Chapter Two

Data Models

Unit Objectives

After completing this unit, you will be able to:

- Understand geographic phenomena and types of geographic phenomena.
- Understand how geographic phenomena are represented in a computer.
- Differentiate types of GIS data models
- Define various GIS data models.
- Describe common spatial data formats used in GIS.
- Describe the vector and raster data models used in GIS and give examples.
- Discus structural capabilities of vector and raster data
- List the advantages and disadvantages of vector and raster data models.
- Explain what topology is and how is it stored in the computer.
- Understand topology and rules of topological consistency, and topology errors.

2.1 Spatial and geospatial data

The terms spatial and geospatial are often used almost interchangeably with geographical phenomena. Geospatial data is factual information related to location on the surface of the Earth. However, spatial data is used to refer more generally to any two-or three-dimensional data whether or not it relates directly to the surface of the Earth.

2.2. Geographic phenomena

A geographic phenomenon is, as a manifestation of an entity or process of interest that can be named or described, georeferenced, and can be assigned a time (interval) at which it is/was present. It is described as triplets: what is it? where is it? and when was it?

2.3. Types of geographic phenomena

Geographic phenomena are divided in to two: geographic fields and geographic objects.

2.3.1. Geographic fields

It is a geographic phenomenon for which, for every point in the study area, a value can be determined. E.g. temperature, rainfall, barometric pressure and elevation. Most natural made features are geographic fields and they have fuzzy boundary. Fields can be discrete or continuous. These fields differ in their type of cell values.

2.3.1.1. Continuous field

In a continuous field, the field values along any path through the study area do not change abruptly, but only gradually. Good examples of continuous fields are air temperature, and elevation. Continuous field can even be differentiable, meaning that we can determine a measure of changes in the field value per unit of distance. For example, if the field is elevation, this measure would be slope, i.e. the change of elevation per meter distance. Various geographic phenomena have the characteristics of continuous functions in space. For example, an elevation field can be measured at many locations, and each location may be given different value. Interpolation allows us to infer a reasonable elevation value for locations by a principle of **spatial autocorrelation** based on **'Tobler's first law of geography',** which states that "locations that are closer together are more likely to have similar values than locations that are further apart." A continuous raster is called a 'floating point' raster.

2.3.1.2. Discrete fields

Discrete fields divide the study space in mutually exclusive, bounded parts, with all locations in one part having the same field value. Typical examples are land classifications, using either geological classes, soil type, land-use type, crop type or natural vegetation type. Discrete fields can easily be converted into polygons, since it is easy to draw boundary lines around a group of cells of the same value. Discrete fields, store cell values of type 'integer', called an integer raster.

2.3.1.3. Spatial data types and values

Different kinds of data values may represent spatial 'phenomena'. Some of these data types limit the types of analyses that can be done on the data itself. There are four data values:

1) **Nominal (categorical) data** values provide a name or identifier to discriminate between different values. For example, soil type of a given area as belonging to a certain category. We cannot do true computations with these values.

2) **Ordinal data** values are numerical data that can be put in a natural sequence; but do not allow any other type of computation (household income, classified as 'low', 'average' or 'high').

3) **Interval data** values are quantitative, allow simple forms of computation like addition and subtraction. However, interval data has no arithmetic zero value (a temperature).

4) **Ratio data** values are numerical data allow most, if not all forms of arithmetic computation; and have a natural zero value (e.g. distances). Continuous fields can be expected to have ratio data values; hence we can interpolate them.

Nominal or categorical data values are 'qualitative' data because they are limited to make computations. Interval and ratio data are known as 'quantitative' data as they refer to quantities. However, ordinal data do not fit either of these data types. Often, ordinal data refer to a ranking scheme or a kind of hierarchical phenomena. Road networks, are made up of motorways, main roads and residential streets.

2.3.2. Geographic objects

Geographic objects are physical objects that can be georeferenced and represented graphically. Usually, geographic objects are not studied in isolation; instead they are looked as a collection of objects and considered as a unit (aggregated level). Therefore, they can be easily distinguished, named, and their position in space is determined by a combination of one or more of the following parameters:

- Location (where is it?),

- Shape (what form does it have?),
- Size (how big is it?) and
- Orientation (in which direction is it facing?).

Most manmade phenomena are geographic objects; they have crisp/sharp boundary.

2.3.3. Boundaries

There are two boundary concepts: crisp and fuzzy. A crisp boundary can be determined at an almost arbitrary level of precision, dependent only on the data-acquisition technique applied. Fuzzy boundaries do not have a precise line or intact, but they are areas of transition. As a rule-of-thumb, crisp boundaries are more common in manmade phenomena; whereas fuzzy boundaries are more common in natural phenomena. In many cases, boundaries that are fuzzy in reality represented as crisp boundaries in GIS. Both geographic objects and discrete fields have boundaries.

Figure 2.1: Two types of boundaries: *crisp* boundaries (left) and *fuzzy* boundaries (right)

2.4 Computer representations of geographic data

In GIS, the most common types of geographic data representations are:

- Raster data for representing images, gridded thematic data, and surfaces
- Vector data for representing geographic objects or features
- Triangulated irregular networks (TINs) for representing surfaces

- GIS geodatabase stores these geographic data representations in a structured and organized manner, called *data structure.* Data structure allows to associate spatial data, and non-spatial (attribute) data.

2.4. Concepts of Raster and Vector Data /GIS Data Models

Data models are a set of rules and/or constructs used to describe and represent aspects of the real world in a computer. Data in a GIS represents in simplified view of physical entities or phenomena called **MODEL**. The purpose of spatial data model is to provide a formal means of representing and manipulating spatially referenced information. The two types of spatial data models are: Raster Data model and Vector data model.

2.4.1. Raster data model

R**aster data model** – identifies and represents grid cells for a given region of interest. Raster cells are arrayed in a row and column pattern. Raster model is used most commonly with variables that change continuously across a region. E.g. Elevation, mean-temperature, slope, average rainfall, and soil moisture. A raster data type is, in essence, any type of digital image. Anyone who is familiar with digital photography will recognize the pixel as the smallest individual unit of an image. A combination of these pixels will create an image, distinct from the commonly used scalable vector graphics, which are the basis of the vector model. While a digital image is concerned with the output as representation of reality, in a photograph or art transferred to computer, the raster data type will reflect an abstraction of reality. Aerial photos are one commonly used form of raster data, with only one purpose, to display a detailed image on a map or for the purposes of digitization. Other raster data sets will contain information regarding elevation, a DEM (digital Elevation Model), or reflectance of a particular wavelength of light. The resolution of the raster data set is its cell width in ground units. Raster data models define the world as a regular set of cells in a grid pattern with associated attributes. It is typically, square & evenly spaced in X & Y direction.

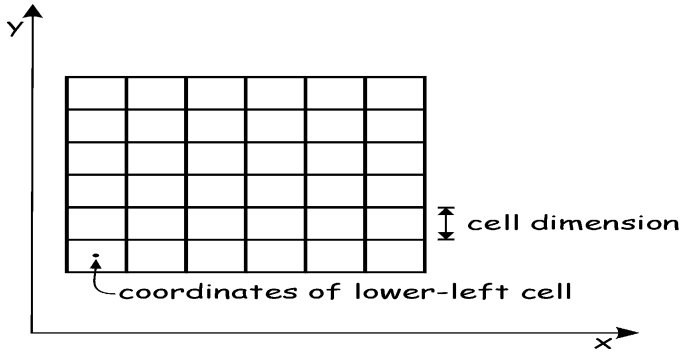


Figure 2.2: Raster representation

Raster data is stored in various formats; from a standard file-based structure of TIF, JPEG formats to binary large object (BLOB) data stored directly in a relational database management system (RDBMS) similar to other vector-based feature classes. Database storage, when properly indexed, typically allows for quicker retrieval of the raster data but can require storage of millions of significantly sized records. Geographic variation is expressed by assigning values or attributes to cells or pixels (picture elements). Raster cells represent a two-dimensional matrix organized in rows and columns as shown in (figure 2.3):

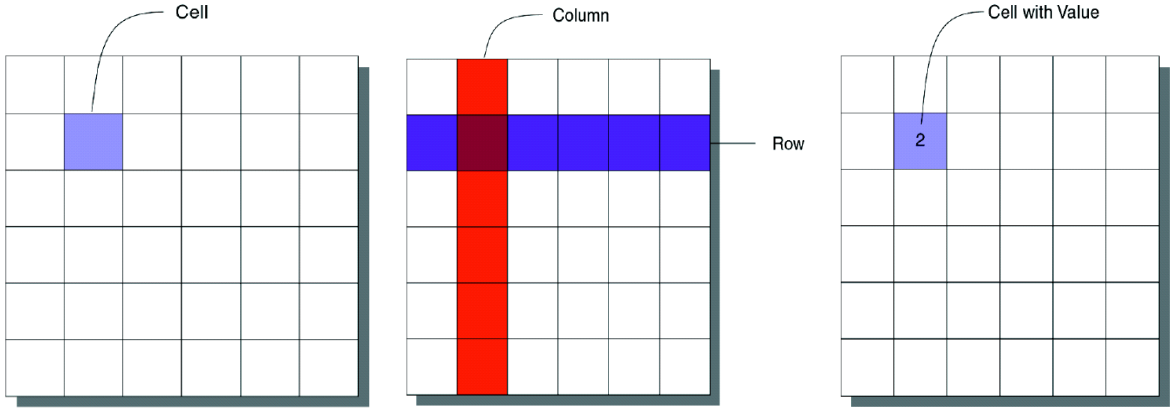


Figure 2.3: Cell, rows & columns and cell values

Raster cell values are stored with one of the following types:

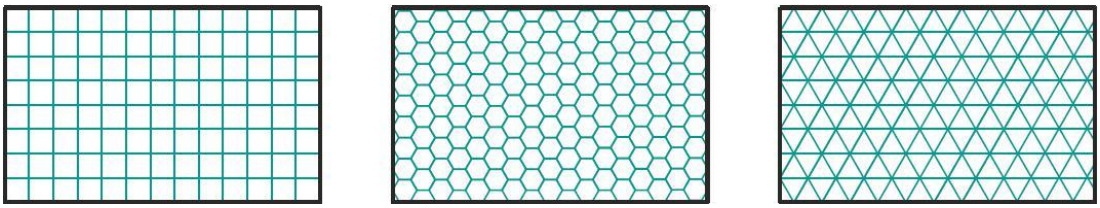
1) **Binary:** -a value, which indicates the presence or absence of a feature. E.g., in a layer of roads, we might use 1 for pixels containing part of the road, and 0 for pixels which did not.

2) **Enumeration**: - a value from classification. E.g., a soil layer might contain codes representing the different soil types–1 for alluvial, 2 for red soil, etc. Since the values are not directly related to the soil type, there should be a key to indicate the meaning of each value.

3) **Numerical**: - integer or floating point numbers recording value of a geographic phenomenon. E.g., values represent the height of the land surface - Digital Elevation Model (DEM).

2.4.1.1. Regular tessellation

Tessellation (or tiling) is partitioning of a space into *mutually exclusive* cells, each cell assigned field value. There are three types of regular tessellations with the **same shape** and **size.** A rectangular or square raster unit is the simplest, represented in 2D as an array of **n × m** raster cell elements. The field value of a cell is discrete, or differentiable. There are conventions to state the cell boundaries. With square cells, this convention says, "the lower and left boundaries belong to the cell". The advantage of regular tessellations is that it is simple structures with straightforward algorithms (i.e., we can make our computations specific to a particular partition). The disadvantage is that they may not be adaptive to the spatial phenomenon they represent since the cell boundaries are fixed.



a) Square cells b) Hexagonal cells c) Triangular cells

Figure 2.4: The three common types of regular tessellations: square cells, hexagonal cells and triangular cells

2.4.1.2. Irregular tessellation

Irregular tessellation is also partitioning of spaces into mutually exclusive cells. The cells may **vary in size** and **shape**, allowing them to be adaptive to the phenomena they represent. However, they are more complex than regular tessellations, the method called *quad-tree.* To construct a quad tree, the field is successively split into four quadrants until all parts have the same field value. The result is an upside-down tree-like structure, hence 'quad-tree'. Tessellations do not explicitly store georeferences of the phenomena that they represent; instead, they provide a georeference for all cells of a raster.

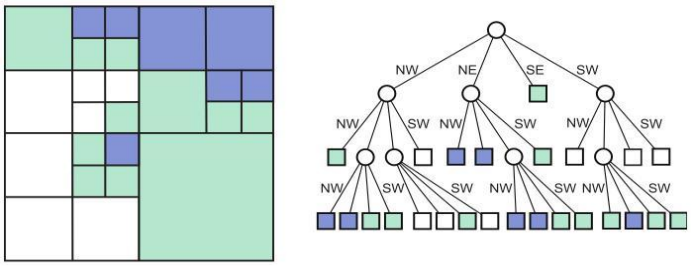


Figure 2.5: An 8×8, three-valued raster (with colors) and its representation as a region quad-tree

2.4.1.3. Tessellations for representing geographic fields

Geographic fields can be represented by means of a tessellation, a TIN or a vector. The choice depends on the application requirements in mind. Raster commonly used to represent geographic fields. **Figure 2.6** illustrates a raster representation of a continuous elevation field. The right side is a zoomed-in part of the left side figure. Different shades indicate different elevation values stored in the raster. The darker areas indicate higher elevations and the lighter ones indicate lower elevation.

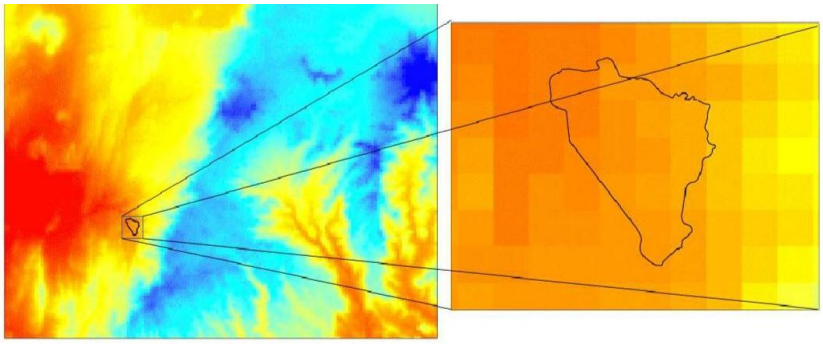


Figure 2.6: Raster representation of elevation fields.

2.4.1.4 Tessellations for representing geographic objects

For remote sensing images data sources, there are various techniques to process and classify images to store in a GIS as a raster. Image classification characterizes each pixel into one of a finite list of classes, thereby obtaining an interpretation of the contents of the image. The recognized classes can be crop types as in the case of Figure 2.7 or urban land-use classes as in the case of Figure 2.8. These Figures illustrate unprocessed images (a) and classified version of the image (b).

Figure 2.7: Unprocessed digital image (a) and classified raster (b) of an agricultural area

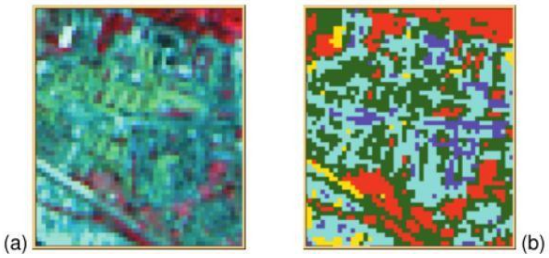


Figure 2.8: Unprocessed digital image (a) and classified raster (b) of an urban area

2.4.2. Vector Data model

Vector data model is a model that uses discrete elements such as points, lines and polygons to represent the geometry of real world entities. In a GIS, geographical features are often expressed as vectors, by considering those features as geometrical shapes. In the popular ESRI Arc series of programs, these are explicitly called shape files. A vector data model uses sets of coordinates

and associated attribute data to define discrete objects. There are three basic types of vector objects: These are point, line and polygon as shown in figure 2.9.

Points: Zero-dimensional points are used for geographical features that can best be expressed by a single grid reference; in other words, simple location. For example, the locations of wells, peak elevations, feature of interest or trailheads. Points convey the least amount of information of these file types.

Lines or polylines: One-dimensional lines or polylines are used for linear features such as rivers, roads, railroads, trails, and topographic lines. **Linear Features** – often referred to as **arcs**, are represented as **lines**. Lines typically have starting, ending and intermediate points to represent the shape of the linear entity. Starting points and end points sometimes referred to as **nodes**. Intermediate points in a line are referred to as **vertices**.

Polygons: Two-dimensional polygons are used for geographical features that cover a particular area of the earth's surface. Such features may include lakes, park boundaries, buildings, city boundaries, or land uses. Polygons convey the most amount of information of the file types.

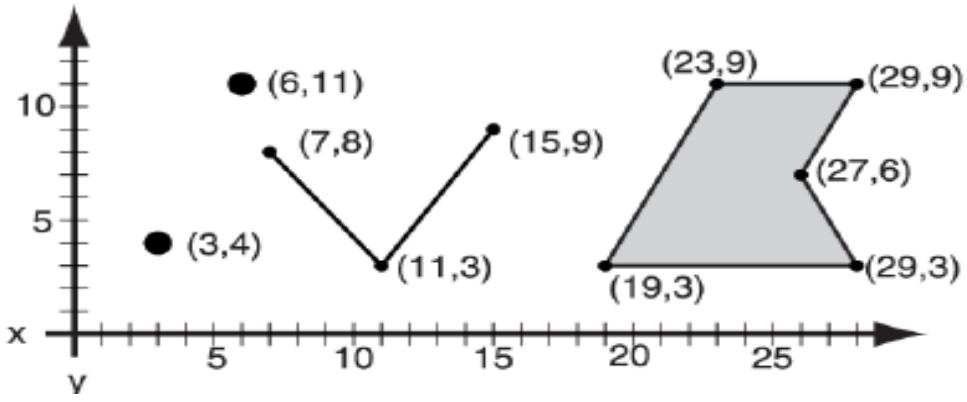


Figure 2.9: Point, Line and Area Representations

N.B. There is no uniformly superior way to represent features. The representation depends as much on the **detail, accuracy,** and **intended use of the data set** as our common **conception** and

general shape of the objects. E.g. Buildings may be represented by either point, line, or polygon features.

2.4.2.1 Types of Vector Data model

Two main types of vector data models are:

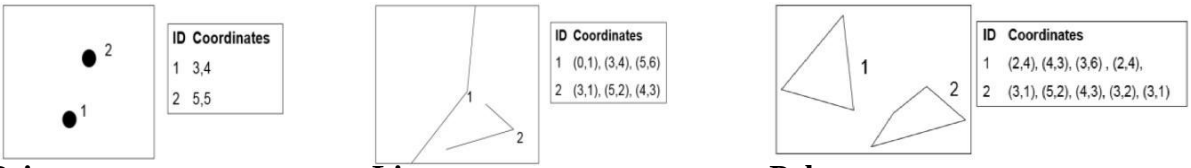
I. Spaghetti vector data model

II. Topological vector data model

I. The Spaghetti Model

- The spaghetti model is the most simple vector data model
- The model is a direct representation of a graphical image
- Collection of coordinate strings with no structure
- No spatial relationships stored
- Creates some redundancies within the data model and therefore reduces efficiency.
- Inefficient data storage technique
- Easy to implement, good for fast drawing
- Storage and searches are sequential, storage of attribute data

In spaghetti model, common boundaries of adjacent polygons are recorded twice, without encoding spatial relationships (Figure 2.10). This is the weakness of the model to perform spatial analysis. However, the model can efficiently reproduce digital maps since extraneous information to the plotting process is not stored.



Points **Lines** **Polygons**

Figure 2.10: Spaghetti data model: each point, line, or polygon with coordinates list to define geometry

II. Topology

Topology is the branch of mathematics; explicitly define spatial relationships between points, lines, and polygons, to determine how they share geometry (ESRI, 1999). The inclusion of

topology into the data model allows a single line to represent the shared boundary to denote which side of the line belongs with which polygon. For example, when you stand on a hill and look down on a landscape, you can identify intersecting streets and adjacent properties. A computer uses topology to identify these relationships. The geometric relationship between features and the corresponding attributes are crucial for spatial analysis and integration of data in a GIS (M. Anji Reddy, 2008). Topology can be 'stored' as topological data model, and used for analysis of non-topological data. Topology consists of three elements: *adjacency*, *containment*, and *connectivity.* Adjacency and containment describe the geometric relationships between area features. Containment is an extension of adjacency to describe area features, which may be wholly contained within another area feature.

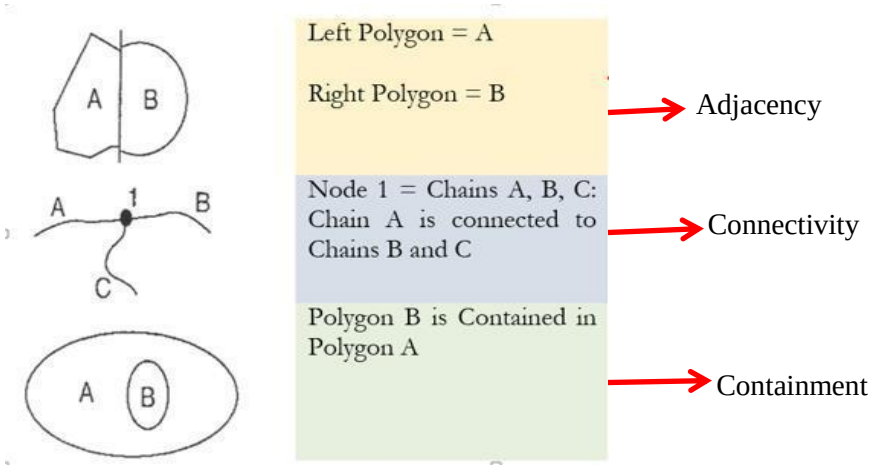


Figure 2.11: Topological spatial relationships

1) Adjacency**,** also called **contiguity** is a topologic concept that allows the vector data model to determine shared boundary of features. Areas are described as being adjacent when they share a common boundary. Arcs have direction, left and right sides to determine polygons on its left and right sides. *Adjacency answers, "Which polygons are adjacent to which?"* and used spatial analysis.
2) **Connectivity** describes linkages between line features, like road network. It connects streams to rivers, or follows a path from the water treatment plant to a house. In the arc-node data structure, arcs connect to each other at nodes and have both a from-node,

indicating where the arc begins, and a to-node, indicating where the arc ends, which is called an arc-node topology, and supports an arc-node list. Connected arcs are determined by searching through the list for common node numbers. *Connectivity answers which line segments are connected?" for network analysis.*

3) **Containment or area definition**, described as an arc that connects to surround an area to define a polygon, also called polygon-arc topology. Arcs construct polygons, and each arc is stored only once. In this case, area features which may be wholly contained within another area feature. For instance, an island defines an inner boundary (or hole) of a polygon. Containment answers, *"Which spatial features are contained within which?"* and used for selection or geocoding.

These three topological relationships will 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

Rules of topological consistency

There are five rules of topological consistency:

1. Rule 1: Every arc must be bounded by two 'nodes', namely its begin and end node
2. Rule 2: Every arc borders two 'polygons', namely its 'left' and 'right' polygons
3. Rule 3: Every polygon has a closed boundary consisting of cyclic sequence of nodes and arcs.
4. Rule 4: Around every node exists an alternating (and cyclic) sequence of arcs and polygons.
5. Rule 5: -Arcs only intersect at their (bounding) nodes

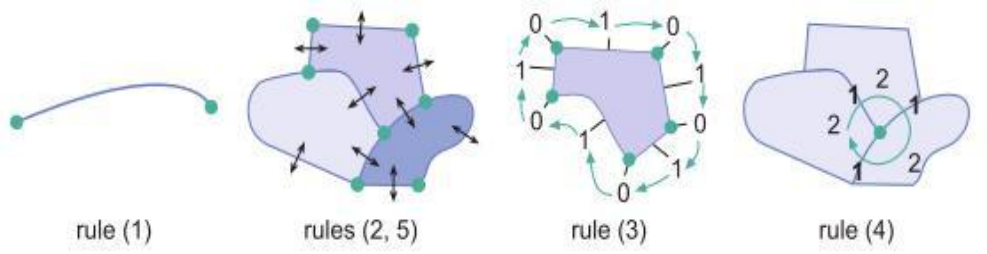


Figure 2.12: The five rules of topological consistency in two-dimensional space

Rules of topological relationships, presented and illustrated in Figure 2.12, ensure topological consistency of all features in 2D space. Such spatial relationships, along with the rules for the behavior of their representations, each define a potential case in order to maintain **data integrity.**

- Parcels must not overlap.
- Adjacent parcels have shared boundaries.
- Streamlines cannot overlap and must connect to one another at their endpoints.
- Adjacent counties have shared edges.
- Counties must completely cover and nest within states.
- Blocks must completely cover and nest within Block Groups.
- Road centerlines must connect at their endpoints.

There are eight spatial relationships: disjoint, meets, equals, inside, covered by, contains, covers, and overlaps that can be used, for instance, in queries on a spatial database. For example;

- Find all light poles that are inside pasture
- Find all the plots adjacent to a river

Topology in a geodatabase

In ArcGIS geodatabase, every topology is associated with a set of topology rules**.** Topology rules can be defined for the features within a *feature class* or for features between *two feature classes.* When a topology is *validated*, the rules are tested. ArcGIS topologies have the following characteristics:

- They exist within feature datasets
- All participating feature classes have the same spatial reference.
- There can be multiple topologies within a feature dataset.
- Feature classes can only participate in one topology.
- Feature classes cannot participate in both a topology and a geometric network.
- A topology can contain multiple point, line, and polygon feature classes.

Advantages of creating and storing topological relationships in geodatabase:

- Data is stored efficiently, and can be processed quickly.
- Ensuring geometric correctness of the data
- Improved data quality - detecting and correcting errors and validates data.
- Carrying out some types of spatial analysis (such as selections, and network analysis).

GIS queries can be optimized by storing information about topological relationships, such as:
Polygon adjacency (e.g., *who owns the parcels adjoining* those owned by another owner?).
Containment (e.g., which manholes lie within the pavement area of a given street?).

Common topological errors in GIS

Topological information permits an automatic verification of data consistency and detects errors. Such topological errors, as shown in Figure 2.13, are:

1. Incomplete closing polygons during the encoding process,
2. Open or unclosed polygons, which occur when an arc does not loop back;
3. Unlabeled polygons, occur when an area does not contain any attribute information,
4. Sliver polygons, where the shared boundary of the two polygons does not meet exactly.
5. Line feature error types, such as undershot and overshoot of lines.

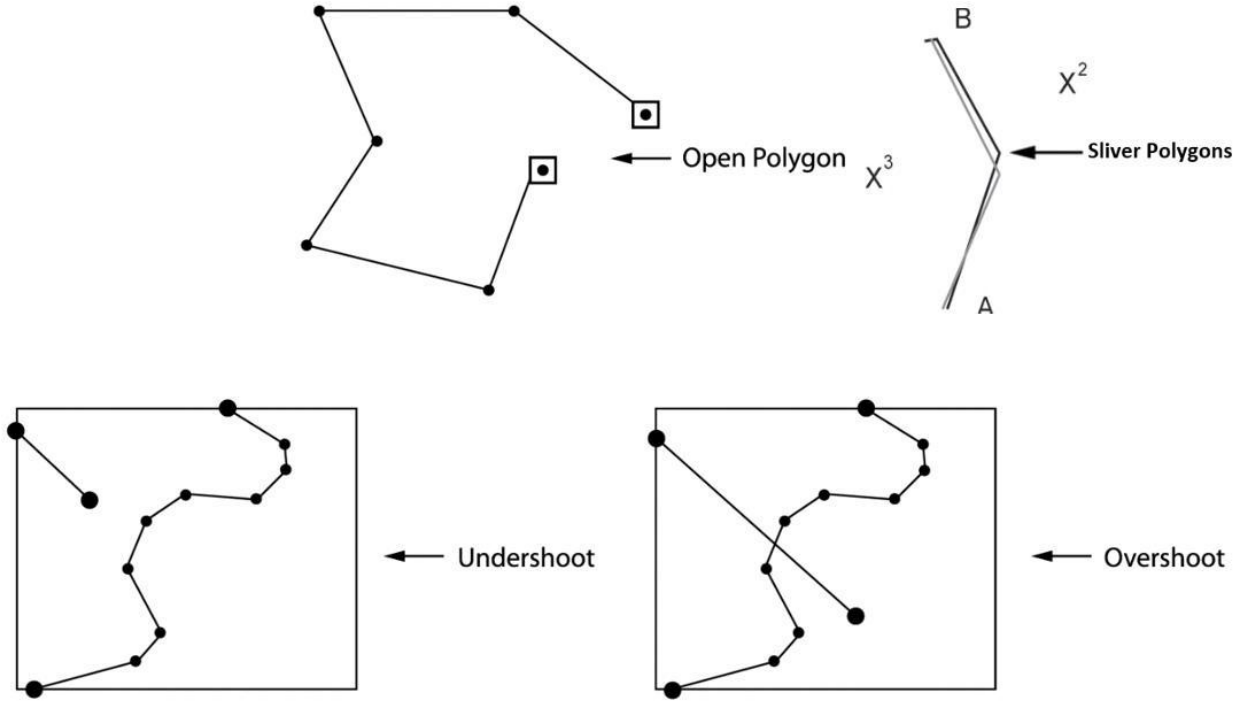


Figure 2.13: Common topological errors

Triangulated irregular networks (TIN)

TIN representation of geographic fields is a hybrid between tessellations and vector. It is an efficient and accurate model for representing **continuous surfaces**. TIN commonly used to represent Digital Terrain Models (DTMs). The principle behind a TIN is simple:

- It is constructed from a set of location measurements, such as elevation.
- Any location together with its elevation value can be viewed as a point in 3D space.
- In a 3D space, three points uniquely determine a plane, as long as they are not collinear.

From these points, we can construct an irregular tessellation made of triangles (Figure 2.15 A).

A plane fitted through these points has a fixed aspect and gradient used to compute other elevation values, such as P (Figure 2.14).

The location P is an arbitrary location that has no associated elevation measurement.

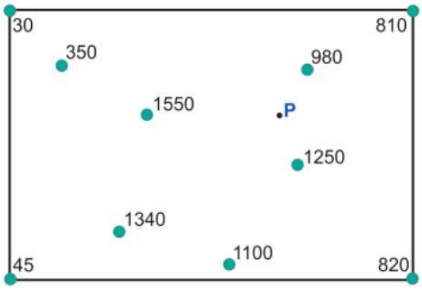


Figure 2.14: Input locations and their values for a TIN construction.

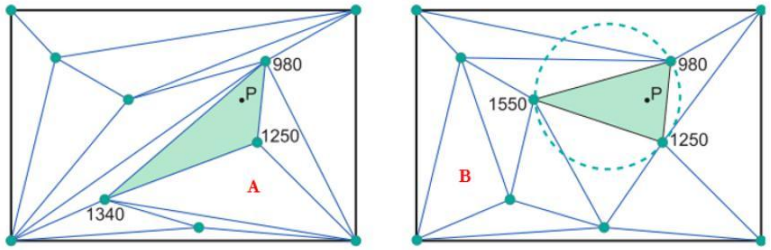


Figure 2.15: Two triangulations based on the input locations of Figure 2.14. The left one with many 'stretched' triangles; (right) the triangles are more 'equilateral', known as *Delaunay triangulation.*

Hence, by restricting the use of a plane to the triangular area 'between' three **anchor points,** we can obtain a triangular tessellation of the complete study space, which is a **Delaunay triangulation** or an optimal triangulation ('equal-sided' triangle).

2.5.Advantages and Disadvantages of Vector Models

Advantages of vectors	Disadvantages
➢ Precise location of features ➢ Storing many attributes ➢ Flexible for cartography ➢ Individual objects can be treated as separate units ➢ Ideally suited for certain types of analysis, especially areas, lengths, Connections ➢ Good representation of reality ➢ Compact data structure ➢ Topology can be described in a network ➢ high-quality graphics and precise drawing ➢ simplicity in searching and editing of geoelements ➢ Adapts well to scale changes	➢ Complex data structures ➢ not suitable for modelling continuous surfaces ➢ Some spatial analysis is difficult or impossible to perform ➢ Difficult to compatible with remote sensing data ➢ complexity of calculations in some analytical operations ➢ Inefficient for image processing ➢ More update-intensive

2.6.Advantages and Disadvantages of Raster data model

Advantages	Disadvantages
• Simple data structure • Easy overlay • Various kinds of spatial analysis • Uniform size and shape • Technology is cheap • Covers vast area • Efficient for image processing	• Large amount of data • Projection transformation is difficult and time consuming • Different scales between layers can be difficult • May lose information • Crude raster maps are considerably less beautiful than line maps

2.7. Comparison of Raster and Vector data models

• **Data structure** Raster- usually simple Vector- usually complex	• **Accessibility** Raster- easy to modify or program due to simple data structure Vector- often complex
• **Storage requirements** Raster- large for most data sets Vector- small for most data sets	• **Display and output** Raster- good for images not for discrete features Vector- map like, poor for images
• **Analysis** Raster- easy for continuous data, simple for many layer combinations Vector- preferred for network analyses, complex for other spatial operations	• **Coordinates conversion** Raster- may be slow depending on size of data sets Vector- simple
• **Positional precision** Raster- floor set by pixel size Vector- limited only by quality of positional measurements	