Data Management

22-11-2018

Ref: Lec. Notes of Dr. Nitin Tripathai, AIT Thailand.

Data Management

- Core of GIS----Data and Data Management
- It makes information available to user
  - User need not to be familiar
- Data Acquisition and Preprocessing
  - Preparing data for storage and use
- DATABASE: A structured collection of information
- Data Management tools (the software or DBMS) provide safe and efficient access to the data.

Logical & Physical Data

- Logical:
  - The way data appears to the user is logical data
- Physical:
  - The details of the data and its organization as stored in the storage media is Physical Data
- Normally Physical data is kept hidden from user, and only logical data is shown to the user
- Logical data should be dynamic, and should suit to the user, and easy to update

DBMS and its Features

- Data Base Management System (DBMS) are high level computer languages that permit user to efficiently store, analyze and manipulate data.
- Essential Features which must be available in any DBMS:
  - Insert and Delete Data
  - Modify database
  - Query and search data required

Defining a new Database

- Define the format of data
  - Integer or Binary
  - Floating point number or exponential number
  - String or Character
- Data Content Definition
  - Legal Contents should be provided. E.g. if data is about latitude and longitudes then text or road name should not be provided.
- Value Restriction
  - DBMS should put some constraint on Data to reduce input errors.

All the above info are placed in separate place called Data Dictionary or **Metadata** or Data about Data.

Qualities of a GIS Data Base

- Efficiency
  - Storage, retrieval, deletion, insertion, and updating large data sets should be efficient
- Network Capability
  - Multiple User Capability
  - Access to Database at different location
- Lack of data redundancy
  - Redundancy is presence of useless data
  - It may cause data corruption
- Data Independence
  - Data and application program should function independently
- Security
  - Protection against unauthorized modification
- Integrity
  - Ability to protect data from system's problems through variety of tools, such as range checking, backups, and recovery.

These qualities are guidelines, and there can tradeoff depending upon user preference and data type.

Data Structures in GIS

Data Structures

- Organization of the data in an information system is referred as data structure
- Organization of the data should be well planned before commencement of the processing
- Most of GIS have inbuilt capabilities of a good DBMS (Database Management System)
- TWO Main Kinds of DATA Structures
  - RASTER
  - VECTOR

Raster Data Structure

- Cellular organization of spatial data
- The images are arranged in form of ‘cells’ at regular interval
- Arranged in rows and columns
- Origin of raster image is generally at top left corner: position (0,0) or (1,1)
- Distance between cells in rows and column is constant
- Most popular cell structure is ‘Rectangle’
- Limitation of specifying the exact location
- Large Volume of Data (large storage requirements)

Tessellations

- Geometrical figures that completely cover a flat surface
  - e.g. Square, triangles, hexagons
- Problems in Triangular and Hexagon Tessellations is that:
  - Cannot be divided into small sizes of same shape
  - Numbering system becomes cumbersome

Data Compression Techniques

- Run Length Encoding
  - Original data is replaced with data pairs or tuples
  - Example

    Original data: 12,12,15,15,15,15,17,17,17,17,17

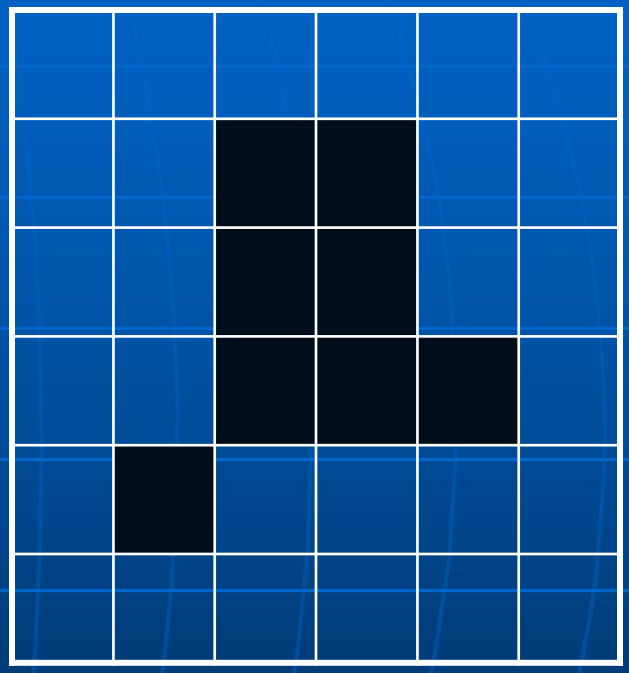
    Encoded data:

    (2,12), (4, 15), (5,17)
  - Benefit: Reduction from 11 elements to 6
  - Good compression if repeating data is available

Data Compression Techniques

- Chain Encoding
  - Represents the boundary of a region by using a series of cardinal direction and cells
  - Example to encode the image.

    N1 E1 N1 E1 N3 E2 S2 E1 S1 W3 S1 W1
  - Starting at lower left cell of the region, the chain code records the regions boundary by using the principle direction and number of cells.
  - Clockwise rotation



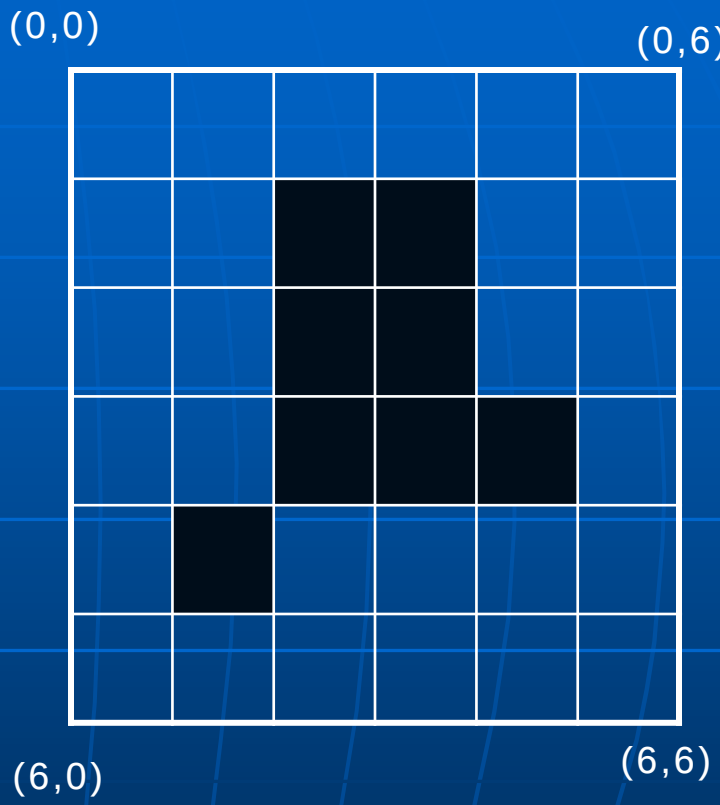
Data Compression Techniques

- Block Code
  - Represents the boundary of a region by using square blocks

  Example to encode the image.

  One four square (1,2)

  Four One square (3,2;3,3; 3,4; 4,1)
  - Specify how many square blocks are there
  - And specify the upper left corner of each square block

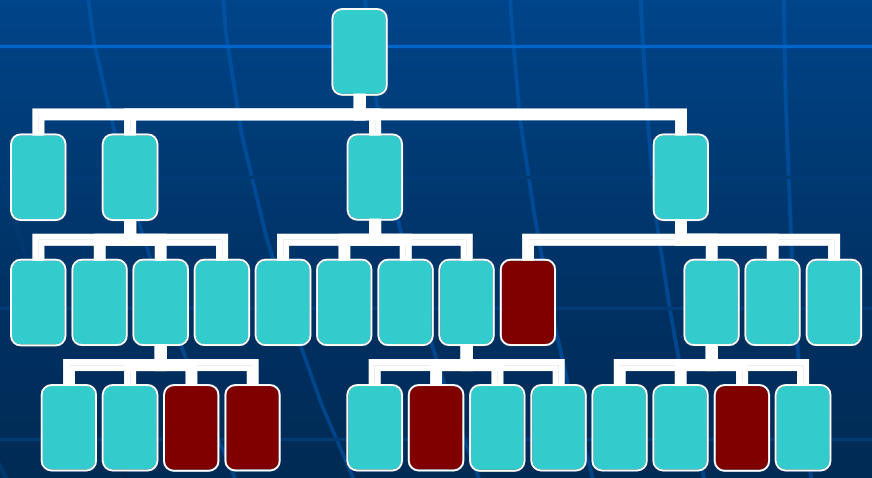


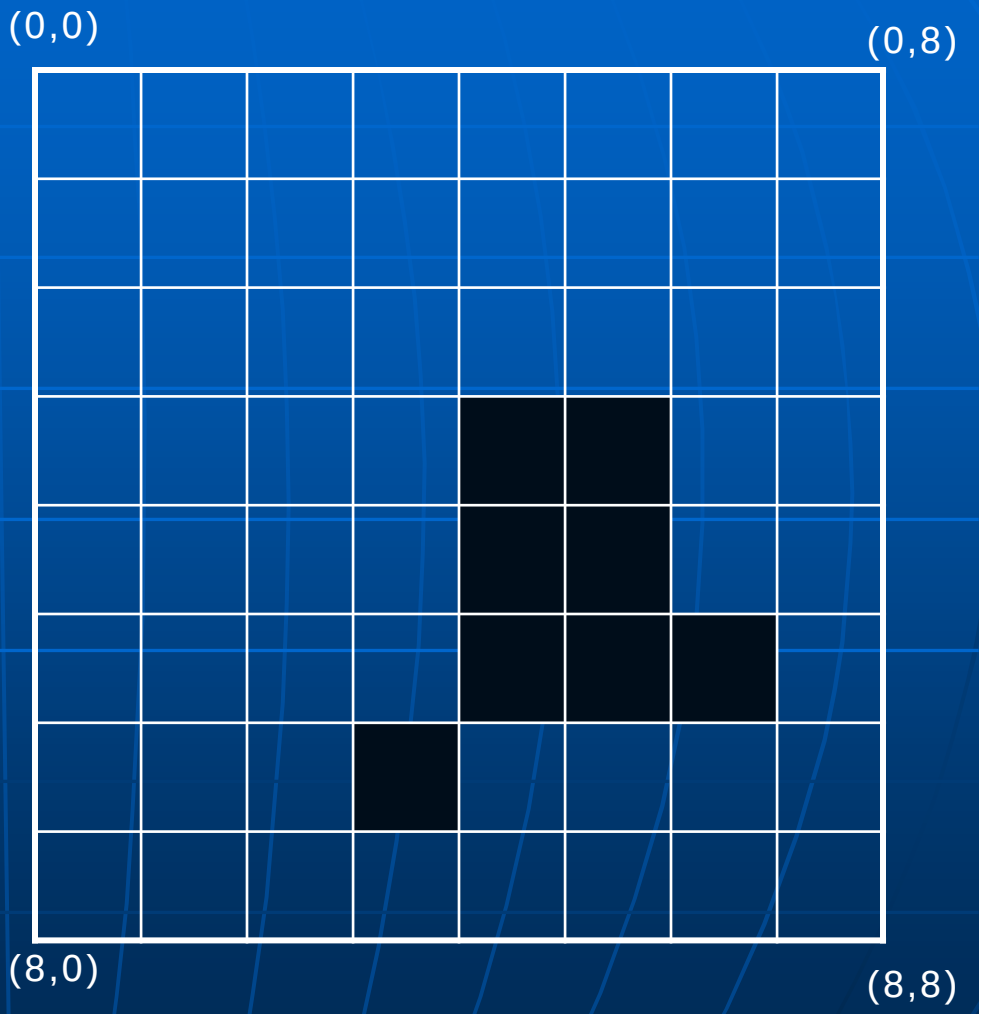
Data Compression Techniques

- Regional Quad Tree
  - Recursive decomposition to divide a grid into hierarchy of quadrants
  - A quadrant with cells having same value will not be subdivided.

Example:

Divide the grid into hierarchy of quadrant. The division stops if the quadrant has all the cells of same value





Data Compression Techniques

- Grass and IDRISI use Run Length Encoding (RLE)
- SPANS uses a Quad Tree Structure
- ARC/Info grids uses a Hierarchical Tile Block Cell Data Structure
  - Each grid is divided into a number of tiles
  - Tiles are divided into series of rectangular blocks
  - Blocks are divided into cells if required.

Other Techniques

- Tiff, GIF can compress data without loosing info
- JPEG however can compress more but loose some info
- MrSid (Multi-resolution Seamless Image) is a techniques which stores image at different resolution within image. Encoding of high detail parts of the image are done with high resolution and other at low resolution.
  - Good compression w/o loosing info

Vector Data Structure

- A Vector is defined by its starting point and its displacement and associated direction
- All objects in Vector models are represented by POINT, LINE or POLYGONS.
- Most of CAD use Vector data structure

Common Vector Data Structures

- Whole Polygon Structure
- Dual Independent Map Encoding
- Arc-Node Structure
- Relational Structure
- Digital Line Graph

Whole Polygon Structure

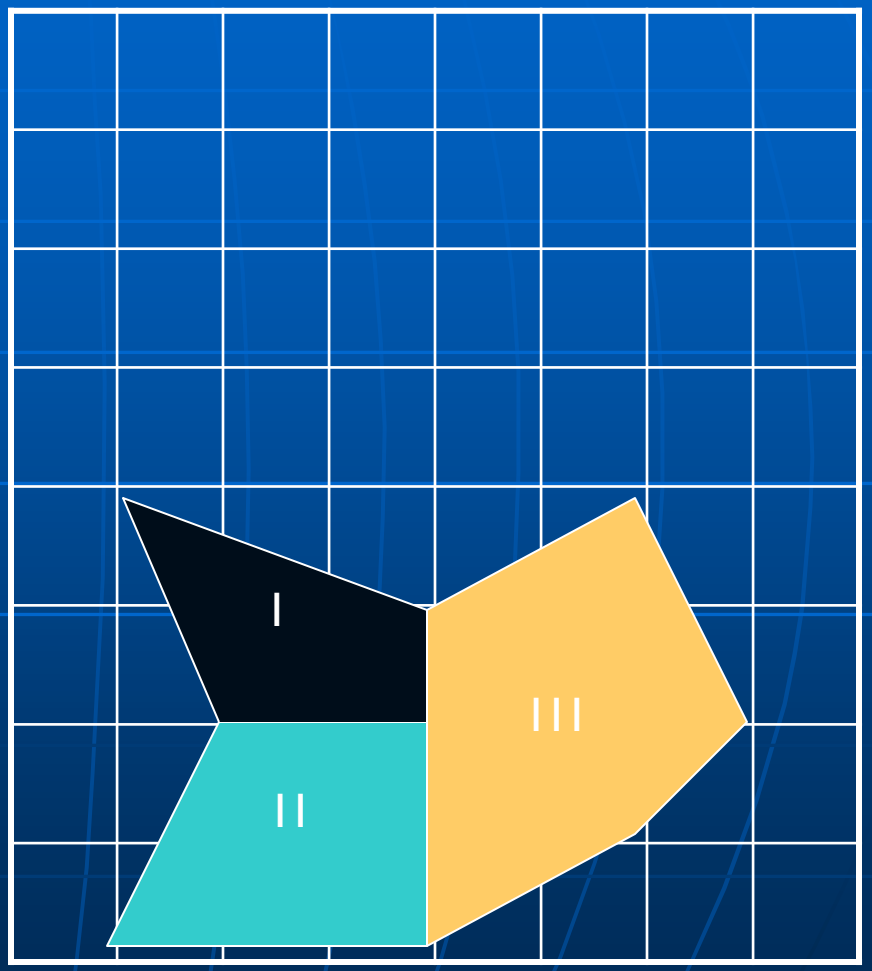
- Each layer in the database is dissolved in number of polygons
- Each polygon is encoded as sequence of locations (points) that define the boundaries of each closed area.
- Each polygon is stored as a separate entity
- No explicit mean to define neighborhood
- Attributes of each polygon can be stored with the list of the coordinates
- TOPOLOGY: The relationship between different spatial objects. (Which polygon share which boundary, which point exist on edge of a polygon, etc.)
- Several points that are shared by the polygons are repeated, as such enhancing the redundancy in data storage

Whole Polygon Structure

Polygon I
1,4
4,3
4,2
2,2

Polygon II
2,2
4,2
4,0
1,0

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



DIME Structure

- US Bureau of Census Developed it
- Dual Independent Map Encoding
- Designed to incorporate topological information about urban areas
- Used to archive data, as it is used for data exchange in various systems
- Basic element is LINE, defined by two points and information about **To** and **From** Node
- Lines are assumed to be straight
- Curved lines are drawn by combination of straight lines
- Additional attributes can be coded in DIME
- Disadvantage: Difficulty in manipulating complex lines
- Advantage: it has capability to match addresses of objects in multiple files, as addresses are explicitly stored in DIME

DIME

Segmental Code									
Segment Name	Nodes		Polygon		Addresses				
	From	To	Left	Right	Left		Right		Header
					Low	High	Low	High	
Brick St	1	2	-	Smith Est	101	175	102	178	1000
Cherry St	3	4	-	Smith Est	103	177	104	180	1000
Rutger St	4	1	-	Smith Est	8602	8686	8603	8685	1000

Table for node Location:

Node	Easting	Northing
1	127.251	1340.6
2	...	
3	..	
4		

ARC Node Structure

- OBJECTS in the database are structured hierarchically
- Points are the basic elemental components

ARC Node Structure

Table for node Location:

Node	Easting	Northing
1	127.251	1340.6
2	…	
3	..	
4		

ARCS				
Number	Nodes		Attributes	
	From	To	Length	Type of Pavement
I	4	1		
II	1	2		
III	2	3		
IV	3	4		

ARC Node Structure

Polygon		Attributes	
Name	Arcs	Owner	Type
A34	I, II, III, IV	Jhon Smith	Commercial
A35	....		

Relational data structure

- It is another form of arc node vector structure
- In it attributes information is kept separately
  - In ARC-Node structures attributes are stored with the topological data
- Used by many software
- Basic units are point, arcs and polygons
- No redundancy
  - Points are stored only once

Digital Line Graph

- Developed by USGS
- Use codes to distinguish features
- E.g.

Major Codes
020 for Hypsography,050 for Hydrograph, 070 Surface cover
Minor:
001-099 for nodes
100-199 for areas
200-299 for lines
300-399 for degenerated lines
400-499 for general purpose codes
.....

- 050-0001 may show Upper end of a stream
- 050-0101 may show a reservoir
- 050-0200 may show a shore line

Thanks