Data Structure in GIS

Dr.Weerakaset Suanpaga (D.ENG)

Department of Civil Engineering
Faculty of Engineering , Kasetsart University
Bangkok, Thailand

http://pirun.ku.ac.th/~fengwks/gis/**lecture/2datastructure.pdf**

Data Structure in GIS

- **Organization of data in an information system is referred to as data structure**
- **Data must be organized in a well planned structure in a GIS before commencement of processing**
- **Different kind of spatial data**
  - Theme or
  - Datalayer or
  - Dataplane

- Three geometrical entities in each datalayer
  - Points
  - Lines
  - Polygons
- Points
  - Location of Tube-wells, Water Tanks, Sampling Stations of Rain Gauge etc..
- Lines
  - Roads, Canals, Streams etc..
- Polygons
  - Reservoir, Lake, District, State, Country Boundaries etc..

Essential Spatial Information

- Attribute Data or Ancillary Information
- For a Tube-Well
  - Essential information is its location : **Geodetic Co-ordinates (x, y)**
  - Attributes : Ownership, Depth (deep/shallow, quantitatively), Quality of Water, Pumping volume and Rate, Date of Boring, Expected Life, etc.

DBMS

- Most of the Geographic Information Systems have the inbuild capabilities to store and manipulate the attribute data in addition to spatial information.
  - Database Management System (DBMS)
    - Attribute Table may be generated for further processing, linking and analysis.

Raster Data Structure

- Cellular Organization of Spatial Data
- The image is arranged in form of 'cells' at regular interval
  - The parameters of interest are arranged in these cells.

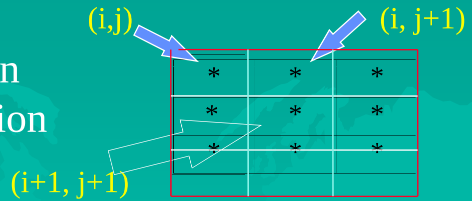
Simple Raster Arrays

- Arranged in Rows and Columns
  - Rows : in East - West direction
  - Columns : in North - South direction
- Origin of Raster Image is generally at top left corner : Position (0, 0) or (1,1)
- Distance between cells in row and column directions is constant
- Most popular cell structure : Rectangle

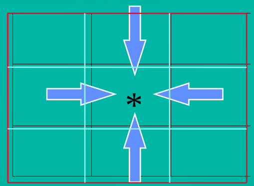
Limitations

- Limitaion of specifying location
- Adjoining cells may not be evenly spaced

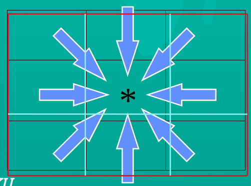
Location Limitation



Four-Connected Evenly Spaced



Eight-Connected Uequally Spaced

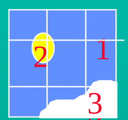


Raster Cell Size

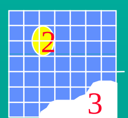
- Minimum Mapping Unit (MMU)
- Raster Cell Size
- These two are not same
- They are in terms of appropriate MMU or Resel

Minimum Mapping Unit

1.Green Land
2.Coconut Tree
3.Palm Tree



a) Vegetation Map
b) Raster Map (Poor)
c)Raster Map (Better)



•Tessellations

- Geometrical figures that completely cover a flat surface
  - Examples: Square, Triangular, Hexagonal
- Problems in Triangular and Hexagonal Tessellations
  - can not be divided into smaller sizes of same shape
  - numbering system becomes cumbersome

Volume of Data

- Raster Cells or **Pixels** (Picture Elements)
- In satellite data of IRS-1C PAN 2048X2048 pixels cover an area of 5.8X2048X2048 m^2
- Data Compression Techniques
  1. Run Length Encoding
  2. Chain Coding

◆ Run Length Encoding :

Original data is replaced by data pairs or tuples

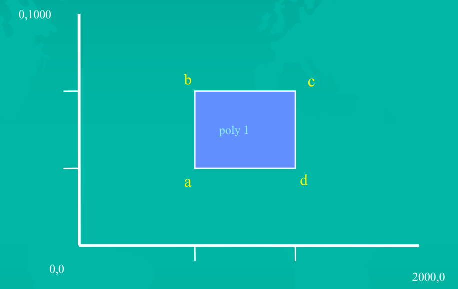
- 12, 12, 15, 15, 15, 15, 17, 17, 17, 17, 17, 17, 17,
- Encoded as (2, 12) (4, 15) (7, 17)
- Reduction from 13 elements to 6 elements
- Good compression when repeating data are available

hain Code :

lap is considered as spatially referenced bject placed on a back ground

Storing the Areas :

Record the starting point on the border of the object. Sequence of cardinal directions of the cells make up the boundary



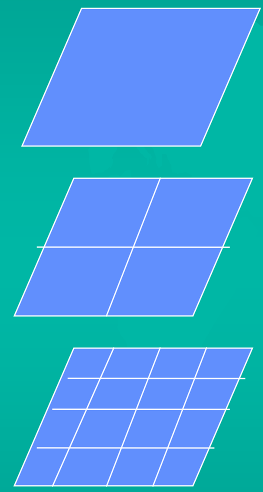
Chain code method

Coordinates of a,b,c,d - poly 1

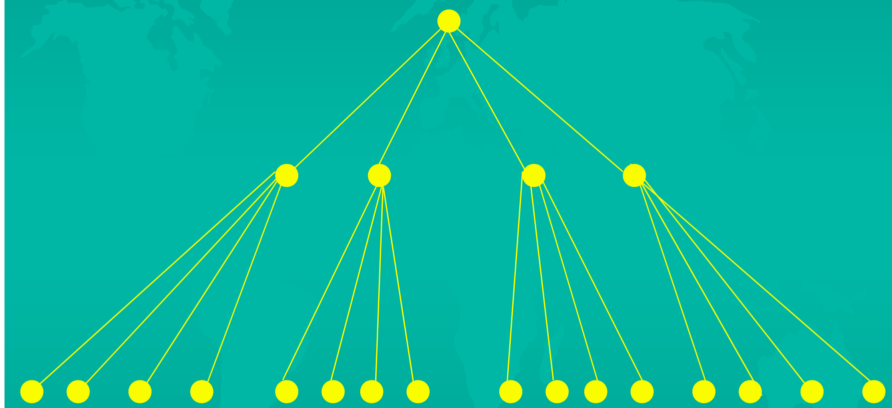
Hierarchical Raster Data Structure

- A modified Raster Structure
- Information is stored in inter-related multiple layers
- Also understood as **PYRAMIDAL** Data Structure
- A particular form is **QUADTREE** Data Structure

Raster Layers of Different Cells Sizes



Tree Transitions



Quadtree Data Structure

- Higher level pixels has twice the width & height of the previous level (area is four times)
- Four-fold reduction in number of pixels in each layer (see next slide)
- Structure represents a TREE
- Advantage:
  - Sorting & Searching is facilitated
  - Saving of computational time as some processing steps are not required
- Disadvantage: More space is required by the dat

Cosider a raster with 32 size. This raster requires 32X32 = 1024 cells. Considering higher cells, we require:

Layer	Width in Cell	Total Cells
1	32	1024
2	16	256
3	8	64
4	4	16
5	2	4
6	1	1

Vector Data Structure

- A vector is defined with reference to its
  - Starting point, Associated Displacement and Direction (or bearing)
- In Spatial Database object is defined as
  - Point, Line, Circle, Polygon (with its co-ordinate or description)
- For Example,
  - Circle is specified by co-ordinate of center and radius

Vector Data Structure

- In Raster Structure, plane of image is decomposed in smaller cells
  - Circle is represented by pixel occupying perimeter
  - Problem in deciding cell size
- Most of the computer graphics and CAD systems are using Vector Data Stucture
- In GIS spatial data is encoded and processed as Vector Data Structure

Common Vector Data Structures

- Whole Polygon System
- Dual Independent Map Encoding (DIME) file structure
- Arc-Node Structure
- Relational Structure
- Digital Line Graph

Whole Polygon Structure

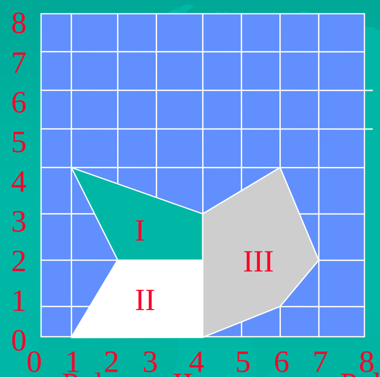
- Each Layer in database is dissolved in number of polygons
- Each polygon is encoded as a sequence of locations that define the boundaries of each closed area in a specified co-ordinate system
- Each polygon is then stored as an independent feature
- No explicit means to define adjacent area

- Attribute of each polygon may be stored with the coordinated list
- Topographical organization is missing

TOPOLOGY : The relationship between different spatial objects e.g. which polygons share a boundary, which points fall along the edge of a particular polygon etc.

- Several lines and points which are shared by adjacent polygons are recorded more than once
  - This creates problem during processing and data get corrupted

Whole Polygon Structure



Polygon I	Polygon II	Polygon III
1, 4	2, 2	6, 4
4, 3	4, 2	7, 2
4, 2	4, 0	6, 1
2, 2	1, 0	4, 0
		4, 3
		4, 3

Nodes that define each polygon are stored separately.

DIME Structure

- US Bureau of Census developed Dual Independent Map Encoding (DIME) File Structure
- Designed to incorporate topological information about urban areas for use in demographic analysis
- This format may be a basic data format while processing in GIS but is used to archive the spatial/topological data

- This facilitates the data exchange in different systems
- The basic element is line segment defined by two end points or nodes
- The line segments and nodes are shared by adjacent polygon units
- Line segments are assumed to be straight
- Curved lines are assumed to be represnted by multiple straight line segments

- Each line segment is stored with three essential component
  - A segment name (name of street)
  - Node identifiers ('From' and 'To' end points)
  - Identifiers for polygons on left and right side of the segment
- A number of additional attributes may be coded in DIME structure

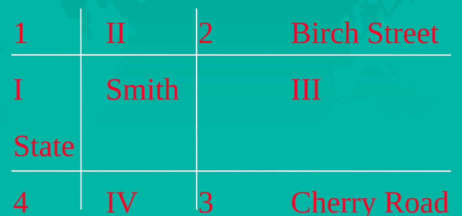
- Disadvantage: Difficulty in manipulating complex lines as in functions that require search along streets
  - Since streets are broken into smaller street segments by the cross-streets, it is a significant computational effort to follow the segments in the sequence when required
- Advantage: It has the ability to match addresses of spatial objects in multiple files since addresses are explicitly stored in the DIME file

Arc-Node Structure

1. Objects in the database are structured hierarchically

1. Points are the basic elemental component

Arc Node Structure



Nodes:

Number	Easting	Northing	TrafficControl	Crosswalk
1	126.5	578.2	Light	Yes
2	218.6	581.9	Sign	Yes
3	224.2	470.4	None	No
4	129.1	471.9	Sign	No

Arcs :

Polygons :

Relational Data Structure

- It is another form of arc-node vector data structure
- In this, attribute information is kept separately
  - In previous Arc-Node example, data attributes were stored with topological information

- This has been adopted by many GIS packages
- Relational Database Management System (RDBMS) softwares are available and can be integrated with some GIS to achieve added flexibility and portability

Relational Data Structure

Nodes:

Number	Easting	Northing	Number	TrafficControl	Crosswalk
1	126.5	578.2	1	Light	Yes
2	218.6	581.9	2	Sign	Yes
3	224.2	470.4	3	None	No
4	129.1	471.9	4	Sign	No

- ARCS: The individual line segments which are defined by a series of (x,y) co-ordinates
- NODES: Intersection of arcs and also terminal points of arcs
- POLYGONS: Areas that are completely bounded by a set of arcs
- Nodes are thus, shared by both arcs and adjacent polygons
- Encoding the geometry with no redundancy
  - Points are stored only once and are reused as often as necessary

Digital Line Graph (DLG)

- Developed by US Geological Survey
- The data contents of the DLG files are subdivided into different thematic layers
  - First layer consists of boundary information including both political and administrative
  - Second layer : Hydrographic Features
  - Third layer : Transportation Network
  - Fourth layer : It is based on Public Land Survey System (US Bureau of Land Management)

- The essential elements of the DLG Level 3 structure are same as discussed in earlier ones
  - Whole nodes are intersection of points or end points
  - Additional features are defined as points on lines
  - Lines have starting and ending nodes
  - These help in specifying direction along the line as well as areas on both the sides (left and right)

- A **degenerate line** defined as a line of zero length and is used to define features that are indicated as a point on the map
  - It has same starting and ending point
- Areas in DLG are completely bounded by lines
- Each area may have an associated point representing the characteristics of the area
- The POINT, LINE and AREA elements provide information about topology and location

Attribute Codes

- Major Code
  - Three digit long
    - First two digits represent general category of elements
    - Third digit represents additional information
- Minor Code
  - Four digit long
    - First digit is generally zero
    - Remaining digits represent details

- To illustrate the details that are stored in DLG format, a few of the codes from the hydrography DLG data layer are prescribed here:

Nodes

050	0001	Upper end of stream
050	0004	Stream entering water body
050	0005	Stream leaving water body

Areas

050	0101	Reservoir
050	0103	Glacier
050	0106	Fish Hatching

Lines

050	0200	Shore Line
050	0201	Man-made shore Line

Degenerate Lines	050 0300 springs 050 0302 flowing well
General Purpose Attributes	050 0400 rapids 050 0401 falls
General Descriptive Attributes	050 0601 underground 050 0604 tunnel

- A DLG Level 3 data file contains a number of header record followed by the data record
- HEADER RECORD : Provides information about date of creation of file, map projection and co-ordinate system; and the number of points, lines and areas stored in the file
- Data Record for NODES include
  - Node Location
  - Major and Minor Attribute Codes
  - Text String

- Data Record for AREA include
  - Description (co-ordinates of points)
  - Attribute Code and Associated Text String
- Data Record for LINES include
  - Description
    - Starting and Ending Nodes
    - Areas on the Left and Right
  - An ordered sequence of (x, y) co-ordinates
  - Attribute Code and Associated Text String

That's all about Data Structure!