/

UNIT - II

Spatial Data Models

Database Structures - Relational, Object Oriented - ER diagram - spatial Data Models - Raster Data Models - Raster Data Compression - Vector Data structures - Raster vs vector Models - TIN and GRID data models - OGC standards - Data Quality.

Database Model:

Data model defines the logical structure of a database. Data models are fundamental entities to introduce abstraction in a DBMS. Data models define how data is connected to each other and how they are processed and stored inside the system. There are a number of different database data models. Amongst those that have been used for attribute data in GIS are the hierarchial, network, relational, object - relational and object - oriented data models, Of those the relational data model has become the most widely used model.

Relational Data Model:

Data are organized in a series of 2-D tables, each of which contains records for one entity. These tables are linked by common data known as keys.

Queries are possible on individual tables groups of tables. For the happy Valley data, the below figure illustrates an example of one such table.

Hotel ID	Name	Address	No of Rooms	Standard
001	Mountain View	23, High Street	15	Budget
002	Palace Deluxe	Pine Avenue	12	Luxuary
003	Ski Lodge	0, Ski School Road	40	Standard.

The data in a relational database are stored as a set of base tables with the characteristics described above. Other tables are created as the database is queried and these represent virtual views. The table structure is extremely flexible and allows a wide variety of queries on the data. Queries are possible on one table at a time, or on more than one table by linking through key fields. Queries generate further tables, but these new tables are not usually stored. There are few restrictions on the types of query possible.

With many relational databases querying is facilitated by menu systems and icons or 'query by example' systems. Frequently, queries are built up of expressions based on relational algebra, using commands such as SELECT, PROJECT, JOIN. SQL has been developed to facilitate the querying of relational databases.

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 type, the constraints and the relationships for the description or storage of data. Following are most often used data models.

i) Hierarchical Data Structure Model

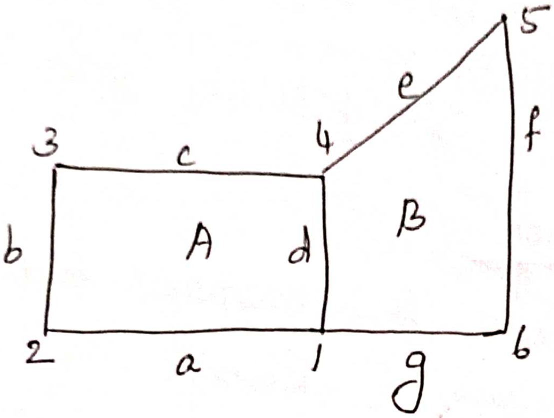
ii) Network Data Structure Model

iii) Relational Data Structure Model

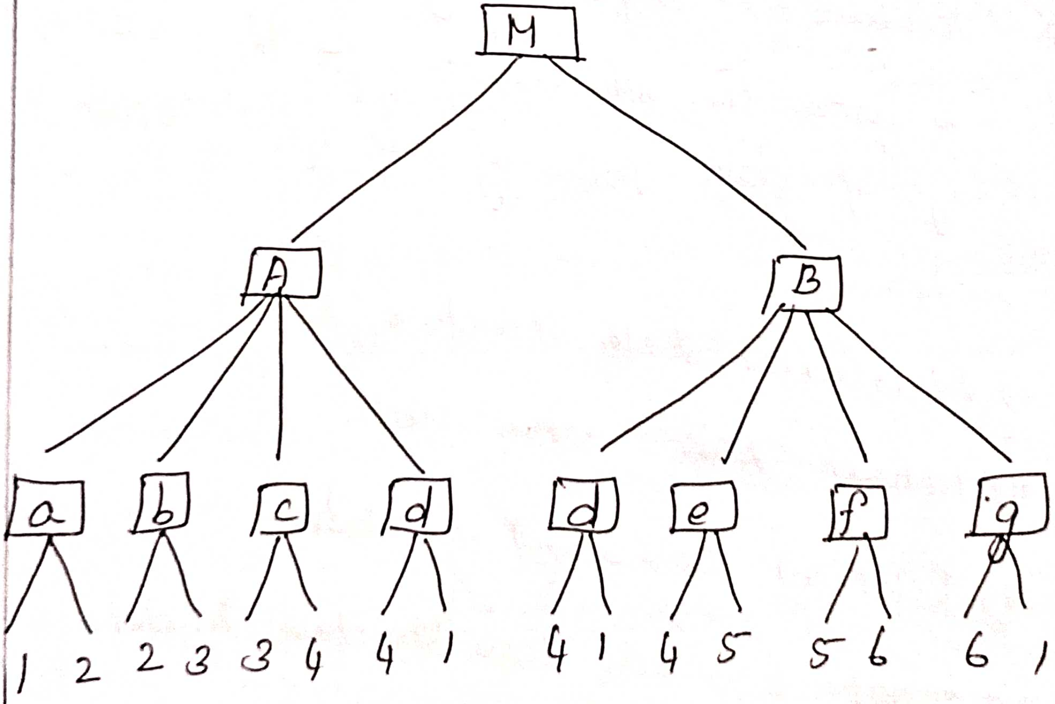
iv) Object Oriented Data Structure Model.

Hierarchial 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. Records are connected through pointers that store the address of the related work.



Map.



Each pointer establishes a parent-child relationship where a parent can have more than one child but a child can have only one parent.

There is no connection between the elements at the same level. To locate a record, you have to start at the top of the tree, with a parent record and trace down the tree to the child.

Advantages:

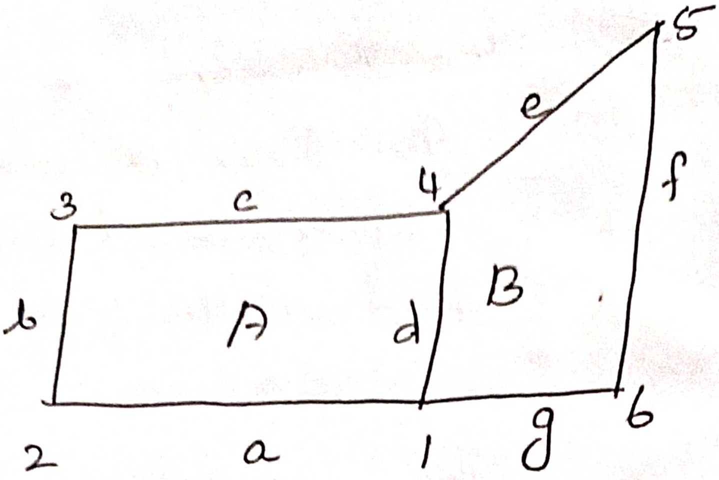
* Easy to understand
* Accessing records and updating records are very fast.

Disadvantages:

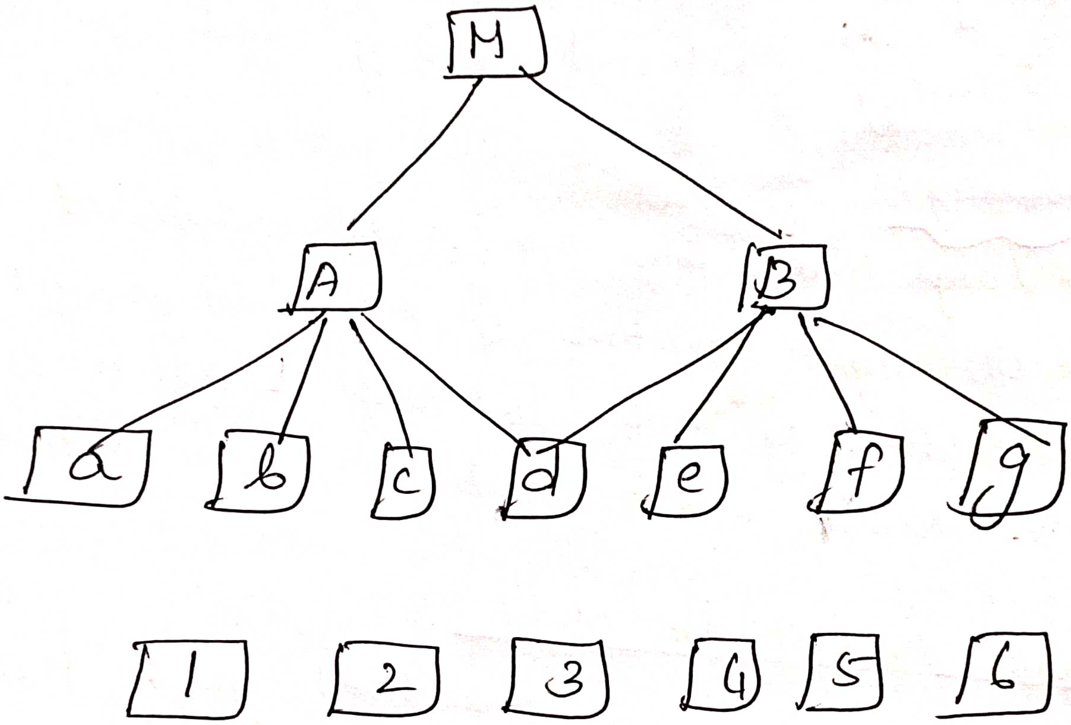
* Large index files are to be maintained.
* The rigid structure of this model doesn't allow alteration of tables, therefore to add a new relationship entire database is to be redefined.

Network Data Structure Model:

A network is a generalized graph that captures relationships between objects using connectivity. 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 ~~ea~~ each record to have many parents and many children thus allowing a natural model of relationships b/w entities.



Map



Advantages:

* Many to many relationships are easily implemented in a network data model.

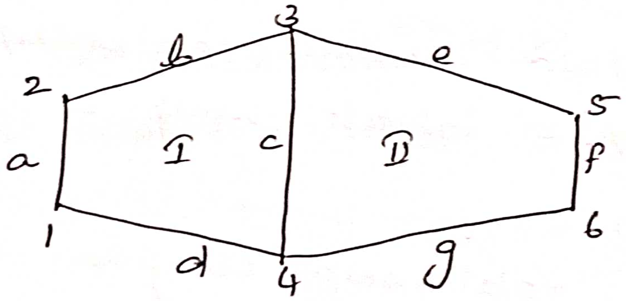
* Data access and flexibility in network model is better than that in hierarchical model.

* Eliminates redundancy.

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. 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.



Maps:

	P1	P2
M	I	II

Polygons:

I	a	b	c	d
II	e	c	f	g

Lines:

a	1	2
b	2	3
c	3	4
d	4	1
e	3	5
f	5	6
g	6	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:

General Properties: there should be an inheritance relationship.

Abstraction: Objects, Classes, and super classes are to be generated by classification, generalization, association and aggregation.

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

ER Diagram:

An entity-relationship diagram 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 a ERD are:

- Entities
- Relationship
- Attributes.

Steps involved in creating ERD include:

i) Identifying and defining the entities

ii) Determining all interactions between the entities.

iii) Analyzing the nature of interactions/determining the cardinality of the relationships

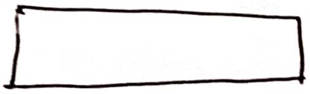
iv) Creating ERD.

Entity-Relationship Diagram:

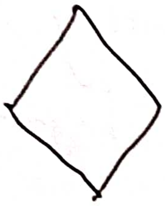
An ERD is crucial to creating a good database design. It is used as a high-level logical data model, which is useful in developing a conceptual design for databases.

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.

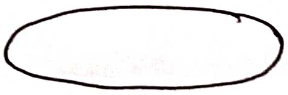
An attribute of an entity is a partic
property that describes the entity. A relationship
the association that describes the entity. A
relationship is the association that describes the
interaction between entities. Cardinality, in the
context of ERD, is the number of instances of
one entity that can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



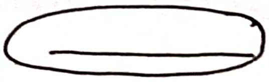
Entity



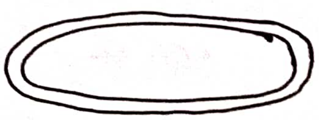
Relationship



Single Valued Attribute



Primary Key



Multivalued attribute

M-1 Many to one cardinality

1-M One to many cardinality.

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 ~~about~~ 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 data, unless the data are organized primarily by spatial attributes.

* 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.

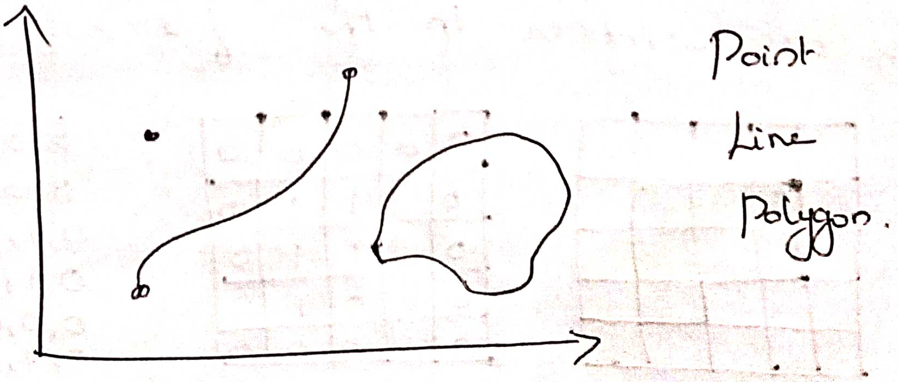
Traditionally spatial data has been stored and presented in the form of a map. Three basic types of spatial data evolved for storing data digitally. These are referred to as

- Vector
- Raster
- Image.

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 sequencial 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 consists of a string of vertices terminated by a node. A node is defined as vector data formats.

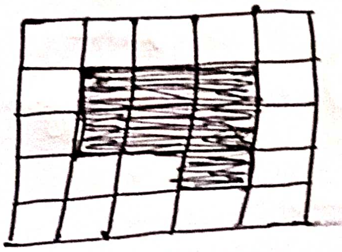


Spatial Data Structures:

Data structures provide the information that the computer requires to reconstruct the spatial data model in digital form. There are many different data structures in use in GIS. This diversity is one of the reasons why exchanging spatial data between different GIS software can be problematic. However, despite this diversity data structures can be classified according to whether they are used to structure raster or vector data.

Raster Data Structures:

In the raster world a range of different methods is used to encode a spatial entity for storage and representation in the computer. The below figure shows the most straightforward method of coding raster data. The cells in each line of the image are mirrored by an equivalent row of numbers in the file structure.



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

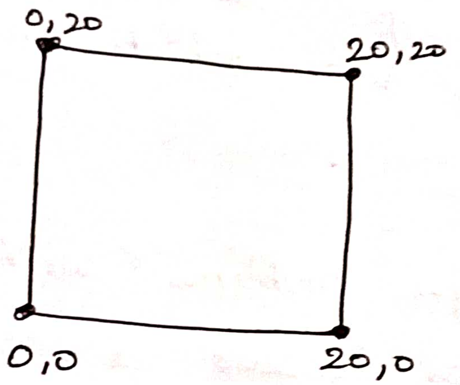
5,5,1
0,0,0,0,0
0,1,1,1,0
0,1,1,1,0
0,0,0,1,0
0,0,0,0,0

In a simple raster data structure, as illustrated above, different spatial features must be stored as seperate data layers. Thus to store more ~~d~~ raster entities, seperate data files would be required, each representing a different layer of spatial data. However, if the entities donot occupy the same geographic location, then it is possible to store them all in a single layer, with an entity code given to each cell. This code informs the user which entity is present in which cell.

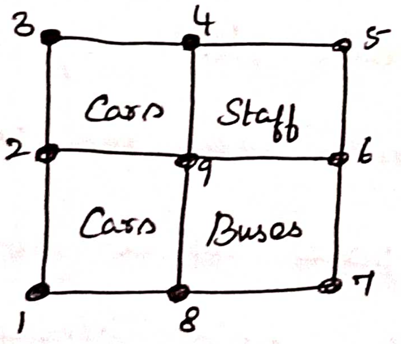
One of the major problems with raster data set is their sizes, because a value must be recorded and stored for each cell in an image. Thus, a complex image made up of a mosaic of different features requires the same amount of storage space as a similar raster map showing the location of a single forest. To address this problem a range of data compression methods have been developed.

Vector Data Structure:

There are many potential vector data structures that can be used to store the geometric representation of entities in the computer. The simplest vector data structure that can be used to represent a geographical image in the computer is a file containing (x,y) co-ordinate pairs that represent the location of individual point features.



x	y
0	0
0	20
20	20
20	0
0	0



ID	Points
Cars	1,2,3,4,9,8
Staff	4,5,6,9
Buses	6,7,8,9

Polygon file.

Id	x	y
1	0	0
2	0	10
3	0	20
4	10	20
⋮	⋮	⋮
9		
10	10	10

The above figure shows such a vector data structure for the Happy Valley car park. Note how a closed ring of co-ordinate pairs defines the boundary of the polygon. The limitations of simple vector data structure start to emerge when more complex spatial entities are considered. For eg) consider the Happy Valley car park divided into different parking zones. The car park consists of a number of adjacent polygons. If the simple DS illustrated in figure were used to capture this entity then the boundary line shared between adjacent polygons would be stored twice. This may not appear too much of a problem in the case of this example, but consider the implications for a map of the 50 states in USA.

There is a considerable range of topological data structures in use. by GIS. All the structures available try to ensure that

* no node or line segment is duplicated.

* Line segments and nodes can be referenced to more than one polygon.

* all polygons have unique identifiers and

* island and hole polygons can be adequately represented.

Raster Data Compression:

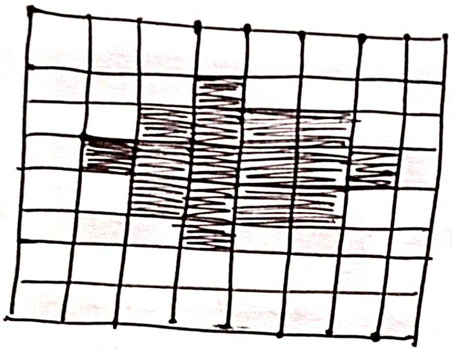
Data compression refers to the reduction of data volume, a topic particularly important for data delivery and web mapping. Data compression is related to how raster data are encoded. Quadtree and RLE, because of their efficiency in data encoding, can also be considered as data compression methods.

A variety of techniques are available for data compression. They can be lossless or lossy. A lossless compression preserves the cell or pixel values and allows the original raster data that are used for analysis or deriving new data. RLE is an example of lossless compression. Other methods include LZW and its variations.

A lossy compression cannot reconstruct fully the original Image but can achieve higher compression ratios than a lossless compression. Lossy compression is therefore useful for raster data that are used as background images rather than for analysis. Image degradation through lossy compression can affect GIS related tasks such as extracting ground control points from aerial photographs or satellite images for the purpose of georeferencing.

1) Run Length Encoding:

Run length encoding stores cells on a row-by-row basis. Instead of recording each individual cell's values, run length encoding groups cell values by row.

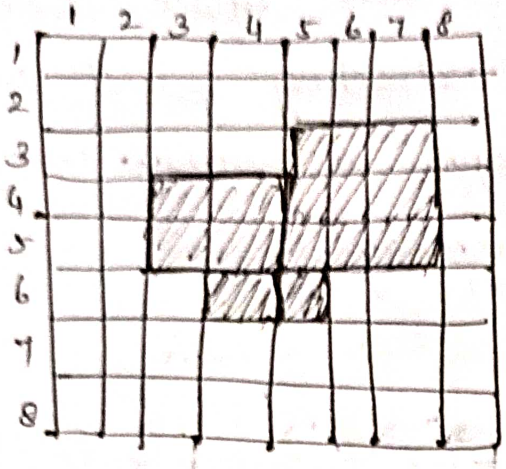


(8, 8, 1)
(0, 8)
(0, 3) (1, 1) (0, 4)
(0, 2) (1, 4) (0, 2)
(0, 1) (1, 6) (0, 1)
(0, 2) (1, 4) (0, 2)
(0, 8)
(0, 8)

2) Block Encoding:

The block encoding raster storage technique assigns areas that are blocks to reduce redundancy.

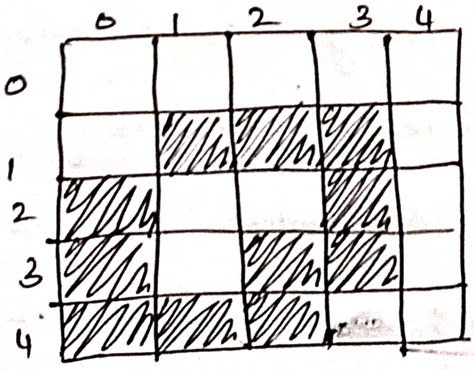
The block coding raster image compression method subdivides an entire raster image into hierarchial blocks. It's an extension of the run length encoding technique, but extends it to two dimensions.



Block size: 9
Count : 1
Coordinate : (5,3)
Block Size: 4
Count : 1
Coordinate : (3,4)
Block Size: 1
Count : 2
Co-ordinate : (4,6) (5,6).

Chain Coding:

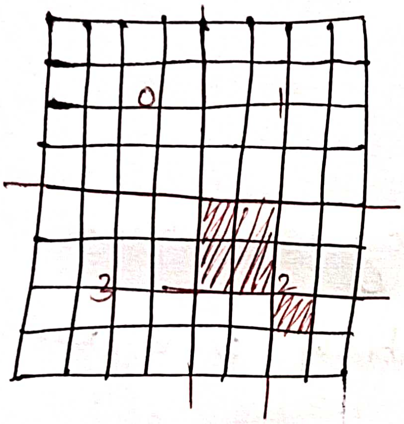
Chain coding defines the outer boundary using relative positions from a start point. The sequence of exterior is stored where the endpoint finishes at the start point. During the encoding, the direction is stored as an integer. However, in this example we can cardinal directions for simplicity. ~~Par~~.

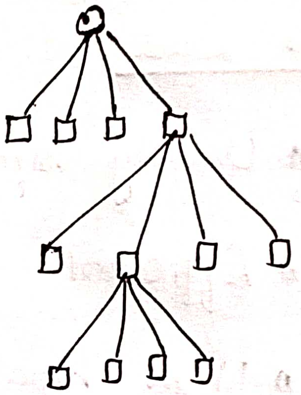


(1,1)
E3, S2, W1, S1, W2, N2.

4) Quad Tree Encoding:

Quad trees are raster data structures based on the successive reduction of homogenous cells. It recursively subdivides a raster image into quarters. The subdivision process continues until each cell is classed.





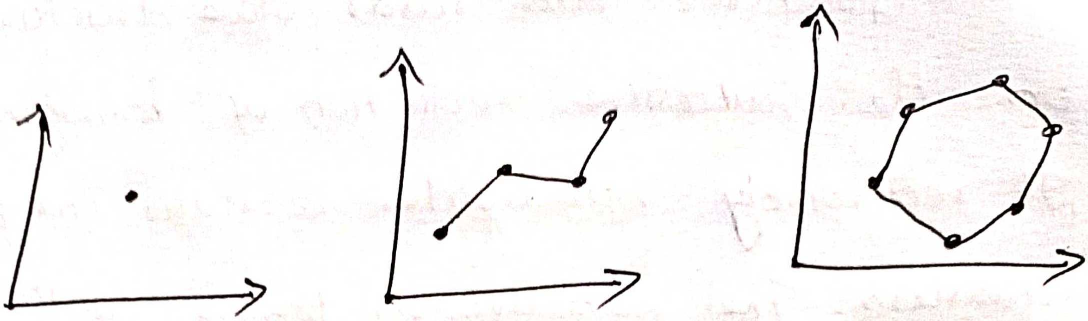
Vector Vs Raster Data Set:

Vector:

- * Usually complex
- # Difficult for overlay operation
- # High spatial variability is inefficiently represented.
- # Small file size
- # Vector model is often used for representing

discrete features with definable boundaries

Eg)

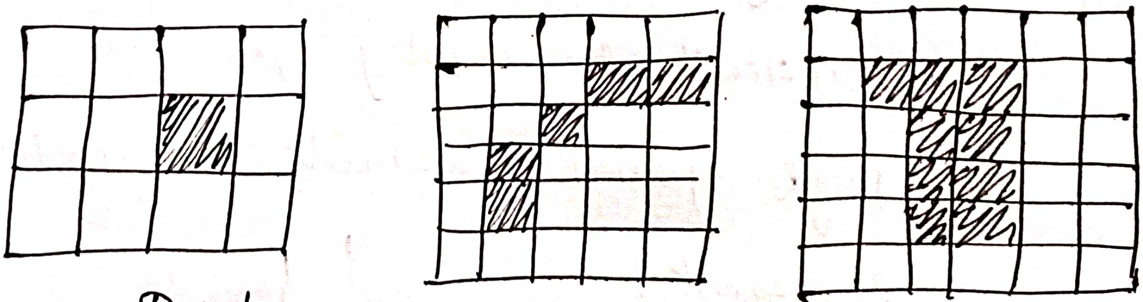


Point Line Polygon

Raster Data :

* Usually simple
* Efficient for overlay operations
* High spatial variability is efficiently represented.
* Large file size
* Raster data model is widely used for representing continous spatial features.

Eg)



Point Line Polygon

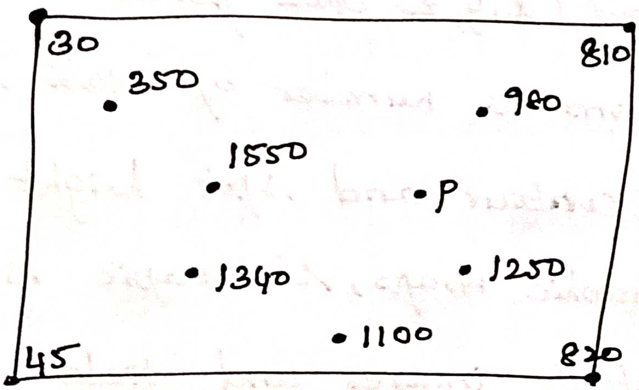
Digital Terrain Modelling:

The abbreviation DTM is used to describe a digital data set which is used to model a topographic surface. To model a surface accuratly it would be necessary to store an almost infinite number of observations. Since this is impossible, a surface model approximate a continous surface using a finite number of observations. Thus, an appropriate number of observations must be selected, along with their geographical location.

The resolution of a DTM is determined by the frequency of observations used. DTMs are created from a series of either regularly or irregularly spaced (x, y, z) data points. DTMs may be derived from a number of data sources. These include contour and spot height information found on topographic maps, stereoscopic aerial photography, satellite images and field surveys.

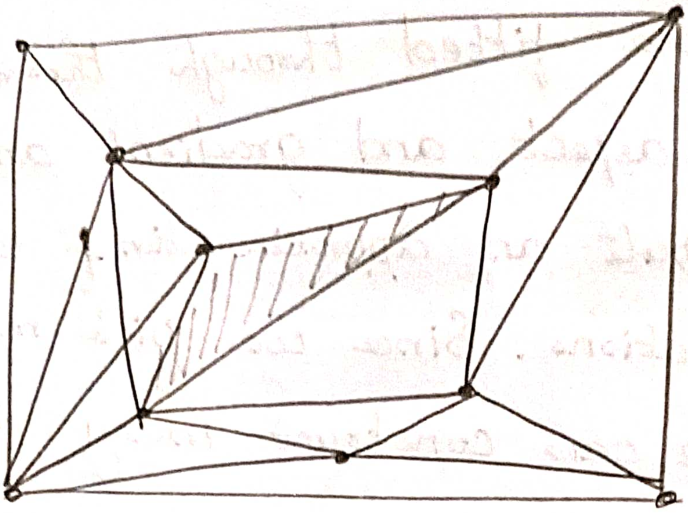
Triangulated Irregular Networks:

A commonly used data structure in GIS is triangulated irregular network. It is on the standard implementation techniques for digital terrain models, but it can be used to represent any continous field. The principles behind a TIN are simple. It is built from a set of locations for which we have a measurment for instance an elevation. The locations can be arbitrarily scattered in space and are usually not on a nice irregular grid. Any location together with its elevation value can be viewed as a point in three dimensional space. This is illustrated in below figure. From these 3D points, we can construct an irregular tessellation made of triangles.

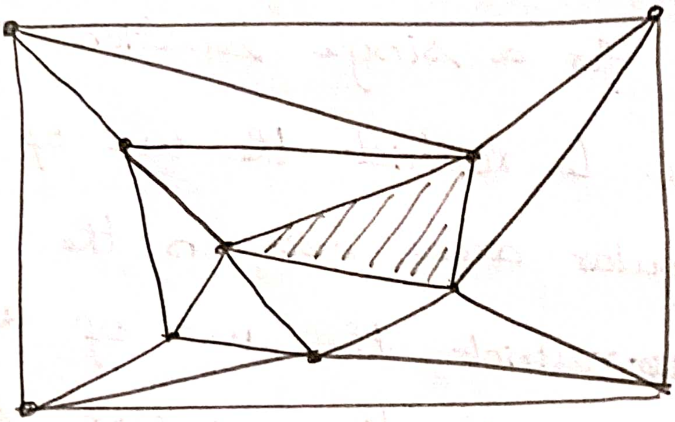


In 3D space, three points uniquely determine a plane; as long as they not collinear, ie) they must not be positioned on the same line. A plane fitted through these points has a fixed aspect and gradient and can be used to compute an approximation f elevation of other locations. Since we pick many triples of points, we can construct many such planes and therefore we can have many elevation approximations for a single location such as 'p'. So it is wise to restrict the use of a plane to the triangular area between the three points.

If we restrict the use of a plane to the area between its three anchor points, we obtain a triangular tessellation of the complete study space. Unfortunatly, there are many different tessellations for a given input set of anchor points. Some tessellations are better than others, in the sense that they make smaller errors of elevation approximations. For instance, it we base our elevation computation for location p' on the left hand shaded triangle, we will get another value than from the right hand shaded triangle.



(a)



(b)

The second will provide a better approximation becaus the average distance from p' to the three triangle anchors is smaller. The triangulation shown in below figure happens to be a Delaunay triangulation, which in a sense is an optimal triangulation. There are multiple ways of defining what such a triangulation is, but we suffice here to state two important properties. The first is that the triangles are as equilatiral as they can be, given the set of anchor points. The second property is that for each triangle, the circumcircle through its three anchor points does not contain any other anchor point. One such circumcircle is depicted on the right figure.

A TIN clearly is a vector representation; each anchor point has a stored georeference. Yet, we might also call it an irregular tessellation, as the chosen triangulation provides a partitioning of the entire study space. However, in this case, the cells do not have an associated stored value as it typical of tessellation, but rather a simple interpolation function that uses the elevation values of its three anchor points.

GIS Data Standards:

The number of formats available for GIS data is almost as large as the number of GIS packages on the market. This makes the sharing of data difficult and means that data created on one system is not easily read by another system. This problem has been addressed in the past by including data conversion functions in

GIS software. These conversion functions adop commonly used exchange formats such as DXF and E00.

Open Geospatial Consortium (OGC):

There is still no universally accepted GIS data standard, although the Open Geospatial Consortium (OGC,) formed in 1994 by a group of leading GIS software and data vendors, is working to deliver spatial interface specifications that are available for global use. The OGC has proposed the Geography Markup Language as a new GIS data standard.

The Geography Markup Language is a non-proprietary computer language designed specifically for the transfer of spatial data over the internet. GML is based on XML, the standard language of the internet.

GML has proposed by the OGC as an universal spatial data standard. GML is likely to become very widely used because it is:

i) Internet friendly

ii) not tied to any proprietary GIS

iii) Specifically designed for feature-based spatial data

iv) Open to use by anyone

v) Compatible with industry-wide to IT standards.

It is also likely to set the standard for the delivery of spatial information content to PDA and WAP devices, and so form an important component of mobile and location-based GIS technologies. The collection of geoportals and various other complimentary services, create a Spatial Data Infrastructure (SDI).

Spatial Data Infrastructure:

An SDI is used to represent all the components that enable access to spatial data including relevant technologies, policies and instituitional arrangements. Using electronic media, SDI connect nationally distributed repositories of geospatial information and make them availability on a device through a single entry point often reffered to as a 'geoportal'. They facilitate data providers and users to participate in the digital spatial community at a national scale and provide a basis for spatial data discovery, elevation and application for users within government, commercial and non-profit sectors, and academia and by citizens in general. The Global Spatial Data Infrastructure Association links national SDIs to establish a connection for all users to share and reuse the available datasets.