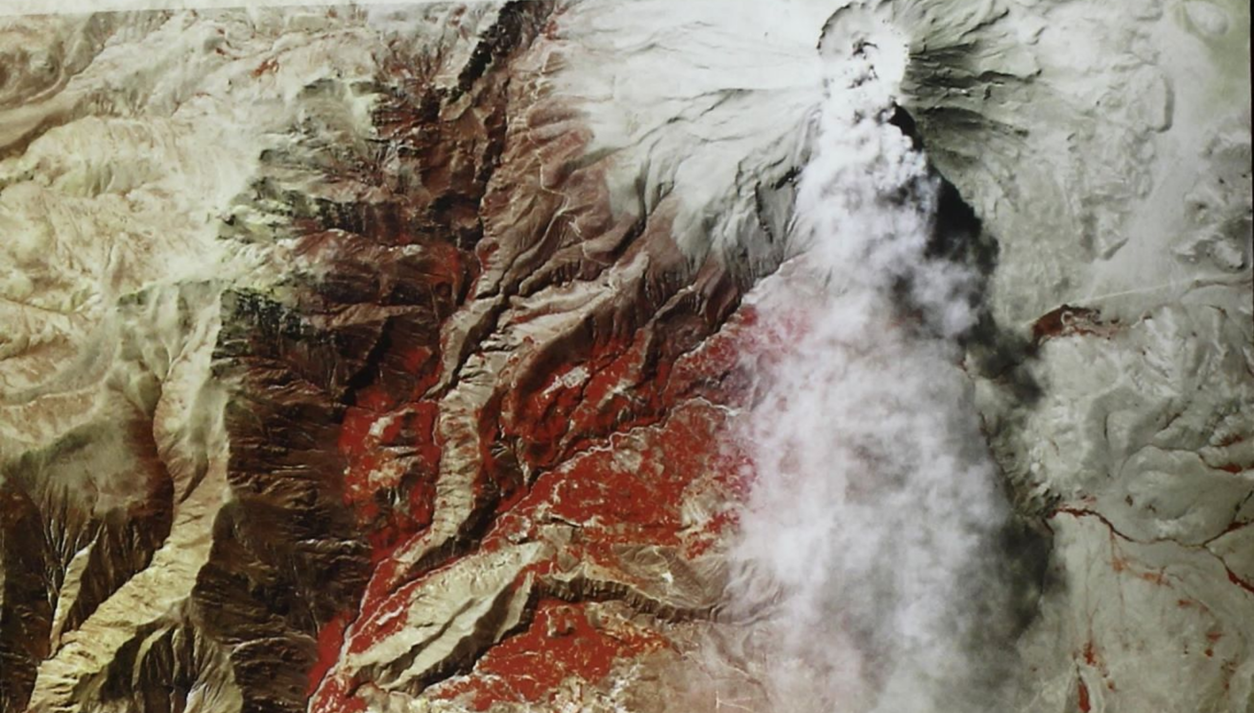
SIXTH EDITION

GIS Fundamentals

A First Text on Geographic Information Systems



Paul Bolstad

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.

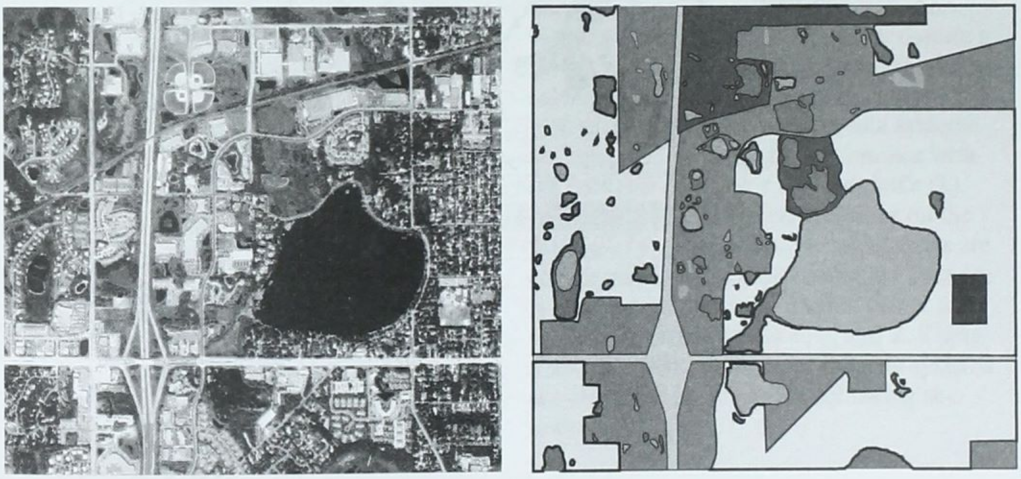
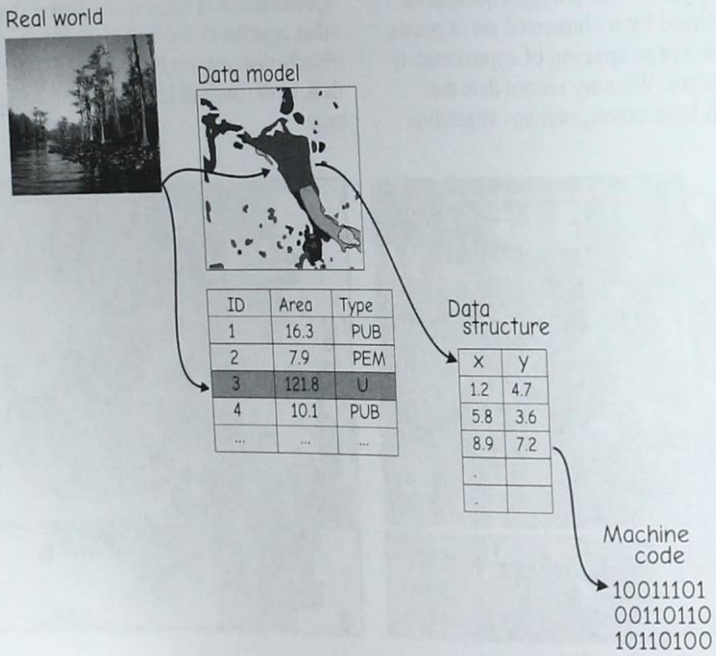


Figure 2-1: A physical entity is represented by a spatial object in a GIS. Here, lakes (dark areas in the photograph) and other land cover types are represented by polygons in the data layers on the right.

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



ID	Area	Type
1	16.3	PUB
2	7.9	PEM
3	121.8	U
4	10.1	PUB
...	...	...

x	y
1.2	4.7
5.8	3.6
8.9	7.2
.	
.	

10011101
00110110
10110100

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.

igin when measured along standard direc-
ons. Single or groups of coordinates are
ganized to represent the shapes and bound-
ies that define objects. Coordinates are
sually based upon standardized map projec-
ons (discussed in Chapter 3). Each projec-
on unambiguously defines the coordinate
alues for every point in an area.

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

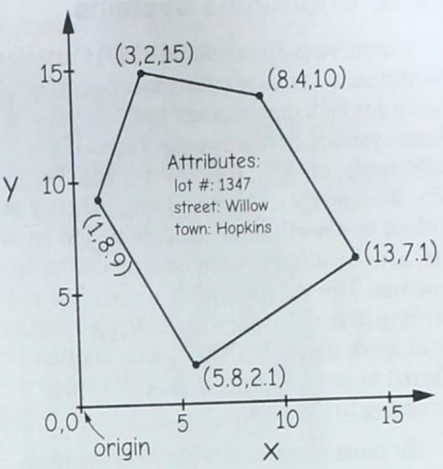


Figure 2-4: Coordinate and attribute data are used to represent entities.

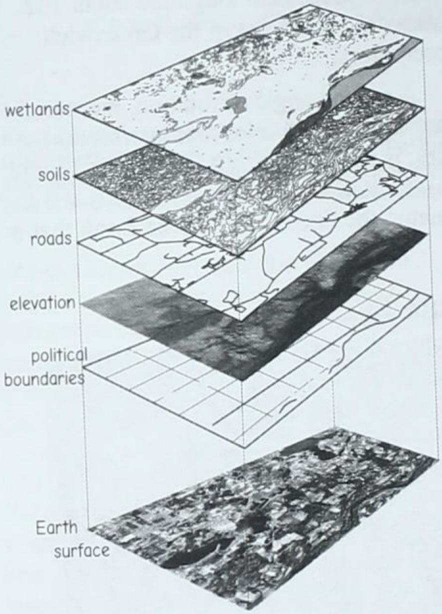


Figure 2-3: Spatial data are often stored as separate thematic layers, with objects grouped based on a set of properties, e.g., water, roads, or land cover, or some other agreed-upon set.

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 (ϕ), longitude (λ) 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^{o}), 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

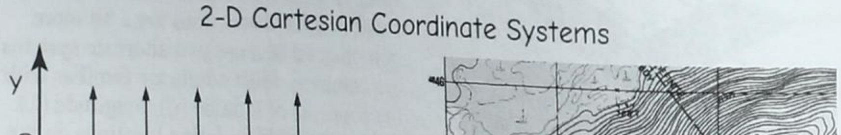


Figure 2-5: A 2-D coordinate system defines X and Y axes (left panel in figure above), and specify coordinate locations by these X-Y pairs. Coordinate values increase in rightward (X) and upward (Y) directions, and lines of constant X or Y values may be used to aid in location on maps (right, above).

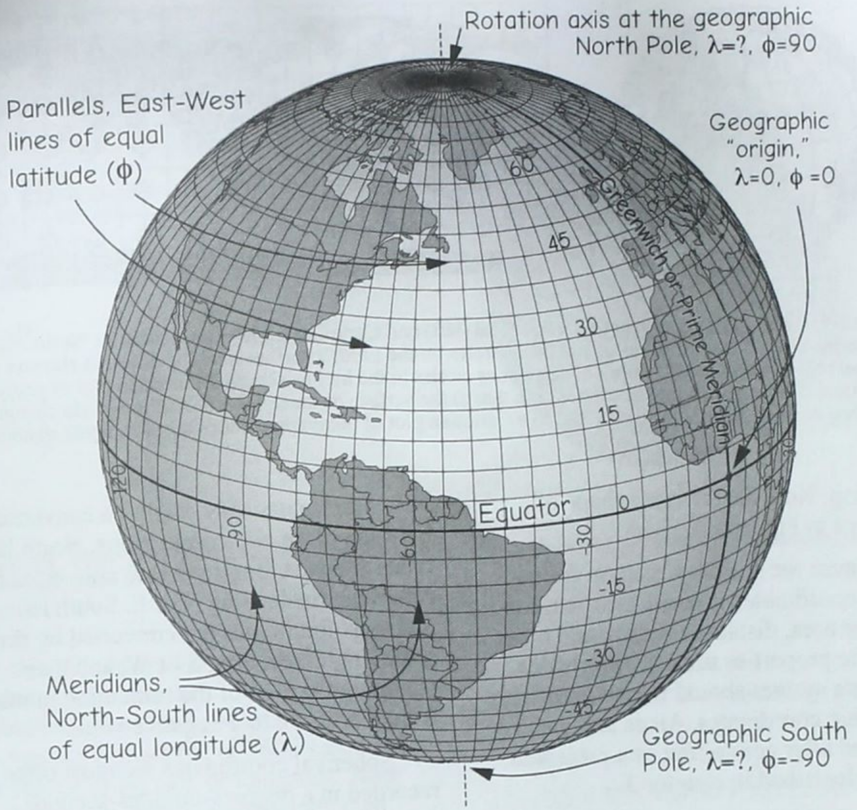


Figure 2-6: Conventions when referring to geographic latitudes and longitudes. Meridians are lines running north-south that have constant longitudes. Parallels are lines running east-west that have constant latitudes. Latitude is zero on the Equator. Longitude is zero on the Greenwich Meridian and undefined at the poles, because all longitudinal meridians intersect there (λ = ? in the figure).

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 longitude varies from approximately 111.3 kilometers at the Equator, to 0 kilometers at the poles. In contrast, the ground distance for 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^{o}, 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-

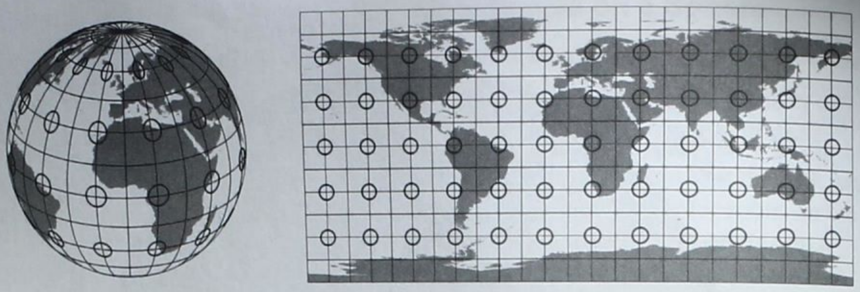


Figure 2-7: Geographic coordinates on a spherical (left) and Cartesian (right) representation. Notice that "circles" defined by a 5 degree radius do not form circles on the Earth's surface near the poles, as shown on the spherical representation (left figure), but appear as circles in the highly distorted Cartesian plot of geographic coordinates (right). This figure illustrates both that a) the surface distance for a unit of longitude changes depending on your location on Earth, and b) a Cartesian plot of geographic coordinates is highly distorted.

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-

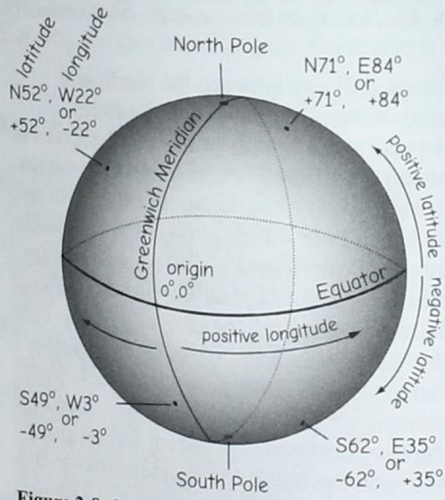


Figure 2-8: Spherical coordinates of latitude and longitude are most often expressed as directional (N/S, E/W), or as signed numbers. Latitudes are positive north, negative south; longitudes are positive east, negative west.

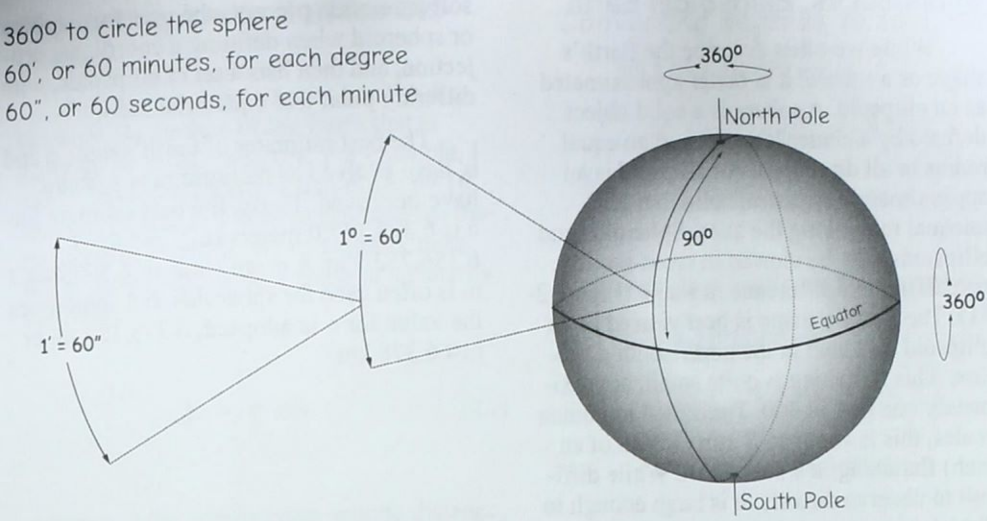

360° to circle the sphere
60', or 60 minutes, for each degree
60", or 60 seconds, for each minute
1° = 60'
1' = 60"
360°
North Pole
90°
Equator
360°
South Pole

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.

tude. In DMS, each degree is made up of 60 minutes of arc, and each minute is in turn divided into 60 seconds of arc (Figure 2-9). This yields 60 times 60, or 3600 seconds for each degree of latitude or longitude. Note that the ancient Babylonians established these splits almost 4,000 years ago, defining 360 degrees for a complete circle, and we've carried this convention down to today.

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.

DD from DMS

$DD = D + M/60 + S/3600$

e.g.

$DMS = 32^\circ\ 45'\ 28''$

$DD = 32 + 45/60 + 28/3600$

$= 32 + 0.75 + 0.0077778$

$= 32.7577778$

DMS from DD

D = integer part

M = integer of decimal part × 60

S = 2nd decimal × 60

e.g.

$DD = 24.93547$

$D = 24$

M = integer of first decimal × 60

$= 0.93547 \times 60$

= integer of 56$\underline{.1282}$

$= 56$

S = 2nd decimal × 60

$= 0.1282 * 60 = 7.692$

so DMS is

$24^\circ\ 56'\ 7.692''$

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

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.

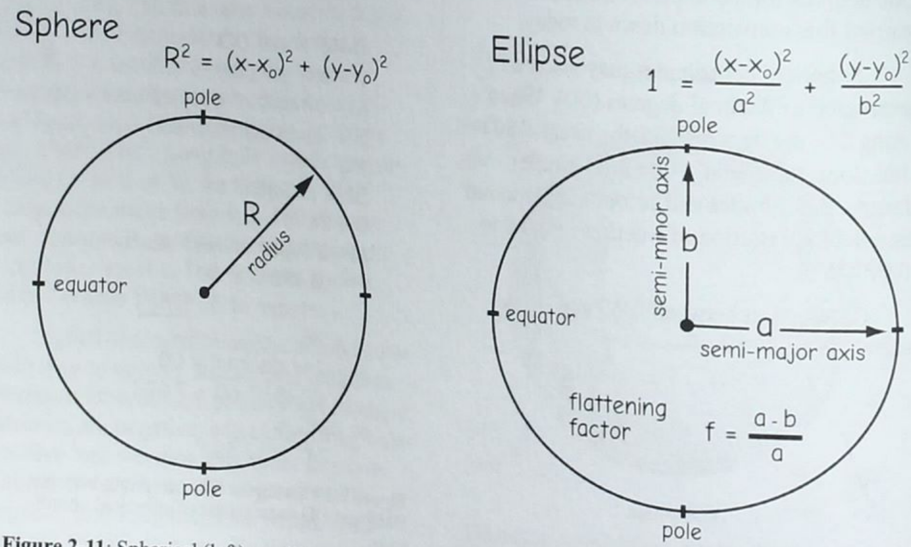


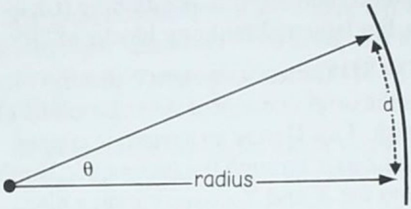
Figure 2-11: Spherical (left) vs. ellipsoidal (right) approximations of the Earth's shape. A sphere has a single radius, while an ellipse has different radii along the semi-major and semi-minor axes. The spheroid and ellipsoid can be thought of as rotating these two basic shapes around the polar axis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

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:

$$d = r \cdot \theta \qquad (2.1)$$

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



d = radius • θ

where θ is measured in radians, with
1 radian = 57.2957°

Given an Earth radius of 6,378,137 m, how much distance is spanned by 10" of arc?

Arc= 10"/3600"/1° = 0.00277778°

=0.00277778°/57.2957 degrees per radian
= 0.000048481435 radians

d = 6378137m • 0.000048481435
= 309.2 meters

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

Converting degrees to radians:

30.1487 degrees is

30.1487 / 57.2957795

= 0.52619 radians

Converting radians to degrees:

1.284 radians is

1.284 x 57.2957795

= 73.5678 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 longitudal convergence. The formula is best used as a first approximation of distance spanning generic arcs, and not using longitudinal coordinates.

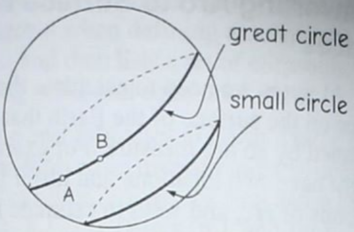
Note that the angle should be specified in radian measure, defined as 2π radians per the 360 degrees, or approximately 57.2957795 degrees per radian. Radian measures are an alternative to degrees, and scale the rotation by the radius of the circle. You may easily convert between radian and degree units (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 (ϕ_A, λ_A), and
B with latitude, longitude of (ϕ_B, λ_B)



The great circle distance between points on a sphere is given by the formula:

$$d = r \cdot 2\sqrt{\sin^{-1}[(\sin^2(\tfrac{\Delta\phi}{2})) + \cos(\phi_A) \cdot \cos(\phi_B) \cdot \sin^2(\tfrac{\Delta\lambda}{2})]}$$

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 $\frac{\Delta\phi}{2}$, $\frac{\Delta\lambda}{2}$ 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°, 2.349014°
Latitude, longitude of Seattle, USA = 47.655548°, -122.30320°

$$d = 6378 \cdot 2\sqrt{\sin^{-1}[(\sin^2(0.604584)) + \cos(48.864716) \cdot \cos(47.655548) \cdot \sin^2(62.36107)]}$$
$$= 8{,}034.8391\text{km}$$

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 it's 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

3-D Cartesian Coordinate System

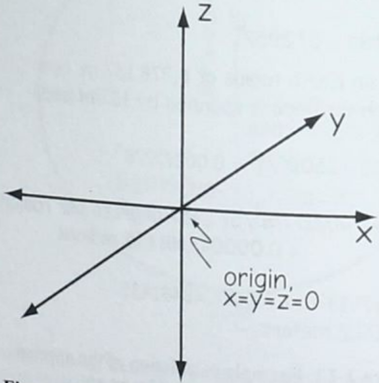
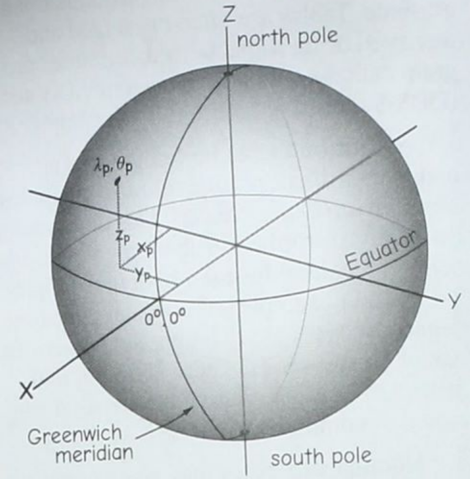


Figure 2-15: A 3-D Cartesian coordinate system.



longitude, latitude from known 3-D Cartesian

$\lambda_p, \theta_p = F(x_p, y_p, z_p)$

3-D Cartesian from known latitude, longitude

$x_p, y_p, z_p = G(\lambda_p, \theta_p)$

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).

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

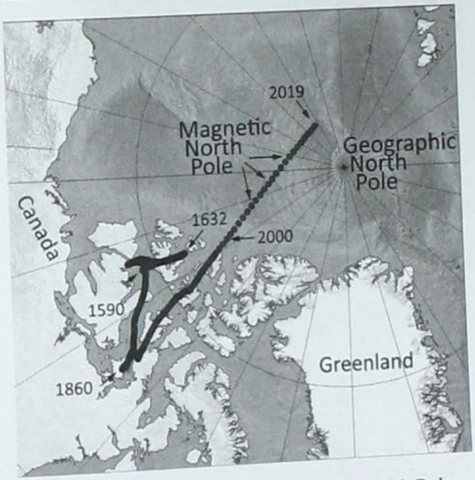


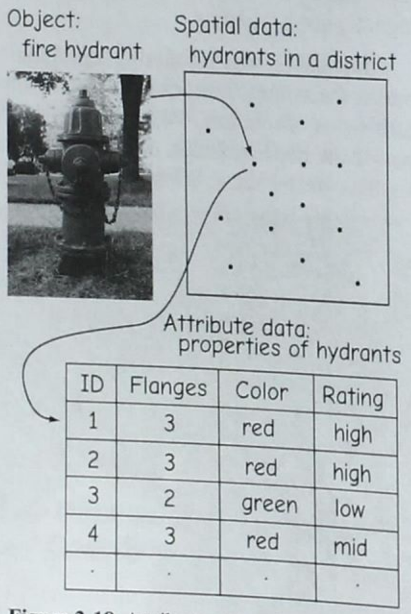
Figure 2-17: Magnetic and geographic North Poles. Year dates show how the Magnetic North has wandered through time, increasing in velocity over the past few decades.

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).



ID	Flanges	Color	Rating
1	3	red	high
2	3	red	high
3	2	green	low
4	3	red	mid
.	.	.	.

Figure 2-18: Attributes are typically envisioned in a table, with objects arranged in rows and attributes aligned in columns.

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.