GIS Fundamentals:

A First Text on Geographic Information Systems

6th Edition

Paul Bolstad
University of Minnesota - Twin Cities
2 Data Models

Introduction

Data in a GIS represent a simplified view of physical entities – the roads, mountains, accident locations, or other features we care about. Data include information on the spatial location and nonspatial properties of entities.

Each entity is represented by a spatial object in a GIS, defining an entity-object correspondence. Because every computer system has limits, we can’t save the exact boundary or all characteristics of features. As illustrated in Figure 2-1, we may represent land cover by a set of polygons. The polygon boundaries may be defined by a connected set of points, e.g., at an average spacing of approximately every 3 meters. We may record data that define each land cover, perhaps vegetation type, ownership, and landuse. Edge details smaller than 3 m and unrecorded characteristics such as value are not included in this representation.

The spatial detail and essential characteristics are subjectively chosen by the data developer. The density of points required by a surveyor will be different than that for a land use planner. The essential characteristics of a forest would be different in the eyes of a logger than those of a hunter or hiker. No one representation is universally better than any other, and the GIS developer seeks to define objects that support the intended use of the data, at the desired level of detail and accuracy.

Figure 2-1: A physical entity is represented by a spatial object in a GIS. Here, lakes (dark areas in the photograph) and other land cover types are represented by polygons in the data layers on the right.
A spatial data model (Figure 2-2) may be defined as the objects in a spatial database plus the relationships among them. The term “model” is fraught with ambiguity because it is used in many disciplines to describe many things. Here, a spatial data model provides a formal means of representing and manipulating spatially referenced information. In Figure 2-1, our data model consists of two parts. The first is a set of polygons recording the edges of distinct land uses, and the second part (not shown in the figure) is a set of numbers, letters, or words associated with each polygon. The data model is the most recognizable level in our computer abstraction of the real world. Data structures (how we organize the information in the computer) and binary machine code (how we record it), are successively less recognizable but more computer-compatible forms of the spatial data (Figure 2-2).

Most GIS store our data as a set of layers (Figure 2-3). Each layer organizes the spatial and attribute data for a kind of cartographic object, and are often referred to as thematic layers. As an example, consider a GIS database that includes a soils data layer, a population data layer, an elevation data layer, and a roads data layer. The roads layer contains only roads data, including the location and properties of roads in the analysis area. Information on soils, political boundaries, and elevation are contained in their respective data layers. Through analyses we may combine data to create a new data layer; for example, we may identify areas that have high elevation and join this information with the soils data. This combination may create a new data layer with a composite soils-elevation variable.

Coordinates are used to define the spatial location and extent of geographic objects (Figure 2-4). A coordinate most often consists of a pair or triplet of numbers that specify location in relation to an origin. The coordinates quantify the distance from the

![Figure 2-2: Levels of abstraction in the representation of spatial entities. The real world is represented in successively more machine-compatible but humanly obscure forms.](image)
origin when measured along standard directions. Single or groups of coordinates are organized to represent the shapes and boundaries that define objects. Coordinates are usually based upon standardized map projections (discussed in Chapter 3). Each projection unambiguously defines the coordinate values for every point in an area.

Typically, attribute data complement the coordinate data for cartographic objects (Figure 2-4). These attribute data record the non-spatial components of an object, such as a name, color, pH, or cash value. Keys, labels, or other indexes are used so that the coordinate and attribute data may be viewed, related, and manipulated together.

Coordinate Data

Coordinates define location in two- or three-dimensional space. Spatial data in a GIS most often use coordinate pairs, $x$ and $y$, in a Cartesian coordinate system, named after Rene Descartes, the system’s originator. These pairs define data on a flat, two-dimensional surface, and define the locations of features in our data layers. When working over large areas, we often require a three-dimensional representation. Coordinates in three dimensions are a bit more complicated because two alternate systems are common. Most adults are familiar with the concepts of latitude ($\phi$), longitude ($\lambda$) and an elevation to define locations on the surface of the Earth. Spatial calculations are often easier in a three-dimensional Cartesian system starting near the Earth’s center and using coordinate triplets $X$, $Y$, and $Z$. These alternate conventions for coordinate systems are described in turn in the following sections.
Planar Coordinate Systems

Planar, two-dimensional (2-D) Cartesian coordinate systems are the most common choice for GIS data storage and analysis. These systems define two orthogonal axes (right angle, or 90°), forming a plane (Figure 2-5). We specify a Y-axis, usually aligned at or close to a north-south direction, and an X-axis, usually aligned at or near an east-west direction. The Y-axis is often referred to as a northing axis and values increase upwards in a grid north direction. The X-axis is often referred to as an easting axis with values increasing to the right.

We must be careful when making measurements on our flat, 2-D data. When we display geographic data on a flat surface, we unavoidably distort relative locations, because the Earth’s true surface is curved. Distance or area measurements are not the same on our imaginary flat surface as on the Earth’s surface. We typically introduce small errors when we ignore the Earth’s curvature, and we can keep errors below acceptably small values by limiting the area over which we use our flat 2-D model. As the mapped distance increases, the error increases to magnitudes we usually can’t ignore. Specific methods for managing distortion in this curved to flat surface conversion are discussed in Chapter 3.

Coordinates on a Sphere

When we map over larger areas or when we need the highest precision and accuracy, we often use a three-dimensional, spherical coordinate system. Hipparchus, a Greek mathematician of the 2nd century B.C., was among the first to specify locations on the Earth using angular measurements on a sphere. A common spherical system uses two angles of rotation on a sphere with a fixed radius, R, to specify locations on Earth (Figure 2-6). The first angle of rotation, the longitude (λ), measures east-west distances around the polar, rotational axis of Earth. Zero is set for a line that passes near the Greenwich Observatory in England, and the distance angle is positive eastward and negative westward (Figure 2-6). The zero longitude, also known as the Prime Meridian or the Greenwich Meridian, was first specified through the Royal Greenwich Observatory in England, but measurement improvements, crustal movements, and changes in conventions now place zero longitude about 102 meters (335 feet) east of the Greenwich Observatory.

A second angle of rotation, measured along north-south lines that intersect the poles, is used to define a latitude (φ, Figure 2-6). Latitudes are specified as zero at the Equator, the line encircling the Earth that is...
always halfway between the North and South Poles. By convention, latitudes increase to maximum values of 90 degrees in the north and south, or, if a sign convention is used, from -90 at the South Pole to 90 at the North Pole. Lines of constant longitude are called meridians, and lines of constant latitude are called parallels (Figure 2-6). Because the meridians converge, geographic coordinates do not form a Cartesian system. A Cartesian system defines lines on a right-angle, planar grid. Geographic coordinates occur on a curved surface, and the longitudinal lines cross at the poles. This convergence means the distance spanned by a degree of latitude varies only slightly, from 110.6 kilometers at the Equator to 111.7 kilometers at the poles. The slight difference with latitude is due to a non-spherical Earth, something we’ll describe a bit later.

Convergence causes distortion because a degree of latitude spans a greater distance near the poles than a degree of longitude. For example, “circles” with a fixed radius in geographic units, such as 5°, are not circles on the surface of the globe, with distortion greatest at the poles (Figure 2-7, left). They may appear as circles when the Earth’s surface is “unrolled” and plotted on a flat map (Figure 2-7, right), but treating spherical coordinates (latitudes/longitudes) as Cartesian coordinates creates an inherently dis-
torted map. Note the distorted shape of Antarctica in Figure 2-7, right.

Because the spherical system for geographic coordinates is non-Cartesian, formulas for area, distance, angles, and other geometric properties used in a Cartesian coordinate system should not be used with geographic coordinates. Areas are usually calculated after converting to a projected system, described in chapter 3.

There are two primary conventions used for specifying latitude and longitude (Figure 2-8). The first uses a leading letter, N, S, E, or W, to indicate direction, followed by a number to indicate location. Northern latitudes are preceded by an N and southern latitudes by an S, for example, N90°, S10°. Longitude values are preceded by an E or W, for example W110°. Longitudes range from 0 to 180 degrees east or west. Note that the east and west longitudes meet at 180 degrees, so that E180° equals W180°.

Signed coordinates are the second common way to specify latitude and longitude. Northern latitudes are positive and southern latitudes are negative, and eastern longitudes positive and western longitudes negative. Latitudes vary from -90 degrees to 90 degrees, and longitudes vary from -180 degrees to 180 degrees. By this convention, the longitudes "meet" at the maximum and minimum values, so -180° equals 180°.

Coordinates may easily be converted between these two conventions. North latitudes and east longitudes are converted by removing the leading N or E. South latitudes and west longitudes are converted by first removing the leading S or W, and then changing the sign of the remaining number from a positive to a negative value.

Spherical coordinates are most often recorded in a degrees-minutes-seconds (DMS) notation: N43° 35′ 20″ for 43 degrees, 35 minutes, and 20 seconds of lati-
360° to circle the sphere
60', or 60 minutes, for each degree
60", or 60 seconds, for each minute

Spherical coordinates may also be expressed as decimal degrees (DD). When using DD, the degrees take the usual -180 to 180 (longitude) and -90 to 90 (latitude) ranges, but minutes and seconds are reported as a decimal portion of a degree (from 0 to 0.99999...).

Conversion between DMS and DD is shown in Figure 2-10.

**Figure 2-9:** There are 360 degrees in a complete circle, with each degree composed of 60 minutes, and each minute composed of 60 seconds.

**Figure 2-10** Examples for converting between DMS and DD expressions of spherical coordinates.

### DD from DMS

\[
DD = D + M/60 + S/3600
\]

**e.g.**

\[
DMS = 32° 45' 28"
\]

\[
DD = 32 + 45/60 + 28/3600
\]

\[
= 32 + 0.75 + 0.0077778
\]

\[
= 32.7577778
\]

### DMS from DD

\[D = \text{integer part}\]

\[M = \text{integer of decimal part} \times 60\]

\[S = \text{2nd decimal} \times 60\]

**e.g.**

\[DD = 24.93547\]

\[D = 24\]

\[M = \text{integer of first decimal} \times 60\]

\[= 0.93547 \times 60\]

\[= \text{integer of 56.1282}\]

\[= 56\]

\[S = \text{2nd decimal} \times 60\]

\[= 0.1282 \times 60 \approx 7.692\]

so DMS is

\[24° 56' 7.692"\]
Spherical vs. Ellipsoidal Earth

While we often describe the Earth’s shape as a sphere, it is better approximated as an ellipsoid. A sphere is a solid object defined by a center location and an equal radius in all directions. An ellipsoid is an approximately spherical solid, but with unequal radii along the axes. Spheroids and ellipsoids may be viewed in cross-section, revealing their difference in shape (Figure 2-11). The Earth’s shape is best viewed as an ellipsoid flattened in the north-south direction. This flattening is quite small, approximately one part in 300. Translated to human scales, this is about an 8 mm (1/30th of an inch) flattening in a basketball. While difficult to observe directly, it is large enough to distort common geodetic measurements and navigation on the surface of the Earth. Many navigation and measurement estimates have two sets of formulas, one an approximation based on a purely spherical globe, and a more complicated and precise set based on an ellipsoidal shape.

Note that the words spheroid and ellipsoid are often used interchangeably. GIS software often prompts the user for a sphere or spheroid when defining a coordinate projection, and then lists a set of ellipsoids, with differing polar and equatorial radii.

The best estimates of Earth’s radii, \( a \) and \( b \), have evolved as measurement systems have improved. Today, the best estimate for \( a \) is 6,378,137.0 meters (m), and for \( b \) 6,356,752.3 m. A mean value of 6,367,444.7 m is often used for spheroids, but sometimes the value for \( a \) is adopted, 6,378,137 m, or just 6,378 km.

![Figure 2-11: Spherical (left) vs. ellipsoidal (right) approximations of the Earth’s shape. A sphere has a single radius, while an ellipse has different radii along the semi-major and semi-minor axes. The spheroid and ellipsoid can be thought of as rotating these two basic shapes around the polar axis to create solid figures.](image-url)
Converting Arc to Surface Distances

At times we need to calculate the distance on the surface of the Earth that is spanned by an arc measure. For example, I might have two locations that differ by 10 seconds of arc, and wish to estimate the distance between them. We can approximate the surface distance on a circle or sphere by the formula:

\[ d = r \cdot \theta \]  

(2.1)

where \( d \) is the approximate ground distance, \( r \) is the radius of the circle or sphere, and \( \theta \) is the angle of the arc. There is a more complicated formula for ellipsoidal surfaces, but the above formula is acceptable for most applications.

**Figure 2-12**: Example calculation of the approximate surface distance spanned by an arc.

**Converting degrees to radians:**

30.1487 degrees is

\[
\frac{30.1487}{57.2957795} = 0.52619 \text{ radians}
\]

**Converting radians to degrees:**

1.284 radians is

\[
1.284 \times 57.2957795 = 73.5678 \text{ degrees}
\]

**Figure 2-13**: Conversion between radian and degree angle units.

Figure 2-12 shows an example calculation of arc length, using the average radius for Earth. Note that equation (2.1) applies to a generic arc angle, measured in the direction of the spanned arc, without regard to the latitude/longitude system. Substituting latitude values will result in a reasonably accurate answer, but substituting longitude values anywhere but along the Equator will result in an error, largest near the poles, due to longitudinal convergence. The formula is best used as a first approximation of distance spanning generic arcs, and not using longitudinal coordinates.

Note that the angle should be specified in radian measure, defined as \( 2\pi \) radians per the 360 degrees, or approximately 57.2957795 degrees per radian. Radian measures are an alternative to degrees, and scale the rotation by the radius of the circle. You may easily convert between radian and degree units (Figure 2-13). Many spreadsheet, online, and app programs by default use radian measure, and substituting degrees will lead to errors.
Great Circle Distance
Spherical approximation
Consider two points on the Earth’s surface, 
A with latitude, longitude of \((\phi_A, \lambda_A)\), and 
B with latitude, longitude of \((\phi_B, \lambda_B)\).

The great circle distance between points on a sphere is given by the formula:
\[
d = r \cdot 2 \sin^{-1}\left(\frac{\sin^2\left(\frac{\Delta \phi}{2}\right)}{\sin^2\left(\frac{\Delta \lambda}{2}\right)}\right)
\]
where \(d\) is the shortest distance on the surface of the Earth from A to B, 
r is the Earth’s radius, approximately 6378 km, and \(\Delta \phi, \Delta \lambda\) are the differences between point latitudes and longitudes, divided by two.

As an example, the distance between Paris, France, and Seattle, USA, is:
Latitude, longitude of Paris, France = 48.864716\(^o\), 2.349014\(^o\)
Latitude, longitude of Seattle, USA = 47.655548\(^o\), -122.30320\(^o\)
\[
d = 6378 \cdot 2 \sin^{-1}\left(\frac{\sin^2(0.604584)}{\sin^2(62.36107)}\right) = 8,034.8391\text{km}
\]

Figure 2-14: Calculation of the great circle distance between points.

The great circle distance formula should be used to estimate the surface distance between two points when using latitudes/longitudes (Figure 2-14). A great circle is defined by any plane that intersects a globe and passes through its center. The Equator and meridians are great circles, while lines of equal latitude other than the Equator are not great circles. A great circle distance is the shortest path on the Earth’s surface between two points, and long-distance airline routes approximate great circles. As with all trigonometric formulas, you should know if your calculations expect degree or radian measures as input, and convert accordingly.

Three-Dimensional, Earth-Centered Coordinates
We noted an alternate, three-dimensional (3-D) Cartesian representation of coordinates for locations, typically in, on, or near the Earth (Figure 2-15). This is commonly used in geodesy, the science of the Earth’s shape, size, and physical dynamics, that underpins all coordinate measures.

Geodesy is at the heart of map projections (Chapter 3) and satellite positioning (Chapter 5), fundamental building blocks of GIS.

The 3D Cartesian system typically places the origin near or at the mass center of the Earth. This Cartesian system is aligned with the Z axis through the geographic North Pole and the X and Y axes forming a plane.

Figure 2-15: A 3-D Cartesian coordinate system.
on the Equator (Figure 2-16). The positive X-axis intersects the ellipsoid where latitude and longitude values are both zero, and the positive Y-axis intersects the ellipsoid at a longitude of 90 and latitude of 0.

Mathematical formulas allow us to calculate any X, Y, and Z given any latitude, longitude, and Earth radii (Figure 2-16). Each latitude/longitude/radius coordinate in the geographic system corresponds to an X-Y-Z triplet in the 3-D Cartesian coordinate system. These formulas are commonly used by geodesists in the most precise surveys, but are also embedded in many softwares that convert between different versions of our coordinate data.

There are two different sets of equations, one assuming a spherical Earth, and a more accurate one assuming an ellipsoidal Earth. A detailed discussion of these is best left for an advanced course, so formulas are included in Appendix C for reference.

**Geographic and Magnetic North**

There is often confusion between magnetic north and geographic north. Magnetic north and the geographic north do not coincide (Figure 2-17). Magnetic north is the location towards which a compass points. The geographic North Pole is the average northern location of the Earth’s axis of rotation. If you were standing on the geographic North Pole with a compass, it would point approximately in the direction of the Bering Straits, and some 200 kilometers away. In addition, Magnetic North “wanders” through time, and has recently increased it’s rate of shift (Figure 2-17).

Because magnetic north and the geographic North Pole are not in the same place, a compass does not point towards geographic north when observed from most places on Earth. The compass will usually point east or west of geographic north, defining an angular difference called the magnetic declination. Declination varies across the globe, and also has varied through time as magnetic north wanders.

Note that our definition of geographic north is the average northern location of the Earth’s axis of rotation. We say average because the Earth wobbles, or nutates, on its axis. This means the axis location varies

---

**Figure 2-16**: Formulas exist to convert between known spherical geographic coordinates (latitude and longitude on a spheroid) and corresponding 3-D Cartesian coordinates (see appendix C).

**Figure 2-17**: Magnetic and geographic North Poles. Year dates show how the Magnetic North has wandered through time, increasing in velocity over the past few decades.
slightly, within a circle about 9 meters (30 feet) across, so the northern pole location is always within this circle. The nutation has a period of 433 days, with the pole returning back to its original location over that time.

### Attribute Data and Types

Attribute data are used to record the non-spatial characteristics of an entity. Attributes, also called *items* or *variables*, may be envisioned as a list of characteristics that describe features. Color, depth, weight, owner, vegetation type, or land use are examples of variables that may appear as attributes. Attributes record values; for example, a fire hydrant may be colored red, yellow, or orange, have 1 to 4 flanges, and a pressure rating of any real number from 0 to 12,000.

Attributes are often presented in tables and arranged in rows and columns (Figure 2-18). Each row corresponds to a spatial object, and each column corresponds to an attribute. Tables are often organized and managed using a specialized computer program called a database management system (DBMS, described more fully in Chapter 8).

All attributes can be categorized as nominal, ordinal, or interval/ratio attributes. *Nominal attributes* are variables that provide descriptive information about an object. The color is recorded for each hydrant in Figure 2-18. Other examples of nominal data are vegetation type, a city name, the owner of a parcel, or soil series. There is no implied order, size, or quantitative information contained in nominal attributes.

Nominal attributes may also be images, film clips, audio recordings, or other descriptive information, for example, GIS for real estate often have images of the buildings as part of the database. Image, video, or sound recordings stored as attributes are sometimes referred to as “BLOBs” for *binary large objects*.

*Ordinal attributes* imply a ranking by their values. An ordinal attribute may be descriptive, such as high, mid, or low, or it may be numeric; for example, an erosion class with values from 1 to 10. The order reflects only rank, and not scale. An ordinal value of four has a higher rank than two, but we can’t infer that the attribute value is twice as large.

*Interval/ratio attributes* are used for numeric items where both rank order and absolute difference in magnitudes are represented, for example, the number of flanges in the second column of Figure 2-18. These data are often recorded as real numbers on a linear scale. Area, length, weight, height, or depth are a few examples of attributes that are represented by interval/ratio variables.

Items have a *domain*, a range of values they may take. Colors might be restricted to red, yellow, and green; cardinal direction to north, south, east, or west; and size to all positive real numbers.
Chapter 2: Data Models

Common Spatial Data Models

All spatial data models are based on a conceptualization. As an example, consider a regional map that defines roads as lines. We conceive of each road as a linear feature that fits into a small number of categories. These lines connect cities and towns that are shown as discrete points or polygons on the map. Road properties may include only the road type, e.g., highway or local road. The roads have a width represented by a line symbol on the map; however, the scaled road width may not represent the true road width. All state highways are represented equally although they may vary. Some may have wide shoulders, others not, or dividing barriers of concrete, versus a broad vegetated median, but we may choose to omit this variation, fitting all highways into one class.

There are two main conceptualizations used for digital spatial data. The first defines discrete objects using a vector data model. This model uses discrete elements such as points, lines, and polygons to represent the geometry of real-world entities (Figure 2-19, left). Farm fields, roads, wetlands, cities, and census tracts are examples of entities that are often represented by discrete vector objects. Points are often used to define the locations of “small” objects such as wells, buildings, or ponds. Lines may be used to represent linear objects, for example, rivers or roads, or to enclose polygons, which identify area objects. Starting points and ending points for a line are sometimes referred to as nodes, while intermediate points in a line are referred to as vertices.

Vector objects are discrete. A forest may share an edge with a pasture, and this boundary is represented by lines. In truth, a forest edge may grade into a mix of trees and shrubs, then shrubs and grass, then pure grass; however, in the vector conceptualization, a line between two land cover types will be drawn to indicate a discrete, abrupt transition. Lines and points have coordinate locations, but points have no dimension, and lines have no dimension perpendicular to their direction. Area features are defined by a closed, connected set of lines.

The second common conceptualization identifies and represents grid cells for a given region of interest. This conceptualization employs a raster data model (Figure 2-19, right). Raster cells are arrayed in a row and column pattern to provide “wall-to-wall” coverage of a study region. Cell values are used to represent the type or quality of

![Figure 2-19: Vector and raster data models.](image-url)
mapped variables. Raster models are often used with variables that may change continuously across a region. Elevation, mean temperature, slope, average rainfall, cumulative ozone exposure, or soil moisture are examples of phenomena that are often represented as continuous fields. Raster representations are also sometimes used to represent discrete features, for example, class maps of vegetation or political units.

Data models are often interchangeable in that many phenomena may be represented by many data models. For example, elevation may be represented as a raster surface (continuous field) or as a series of lines representing contours of equal elevation (discrete objects). Data may be converted from one model to another; for example, the location of contour lines may be determined by evaluating the raster surface, or a raster data layer may be derived from a set of contour lines. These conversions entail some costs both computationally and perhaps in data accuracy.

The decision to use either a raster or vector model often depends on our conceptualization of the objects and the most frequent operations performed. We think of elevation as a continuous variable and slope is more easily determined when elevation is represented in a raster data set. However, discrete contours are often the preferred format for printed maps, so the discrete conceptualization of a vector data model may be preferred in some cases. The best data model for a given application depends on the most common operations, the experiences and views of the GIS users, the form of available data, and the influence of the data model on data quality.

Other, less common data models are sometimes used. A triangulated irregular network (TIN) is one such model, employed to represent surfaces such as elevations, through a combination of point, line, and area features. We will introduce and discuss less common data models later in this chapter.

**Vector Data Models**

A vector data model uses sets of coordinates and associated attribute data to define discrete objects. Groups of coordinates define the location and boundaries of discrete objects, and these coordinate data plus their associated attributes are used to create vector objects representing the real-world entities (Figure 2-20). In the most common vector models, there is an attribute table associated with each vector layer, and a single row in the table corresponding to each feature in the data layer. These vector layers are said to contain *single-part features*, because there is a single geographic object for each row in the table, with one to several columns in each row. All values in a column have the same type, so for any given column, all entries might be ordinal, or interval ratio, or a BLOB, or some other defined type. An identifier value, or ID, is typically included, and this value is often unique within the table, with an unrepeated value assigned for each row and corresponding feature.

There are three basic types of vector objects: points, lines, and polygons (Figure 2-20, top). A point uses a single coordinate pair to represent the location of an entity that is considered to have no dimension. Gas wells, light poles, accident location, and survey points are examples of entities often represented as point objects. Some of these have real physical dimension, but for the purposes of the GIS users they may be represented as points. In effect, this means the size or dimension of the entity is not important, only its location.

Attribute data are attached to each point, and these attribute data record the important non-spatial characteristics of the point entities (Figure 2-20). When using a point to represent a light pole, important attribute information might be the height of the pole, the type of light and power source, and the last date the pole was serviced.

Linear features are represented as lines in vector data models (Figure 2-20, mid). Lines are most often represented as an ordered set of coordinate pairs. Each line is
made up of line segments that run between adjacent coordinates in the ordered set. Attributes in a table correspond to line segments (Figure 2-20, mid). Curved linear entities are most often represented as a collection of short, straight, line segments, although curved lines are at times represented by a mathematical equation describing a geometric shape. The line starting and ending points are often called nodes, and intermediate points used to represent the line shape are called vertices.

Area entities are most often represented by closed polygons (Figure 2-20, bottom). These polygons are formed by a set of connected lines, either one line with an ending point that connects back to the starting point, or as a set of lines connected start-to-end. Polygons have an interior region and may entirely enclose other polygons in this region. Polygons may be adjacent to other polygons and thus share “bordering” or “edge” lines with other polygons. Attribute data such as area, perimeter, land cover type, or county name may be linked to each polygon (Figure 2-20, bottom).
Note that there is no uniformly superior way to represent features, and we may represent the same features as points, lines, or polygons (Figure 2-21). Some feature types may appear to be more "naturally" represented one way: manhole covers as points, roads as lines, and parks as polygons. However, in a very detailed data set, the manhole covers may be represented as circles, and both edges of the roads may be drawn and the roads represented as polygons. The best representation depends on the detail, accuracy, and intended use of the data set.

Vector layers sometimes have a many-to-one relationship between geographic features and table rows (Figure 2-22), defining multi-part features. In these instances, many spatially distinct features are matched with a row, and the row attributes apply to all the distinct features. This is common when representing islands, groups of buildings, or other clusters of features that make up a perceived whole thing. These multi-part features may have multiple geographic objects that correspond to one row.

Multi-part features may also be used for large data sets, for example, when millions of point observations are collected automatically with laser scanners. Tables are often slower to process than point geographies, and so reducing the table size by grouping points into multi-part features may shorten many operations.

**Figure 2-21:** The same objects may be represented by points, lines, or polygons, depending on our view of and intended use for the data.
**Figure 2-22**: Example of multi-part and single-part features. Here, counties for Rhode Island, a state in the Eastern U.S.A., are shown with one table entry for each county (top), with multi-part features in a layer, and with one table entry for each distinct polygon (bottom), with only single-part features in the layer. Note that calculations, analysis, and interpretation may differ for multi-part v.s. single-part features.
Care is warranted when converting multi-part features to single-part features. The most common problems arise for aggregate variables in polygon layers, such as total counts. For example, population data are often delivered by census areas such as states. Many states, e.g., Hawaii, have several parts and are represented by a multi-part shape. The population is associated with the aggregated set of polygons comprising the state (Figure 2-23). When converted to single-part shapes, the attributes are often copied for each component polygon. In our example, all single-part polygons will be assigned the attribute values for the multi-part feature, in effect repeating counts for each part. Subsequent aggregation or calculation across the population column may result in error.

Attributes for converted shapes may be corrected. If component data are available, they can be assigned to each of the single-part features. If not, then some weighting scheme may be available, for example, if there is a correlation between area and count. Until they are reviewed and appropriately adjusted, single-part attributes derived from multi-part features should be used with caution.

**Polygon Inclusions and Boundary Generalization**

Vector data frequently exhibit two characteristics: polygon inclusions and boundary generalization. These characteristics are often ignored, but may affect the use of vector data. These concepts must be understood, their presence evaluated, and effects weighed in the use of vector data sets.

**Polygon inclusions** are areas in a polygon that are different from the rest of the polygon, but still part of it. Inclusions occur because we typically assume an area represented by a polygon is homogeneous, but this is often untrue, as illustrated in Figure 2-24. The figure shows a vector polygon layer representing raised landscaping beds (a). The general attributes for the polygon may be coded; for example, the surface type may be recorded as cedar mulch. The area noted in Figure 2-24b shows a walkway that is an inclusion in a raised bed. This walkway has a concrete surface. Hence, this walkway is an unresolved inclusion within the polygon.

---

**Figure 2-23:** multi-part to single-part conversion may lead to errors in subsequent analysis because attributes may be copied from the original, multi-part cluster (left, above), to each single-part component (right, above). Density, sums, or other derived variables often should be re-calculated for single-part features, but often are not, resulting in errors.
One solution creates a polygon for each inclusion. This often is not done because it may take too much effort to identify and collect the boundary location of each inclusion, and there typically is some lower limit, or *minimum mapping unit*, on the size of objects we care to record in our data. Inclusions are present in some form in many polygon data layers.

*Boundary generalization* is the incomplete representation of boundary locations. This problem stems from the typical way we represent linear and area features in vector data sets. As shown in Figure 2-24c, polygon boundaries are represented as a set of connected straight-line segments. The segments are a means to trace the boundaries separating different area features. For curved lines, these straight line segments may be viewed as a sampling of the true curve, and there is typically some deviation of the line segment from the “true” curved boundary. The amount of generalization depends on many factors, and should be so small as to be unimportant for any intended use of the spatial data. However, since many data sets may have unforeseen uses or may be obtained from a third party, the boundary generalization should be recognized and evaluated relative to the specific requirements of any given spatial analysis. There are additional forms of generalization in spatial data, and these are described more thoroughly in Chapter 4.

**Figure 2-24**: Examples of polygon inclusions (sidewalk inclusion in flower bed shown in a and b), and boundary generalization (c) in a vector data model. These approximations typically occur as a consequence of adopting a vector representation, and their impacts must be considered when using vector data.
Vector Topology

Vector data often contain *vector topology*, enforcing strict connectivity and recording adjacency, and planarity. Early systems employed a spaghetti data model (Figure 2-25a), in which lines may not intersect when they should, and may overlap without connecting. The spaghetti model severely limits spatial data analysis and is little used except for very basic data entry or translation.

Topological models (Figure 2-25b) create an intersection and place a node at each line crossing, record connectivity and adjacency, and maintain information on the relationships between and among points, lines, and polygons in spatial data. This greatly improves the speed, accuracy, and utility of many spatial data operations.

Topological properties are conserved when converting vector data among common coordinate systems, a common practice in GIS analysis (described in Chapter 3). Polygon adjacency is an example of a topologically invariant property, because the list of neighbors for any given polygon does not change during geometric stretching or bending (Figure 2-25, b and c). These relationships may be recorded separately from the coordinate data.

Topological vector models may vary, and enforce particular types of topological relationships. *Planar topology* requires that all features occur on a two-dimensional surface. There can be no overlaps among lines or polygons in the same layer (Figure 2-26). When planar topology is enforced, lines may not cross over or under other lines. At each line crossing there must be an intersection.

The left side of Figure 2-26 shows non-planar graphs. In the top left figure, four line segments coincide. At some locations the lines intersect at a node, shown as white-filled circles, but at some locations a line passes over or under another line segment. These lines are nonplanar. The top right of Figure 2-26 shows planar topology enforced for these same four line segments. Nodes are found at each line crossing.

Polygons can also be nonplanar, as shown at the bottom left of Figure 2-26. Two polygons overlap slightly at an edge. This may be due to an error; for example, the two polygons share a boundary but have been recorded with an overlap, or there may be two areas that overlap in some way. If topological planarity is enforced, these two polygons must be resolved into three separate, nonoverlapping polygons. Nodes are placed at the intersections of the polygon boundaries (lower right, Figure 2-26).

Figure 2-25: Spaghetti (a), topological (b), and topological warped (c) vector data. Figures b and c are topologically identical because they have the same connectivity and adjacency.
There are additional topological constructs besides planarity that may be specified. For example, polygons may be exhaustive, in that there are no gaps, holes, or “islands” allowed. Line direction may be recorded, so that a “from” and “to” node are identified in each line. Directionality aids the representation of river or street networks, where there may be a natural flow direction.

There is no uniform set of topological relationships that are included in all topological data models. Different vendors have incorporated different topological information in their data structures. Planar topology is often included, as are representations of adjacency (which polygons are next to which) and connectivity (which lines connect to which).

Some GIS software create and maintain detailed topological relationships in their data. This results in more complex and perhaps larger data structures, but access is often faster, and topology provides more consistent, “cleaner” data. Other systems maintain little topological information in the data structures, but compute and act upon topology as needed during specific processing.

Topology may also be specified between layers, because we may wish to enforce spatial relationships between entities that are stored separately. As an example, consider a data layer that stores property lines (cadastral data), and a housing data layer that stores building footprints (Figure 2-27). Rules may be specified that prevent polygons in the housing data layer from crossing property lines in the cadastral data layer. This would indicate a building that crosses a property line. Most such instances occur as a result of small errors in data entry or misalignment among data layers. Topological restrictions between two data layers avoid these inconsistencies. Exceptions may be granted in those few cases when a building truly does cross property lines.

Figure 2-26: Nonplanar and planar topology in lines and polygons.
There are many other types of topological constraints that may be enforced, both within and between layers. **Dangles**, lines that do not connect to other lines, may be proscribed, or limited to be greater or less than some threshold length. Lines and points may be required to coincide, for example, water pumps as points in one data layer and water pipes as lines in another, or lines in separate layers may be required to intersect or be coincident. While these topological rules add complexity to vector data sets, they may also improve the logical consistency and value of these data.

Topological vector models often use codes and tables to record topology. As described above, nodes are the starting and ending points of lines. Each node and line is given a unique identifier. Sequences of nodes and lines are recorded as a list of identifiers, and point, line, and polygon topology recorded in a set of tables. The vector features and tables in Figure 2-28 illustrate one form of this topological coding.

Many GIS software systems are written such that the topological coding is not visible to users, nor directly accessible by them. Tools are provided to ensure the topology is created and maintained, that is, there may be directives that require that polygons in two layers do not overlap, or to ensure planarity for all line crossings. However, the topological tables these commands build are often quite large, complex, and linked in an obscure way, and therefore hidden from users.

Point topology is often quite simple. Points are typically independent of each other, so they may be recorded as individual identifiers, perhaps with coordinates included, and in no particular order (Figure 2-28, top).

Line topology typically includes substantial structure and identifies at a minimum the beginning and ending points of each line (Figure 2-28, middle). Topology may be organized in tables, including line identifiers, starting nodes, and ending nodes for lines. Lines may be assigned a direction, and the polygons to the left and right of the lines recorded.

Polygon topology may also be defined by tables (Figure 2-28, bottom). The tables may record the polygon identifiers and the ordered list of connected lines that define the polygon. The lines for a polygon form a closed loop, so the starting node of the first line in the list also serves as the ending node for the last line in the list.

Topological models greatly enhance many vector operations. Adjacency analyses are reduced to a “table look-up”, a quick and easy operation in most software systems. Assume the city is represented as a single polygon, and we seek all neighboring polygons. Adjacency analysis reduces to 1) scanning the polygon topology table to find the city polygon and reading the list of lines that bound the polygon, and 2) scanning this list of lines, accumulating a list of all left and right polygons. Polygons adjacent to the city may be identified from this list. List searches on topological tables are typically much faster than searches involving coordinate data.

Topological data models often have an advantage of smaller file sizes, largely because coordinate data are recorded once. For example, a nontopological approach
often stores polygon boundaries twice. Lines 52 and 53 at the bottom of Figure 2-28 will be recorded for both polygon A and polygon B. Long, complex boundaries in polygon data sets may double their size. This increases both storage requirements and processing.

There are limitations and disadvantages to topological vector models. First, there are computational costs in defining the topological structure of a vector data layer. Software must determine the connectivity and adjacency information, assign codes, and build the topological tables. Computational costs are typically quite modest with current computer technologies.

Second, the data must be very “clean”, in that all lines must begin and end with a node, all lines must connect correctly, and all polygons must be closed. Unconnected lines or unclosed polygons will cause errors during analyses. Significant human effort may be required to ensure clean vector data because each line and polygon must be checked. Software may help by flagging or fixing “dangling” nodes that do not connect to other nodes, and by automatically identifying all polygons. Each dangling node and polygon may then be checked, and edited as needed to correct errors.

Limitations and the extra editing are far outweighed by the gains in efficiency and analytical capabilities provided by topological vector models. Many current vector GIS packages use topological vector models in some form.
Vector Features, Tables, and Structures

As described earlier, geographic features are associated with nonspatial attributes in vector models; tables are used to organize the attributes. In most GIS software, we can most easily view the tables and a graphic representation of the spatial data as a linked table and digital map (Figure 2-29, top).

Most GIS employ underlying file structures to organize components of the spatial data. An example organization is shown in the bottom half of Figure 2-29, where the topological elements are recorded in a linked set of tables, in this example one for each of the polygons, lines, and nodes and vertices. Most GIS maintain the spatial and topological data as a single or cluster of linked files. This internal file structure is often insulated from direct manipulation by the GIS user, but underlies nearly all spatial data manipulations. A user may directly edit or otherwise manipulate table values, usually with the exception of the ID, and the underlying topology and coordinate data are accessed via requests to display, change, or analyze the spatial data components. Data layers may also include additional information (not shown) on the origin, region covered, date of creation, edit history, coordinate system, or other characteristics of a data set.

Note that not all GIS store coordinate and topological data in non-tabular file structures. Coordinates, points, lines, polygons, and other composite features may be stored in tables similar to attribute tables. It is premature to discuss the details of these spatially enabled databases, because they are based on something called a relational data model, described in detail in Chapter 8. Faster computers support this generally more flexible approach, allowing simpler and more transparent access across different types of GIS software.

![Image: Figure 2-29: Features in a topological data layer typically have a one-to-one relationship with entries in an associated attribute table. The attribute table typically contains a column with a unique identifier, or ID, for each feature. Topology and coordinate data are often hidden from the user, but linked to the attribute and geographic features through pointers and index variables, described in the Data and File Structures section, later in this chapter.]
Suggested Reading


Study Questions

2.1 - How is an entity different from a cartographic object?

2.2 - Describe the successive levels of abstraction when representing real-world spatial phenomena on a computer. Why are there multiple levels, instead of just one level in a spatial data representation?

2.3 - Define a data model and describe three primary differences between the two most commonly used data models.

2.4 - Characterize the following lists as nominal, ordinal, or interval/ratio:
   a) 1.1, 5.7, -23.2, 0.4, 6.67
   b) green, red, blue, yellow, sepia
   c) white, light grey, dark grey, black
   d) extra small, small, medium, large, extra large
   e) forest, woodland, grassland, bare soil
   f) 1, 2, 3, 4, 5, 6, 7

2.5 - Characterize the following lists as nominal, ordinal, or interval/ratio:
   a) Spurs, Citizens, Reds, Hornets, Baggies, Toffees, Potters
   b) pinch, handful, bucket, bushel, truckload
   c) 6.2, 7.8, 1.1, 0.5, 19.3
   d) gram, kilogram, metric ton
   e) Mexico, Canada, Argentina, Guyana, Martinique
   f) small, smaller, smallest

2.6 - Indicate which of the following are allowable geographic coordinates:
   a) N45° 45’ 45”   b) longitude -127.34795°   c) S96° 12’ 33”
   d) E 66° 15’ 60”   e) W -12° 23’ 55”   f) N 56.9999°

2.7 - Indicate which of the following are allowable geographic coordinates:
   a) N145° 45’12”   b) latitude -62.34795°   c) S110° 52’ 43”
   d) S 49° 15’ 60”   e) N 89° 59’ 59”   f) S -46.6000°
2.8 - Convert the following degree measures to radians:
   a) 47.2837°      b) 155.724°      c) -111.2045°

   Convert the following radian measures to degrees:
   d) 0.0042       e) -1.26       f) 2.25037

2.9 - Convert the following degree measures to radians:
   a) 102.83°      b) -21.533°      c) 92.045°

   Convert the following radian measures to degrees:
   d) 1.52       e) 0.014       f) 0.37

2.10 - Complete the following coordinate conversion table, converting the listed points from degrees-minutes-seconds (DMS) to decimal degrees (DD), or from DD to DMS. See Figure 2-10 for the conversion formula.

<table>
<thead>
<tr>
<th>Point</th>
<th>DMS</th>
<th>Decimal Degrees</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36°45'12&quot;</td>
<td>36.75333</td>
</tr>
<tr>
<td>2</td>
<td>114°58'2&quot;</td>
<td>14.00917</td>
</tr>
<tr>
<td>3</td>
<td>85°19'7&quot;</td>
<td>275.000001</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.99528</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>183°19'22&quot;</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.11 - Complete the following coordinate conversion table, converting the listed points from degrees-minutes-seconds (DMS) to decimal degrees (DD), or from DD to DMS. See Figure 2-10 for the conversion formula.

<table>
<thead>
<tr>
<th>Point</th>
<th>DMS</th>
<th>Decimal Degrees</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>97°45'10&quot;</td>
<td>97.75278</td>
</tr>
<tr>
<td>2</td>
<td>122°10'2&quot;</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>15°0'12&quot;</td>
<td>322.19861</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>152.65583</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>5.75</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>23°12'50&quot;</td>
<td></td>
</tr>
</tbody>
</table>

2.12 - Assume a spherical Earth with a radius of 6378.0 km. Calculate the great circle distances from St. Paul, Minnesota, latitude 44.9537°, longitude -93.09° to the following points:
   a) Chicago, latitude 41.8781°, longitude -87.6298°
   b) Reykjavik, latitude 64.1265°, longitude -21.8174°
   c) Buenos Aires, latitude -34.6037°, longitude -58.3816°

2.13 - Assume a spherical Earth with a radius of 6378.0 km. Calculate the great circle distances from St. Paul, Minnesota, latitude 44.9537°, longitude -93.09° to the following points:
   a) New York, latitude 40.7128°, longitude -74.0059°
   b) Paris, latitude 48.8566°, longitude 2.3522°
   c) Tokyo, latitude 35.6895°, longitude 139.6917°

2.14 - What is vector topology, and why is it important? What is planar topology, and when might nonplanar be more useful than planar topology?
2.15 - Identify the number of times each of the following topological rules are broken for the building outlines (red/darker) and property parcels (green/light) polygon layers. Note that all layers are semi-transparent so that you may identify overlaps.

- a) Buildings must not overlap.
- b) Parcels must not overlap.
- c) Parcels must not have gaps.
- d) Buildings must be entirely within the parcel layer.
- e) A building must not span a parcel boundary.
2.16 - Identify the number of times each of the following topological rules are broken for the building outlines (red/darker) and property parcels (green/light) polygon layers. Note that all layers are semi-transparent so that you may identify overlaps.

a) Buildings must not overlap.
b) Parcels must not overlap.
c) Parcels must not have gaps.
d) Buildings must be entirely within the parcel layer.
e) A building must not span a parcel boundary.