What is a GIS?

A Geographic Information System is a multi-component environment used to create, manage, visualize and analyze data and its spatial counterpart.

A geographic information system (GIS) is a system designed to capture, store, manipulate, analyze, manage, and present all types of geographical data.

It's important to note that most datasets you will encounter in your lifetime can all be assigned a spatial location whether on the earth's surface or within some arbitrary coordinate system. So in essence, any dataset can be represented in a GIS: the question then becomes "does it need to be analyzed in a GIS environment?" The answer to this question depends on the purpose of the analysis. If, for example, we are interested in identifying the ten African countries with the highest conflict index scores for the 1966-78 period, a simple table listing those scores by country is all that is needed.

DATA is defined as a body of facts or figures which have been gathered systematically for one or more specific purposes. Data is a plural and in a broad sense it can include things such as pictures (binary images), programmes and rules. Informally, data are the things we want to store in a database.

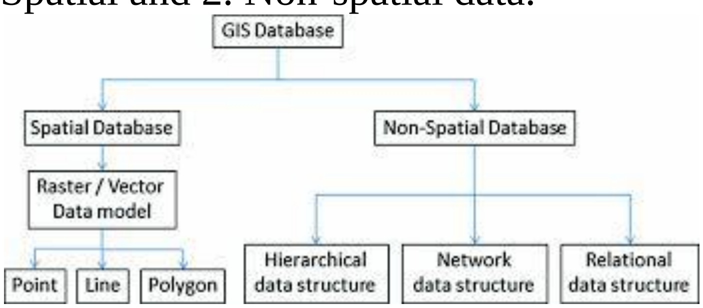
It can exist in the following forms: z linguistic expressions (e.g., name, age, address, date, ownership) z symbolic expressions (e.g., traffic signs) z mathematical expressions (e.g., E = mc2) and z signals (e.g., electromagnetic waves)

A DATABASE is defined as an automated, formally defined and centrally controlled collection of persistent data used and shared by different users in an enterprise. The term 'centrally controlled' means that databases tend to be physically distributed in different computer systems in the same time at different locations. A database is set-up to serve the information needs of an organisation. The sharing of data is the key to the concept of a database. Data in a database are described as 'permanent' in the sense that they are different from 'transient' data such as input to and output from an information system. The data usually remain in the database for a considerable length of time, although the actual content of the data can change very frequently. The use of database does not mean the demise of data files; data in a database are still organised and stored as data files. The use of database represents a change in the perception of data, mode of data processing and purposes of using data rather than physical storage of the data.
A database is a collection of related data which represents some aspect of the real world. A database system is designed to be built and populated with data for a certain task.

Types of Database

Database is primarily of two types:

1. Spatial and 2. Non-spatial data.



Spatial Data:

1. Spatial data is geographical representation of features. In other words, spatial data is what we actually see in the form of maps (containing real-world features) on a computer screen.

2. Spatial data, also known as geospatial data, is information about a physical object that can be represented by numerical values in a geographic coordinate system.

It includes location, shape, size and orientation information of features or objects. For example, a particular square in which its center (the intersection of its diagonals) specifies its location; its shape is a square; length of one of its sides specifies its size and angle its diagonals e.g., the x-axis specifies its orientation. Spatial data includes spatial relationships, for example, the arrangement of three stumps in a cricket ground.

Spatial data can further be divided into two types- **vector** and **raster data**.

VectorData

Vector data represents any geographical feature through point, line or polygon or combination of these.

1.Point

A point in GIS is represented by one pair of coordinates (x & y). It is considered as dimension-less object. Most of the times a point represent location of a feature (like cities, wells, villages etc.).

2.Line

A line or arc contains at least two pairs of coordinates (say- x1, y1 & x2, y2). In other words a line should connect minimum two points. Start and end points of a line are referred as nodes while points on curves are referred as vertices. Points at intersections are also called as nodes. Roads, railway tracks, streams etc. are generally represented by line.

3.Polygon

In simple terms, polygon is a closed line with area. It takes minimum three pairs of coordinates to represent an area or polygon. Extent of cities, forests, land use etc. is represented by polygon.

RasterData

Raster data is made up of pixels. It is an array of grid cells with columns and rows. Each and every geographical feature is represented only through pixels in raster data. There is nothing like point, line or polygon. If it is a point, in raster data it will be a single pixel, a line will be represented as linear arrangement of pixels and an area or polygon will be represented by contiguous neighbouring pixels with similar values.

In raster data one pixel contain only one value (unlike vector data where a point, a line or a polygon may have number of values or attributes) that's why only one geographical feature can be represented by a single set of pixels or grid cells. Hence a number of raster layers are required if multiple features are to be considered (For example- land use, soil type, forest density, topography etc.).

As discussed earlier digital satellite images are also in raster format.

Non-spatial Data:

Attributes attached to spatial data are referred to as non-spatial data. Whatever spatial data we see in the form of a colourful map on a computer screen is a presentation of information which remains stored in the form attribute tables. Attributes of spatial data must contain unique identifier for each object. There may be other field also containing properties/information related a spatial feature. Attribute table of spatial data also contains 'x' and 'y' location (i.e. latitude/longitude or easting/northing) of features; however in some GIS software these columns may remain 'invisible'

It is also known attribute or characteristic data. It consists of the characteristics of spatial features which are independent of all geometric considerations.

Let us illustrate this with the help of an example. The non-spatial data of town comprise of name of the town, its population, settlement type, means of transportation and communication, administration set-up, education institutions, occupations and facilities. It is important to note that all the above mentioned data of town are not dependent on their location identities.

.

For example- if we are doing demographic analysis of villages then attributes of each point (representing a village) must have a unique village ID and other demographic information like total population, number of males & females, number of children etc.

In another example- if we are doing some GIS analysis related to road then each road must have its unique Road ID. Other attributes may include like road length, road width, current traffic volume, number of stations etc.

Database Management System (DBMS) is a software for storing and retrieving users' data while considering appropriate security measures. It consists of a group of programs which manipulate the database. The DBMS accepts the request for data from an application and instructs the operating system to provide the specific data. In large systems, a DBMS helps users and other third-party software to store and retrieve data.

DBMS allows users to create their own databases as per their requirement. The term “DBMS” includes the user of the database and other application programs. It provides an interface between the data and the software application.

History of DBMS

Here, are the important landmarks from the history:

- 1960 - Charles Bachman designed first DBMS system
- 1970 - Codd introduced IBM'S Information Management System (IMS)
- 1976- Peter Chen coined and defined the Entity-relationship model also know as the ER model
- 1980 - Relational Model becomes a widely accepted database component
- 1985- Object-oriented DBMS develops.
- 1990s- Incorporation of object-orientation in relational DBMS.
- 1991- Microsoft ships MS access, a personal DBMS and that displaces all other personal DBMS products.
- 1995: First Internet database applications
- 1997: XML applied to database processing. Many vendors begin to integrate XML into DBMS products.

Characteristics of Database Management System

- Provides security and removes redundancy
- Self-describing nature of a database system
- Insulation between programs and data abstraction
- Support of multiple views of the data
- Sharing of data and multiuser transaction processing
- DBMS allows entities and relations among them to form tables.

- It follows the ACID concept (Atomicity, Consistency, Isolation, and Durability).
- DBMS supports multi-user environment that allows users to access and manipulate data in parallel.

Popular DBMS Software

Here, is the list of some popular DBMS system:

- MySQL
- Microsoft Access
- Oracle
- PostgreSQL
- dBASE
- FoxPro
- SQLite
- IBM DB2
- LibreOffice Base
- MariaDB
- Microsoft SQL Server etc.

Application of DBMS

Sector	Use of DBMS
Banking	For customer information, account activities, payments, deposits, loans, etc.
Airlines	For reservations and schedule information.
Universities	For student information, course registrations, colleges and grades.
Telecommunication	It helps to keep call records, monthly bills,

	maintaining balances, etc.
Finance	For storing information about stock, sales, and purchases of financial instruments like stocks and bonds.
Sales	Use for storing customer, product & sales information.
Manufacturing	It is used for the management of supply chain and for tracking production of items. Inventories status in warehouses.
HR Management	For information about employees, salaries, payroll, deduction, generation of paychecks, etc.

Advantages of DBMS

- DBMS offers a variety of techniques to store & retrieve data
- DBMS serves as an efficient handler to balance the needs of multiple applications using the same data
- Uniform administration procedures for data
- Application programmers never exposed to details of data representation and storage.
- A DBMS uses various powerful functions to store and retrieve data efficiently.
- Offers Data Integrity and Security
- The DBMS implies integrity constraints to get a high level of protection against prohibited access to data.
- A DBMS schedules concurrent access to the data in such a manner that only one user can access the same data at a time
- Reduced Application Development Time

Disadvantage of DBMS

DBMS may offer plenty of advantages but, it has certain flaws-

- Cost of Hardware and Software of a DBMS is quite high which increases the budget of your organization.
- Most database management systems are often complex systems, so the training for users to use the DBMS is required.
- In some organizations, all data is integrated into a single database which can be damaged because of electric failure or database is corrupted on the storage media
- Use of the same program at a time by many users sometimes lead to the loss of some data.
- DBMS can't perform sophisticated calculations

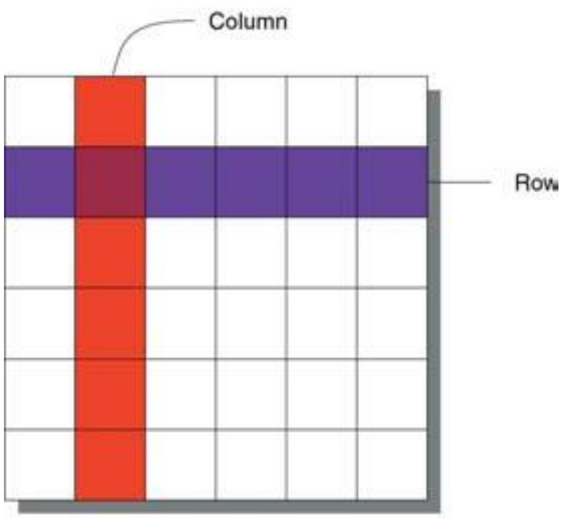
RASTER REPRESENTATION OF DATA

A **raster data** structure is based on a (usually rectangular, square-based) tessellation of the 2D plane into cells. In the example the cells of tessellation A are overlaid on the point pattern B resulting in an array C of quadrant counts representing the number of points in each cell. For purposes of visualization a lookup table has been used to color each of the cells in an image D. Here are the numbers as a simple vector in row/column order:

1 3 0 0 1 12 8 0 1 4 3 3 0 2 0 2 1 7 4 1 5 4 2 2 0 3 1 2 2 2 2 3 0 5 1 9 3 3 3 4 5 0 8 0 2 4 3 2 8 4 3 2 2 7 2 3 2 10 1 5 2 1 3 7

Raster is a method for the storage, processing and display of spatial data. Each area is divided into rows and columns, which form a regular grid structure. Each cell must be rectangular in shape, but not necessarily square. Each cell within this matrix contains location co-ordinates as well as an attribute value. The spatial location of each cell is implicitly contained within the ordering of the matrix, unlike a vector structure which stores topology explicitly. Areas containing the same attribute value are recognised as such, however, raster structures cannot identify the boundaries of such areas as polygons.

Raster data is an abstraction of the real world where spatial data is expressed as a matrix of cells or pixels (see figure 9), with spatial position implicit in the ordering of the pixels. With the raster data model, spatial data is not continuous but divided into discrete units. This makes raster data particularly suitable for certain types of spatial operation, for example overlays or area calculations



ELEMENTS OF RASTER DATA MODEL

1. Cell value. Each cell in a raster carries a value, which represents the characteristic of a spatial phenomenon at the location denoted by its row and column. The cell value can be integer or floating-point.

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

3. Raster bands. A raster may have a single band or multiple bands. 4. Spatial reference. Raster data must have the spatial reference information so that they can align spatially with other data sets in a GIS.

TYPES OF RASTER DATA

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

Raster Data Structure 1. Cell-by-Cell Encoding 2. Run Length Encoding 3. Quad Tree

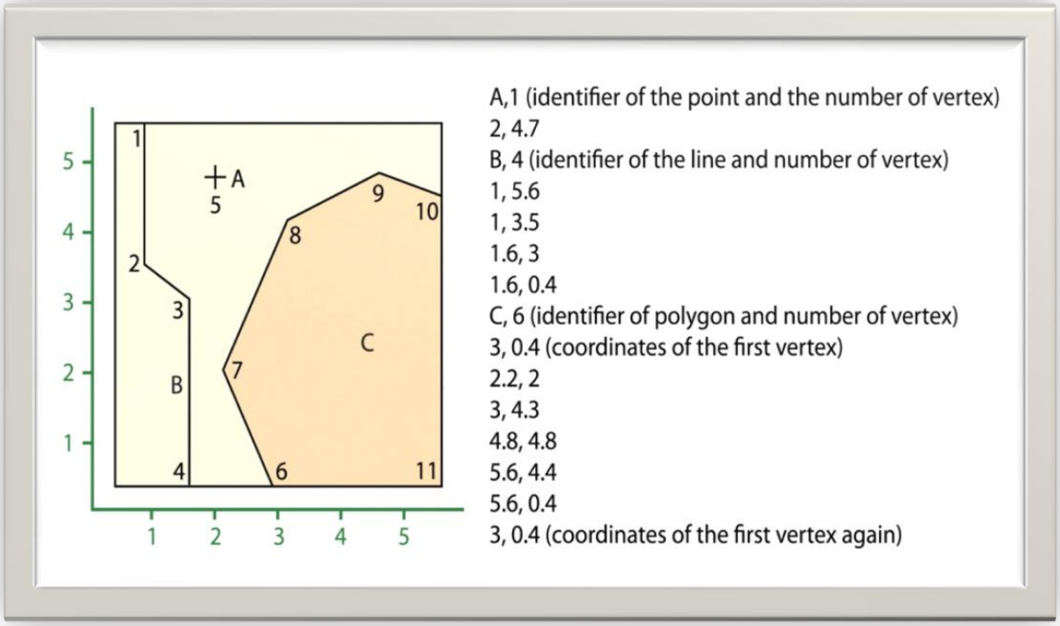
RASTERISATION OF VECTOR DATA

The process of converting vector data, which is a series of points, lines and polygons, into raster data, which is a series of cells each with a discrete value. This process is essentially easier than the reverse process, which is converting data from raster format to vector format.

Vector Data Models Structures

Vector data models can be structured many different ways. We will examine two of the more common data structures here. The simplest vector data structure is called the spaghetti data model (Dangermond 1982).Dangermond, J. 1982. "A Classification of Software Components Commonly Used in Geographic Information Systems." In *Proceedings of the U.S.-Australia Workshop on the Design and Implementation of Computer-Based Geographic Information Systems*, 70–91. Honolulu, HI. In the spaghetti model, each point, line, and/or polygon feature is represented as a string of X, Y coordinate pairs (or as a single X, Y coordinate pair in the case of a vector image with a single point) with no inherent structure (Figure 4.9 "Spaghetti Data Model"). One could envision each line in this model to be a single strand of spaghetti that is formed into complex shapes by the addition of more and more strands of spaghetti. It is notable that in this model, any polygons that lie adjacent to each other must be made up of their own lines, or stands of spaghetti. In other words, each polygon must be uniquely defined by its own set of X, Y coordinate pairs, even if the adjacent polygons share the exact same boundary information. This creates some redundancies within the data model and therefore reduces efficiency.

Figure 4.9 Spaghetti Data Model

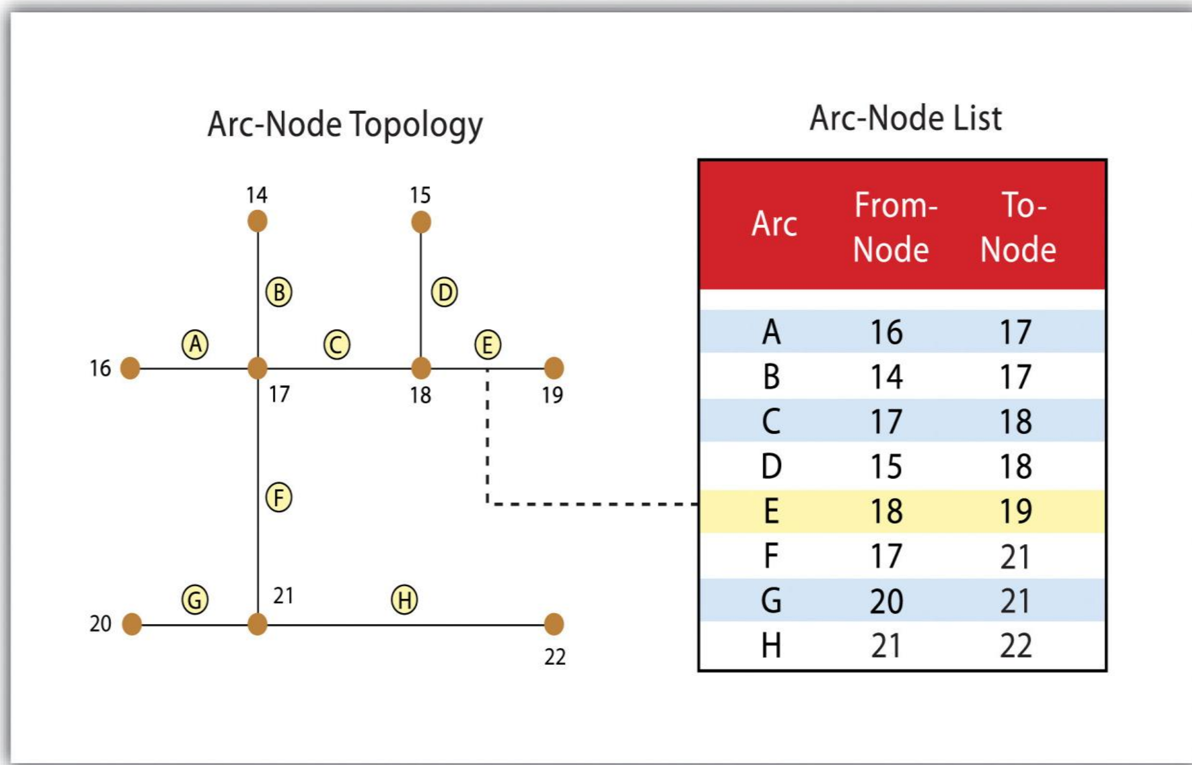


Despite the location designations associated with each line, or strand of spaghetti, spatial relationships are not explicitly encoded within the spaghetti model; rather, they are implied by their location. This results in a lack of topological information, which is problematic if the user attempts to make measurements or analysis. The computational requirements, therefore, are very steep if any advanced analytical techniques are employed on vector files structured thusly. Nevertheless, the simple structure of the spaghetti data model allows for efficient reproduction of maps and graphics as this topological information is unnecessary for plotting and printing.

In contrast to the spaghetti data model, the topological data model is characterized by the inclusion of topological information within the dataset, as the name implies. Topology is a set of rules that model the relationships between neighboring points, lines, and polygons and determines how they share geometry. For example, consider two adjacent polygons. In the spaghetti model, the shared boundary of two neighboring polygons is defined as two separate, identical lines. The inclusion of topology into the data model allows for a single line to represent this shared boundary with an explicit reference to denote which side of the line belongs with which polygon. Topology is also concerned with preserving spatial properties when the forms are bent, stretched, or placed under similar geometric transformations, which allows for more efficient projection and reprojection of map files.

Three basic topological precepts that are necessary to understand the topological data model are outlined here. First, connectivity describes the arc-node topology for the feature dataset. As discussed previously, nodes are more than simple points. In the topological data model, nodes are the intersection points where two or more arcs meet. In the case of arc-node topology, arcs have both a from-node (i.e., starting node) indicating where the arc begins and a to-node (i.e., ending node) indicating where the arc ends (Figure 4.10 "Arc-Node Topology"). In addition, between each node pair is a line segment, sometimes called a link, which has its own identification number and references both its from-node and to-node. In Figure 4.10 "Arc-Node Topology", arcs 1, 2, and 3 all intersect because they share node 11. Therefore, the computer can determine that it is possible to move along arc 1 and turn onto arc 3, while it is not possible to move from arc 1 to arc 5, as they do not share a common node.

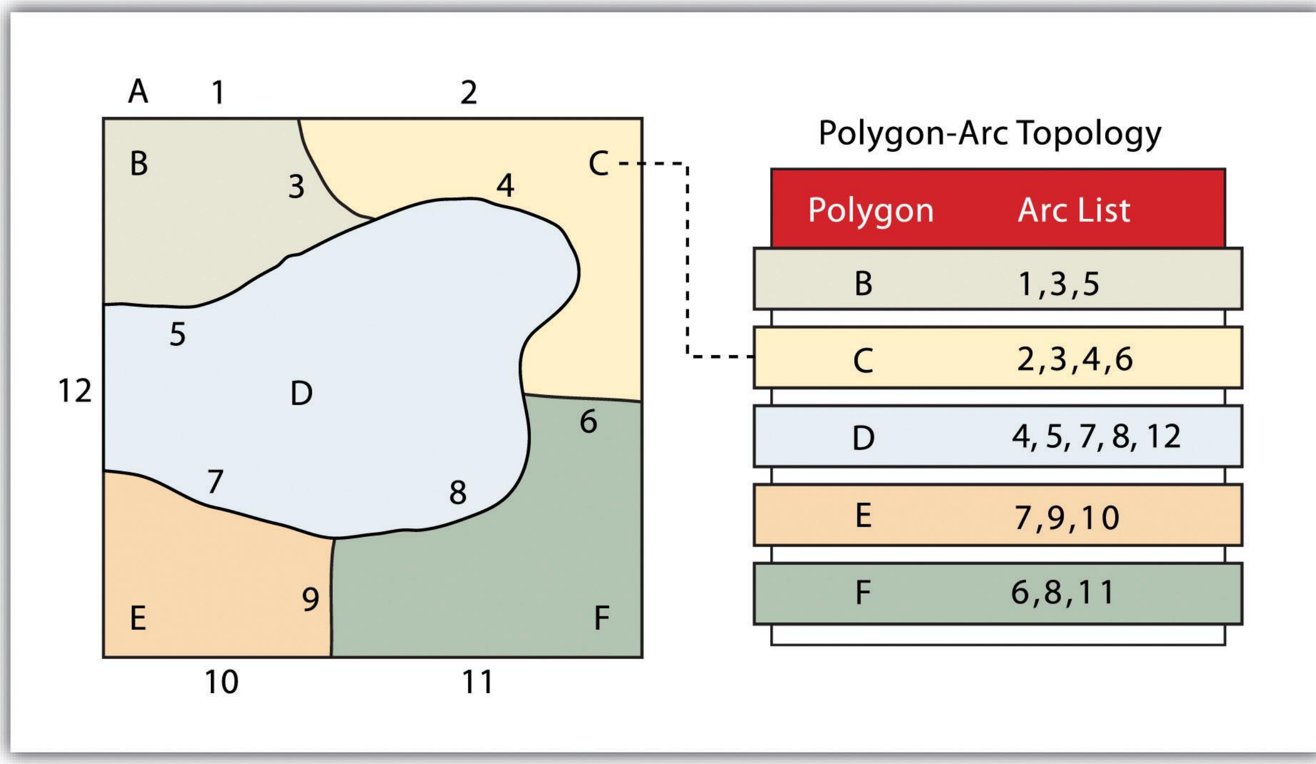
Figure 4.10 Arc-Node Topology



Arc	From-Node	To-Node
A	16	17
B	14	17
C	17	18
D	15	18
E	18	19
F	17	21
G	20	21
H	21	22

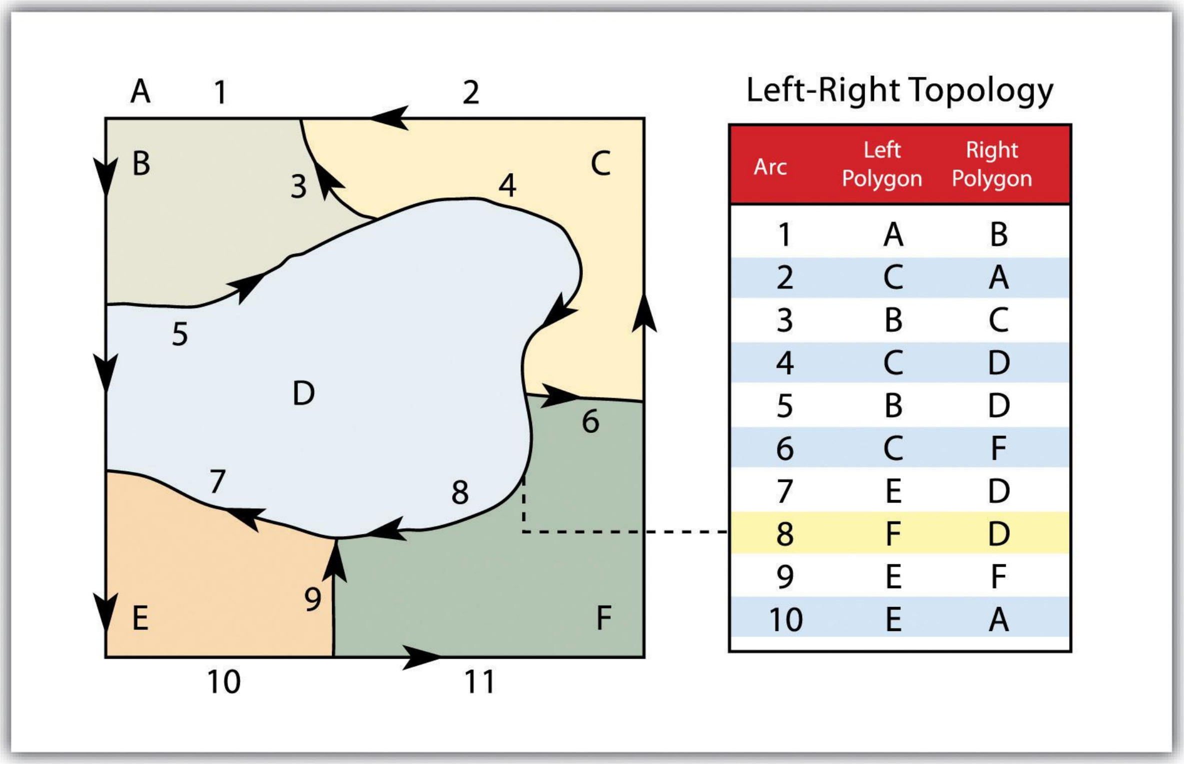
The second basic topological precept is area definition. Area definition states that an arc that connects to surround an area defines a polygon, also called polygon-arc topology. In the case of polygon-arc topology, arcs are used to construct polygons, and each arc is stored only once (Figure 4.11 "Polygon-Arc Topology"). This results in a reduction in the amount of data stored and ensures that adjacent polygon boundaries do not overlap. In the Figure 4.11 "Polygon-Arc Topology", the polygon-arc topology makes it clear that polygon F is made up of arcs 8, 9, and 10.

Figure 4.11 Polygon-Arc Topology



Contiguity, the third topological precept, is based on the concept that polygons that share a boundary are deemed adjacent. Specifically, polygon topology requires that all arcs in a polygon have a direction (a from-node and a to-node), which allows adjacency information to be determined (Figure 4.12 "Polygon Topology"). Polygons that share an arc are deemed adjacent, or contiguous, and therefore the "left" and "right" side of each arc can be defined. This left and right polygon information is stored explicitly within the attribute information of the topological data model. The "universe polygon" is an essential component of polygon topology that represents the external area located outside of the study area. Figure 4.12 "Polygon Topology" shows that arc 6 is bound on the left by polygon B and to the right by polygon C. Polygon A, the universe polygon, is to the left of arcs 1, 2, and 3.

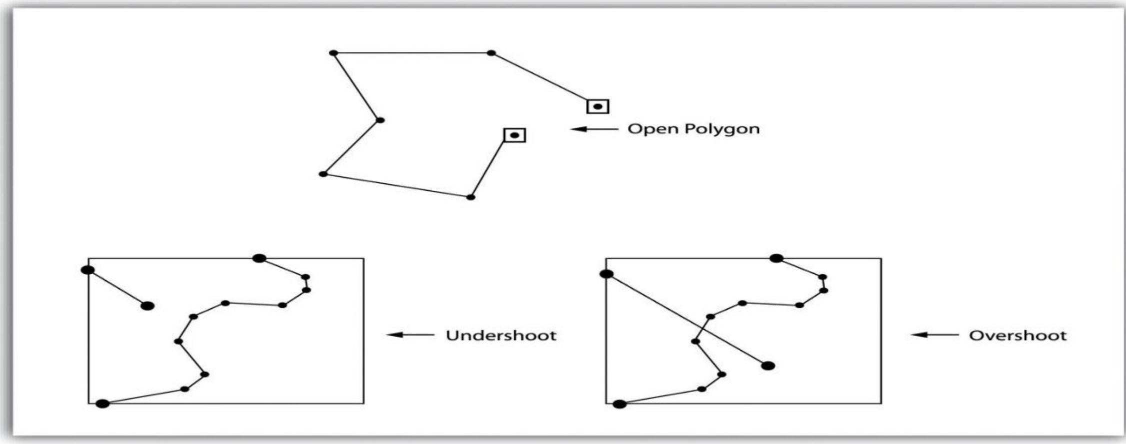
Figure 4.12 Polygon Topology



Topology allows the computer to rapidly determine and analyze the spatial relationships of all its included features. In addition, topological information is important because it allows for efficient error detection within a vector dataset. In the case of polygon features, open or unclosed polygons, which occur when an arc does not completely loop back upon itself, and unlabeled polygons, which occur when an area does not contain any attribute information, violate polygon-arc topology rules. Another topological error found with polygon features is the sliver. Slivers occur when the shared boundary of two polygons do not meet exactly (Figure 4.13 "Common Topological Errors").

In the case of line features, topological errors occur when two lines do not meet perfectly at a node. This error is called an "undershoot" when the lines do not extend far enough to meet each other and an "overshoot" when the line extends beyond the feature it should connect to (Figure 4.13 "Common Topological Errors"). The result of overshoots and undershoots is a "dangling node" at the end of the line. Dangling nodes aren't always an error, however, as they occur in the case of dead-end streets on a road map.

Figure 4.13 Common Topological Errors



Many types of spatial analysis require the degree of organization offered by topologically explicit data models. In particular, network analysis (e.g., finding the best route from one location to another) and measurement (e.g., finding the length of a river segment) relies heavily on the concept of to- and from-nodes and uses this information, along with attribute information, to calculate distances, shortest routes, quickest routes, and so forth. Topology also allows for sophisticated neighborhood analysis such as determining adjacency, clustering, nearest neighbors, and so forth.

Now that the basics of the concepts of topology have been outlined, we can begin to better understand the topological data model. In this model, the node acts as more than just a simple point along a line or polygon. The node represents the point of intersection for two or more arcs. Arcs may or may not be looped into polygons. Regardless, all nodes, arcs, and polygons are individually numbered. This numbering allows for quick and easy reference within the data model.

ADVANTAGES/DISADVANTAGES OF THE VECTOR MODEL

In comparison with the raster data model, vector data models tend to be better representations of reality due to the accuracy and precision of points, lines, and polygons over the regularly spaced grid cells of the raster model. This results in vector data tending to be more aesthetically pleasing than raster data.

Vector data also provides an increased ability to alter the scale of observation and analysis. As each coordinate pair associated with a point, line, and polygon represents an infinitesimally exact location (albeit limited by the number of significant digits and/or data acquisition methodologies), zooming deep into a vector image does not change the view of a vector graphic in the way that it does a raster graphic (see Figure 4.1 "Digital Picture with Zoomed Inset Showing Pixilation of Raster Image").

Vector data tend to be more compact in data structure, so file sizes are typically much smaller than their raster counterparts. Although the ability of modern computers has minimized the importance of maintaining small file sizes, vector data often require a fraction the computer storage space when compared to raster data.

The final advantage of vector data is that topology is inherent in the vector model. This topological information results in simplified spatial analysis (e.g., error detection, network analysis, proximity analysis, and spatial transformation) when using a vector model.

Alternatively, there are two primary disadvantages of the vector data model. First, the data structure tends to be much more complex than the simple raster data model. As the location of each vertex must be stored explicitly in the model, there are no shortcuts for storing data like there are for raster models (e.g., the run-length and quad-tree encoding methodologies).

Second, the implementation of spatial analysis can also be relatively complicated due to minor differences in accuracy and precision between the input datasets. Similarly, the algorithms for manipulating and analyzing vector data are complex and can lead to intensive processing requirements, particularly when dealing with large datasets.

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:

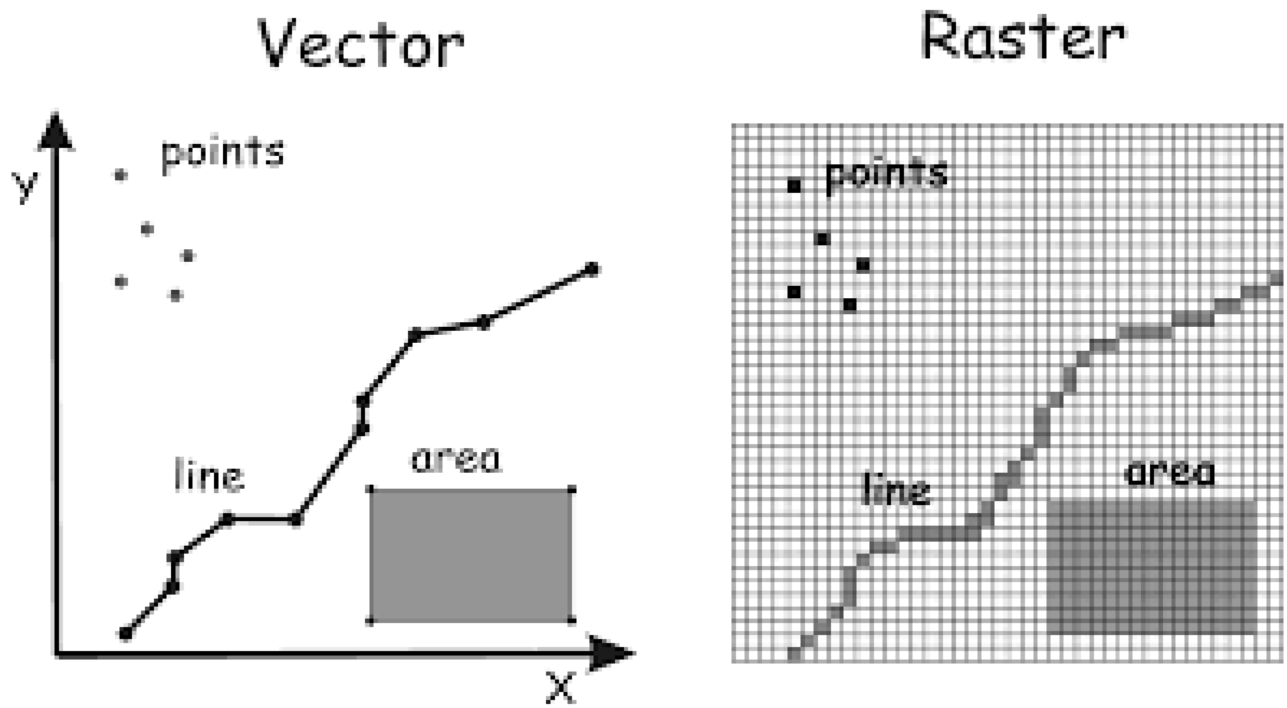
- The cell size determines the resolution at which the data is represented.;

- It is especially difficult to adequately represent linear features depending on the cell resolution. Accordingly, network linkages are difficult to establish.

- Processing of associated attribute data may be cumbersome if large amounts of data exists. Raster maps inherently reflect only one attribute or characteristic for an area.

- Since most input data is in vector form, data must undergo vector-to-raster conversion. Besides increased processing requirements this may introduce data integrity concerns due to generalization and choice of inappropriate cell size.

- Most output maps from grid-cell systems do not conform to high-quality cartographic needs.



ADVANTAGES/DISADVANTAGES OF RASTER AND VECTOR DATA MODELS

	raster	vector
precision in graphics	✗	✓
traditional cartography	✗	✓
data volume	✗	✓
topology	✗	✓
computation	✓	✗
update	✓	✗
continuous space	✓	✗
integration	✓	✗
discontinuous	✗	✓