. They are, in effect, ordered categories where the order indicates the hierarchy of the attribute.

Logical Consistency

Logical consistency refers to how well logical relations among data elements are maintained. It also refers to the fidelity of relationships encoded in the database, they may refer to the geometric structure of the data model (e.g. topologic consistency) or to the encoded attribute information e.g. semantic consistency).

Macro Level Components

Macro level components of data quality pertain to the data set as a whole. They are not generally amenable to testing but instead are evaluated by judgment (in the case of completeness) or by reporting information about the data, such as the acquisition date. Three major macro level components are:

a) Completeness

b) Time

c) Lineage

(a) Completeness

Completeness refers to the exhaustiveness of the information in terms of spatial and attribute properties encoded in the database. It may include information regarding feature selection criteria, definition and mapping rules and the deviations from them. The tests on spatial completeness may be obtained from topological test used for logical consistency whereas the test for attribute completeness is done by comparison of a master list of geo-codes to the codes actually appearing in the database.

There are several aspects to completeness as it pertains to data quality. They are grouped here into three categories: completeness of coverage, classification and verification.

The completeness of coverage is the proportion of data available for the area of interest.

Completeness of classification is an assessment of how well the chosen classification is able to represent the data. For a classification to be complete it should be exhaustive, that is it should be possible to encode all data at the selected level of detail.

Completeness of verification refers to the amount and distribution of field measurements or other independent sources of information that were used to develop the data.

(b) Time

Time is a critical factor in case of any type of data. Some data will be significantly biased depending on the time period over which they are collected.

Example:

Demographic information is usually very time sensitive. It can change significantly over a year. Land cover will change quickly in an area of rapid urbanization.

(c) Lineage

The lineage of a data set is its history, the source data and processing steps used to produce it. The source data may include transaction records, field notes etc. Ideally, some indication of lineage should be included with the data set since the internal documents are rarely available and usually require considerable expertise to evaluate. Unfortunately, lineage information most often exists as the personal experience of a few staff members and is not readily available to most users.

2.14.2. Usage Components

The usage components of data quality are specific to the resources of the organization. The effect of data cost, for example, depends on the financial resources of the organization. A given data set may be too expensive for one organization and be considered inexpensive by another.

Accessibility refers to the ease of obtaining and using the data. The accessibility of a data set may be restricted because the data are privately held. Access to government-held information may be restricted for reasons of national security or to protect citizen rights. Census data are usually restricted in this way. Even when the right to use restricted data can be obtained, the time and effort needed to actually receive the information may reduce its overall suitability.

The direct cost of a data set purchased from another organization is usually well known: it is the price paid for the data. However, when the data are generated within the organization, the true cost may be unknown. Assessing the true cost of these data is usually difficult because the services and equipment used in their production support other activities as well.

The indirect costs include all the time and materials used to make use of the data. When data are purchased from another organization, the indirect costs may actually be more significant than the direct ones.

It may take longer for staff to handle data with which they are unfamiliar, or the data may not be compatible with the other data sets to be used.

2.14.3. Causes of Error

In this section, it is examined when and how errors creep into GIS data. The three major causes of GIS data error are problems found in (i) source data, (ii) data entry and (iii) data analysis.

Errors in Source Data

It has become common now to collect GIS data directly in the field. Data collection can be done using field survey instruments that download data directly into GIS or via GPS receivers that directly interface with GIS software on portable PCs. These techniques can eliminate the need for GIS source data.

But during the last many years, GIS data most often have been digitized from several sources, including hard copy maps, rectified aerial photography and satellite imagery. Hard-copy maps (e.g. paper, vellum and plastic film) may contain unintended production errors as well as unavoidable or even intended errors in presentation. The following are "errors" commonly found in maps.

2.14.4. Map Generalization

Cartographers often deliberately misrepresent map features due to limitations encountered when working at given map scales. Complex area features such as industrial buildings may have to be represented as simple shapes. Linear features such as roads may have to be represented by parallel lines that appear wider on a map. Curvilinear features such as streams may have to be represented without their smaller twists and bends.

Indistinct Boundaries

Indistinct boundaries typically include the borders of vegetated areas, soil types, wetlands and land use areas. In the real world, such features are characterized by gradual change, but cartographers represent these boundaries with a distinct line. Some compromise is inevitable.

Map Scale

Cartographers and photogrammetrists work to accepted levels of accuracy for a given map scale as per National Map Accuracy Standards. Locations of map features may disagree with actual ground locations, although the error likely will fall within specified tolerances. Of course, the problem is compounded by limitations in linear map measurements-typically about 1/100th of an inch on a map scale.

Map Symbology

It is impossible to perfectly depict the real world using lines, colors, symbols and patterns. Cartographers work with certain accepted conventions. As a result, facts and features represented on maps often must be interpreted or interpolated, which can produce errors. For example, terrain elevations typically are depicted using topographic contour lines and spot elevations. Elevations of the ground between the lines and spots must be interpolated. Also, areas symbolized as "forest" may not depict all open areas among the trees.

Errors during Data Entry

GIS data typically are created from hard copy source data. The process often is called "digitization", because the source data are converted to a computerized (digital) format. Human digitization can compound errors in source data as well as introduce new errors. The following are the primary methods of digitizing hard copy source data:

Manual Digitizing

Although manual digitizing is used less often today, it was the predominant digitizing method in the 1980s. Maps are affixed to digitizing tables, registered to a GIS coordinate system and "traced" into a GIS. A digitizing table has embedded in its surface a fine grid of wires that sense the position of a cross hair on a hand held cursor. When a cursor button is pressed, the system records a point at that location in the GIS database. The operator also identifies the type of feature being digitized as well as its attributes.

Photogrammetric mapping also is a manual digitizing process. Through an exacting and rigorous technical process of aerotriangulation, overlapping pairs of aerial photographs are registered to one another and viewed as a 3-D image in a stereoplotter or via special 3-D viewers. In a process called "stereocompilation," a photogrammetrist traces map features that are encoded directly into a database.

Scanning and Keyed Data Entry

In scanning, source data are mechanically read by a device that resembles a large format copy machine. Sensors encode the image as a large array of dots, much like a fax machine scans a letter. High resolution scanners can capture data at 2,000 dots per inch (dpi), but maps and drawings typically are scanned at 100 dpi to 400 dpi. The resulting raster image then is processed and displayed on a computer screen. Further onscreen manual digitizing (i.e. "heads-up digitizing") usually is needed to complete the data entry process. If the source data contain coordinate values for points or the bearings and distances of lines (e.g. parcel lines), then map features can be keyed into a GIS with great precision.

General Data Entry

Accurate digitizing is not easy. It requires certain basic physical and visual skills as well as training, patience and concentration. There also are many opportunities for error, because the process is subject to visual and mental mistakes, fatigue, distraction and involuntary muscle movements. In addition, the "set up" of a map on a digitizing table or a scanned raster image can produce errors. Cell size of a scanned raster image also can affect the accuracy of heads- up digitizing.

A digitizer must accurately discern the centre of a line or point as well as accurately trace it with a cursor. This task is especially prone to error if the map scale is small and the lines or symbols are relatively thick or large. The method of digitizing curvilinear lines also affects accuracy. "Point-mode" digitizing, for example, places sample points at selected locations along a line to best represent it in a GIS. The process is subject to judgment of the digitizer who selects the number and placement of data points. "Stream-mode" digitizing collects data points at a pre-set frequency, usually specified as the distance or time between data points. Every time an operator strays from an intended line, a point digitized at that moment would be inaccurate. This method also collects more data points than may be needed to faithfully represent a map feature. Therefore, post-processing techniques often are used to "weed out" unneeded data points.

Heads-up digitizing often is preferred over table digitizing, because it typically yields better results more efficiently. Keyed data entry of land parcel data is the most precise method. Moreover, most errors are fairly obvious, because the source data usually are carefully computed

and thoroughly checked. Most keyed data entry errors show as obvious mismatches in the parcel "fabric."

GIS software usually includes functions that detect several types of database errors. These error-checking routines can find mistakes in data topology, including gaps, overshoots, dangling lines and unclosed polygons. An operator sets tolerances that the routine uses to search for errors, and system effectiveness depends on setting correct tolerances. For example, tolerances too small may pass over unintentional gaps, and tolerances too large may improperly remove short dangling lines or small polygons that were intentionally digitized.

Errors during Data Analysis

Even if "accurate," the manipulation and analysis of GIS data can create errors introduced within the data or produced when the data are displayed on screen or plotted in hard copy format.

2.14.5. Sources of Possible Errors

Obvious Sources of Error

1) Age of data. With the exception of geological data the reliability decreases with age.
2) Areal coverage: partial or complete. Many countries still have fragmentary coverage of maps at scales of 1:25000 to 1:50000. Moreover, during the last 30 to 40 years, concepts and definitions of map units, the way they should be mapped have changed.
3) Map scale.
4) Density of observation. How dense should observations be "to support a map".
5) Relevance. Not all data used are directly relevant for the purpose for which they are used. Prime example: remotely sensed data.
6) Accessibility. Not all data are equally accessible (e.g. military secrecy). Sometimes data is not available because of inter-department secrecy.
7) Cost.

Errors Resulting from Natural Variations or from Original Measurement

a) Positional accuracy. Topographical data often available with a high degree of positional accuracy. But position of vegetation boundaries etc. often influenced by the subjective judgment of surveyor or by interpretation of remotely sensed data.

b) Accuracy of content: qualitative and quantitative. The problem of whether the attributes attached to points, lines or polygons are correct and free from bias. Sometimes systematic errors occur because of instrument. If pixel is too large then it is not clear that it should be classified as forest, road or camp.

c) Sources of variations in data: data entry, observer bias, natural variation. Data entry error. Field data very much influenced by surveyor (elevations or census takers).

d) Errors arising through processing. Numerical errors in the computer (e.g. joining, matching of a field in GIS software), rounding off errors, truncation etc.

1) Faults arising through topological analyses. Problems associated with map overlay. Digitizing considered infallible. Boundary data is assumed to be sharply defined. All algorithms are assumed to operate in a fully deterministic way.
2) Classification and generalization problems: class intervals, interpolation.

The concepts of complete, compatible, consistent and applicable GIS data previously defined particularly apply to data analysis. Users must consider whether a selected GIS dataset is complete, consistent and applicable for an intended use, and whether it is compatible with other datasets used in the analysis.

The phrasing of spatial and attribute queries also may lead to errors. In addition, the use of Boolean operators can be complicated, and results can be decidedly different, depending on how a data query is structured or a series of queries are executed. For example, the query, "Find all structures within the 100 year flood zone," yields a different result than, "Find all structures touching the 100 year flood zone." The former question will find only those structures entirely within the flood zone, whereas the latter also will include structures that are partially within the zone.

Dataset overlay is a powerful and commonly used GIS tool, but it can yield inaccurate results. To determine areas suitable for a specific type of land development project, one may overlay several data layers, including natural resources, wetlands, flood zones, land uses, land ownership and zoning. The result usually will narrow the possible choices down to a few parcels that would be investigated more carefully to make a final choice. The final result of the analysis will reflect any errors in the original GIS data. Its accuracy only will be as good as the least accurate GIS dataset used in the analysis.

It is also common to overlay and merge GIS data to form new layers. In certain circumstances, this process introduces a new type of error: the polygon "sliver." Slivers often appear when two GIS datasets with common boundary lines are merged. If the common elements have been digitized separately, the usual result will be sliver polygons. Most GIS software products offer routines that can find and fix such errors, but users must be careful in setting search and correction tolerances.

2.14.6. Controlling Errors

GIS data errors are almost inevitable, but their negative effects can be kept to a minimum. Knowing the types and causes of GIS data errors is half the battle; the other half is employing proven techniques for quality control at key stages in the GIS "work flow."

Many errors can be avoided through proper selection and "scrubbing" of source data before they are digitized. Data scrubbing includes organizing, reviewing and preparing the source materials to be digitized. The data should be clean, legible and free of ambiguity. "Owners" of source data should be consulted as needed to clear up questions that arise.

Data entry procedures should be thoroughly planned, organized and managed to produce consistent, repeatable results. Nonetheless, a thorough, disciplined quality review and revision process also is needed to catch and eliminate data entry errors. All production and quality control

procedures should be documented, and all personnel should be trained in these procedures. Moreover, the work itself should be documented, including a record of what was done, who did it, when was it done, who checked it, what errors were found and how they were corrected.

To avoid misusing GIS data and the misapplication of analytical software, GIS analysts including casual users need proper training. Moreover, GIS data should not be provided without metadata indicating the source, accuracy and specifics of how the data were entered.