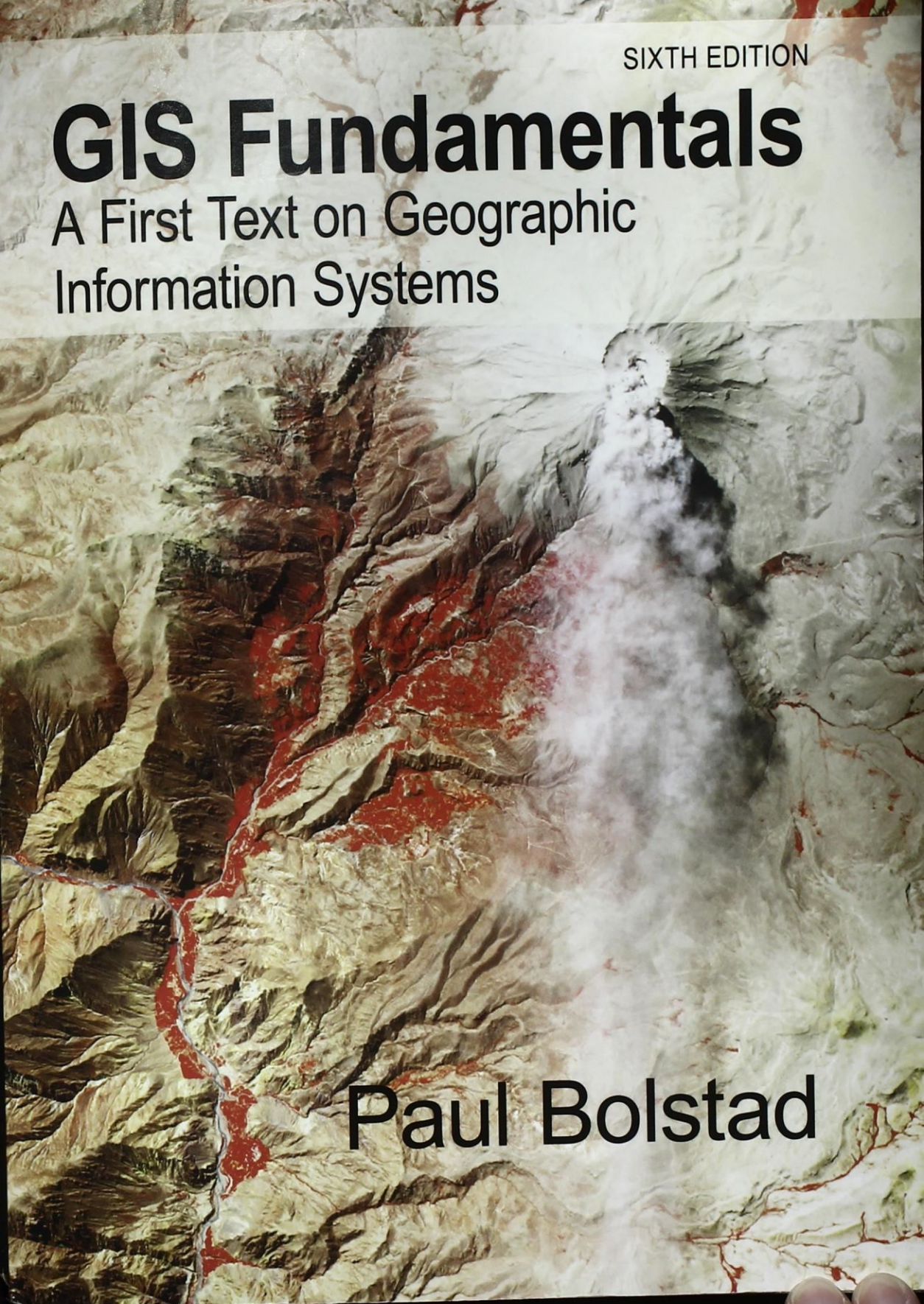
SIXTH EDITION

GIS Fundamentals

A First Text on Geographic Information Systems



Paul Bolstad

3 Geodesy, Datums, Map Projections, and Coordinate Systems

Introduction

Geographic information systems are different from other information systems because they include coordinates that define the location, shape, and extent of geographic objects. For effective GIS use, we must clearly understand how coordinate systems are established for the Earth, how coordinates are measured on the Earth's curving surface, and how these coordinates are converted for use in flat maps, either digital or paper. This chapter introduces *geodesy*, the science of measuring the shape of the Earth, and *map projections*, the transformation of coordinate locations from the Earth's curved surface onto flat maps.

Defining coordinates for the Earth's surface is complicated by four main factors. First, most people view geography on a flat surface. We perceive a flat Earth because the curvature is barely perceptible at human scales. We've used flat maps for more than 40 centuries, and although globes are helpful for visualization at extremely small scales, they are impractical for most purposes.

A flat map must distort geometry in some way because the Earth is curved. When we plot latitude and longitude coordinates on a Cartesian system, "straight" lines will appear bent, and polygons will be distorted. This distortion may be difficult to detect on detailed maps that cover a small area, but the distortions become apparent as the mapped area grows. Because measurements on maps are affected by the distortion, we must use a map projection to reconcile the portrayal of the Earth's curved surface onto a flat surface.

The second main problem in defining a coordinate system results from the irregular shape of the Earth. We learn early on that the Earth is shaped as a sphere. This is a valid approximation for many uses, however, it is only an approximation. Past and present natural forces yield an irregularly shaped Earth. This shape affects how we best map the surface of the Earth, and how we define flat coordinate systems.

Third, our measurements are rarely perfect, and this applies when measuring both the shape of the Earth and the exact position of features on it. All locations depend on measurements that contain some error, and on analyses that require assumptions. Our measurements improve through time, and so does the sophistication of our analysis, so our positional estimates improve; this evolution means our estimates of positions change through time.

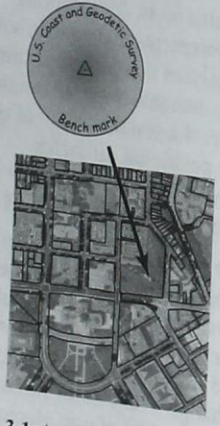
Finally, the physical locations of points on the Earth change through time. Plate tectonics and vertical crustal movements mean the distance from San Francisco to Tokyo changed from 1950 to 2010, and continues to change today. Earth surface rebound from the weight of past glaciers yields elevations in central Canada several centimeters higher than they were a few

decades ago. How do we specify positions through time when the locations aren't truly fixed?

Because of these four factors, we often have several different sets of coordinates to define the same location on the surface of the Earth. Remember, coordinates are sets of numbers that unambiguously define locations, and in a GIS data layer, we usually use an X (easting), Y (northing) and sometimes height value. But each of these values are only "unique" to any given point for a specified set of measurements, calculation assumptions, at a specified time. The coordinates depend on the reference system we use for measuring latitudes and longitudes (which depends on measurement and Earth's shape), how we translate points from a curved Earth to a flat map surface (which depends on how we project), and to what set of measurements we reference our coordinates (our measurement methods and quality), and when (which depends on crustal movement). We may, and often do, address these factors in a number of different ways, and the coordinates for the same point will be different for these different choices. We can translate between these different choices, as long as we are clear in defining them.

An example may help. Figure 3-1 shows the location of a U.S. survey mark, a precisely surveyed and monumented point. Coordinates for this point are maintained by federal and state government surveyors, and resulting coordinates are shown at the top right of the figure. Note that there are three different versions of the latitude/longitude location for this point. Here, the three versions differ primarily due to differences in the measurements, and how measurement errors were adjusted (the third factor, discussed above). The GIS practitioner may well ask, which latitude/longitude pair should I use? This chapter contains the information that should allow you to choose wisely.

Note that there are also several versions of the x and y coordinates for the point in Figure 3-1. The differences in the coordinate values are too great to be due solely to measurement errors. The differences are due primarily to how we choose to project from the curved Earth to a flat map, and in part to the Earth shape we adopt and the measurement system we use.



From Surveyor Data:

	Latitude (N)	Longitude (W)	
NAD83(2007)	44 57 23.23074	093 05 58.28007	
NAD83(1986)	44 57 23.22405	093 05 58.27471	
NAD83(1996)	44 57 23.23047	093 05 58.27944	
	X	Y	
SPC MNS	317,778.887	871,048.844	MT
SPC MNS	1,042,579.57	2,857,766.08	sFT
UTM15	4,978,117.714	492,150.186	MT

From Data Layers:

	X	Y	
MN-Ramsey	573,475.592	160,414.122	sFT
MN-Ramsey	174,195.315	48,893.966	MT
SPC MNC	890,795.838	95,819.779	MT
SPC MNC	2,922,552.206	314,365.207	sFT
LCC	542,153.586	18,266.334	MT

Figure 3-1: An example of different coordinate values for the same point. We may look up the coordinates for a well-surveyed point, and we may also obtain the coordinates for the same point from a number of different data layers. We often find multiple latitude/longitude values (surveyor data, [illegible] x and y values for the same point (surveyor data, or from data layers, bottom).

Whenever we work with spatial data, we must choose how to address the first three factors: projection distortion, an irregularly shaped Earth, and measurement imprecision. If our data are of very high accuracy and precision and we wish to work across time periods, we must address the fourth factor: vertical and horizontal movements of physical locations through time.

It is crucial to realize that different ways of addressing 1) the Earth's curvature, 2) the Earth's deviation from our idealized shape, 3) inevitable inaccuracies in measurement, and 4) physical shifts, will result in different coordinates. These differences are the root of many errors in spatial analysis. As a rule, you should know the coordinate system used for all of your data, and convert all data to the same coordinate system, for the same time epoch, prior to analysis. In some cases the differences when ignoring some of these four factors may be small in relation to the spatial precision required by your analysis, particularly for the fourth factor (time differences between coordinate measurements). As positioning technology improves, we can make increasingly accurate and precise measurements, so in many cases, the epoch of measurement becomes important. This chapter describes how we define, measure, and convert among coordinate systems.

Modern Coordinate Capture, Coordinate Systems, and Datums

Most GIS data collection relies directly or indirectly on satellite-based positioning systems. These systems, described in detail in Chapter 5, allow the rapid, accurate collection of locations. Positions are referenced to Earth-centered, three dimensional, Cartesian coordinate systems – the X, Y, and Z of 3D systems described in Chapter 2. A specific, defined version of a 3D system is called a *datum*. Datums underpin all geographic measurements. The navigation system operated by the U.S. (GPS) provides coordinates in a datum labeled as WGS84(xxxx), where xxxx represents a version number. In most of North America, collected data are often converted to a different datum, labeled as NAD83(yy) system, where yy is a version number. The other satellite positioning systems (GLONASS, BeiDou, Galileo) typically report in a datum labeled ITRF(zzzz), where zzzz is a version number, usually the year of issue. We will describe WGS84, NAD83, and ITRF datums and how they relate to each other in the first half of this chapter.

These various versions of X, Y, and Z Cartesian coordinates are then commonly converted to latitude, longitude, and height coordinates, and subsequently projected to "flat" coordinate values, suitable for layers in a GIS. This process applies for data directly collected with a GPS or other similar satellite-based navigation system, or with data that depend on satellite positioning, such as satellite or aerial images. Since these coordinate systems differ, have changed through time, and data are commonly converted one to another, it is easy to add error to new data so that features don't fall in their true location. Knowledge of the history and technology of datum development helps us understand how to best collect new data, and to integrate older data with newer measurements.

Early Measurements

In specifying a coordinate system, we must first define the size and shape of the Earth. Humans have long speculated on this. Babylonians believed the Earth was a flat disk floating in an endless ocean, while the Greek Pythagoras, and later Aristotle, reasoned that the Earth must be a sphere. They observed that ships disappeared over the horizon, the moon appeared to be a sphere, and that the stars moved in circular patterns, all consistent with a spherical Earth.

The Greeks next turned toward estimating the size of the sphere. They measured locations on the Earth's surface relative to the Sun or stars, reasoning these provided a stable reference frame. This assumption underlies most geodetic observations taken over the past 2,000 years.

decades ago. How do we specify positions through time when the locations aren't truly fixed?

Because of these four factors, we often have several different sets of coordinates to define the same location on the surface of the Earth. Remember, coordinates are sets of numbers that unambiguously define locations, and in a GIS data layer, we usually use an X (easting), Y (northing) and sometimes height value. But each of these values are only "unique" to any given point for a specified set of measurements, calculation assumptions, at a specified time. The coordinates depend on the reference system we use for measuring latitudes and longitudes (which depends on measurement and Earth's shape), how we translate points from a curved Earth to a flat map surface (which depends on how we project), and to what set of measurements we reference our coordinates (our measurement methods and quality), and when (which depends on crustal movement). We may, and often do, address these factors in a number of different ways, and the coordinates for the same point will be different for these different choices. We can translate between these different choices, as long as we are clear in defining them.

An example may help. Figure 3-1 shows the location of a U.S. survey mark, a precisely surveyed and monumented point. Coordinates for this point are maintained by federal and state government surveyors, and resulting coordinates are shown at the top right of the figure. Note that there are three different versions of the latitude/longitude location for this point. Here, the three versions differ primarily due to differences in the measurements, and how measurement errors were adjusted (the third factor, discussed above). The GIS practitioner may well ask, which latitude/longitude pair should I use? This chapter contains the information that should allow you to choose wisely.

Note that there are also several versions of the x and y coordinates for the point in Figure 3-1. The differences in the coordinate values are too great to be due solely to measurement errors. The differences are due primarily to how we choose to project from the curved Earth to a flat map, and in part to the Earth shape we adopt and the measurement system we use.

Coordinates for a Point Location

From Surveyor Data:

	Latitude (N)	Longitude (W)
NAD83(2007)	44 57 23.23074	093 05 58.28007
NAD83(1986)	44 57 23.22405	093 05 58.27471
NAD83(1996)	44 57 23.23047	093 05 58.27944

	X	Y	
SPC MNS	317,778.887	871,048.844	MT
SPC MNS	1,042,579.57	2,857,766.08	sFT
UTM15	4,978,117.714	492,150.186	MT

From Data Layers:

	X	Y	
MN-Ramsey	573,475.592	160,414.122	sFT
MN-Ramsey	174,195.315	48,893.966	MT
SPC MNC	890,795.838	95,819.779	MT
SPC MNC	2,922,552.206	314,365.207	sFT
LCC	542,153.586	18,266.334	MT

Figure 3-1: An example of different coordinate values for the same point. We may look up the coordinates for a well-surveyed point, and we may also obtain the coordinates for the same point from a number of different data layers. We often find multiple latitude/longitude values (surveyor data, [illegible] or x and y values for the same point (surveyor data, or from data layers, bottom).

Whenever we work with spatial data, we ust choose how to address the first three ctors: projection distortion, an irregularly aped Earth, and measurement imprecision. our data are of very high accuracy and pre- sion and we wish to work across time peri- ls, we must address the fourth factor: rtical and horizontal movements of physi- l locations through time.

It is crucial to realize that different ways addressing 1) the Earth's curvature, 2) the arth's deviation from our idealized shape, inevitable inaccuracies in measurement, d 4) physical shifts, will result in different ordinates. These differences are the root of any errors in spatial analysis. As a rule, ou should know the coordinate system used r all of your data, and convert all data to e same coordinate system, for the same me epoch, prior to analysis. In some cases e differences when ignoring some of these our factors may be small in relation to the atial precision required by your analysis, articularly for the fourth factor (time differ- nces between coordinate measurements). s positioning technology improves, we can ake increasingly accurate and precise mea- urements, so in many cases, the epoch of easurement becomes important. This chap- r describes how we define, measure, and onvert among coordinate systems.

Modern Coordinate Capture, Coordinate Systems, and Datums

Most GIS data collection relies directly r indirectly on satellite-based positioning ystems. These systems, described in detail n Chapter 5, allow the rapid, accurate col- ection of locations. Positions are referenced o Earth-centered, three dimensional, Carte- ian coordinate systems – the X, Y, and Z of D systems described in Chapter 2. A spe- ific, defined version of a 3D system is alled a *datum*. Datums underpin all geo- graphic measurements. The navigation sys- em operated by the U.S. (GPS) provides coordinates in a datum labeled as WGS84(xxxx), where xxxx represents a ver- sion number. In most of North America, col- lected data are often converted to a different datum, labeled as NAD83(yy) system, where yy is a version number. The other satellite positioning systems (GLONASS, BeiDou, Galileo) typically report in a datum labeled ITRF(zzzz), where zzzz is a version number, usually the year of issue. We will describe WGS84, NAD83, and ITRF datums and how they relate to each other in the first half of this chapter.

These various versions of X, Y, and Z Cartesian coordinates are then commonly converted to latitude, longitude, and height coordinates, and subsequently projected to "flat" coordinate values, suitable for layers in a GIS. This process applies for data directly collected with a GPS or other similar satellite-based navigation system, or with data that depend on satellite positioning, such as satellite or aerial images. Since these coordinate systems differ, have changed through time, and data are commonly converted one to another, it is easy to add error to new data so that features don't fall in their true location. Knowledge of the history and technology of datum development helps us understand how to best collect new data, and to integrate older data with newer measurements.

Early Measurements

In specifying a coordinate system, we must first define the size and shape of the Earth. Humans have long speculated on this. Babylonians believed the Earth was a flat disk floating in an endless ocean, while the Greek Pythagoras, and later Aristotle, reasoned that the Earth must be a sphere. They observed that ships disappeared over the horizon, the moon appeared to be a sphere, and that the stars moved in circular patterns, all consistent with a spherical Earth.

The Greeks next turned toward estimating the size of the sphere. They measured locations on the Earth's surface relative to the Sun or stars, reasoning these provided a stable reference frame. This assumption underlies most geodetic observations taken over the past 2,000 years.

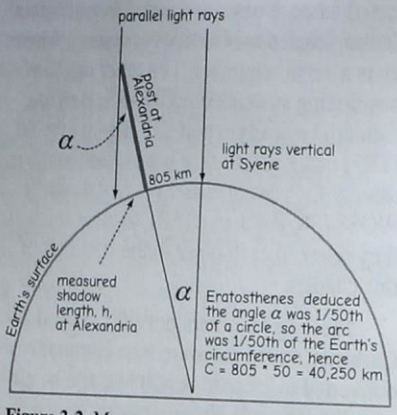


Figure 3-2: Measurements made by Eratosthenes to determine the circumference of the Earth.

Eratosthenes performed early measurements of the Earth's circumference. He noticed that on the summer solstice, the noon sun shone to the bottom of a deep well near the Tropic of Cancer, meaning the sun would be exactly overhead. He also observed that 805 km north, in Alexandria at exactly the same date and time, a vertical post cast a shadow. The shadow/post combination defined an angle that was about 7°12', or about 1/50th of a circle (Figure 3-2).

Eratosthenes deduced that the Earth must be 805 multiplied by 50, or about 40,250 kilometers in circumference. His estimate is within 4% of modern measurements of the Earth's circumference.

Specifying the Ellipsoid

By the 18th century, mathematicians argued that centrifugal forces should cause the equatorial regions of the Earth to bulge. They proposed the Earth would be better modeled by an *ellipsoid*, a sphere slightly flattened at the North and South Poles. Expeditions by the French Royal Academy of Sciences starting in 1730 measured the Earth's shape near the Equator and in the high northern latitudes. Complex, repeated, and highly accurate measurements established that an ellipsoid was the best geometric model of the Earth's surface.

Efforts then focused on precisely measuring the size of the Earth's ellipsoid. As noted in Chapter 2, the ellipsoid has two characteristic dimensions (Figure 3-3): the *semi-major axis*, the radius a in the equatorial direction, and the *semi-minor axis*, the radius b in the polar direction. This difference in polar and equatorial radii is also described as a flattening factor, shown in Figure 3-3.

Celestial observations of the stars (Figure 3-4) are combined with long-distance surface measurements to estimate polar and equatorial radii (Figure 3-5). Measurements are repeated over many different locations, and combined for estimates of the semi-major and semi-minor axes. Because early continental surveys could not span most oceans, ellipsoidal parameters were fit for each country, continent, or comparably large survey area.

Measurement efforts through the 19th and 20th centuries led to a set of official ellipsoids which differed in equatorial and polar radii. The Clarke 1866 ellipsoid was commonly used in North America, and was more flattened than the ellipsoid we use today. The Bessel ellipsoid, common in Europe, also specified radii somewhat different than today's best global estimates. Optical instruments predominated before the

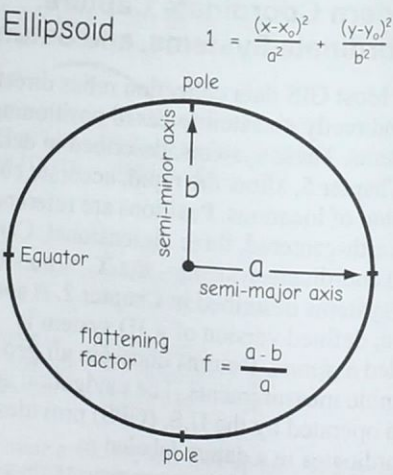


Figure 3-3: An ellipsoidal model of the Earth's shape.

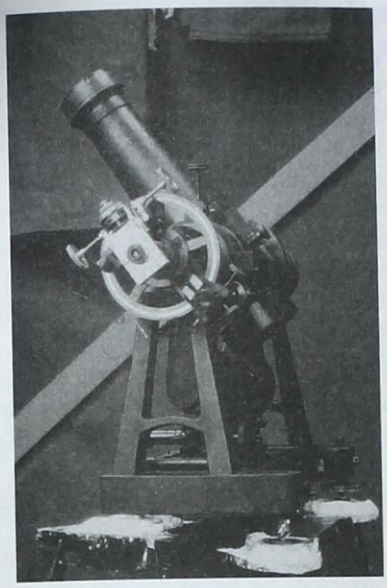


Figure 3-4: An instrument used in the early 1900s for measuring the position of celestial bodies.

early 20th century, and sighting distances were limited by the Earth's curvature. Individual survey legs greater than 50 kilometers (30 miles) were rare, with no good ways to connect surveys across oceans.

Since the 1980s, data derived from satellites, lasers, and broadcast timing signals have been used for extremely precise measurements of relative positions across continents and oceans. Ellipsoids such as the GRS80 provide a "best" overall fit to observed measurements across the globe, and are now preferred and most widely used.

Surface and Ellipsoidal Coordinates

While we make most of our measurements at or near the surface of the Earth, we specify latitudes and longitudes on the ellipsoid, which is usually below the physical surface of the Earth (Figure 3-6). All of our horizontal measurements must be "reduced to," or specified, on the ellipsoid surface. They are mathematically transferred down-

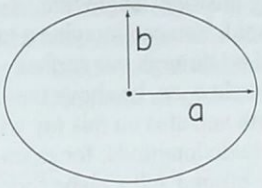
An ellipsoid is defined in part by two radii, a and b

We may use the relationship $d = r \cdot \theta$ to estimate radii:

$$a = \frac{d_1}{\theta_1}$$

$$b = \frac{d_2}{\theta_2}$$

Generally, the measurements are not at the poles and Equator, and the math is more complicated, but the principle is the same.



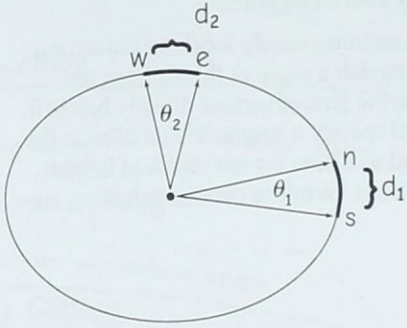


Figure 3-5: Two arcs illustrate the surface measurements and calculations used to estimate the semi-major and semi-minor axes, here for North America. The arc lengths may be measured by surface surveys, and the angles from astronomical observations, as illustrated in Figure 3-2 and Figure 3-3.

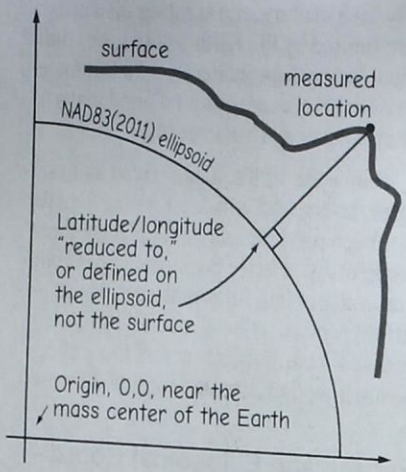


Figure 3-6: Surface measurements are "reduced" downwards onto a chosen ellipsoid directly below the measured location.

ward or upward along a line that is at right angles to the surface of the ellipsoid.

A latitude/longitude location on the ellipsoid is also the latitude/longitude for the surface of the Earth, which may be above or below the ellipsoid. We apply the same latitude and longitude to all points along this line that fall along this right-angle line through the ellipsoid. You can imagine a ray from the ellipsoid up through our surface point. All objects below, on, or above the surface of the Earth and also on this ray will have the same latitude/longitude, for example, a point on the ground and a plane flying directly above that point.

To unambiguously locate an object, e.g., to distinguish a plane in the air from the point on the ground surface directly below it, we must specify a height. We do not use the ellipsoid as a base for our standard heights, and so must introduce another reference surface.

The Geoid

We noted in the introduction that the true shape of the Earth differs slightly from an ellipsoid. Differences in the density of the Earth cause variation in gravitational strength, in turn causing regions to dip below or bulge above a reference ellipsoid (Figure 3-7). This undulating shape is called a *geoid*. In much of the world, including North America, we use a geoid as our zero height.

We define the geoid as the three-dimensional *equipotential surface*, along which the pull of gravity is a specified constant. The geoidal surface may be thought of as an imaginary sea that covers the entire Earth and is not affected by wind, waves, the Moon, or forces other than Earth's gravity. The surface of the geoid extends across the Earth, approximately at mean sea level across the oceans, and continuing under continents at a level set by gravity. The surface is always at right angles to the direction of local gravity.

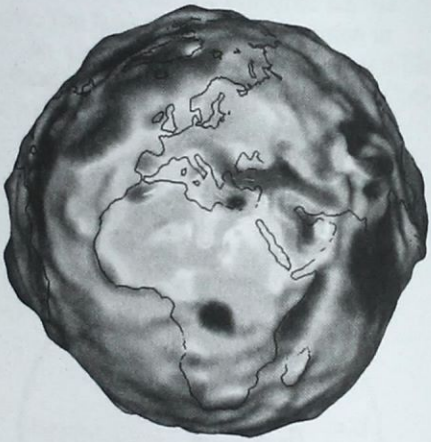


Figure 3-7: Depictions of the Earth's gravity field, as estimated from satellite measurements. These show the undulations, greatly exaggerated, in the Earth's gravity, and hence the geoid (courtesy University of Texas Center for Space Research, and NASA).

We must emphasize that a geoidal surface differs from mean sea level. Mean sea level may be higher or lower than a geoidal surface because ocean currents, temperature, salinity, and wind variations can cause persistent high or low areas in the ocean. These non-gravitational differences can be up to a meter (3 feet), perhaps small on global scale, but large in local or regional analysis. We historically referenced heights to mean sea level, and many believe we still do, but this is no longer true for most spatial data systems.

Because we have two reference surfaces, a geiod and an ellipsoid, we also have two bases from which to measure height. Elevation is typically defined as the distance above a geoid. This elevation above a geoid is also called the *orthometric height* (Figure 3-8), and may be thought of as replacing our older notion of height above mean sea level. Heights above an ellipsoid, or *ellipsoidal heights,* are used in some coordinate system calculations and for some global navigation systems such as GPS, but ellipsoidal heights are not our standard height. These are illustrated in Figure 3-8, with the ellipsoidal height labeled h and orthometric height labeled H. The difference between the ellipsoidal height and orthometric height at any location, shown in Figure 3-8 as N, has various names, including *geoidal height* and *geoidal separation.*

The absolute value of the geoidal height is less than 100 meters over most of the Earth (Figure 3-9). Although it may at first seem difficult to believe, the "average" ocean surface near Iceland is more than 150 meters "higher" than the ocean surface northeast of Jamaica. This height difference is measured relative to the ellipsoid. Since gravity pulls in a direction that is perpendicular to the geoidal surface, the force is at a right angle to the surface of the ocean, resulting in persistent bulges and dips in the mean ocean height. Variation in ocean heights due to swells and wind-driven waves are more apparent at local scales, but are much smaller than the long-distance geoidal undulations.

The geoidal height is quite small relative to the polar and equatorial radii. The Earth's equatorial radius is about 6,780,000 meters, or about 32,000 times the range of the highest to lowest geoidal heights. This small geoidal height is imperceptible at human scales. While relatively small, the geoidal variations in shape must still be considered for accurate vertical and horizontal mapping over continental or global distances.

ellipsoidal height = orthometric height + geoidal height

$h = H + N$

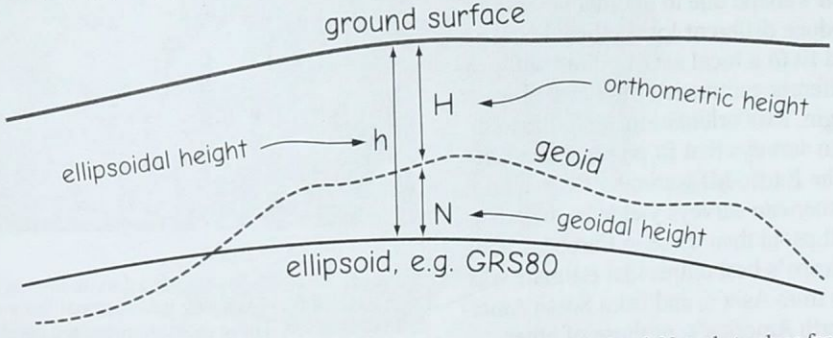


Figure 3-8: Ellipsoidal, orthometric, and geoidal height are interrelated. Note that values for N are highly exaggerated in this figure – values for N are typically much less than H. We often use this formula, e.g., to calculate orthometric height (elevation) when we know the ellipsoidal height (commonly from GPS), and geoidal height (from national models).

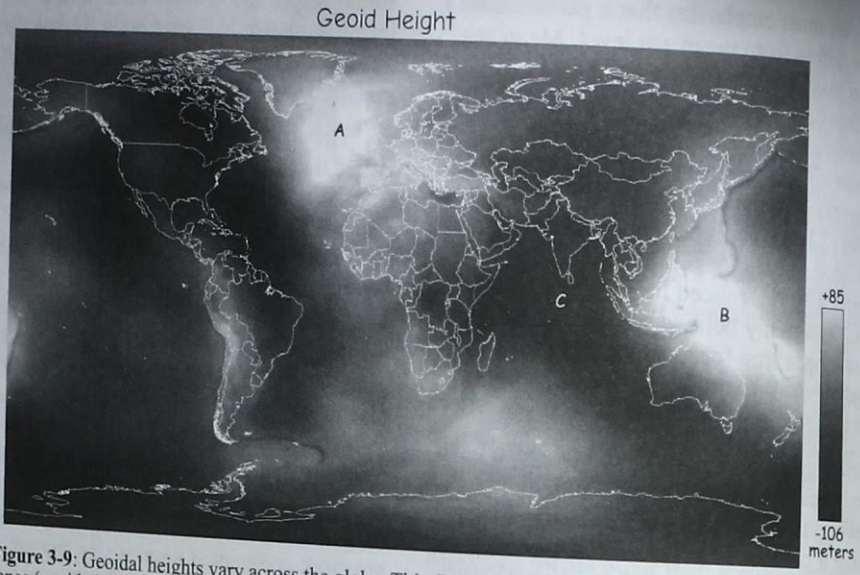


Figure 3-9: Geoidal heights vary across the globe. This figure depicts positive geoidal heights in lighter tones (geoid above the ellipsoid) and negative geoidal heights in darker tones. Note that geoidal heights are positive for large areas near Iceland and the Philippines (*A* and B, respectively), while large negative values are found south of India (*C*). Continental and country borders are shown in white.

The geoid is a measured and interpolated surface; unlike an ellipsoid, the geoidal surface is not defined by a simple mathematical equation. The geoid's surface is measured by a number of methods, initially by a combination of *plumb bob*, a weight suspended by a string that indicates the direction of gravity, and horizontal and vertical distance measurements, and later with various types of *gravimeters* (Figure 3-10), devices that measure the gravitational force.

Figure 3-11 shows how differences in the Earth's shape due to geoidal deviations will produce different local ellipsoids. An ellipsoid fit to a local set of points will produce different estimates of the best ellipsoid origin, axis orientation, and ellipsoid radii than surveys that fit points on another part of the Earth. Measurements based on South American surveys yielded a different "best" ellipsoid than those in Europe. Likewise, Europe's best ellipsoidal estimate was different from Asia's, and from South America's, North America's, or those of other regions. One ellipsoid could not be fit to all the world's survey data because during the

Figure 3-10: A portable field gravimeter, an instrument used for measuring gravitational force at a field location. These measurements are combined with surveying measurements to estimate geoidal surfaces (courtesy National Oceanic and Atmospheric Administration, NOAA).

18th and 19th centuries, there was no clear way to combine a global set of measurements.

Satellite-based measurements in the late 20th century substantially improved the global coverage, quality, and density of geoidal height measurements, aiding the development of globally-accurate geoids and ellipsoids. The GRACE experiment, initiated with the launch of twin satellites in 2002, is an example of such improvements. Distances between a pair of satellites are constantly measured as they orbit the Earth. The satellites are pulled closer or drift farther from the Earth due to variation in the gravity field. Because the orbital path changes slightly each day, we eventually have nearly complete Earth coverage of the strength of gravity, and hence the location of the reference gravitational surface. The ESA GOCE satellite, launched in 2009, uses precision accelerometers to measure gravity-induced velocity change. GRACE and GOCE observations have substantially improved our estimates of the gravitational field and geoidal shape.

Satellite and other observations are used by geodesists to develop geoidal models. These support a series of geoid estimates, for example, by the U.S. NGS with GEOID90 in 1990, with succeeding geoid estimates in 1993, 1996, 1999, 2003, 2009, and 2012. These are called models because we measure geoidal heights at points or along lines at various parts of the globe, but we need geoidal heights everywhere. Equations are statistically fit that relate the measured geoidal heights to geographic coordinates. Given any set of geographic coordinates, we may then estimate the geoidal height. These models provide an accurate estimation of the geoidal heights for the entire globe.

Horizontal Datums

The geographic coordinate system described in Chapter 2 is based on an established zero meridian passing near the Greenwich Observatory, in England. However, this

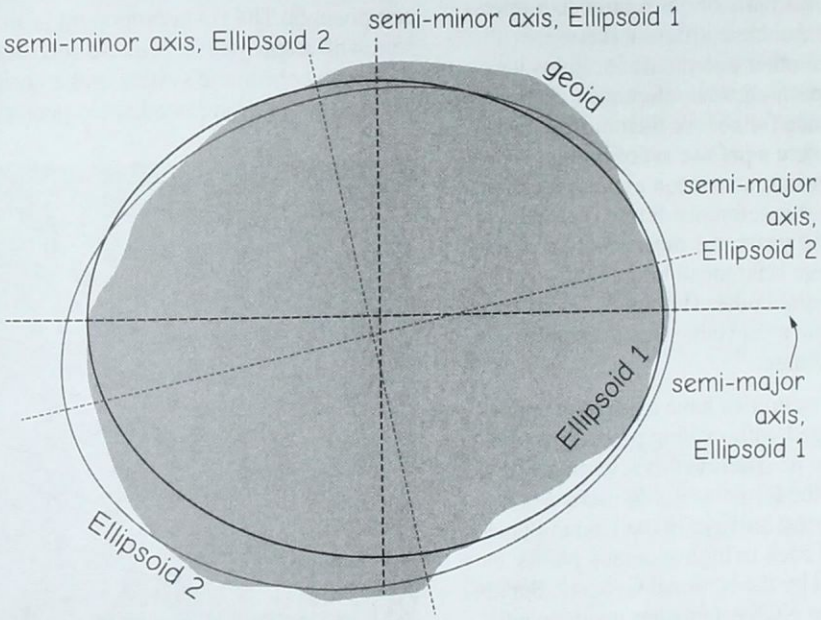


Figure 3-11: Different ellipsoids were estimated due to local irregularities in the Earth's shape. Local best-fit ellipsoids varied from the global best fit, but until the 1970s, there were few good ways to combine global geodetic measurements.

gives us the exact longitude of only one arc, the zero line of longitude. We must estimate the longitudes and latitudes of all other locations through surveying measurements, until recently by observing stars and by measuring distances and directions between points. These surveying methods have since been replaced by modern, satellite-based positioning, but even these new methods are ultimately dependent on astronomical observations. Through these methods, we establish a set of points on Earth for which the horizontal and vertical positions have been accurately determined. These accurately determined points and associated measured and mathematical surfaces are *datums*, references against which we measure all other locations.

These well-surveyed points allow us to specify a *reference frame*, including an origin or starting point. If we are using an ellipsoidal reference frame, we must also specify the orientation and radii of our ellipsoid. If we are using a three-dimensional Cartesian reference frame, we must specify the X, Y, and Z axes, including their origin and orientation. We can choose different values for these various parts of our reference frame, and hence can have different reference frames. All other coordinate locations we use are measured with reference to the chosen reference frame. We then must painstakingly measure a precise set of highly accurate points, so we can express locations relative to this reference frame. For most of the past 150 years, the most accurate observations were referenced to the Sun, stars, or other celestial bodies (Figure 3-12), as they provided the most stable way to establish our reference frame.

Many countries have a government body charged with making precise surveys of points to help define this reference frame, and make the frame useful to users. For example, most surveys in the United States are related back to high accuracy points maintained by the National Geodetic Survey (NGS). The NGS establishes geodetic latitudes and longitudes of known points, most

Figure 3-12: Astronomical observations were used in early geodetic surveys to measure datum locations (courtesy NOAA).

of which are monumented with a metal disk, concrete posts, or other durable markers.

A *geodetic datum* is a reference surface. A geodetic datum consists of two major components. The first component is an ellipsoid with a spherical or three-dimensional Cartesian coordinate system and an origin. Eight parameters are needed to specify the

Figure 3-13: A bronze disk used to monument a survey mark.

ellipsoid: a and b to define the size/shape of the ellipsoid; the X, Y, and Z values of the origin; and an orientation angle for each of the three axes.

A datum includes a set of positions that have been painstakingly surveyed, against which subsequent surveys are referenced. A datum is sometimes defined as a reference surface, and a *realization of a datum* as that surface plus a network of precisely measured points. The measured points describe a *Terrestrial Reference Frame,* or specific measured datum. This clearly separates the theoretical reference surface from a useful terrestrial reference frame, complete with points from which we can survey new points or re-measure old. While this more precise language may avoid some confusion, datum commonly refers to both the defined surface and the various realizations of each datum.

Precisely surveyed points are commonly known as *survey marks* and *bench marks*, with the latter often reserved for precise vertical surveys. Marks often consist of a metal disk embedded in rock or concrete (Figure 3-13), although they also may consist of marks chiseled in rocks, embedded iron posts, or other long-term marks. Due to the considerable effort and cost of establishing the coordinates for each survey mark, they are often redundantly monumented, and their distance and direction from specific local features are recorded. Control survey points are often identified with a number of nearby signs to aid in recovery (Figure 3-14).

Figure 3-14: Signs are often placed near control points to warn of their presence and aid in their location.

The NGS maintains and disseminates information on survey marks in the United States (Figure 3-15), with access via the World Wide Web (http://www.ngs.noaa.gov). Stations may be found based on a station name, a state and county name, a type of station (horizontal or vertical), by survey order, survey accuracy, date, or coordinate location. These stations may be used as reference points to check the

```
National Geodetic Survey,   Retrieval Date = SEPTEMBER 26, 2011
OB0554  DESIGNATION -  CAPE SMALL  OB0554  PID          -  OB0554
OB0554  STATE/COUNTY-  ME/SAGADAHOC   USGS QUAD   -  PHIPPSBURG (1957)
OB0554
OB0554                          *CURRENT SURVEY CONTROL
OB0554
OB0554* NAD 83(1996)     -  43 46 42.87649(N) 069 50 42.26065(W)     ADJUSTED
OB0554* NAVD 88          -         73.      (meters)  240.     (feet)  SCALED
OB0554
OB0554  LAPLACE CORR-              2.33  (seconds)                     DEFLEC99
OB0554  GEOID HEIGHT-            -25.73  (meters)                      GEOID03
OB0554  HORZ ORDER  -  FIRST
```

Figure 3-15: A portion of a National Geodetic Survey control point data sheet.

accuracy of any data collection method, for example, new GPS/GNSS equipment, or as a starting point for additional surveys.

Different datums are specified through time because our realizations, or estimates of the datum, change through time. New points are added and survey methods improve. We periodically update our datum when there are enough new or better measurements of survey points, or when we change the parameters of the reference frame (e.g., origin, ellipsoid shape). We do this by reestimating the coordinates of our datum points after including these changes, thereby improving our estimate of the position of each point.

There are two main eras of datums, those created before satellites geodesy, and those after. Satellite positioning technologies became commonplace in the last decade of the 20th century, and substantially increased the number and accuracy of datum points. Datums and coordinates found today are a mix of those developed under pre-satellite datums, and those referenced to post-satellite datums, so the GIS user should be familiar with both.

Geodetic surveys in the 18th and 19th centuries combined horizontal measurements with repeated, excruciatingly precise astronomical observations. Astronomical observations were typically used at the starting point, a few intermediate points, and near the end of geodetic surveys. Astronomical positioning required repeated measurements over several nights. Clouds, haze, or a full moon often lengthened the measurement times. In addition, celestial measurements required correction for atmospheric refraction, a process that bends light and changes the apparent position of stars.

Historically, horizontal optical surveys were as precise and much faster than astronomical methods when measuring over distances up to several tens of kilometers. These horizontal surface measurements were used to connect astronomically surveyed points and thereby create an expanded, well-distributed set of known datum points. Figure 3-16 shows an example survey, where open circles signify points established by astronomical measurements and filled circles denote points established by surface measurements.

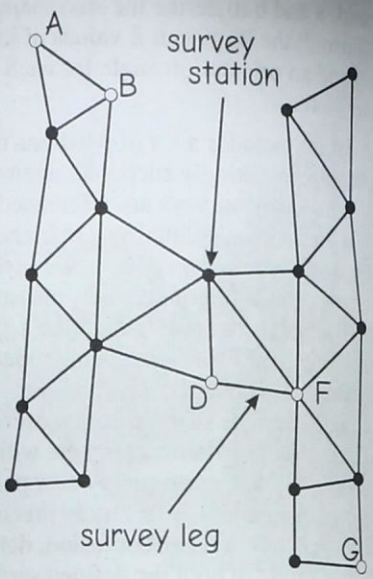


Figure 3-16: A triangulation survey network. Stations may be measured using astronomical (open circles) or surface surveys (filled circles).

Figure 3-16 also illustrates a *triangulation survey*, commonly used prior to satellite positioning. They employ a network of interlocking triangles to determine positions at survey stations. Triangulation surveys were adopted because we can create them through optical angle measurement, with few surface distance measurements, an advantage in the late 18th and early 19th centuries when many datums were first developed. Triangulation also improves accuracy; because there are multiple measurements to each survey station, the location at each station may be computed by various paths.

Triangulation networks spanned long distances, from countries to continents (Figure 3-17). Individual measurements of these triangulation surveys were rarely longer than a few to tens of kilometers; however, each leg of the larger triangles were made up themselves of smaller triangulation traverses.

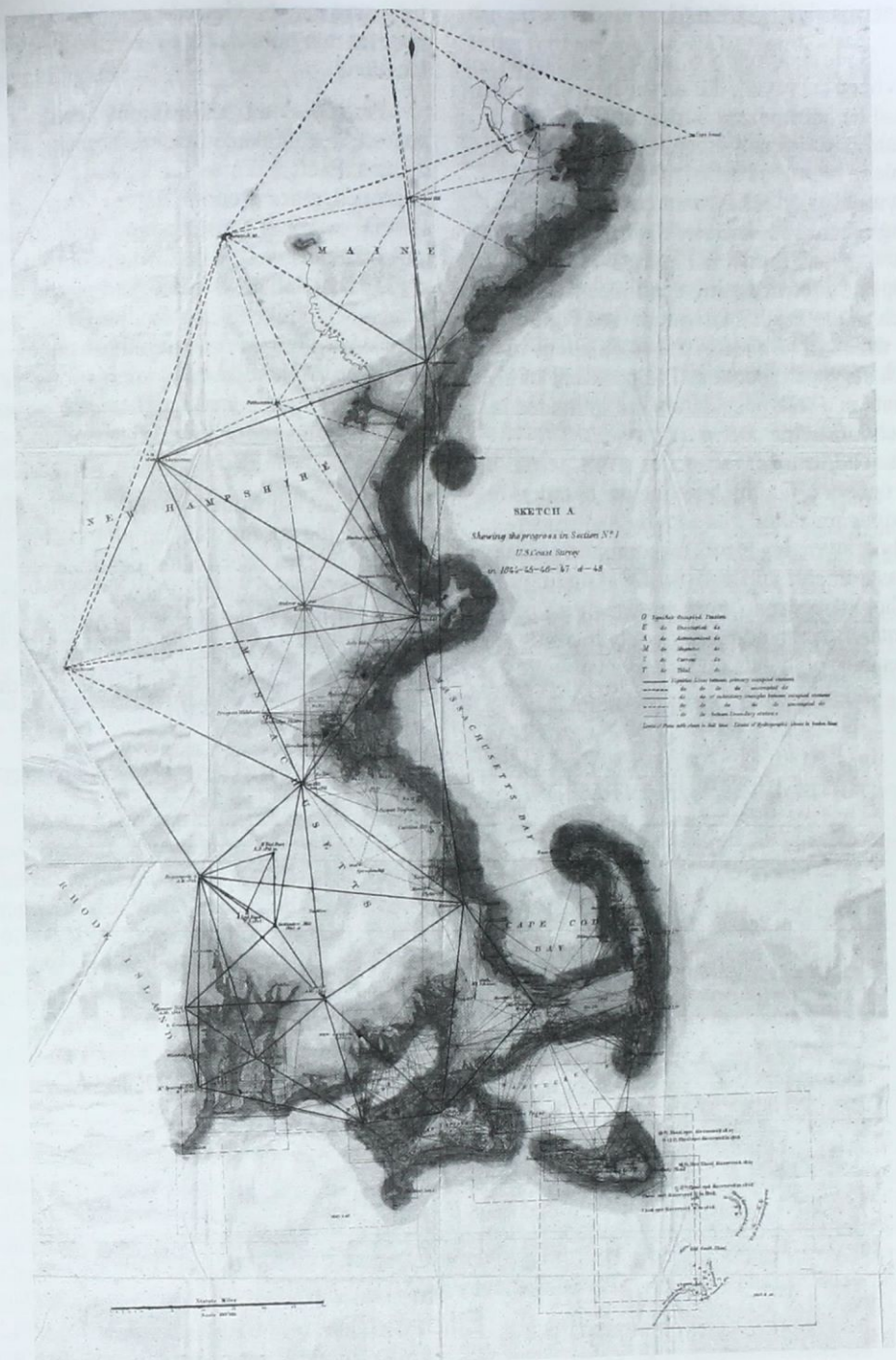


Figure 3-17: A map of the triangulation survey network established the northeastern coast of the United States of America in the early 1800s. Each leg of the triangles, shown here as a single line, is in turn a triangulation survey. This nested triangulation provides reinforcing measurements, thereby increasing the accuracy of the surveyed positions (courtesy NOAA).

Datum Adjustment

Once a sufficiently large set of points has been surveyed, the survey measurements must be harmonized into a consistent set of coordinates. Small inconsistencies are inevitable in any large set of measurements, causing ambiguity in locations. In addition, the long reaches spanned by the triangulation networks, as shown in Figure 3-17, could be helpful in recalculating certain constants, such as the Earth's curvature (see Figure 3-5), which in turn affect the calculations of each surveyed location. The positions of all points in a reference datum are estimated in a network-wide *datum adjustment*. The datum adjustment reconciles errors across the network, first by weeding out blunders or obvious mistakes, and also by mathematically minimizing errors by combining repeat measurements and statistically assigning higher influence to more precise measurements. A datum adjustment only incorporates measurements up to a given point in time, and may be viewed as our best estimate, at that point, of the measured set of locations.

Periodic datum adjustments result in series of regional or global reference datums. Each datum is succeeded by an improved, more accurate datum. This is not a trivial exercise, considering the adjustment may include survey data for tens of thousands of old and newly surveyed points from across the continent, or even the globe. Because of their complexity, these continent-wide or global datum calculations were once infrequent. Computers have improved such that datum adjustments now occur every few years.

A datum adjustment usually results in a change in the coordinates for all existing datum points, as coordinate locations are estimated for both old and new datum points. Our best estimates of the datum point coordinates will change. Differences between the datums reflect differences in the

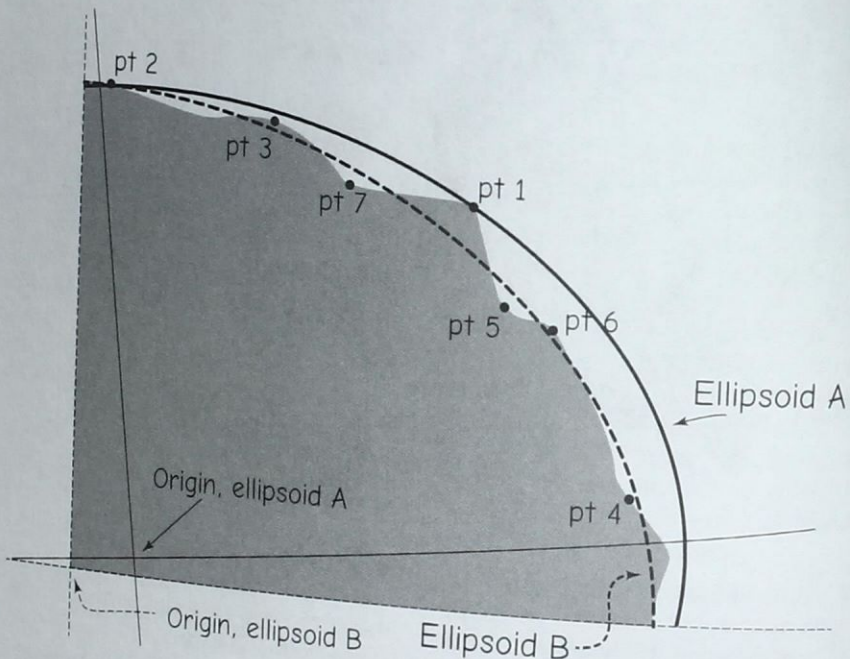


Figure 3-18: An illustration of two datums, one corresponding to Ellipsoid A and based on the fit to pt1 and pt2, and a subsequent datum resulting in Ellipsoid B, and based on a fit of pt1 through pt7. As the number and quality of our survey data improve, subsequent estimates of our best-fitting ellipsoid change.

ontrol points, survey methods, mathemati-al models, and assumptions used in the atum adjustment.

Figure 3-18 illustrates how ellipsoids night change over time, mostly in the origin nd orientation in this example, even for the ame survey region. Ellipsoid A is estimated vith the datum coordinates for pt1 and pt2, vith the shown corresponding coordinate xes, origin, and orientation. Ellipsoid B is ubsequently fit, after pts 3 through 7 have been collected. This newer ellipsoid has a different origin and orientation for its axis, causing the coordinates for pt1 and pt2 to change. The points have not moved, but the best estimate of their locations will have changed, relative to the origin set by the new, more complete set of datum points. You can visualize how the latitude angle from the origin to pt1 will change because the origin for ellipsoid A is in a different location than the origin for ellipsoid B. This apparent, but not real, movement is called the *datum shift*, and is expected with datum adjustments.

Commonly Used Datums

Three main series of horizontal datums have been used widely in North America. The first of these is the NAD series, beginning with the *North American Datum of 1927* (NAD27). NAD27 is a legacy datum, still encountered with some older data. NAD27 was a general least squares adjustment that used the Clarke Ellipsoid of 1866 and held fixed the latitude and longitude of a survey station in Kansas.

The *North American Datum of 1983* (NAD83) is the successor to NAD27. We place a modifier in parentheses after the NAD83 designator, e.g., NAD83(1986) to indicate the year, or version, of the datum adjustment. The original NAD83(1986) included approximately 250,000 stations and 2,000,000 distance measurements. The GRS80 ellipsoid was used, an Earth-centered reference, rather than fixing a surface station as with NAD27. Coordinate shifts from NAD27 to NAD83(1986) were large, often tens to 100 meters. In most instances, the surveyed points physically moved very little, for example, due to tectonic plate shifts, but our best estimates of point location changed.

Precise satellite positioning data became widely available soon after the initial NAD83(1986) adjustment, and were often more accurate than NAD83(1986) position estimates. Between 1989 and 2004, the NGS collaborated with other organizations to create *High Accuracy Reference Networks* (HARNs), also known as *High Precision Geodetic Networks* (HPGN) for most of the U.S. Generally, there is a different NAD83(HARN) for each state or small groups of states.

The HARN and subsequent NAD83 adjustments are largely satellite-based, and mark the transition from physical and optical surveying to GPS/GNSS surveying. They underpin a network of Continuously Operating Reference Stations (CORS, Figure 3-19). The CORS network of satellite observations allowed improved datum realizations, e.g. NAD83(CORS93), NAD83(CORS94), NAD83(CORS96), NAD83(2007), and NAD83(2011). NAD83(2011) is a long-

Figure 3-19: A Continuously Operating Reference Station (CORS), used to collect high-accuracy positional measurements from satellites for modern datum development (courtesy NOAA).

observation adjustment based on CORS stations, with coordinates reestimated for a broad set of survey marks. Both the CORS stations themselves and the bench marks are often used as starting points for more precise local surveys.

The *World Geodetic System of 1984* (WGS84) is a second set of datums developed and primarily used by the U.S. Department of Defense (DOD). It was introduced in 1987 based on Doppler satellite measurements of the Earth, and is used in most DOD maps and positional data. The WGS84 ellipsoid is similar to the GRS80 ellipsoid. WGS84 has been updated with more recent satellite measurements and is specified using a version designator. The update based on data collected up to January 1994 is designated as WGS84(G730). WGS84 datums are not widely used outside of the military because they are not tied to a set of broadly accessible, documented physical points.

There have been several subsequent WGS84 datum realizations. The original datum realization exhibited positional accuracy of key datum parameters to within one to two meters. Subsequent satellite observations improved accuracies. A reanalysis was conducted on data collected through week 873 of the GPS satellite schedule, resulting in the more accurate WGS84(G873). Successive realizations are known as WGS84(G1150), WGS84(1674), and WGS84(G1762), and there will likely be more adjustments in the future.

The third set of datums, commonly used worldwide and increasingly in North America, is known as the *International Terrestrial Reference Frames* (ITRF), with datum realizations of the International Terrestrial Reference System (ITRS). A primary purpose for ITRS is to estimate continental drift and crustal deformation by measuring the location and velocity of points, using a worldwide network of measurement locations. Each realization is noted by the year, for example, ITRF89, ITRF90, ITRF91. Each includes the X, Y, and Z location of each point and the velocity of each point in three dimensions. The European Terrestrial Reference System datum (ETRS89 and frequent updates thereafter) is based on ITRF measurements.

The ITRF and WGS84 datums are maintained by different organizations and based on different sets of measurements, but they have been aligned since 1995, and can be considered equivalent for most purposes, as differences between them since 1995 are generally only a few centimeters.

Although they are both based on modern satellite and other accurate measurements, the ITRF and current NAD83 datums do not align, and coordinates can be off by as much as two meters. Since the WGS84 is aligned with the ITRF series, WGS84 also differ by as much as two meters from NAD83. We should be careful in correctly adjusting for datum shifts between the ITRF/WGS84 and NAD83 datums.

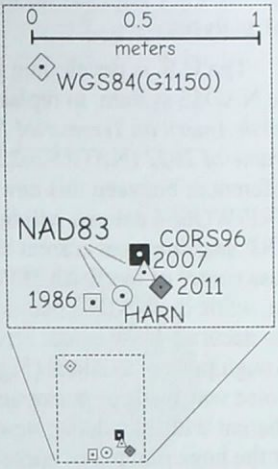
Figure 3-20 illustrates the relative size of datum shifts at an NGS marks between various versions of the NAD, and a WGS84/ITRF, based on estimates provided by the National Geodetic Survey. Notice that the datum shift between NAD27 and NAD83(86) is quite large, approximately 40 meters (130 feet), typical of the up to hundreds of meters of shifts from early, regional datums to modern, global datums. The figure also shows the subsequently smaller shifts for NAD83 datums through time, and relatively larger distance between NAD83 and WGS84/ITRF datums.

A datum shift does not imply that points have moved. Most monumented points are stationary relative to their immediate surroundings. The locations change over time as the large continental plates move, but these changes are small, on the order of a few millimeters per year, except in tectonically active areas such as coastal California; for most locations, it is just our estimates of the coordinates that have changed. As survey measurements improve through time and there are more of them, we obtain better estimates of the true locations of the monumented datum points.

Examples of Datum Shifts

Successive datum transformations for New Jersey control point, Bloom 1

Datum	Longitude (W)	Latitude(N)	Shift(m)
NAD27	74° 12' 3.86927"	40° 47' 0.76531"	
			36.3
NAD83(1986)	74° 12' 2.39240"	40° 47' 1.12726"	
			0.04
NAD83(HARN)	74° 12' 2.39069"	40° 47' 1.12762"	
			0.05
NAD83(CORS96)	74° 12' 2.39009"	40° 47' 1.12936"	
			0.01
NAD83(2007)	74° 12' 2.38977"	40° 47' 1.12912"	
			0.03
NAD83(2011)	74° 12' 2.38891"	40° 47' 1.12839	
			0.98
WGS84(G1150)	74° 12' 2.39720"	40° 47' 1.15946"	



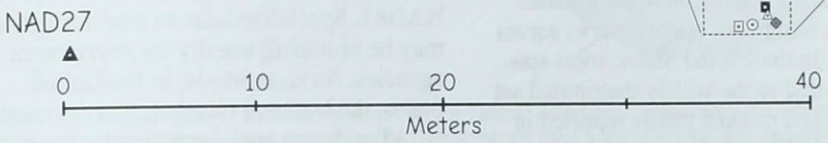


Figure 3-20: Datum shifts in the coordinates of a point for some common datums. Note that the estimate of coordinate position shifts approximately 36 m from the NAD27 to the NAD83(1986) datum, while the shift from NAD83(1986) to NAD83(HARN) then to NAD83(CORS96) are 0.05 m or less. The shift to WGS84(G1150) is also shown, here approximately 0.98 m. Note that the point may not be moving, only our datum estimate of the point's coordinates. Calculations are based on NGS data sheets, NCAT, and HTDP software.

We must emphasize while much data are collected in WGS84/ITRF datums using GNSS (such as GPS), most data are converted to a local or national datum before use in a GIS. In the United States, this typically involves GNSS accuracy augmentation, often through a process called differential correction, described in detail in Chapter 5. Corrections are often based on an NAD83 datum, effectively converting the coordinates to the NAD83 reference, but ITRF datums are also commonly used. Ignorance of this "implicit" conversion among datums is a common source of error in spatial data, and should be avoided.

There are a few points about datums that must be emphasized. First, different datums specify different coordinate systems. You do not expect coordinates for any physical point to be the same when they are expressed relative to different datums.

Second, the version of the datum is important. NAD83(1996) is a different realization than NAD83(2011), and ITRF88 is different than ITRF05. The datum is incompletely specified unless the version is noted. Many GIS software packages refer to a datum without the version, for example, NAD83. This is indeterminate, and confusing, and shouldn't be practiced. It forces the user to work with ambiguity.

Third, differences between families of datums change through time. The NAD83(1986) datum realization is up to two meters different than the NAD83(CORS96), and the original WGS84 differs from the current WGS84 version by more than a meter over much of the Earth. Differences in datum realizations depend on the versions and location on Earth. This means you should assume all data should be converted to the same datum and version before combi-

nation in a GIS. This rule may be relaxed only after you have verified that the datum difference errors are small compared to other sources of error, or small compared to the data accuracy required for the intended spatial analysis.

The U.S. is developing the successor to the NAD83 system, to replace it with the *North American Terrestrial Reference Frame of 2022* (NATRF2022). Most of the differences between this new datum and the ITRF/WGS84 datums will be resolved. The ITRF uses the most current estimates for the mass center of the Earth as the ellipsoid origin, while the NAD83 maintained an earlier, less accurate mass center across the 1986 through present versions (Figure 3-21). This choice was made to postpone the confusion inherent with calculating new coordinates for the huge number of survey marks across the country. In the United States, most spatial data are tied to the widely distributed set of surveyed and marked points reported in the NAD83(CORSxx) datums, and state, county, and local surveys are referenced to these points. The adoption of NATRF2022 will require transforming current NAD83(2011) and earlier data to new coordinates, but will help avoid substantial confusion in position due to datums.

Datum Transformations

Converting coordinates from one datum to another typically requires a *datum transformation*. A datum transformation provides the latitude and longitude of a point in one datum when we know them in another datum; for example, we can calculate the latitude and longitude of a survey mark in NAD83(2011) when we know these geographic coordinates in ITRF08 (Figure 3-22).

Datum transformations are often more complicated when they involve older datums. Many older datums were created piecemeal to optimize fit for a country or continent, so simple formulas often do not exist for transformations involving many older datums, for example, from NAD27 to NAD83. Specialized datum transformations may be provided, usually by government agencies. As an example, in the United States, the National Geodetic Survey created NCAT, a datum transformation tool to convert between various NAD datums.

Transformation among newer datums may use more general mathematical transformations between three-dimensional, Cartesian coordinate systems (Figure 3-22). Transformation equations allow conversion among most NAD83, WGS84, and ITRF

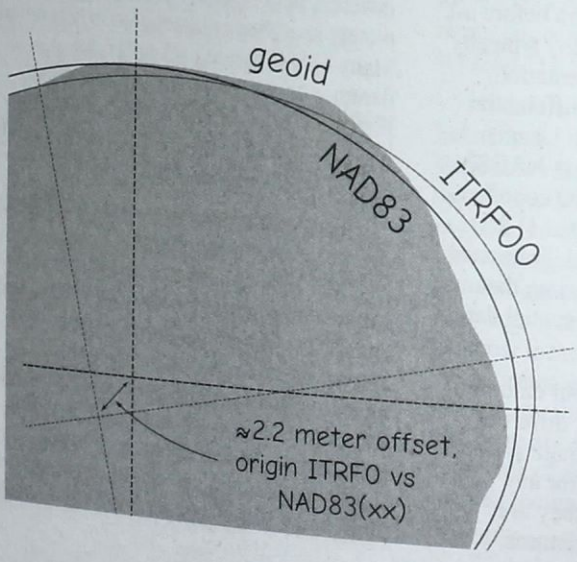


Figure 3-21: The NAD83 and ITRF datums use similar ellipsoid diameters, but different ellipsoid origins and orientations, so coordinates will change when transformed between them. The xx in NAD83(xx) indicates this offset is present through all versions of the NAD83 datum.

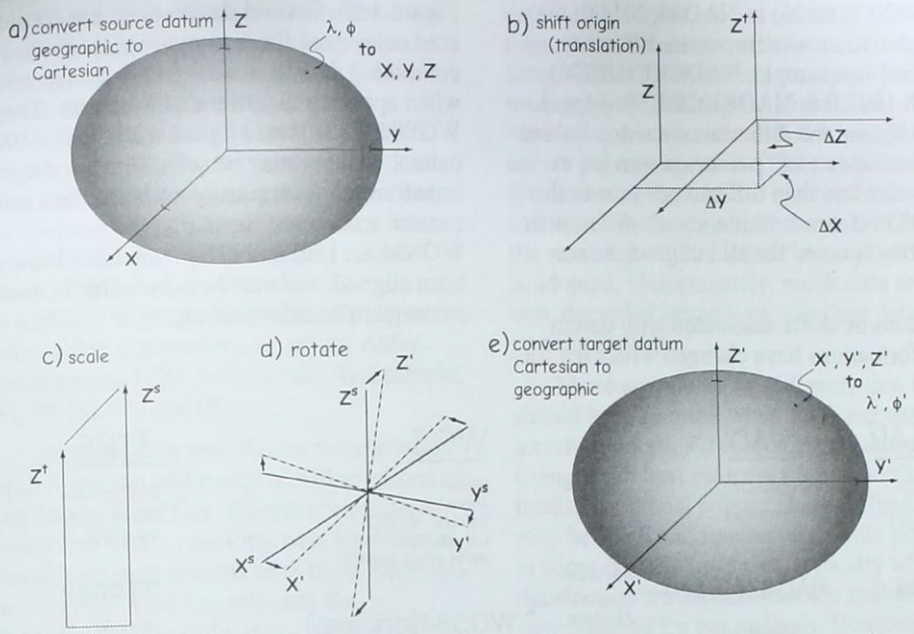


Figure 3-22: Application of a modern datum transformation. Geographic coordinates (longitude, λ, and latitude, ϕ), are transformed to a new datum by a) conversion from geographic to Cartesian coordinates in the old datum (through a set of equations that are not shown), b) applying an origin shift, c) scaling, d) rotating these shifted coordinates, and e) converting these target datum Cartesian coordinates, X', Y', Z', to the longitude and latitude, λ', ϕ', in the target datum.

systems, and are supported in large part by improved global measurements from satellites, as described in the previous few pages. This approach incorporates a shift in the origin, a rotation, and a change in scale from one datum to another.

A datum transformation is typically a multi-step process. In past times, empirical, grid-based methods have been used because many early datums were not strictly derived from coherent mathematical surfaces. Later, a *Molodenski transformation* was common, using a system of equations with three or five parameters. More currently, a *Helmert transformation* is employed using seven or 14 parameters (Figure 3-22). First, geographic coordinates on the source datum are converged from longitude (λ) and latitude (ϕ) to X, Y, and Z Cartesian coordinates. An origin shift (translation), rotation, and scale are applied. This system produces new X', Y', and Z' coordinates in the target datum. These X', Y', and Z' Cartesian coordinates are then converted back to the longitudes and latitudes (λ' and ϕ'), in the target datum.

More advanced methods allow these seven transformation parameters to change through time, to account for tectonic and other shifts, for a total of 14 parameters. These methods are incorporated into software that calculate transformations among modern datums, for example, the Horizontal Time Dependent Positioning (HTDP) tool available from the U.S. NGS (www.ngs.noaa.gov/TOOLS/Htdp/Htdp.shtml). HTDP converts among recent NAD83 datums and most ITRF and WGS84 datums.

Because of tectonic plate movement, the most precise geodetic measurements refer to the epoch, or fixed time period, at which the point was measured or datum fit. The HTDP software includes options to calculate the shift in a location due to different reference datums [for example, NAD83(CORS96) to WGS84(G1150)], the shift due to different

realizations of a datum [for example, NAD83(CORS96) to NAD83(2011)], the shift due to measurements in different epochs [for example, NAD83(CORS96) epoch 1997.0 to NAD83(CORS96) epoch 2010.0], and the differences due to all three factors. Since most points are moving at velocities less than 0.1 mm per year in the NAD83 reference frame, epoch differences are often ignored for all but geodetic surveys.

Datums shifts associated with datum transformations have changed with each successive datum realization, as summarized in Figure 3-23. Several datum pairs are considered equivalent for many purposes when combining data from different data layers, or when applying datum transformations. The WGS84(G730) was aligned with the ITRF92 datum, so these may be substituted in datum transformations requiring no better than centimeter accuracies. Similarly, the WGS84(G1150) and ITRF00 datums have been aligned, and may be substituted in most subsequent transformations.

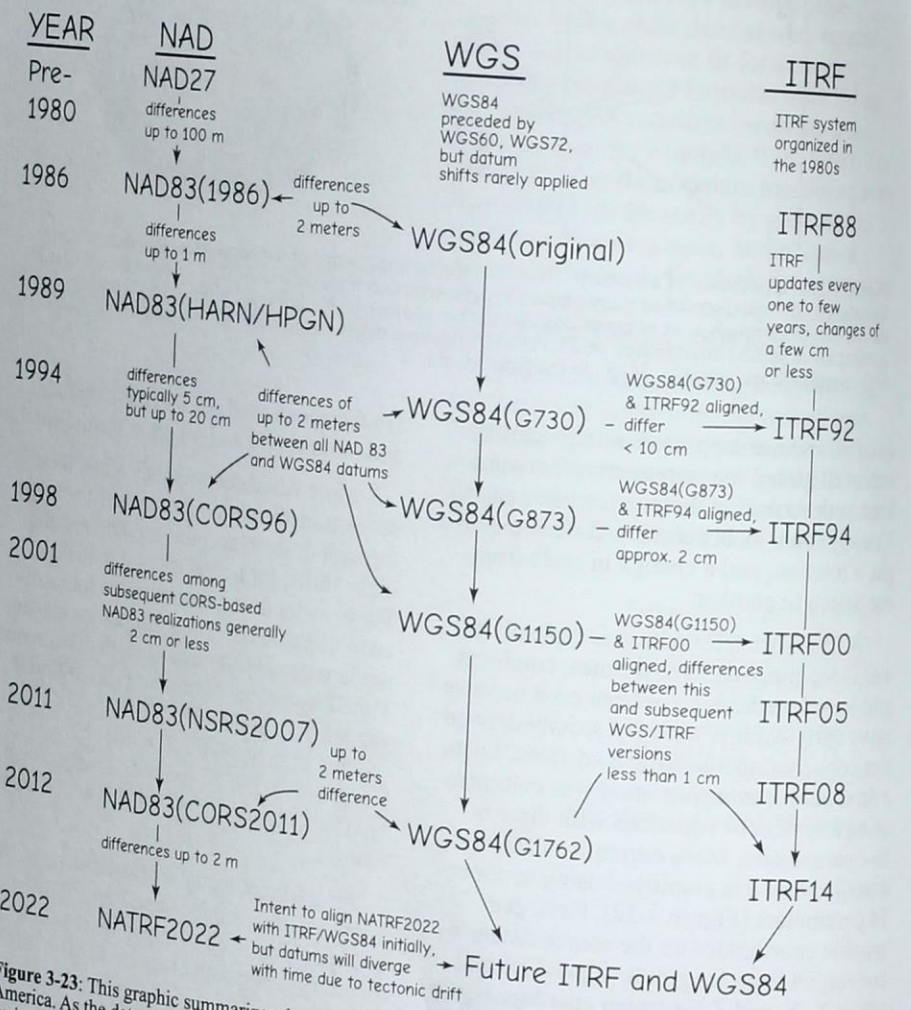


Figure 3-23: This graphic summarizes the evolution of the three main families of datums used in North America. As the datums have been adjusted, horizontal positional differences between survey marks have varied, within the ranges shown. "Aligned" datums (e.g., WGS84(G1150) and ITRF00) may be considered equivalent for most purposes when applying datum transformations.

While locations in the NAD83(xx) and the ITRF/WGS84 datums commonly differ by over a meter, datum shifts internal to these groupings have become small for recent datums. Differences between NAD83(HARN) and NAD83(xx) datums may be up to 20 cm, but are typically less than 4 cm, so these datum realizations may be considered equivalent if accuracy limits are above 20 cm, and perhaps as low as 4 cm. The differences between NAD83(CORS96) and NAD83(2011) are often a few centimeters, as are the differences among ITRF realizations, for example, 91, 94, 00, 05, and 08.

There will be new datum realizations, each requiring additional transformations in the future. The ITRF datums are released every few years, requiring new transformations to existing datums each time. As of this writing, the NGS has released the NAD83(2011) coordinates a nationwide adjustment of passive survey marks and multiyear observations at GNSS/GPS CORS stations.

There is a plan to substantially update the datums used in North America, with the introduction of the *North American Terrestrial Reference Frame of 2022* (NATRF2022). This will initially align official datums for the U.S. with the ITRF and WGS84 datums, removing much of the positional differences for points expressed in these different systems at the time of estimation. It will entail a shift, up to two meters (six feet) in NAD83(2011) coordinates to NATRF2022 coordinates.

Although the NATRF2022 and the ITRF/WGS84 systems will be aligned initially, current plans fix the NATRF2022 to the included tectonic plates, and so will drift from the ITRF positions through time. The transformation will be mathematically simple, using the time since initiation and location, and we expect the US NGS to produce tools to calculate a datum shift given any epoch.

Prior to this decade, differences in datum transformation were usually lower than spatial data error, so it caused few problems. GNSS receivers can now provide centimeter-level accuracy in the field, so what were once considered small datum discrepancies are now apparent. The datum transformation method within any hardware or software system should be documented and the accuracy of the method known before it is adopted. Unfortunately, much data are now degraded because of improper datum transformations.

There are a number of factors that we should keep in mind when applying datum transformations. First, changing a datum changes our best estimate of the coordinate locations of most points. These differences may be small and ignored with little penalty in some specific instances, typically when the changes are smaller than the spatial accuracy required for our analysis. However, many datum shifts are quite large, up to tens of meters. One should know the magnitude of the datum shifts for the area and datum transformations of interest.

Second, datum transformations are estimated relationships that are developed with a specific data set and for a specific area and time. There are spatial errors in the transformations that are specific to the input and datum version. There is no generic transformation between NAD83 and WGS84. Rather, there are transformations between specific versions of each, for example, from NAD83(96) to WGS84(1150).

Finally, GIS projects should not mix datums except under circumstances when the datum shift is small relative to the requirements of the analysis. Unless proven otherwise, all data should be converted to the same coordinate system, based on the same datum. If not, data may misalign.

Vertical Heights and Datums

In its simplest definition, a *vertical datum* is a reference that we use for measuring heights. We commonly specify a vertical datum using a measured, constant gravity (equipotential) surface (Figure 3-24). We then combine these with carefully measured control heights above a specific equipotential surface to define surface heights. As noted in the geoid section on page 92, most government or other organizations use a specific geoid as a reference surface for height, although not everyone adopts the same geoid. Governments adopt "hybrid" geoids that combine their own precise vertical surveys with gravity measurements and models.

Geodesists and surveyors use the term *orthometric heights* to refer to what most of us think of as elevations. This is to clearly refer to our standard heights above our reference surface, different from other height measurements they sometimes use. The orthometric height is the distance from a standard equipotential surface to another level, with the path between the surfaces always at right angle to all intervening gravity surfaces. Orthometric heights have replaced our elevations above mean sea level because, as mentioned earlier, modern vertical heights are referenced to a geoid.

For much of history prior to satellites, *leveling surveys* were used for establishing heights. A standard, seaside bench mark was selected, and distances and elevation differences precisely measured from there to known points. Leveling surveys give the heights of points along their path. Bench marks established at these points were then used to set nearby heights. Early leveling surveys were performed with simple instruments, for example, by *spirit leveling*, using plumb bobs and bubble or tube levels. Horizontal rods were placed between succeeding vertical posts to physically measure height.

The number, accuracy, and extent of leveling surveys increased substantially in the 18th and 19th centuries. Epic surveys that lasted decades were commissioned, such as the Great Arc from southern India to the Himalayas. These surveys were performed at substantial capital and human expense; in one portion of the Great Arc, more than 60% of the field crews died over a six-year period due to illness and mishaps.

Most leveling surveys from the late 1700s through the mid-20th century employed *trigonometric leveling*. This

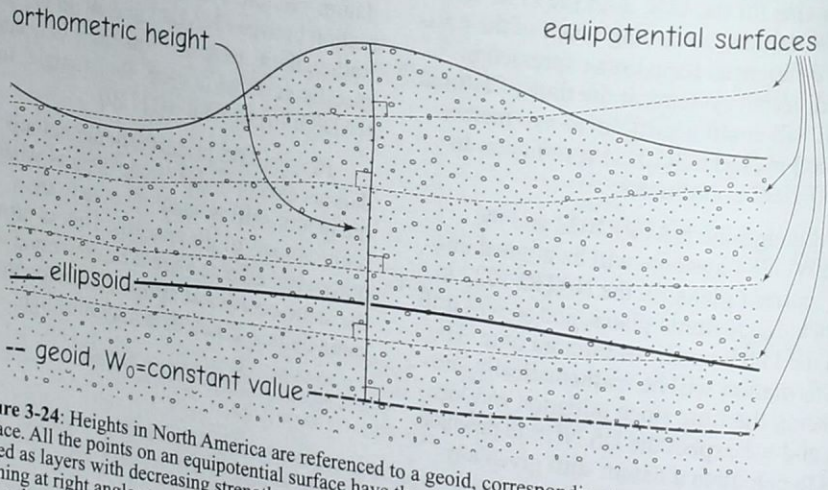


Figure 3-24: Heights in North America are referenced to a geoid, corresponding to a given equipotential surface. All the points on an equipotential surface have the same gravitational pull, and they may be envisioned as layers with decreasing strength at higher levels. Heights are usually specified as orthometric, meaning at right angles to all equipotential surfaces along their path. Because potential surfaces may undulate, orthometric heights may be curved lines, although usually only slightly so.sually only slightly so.

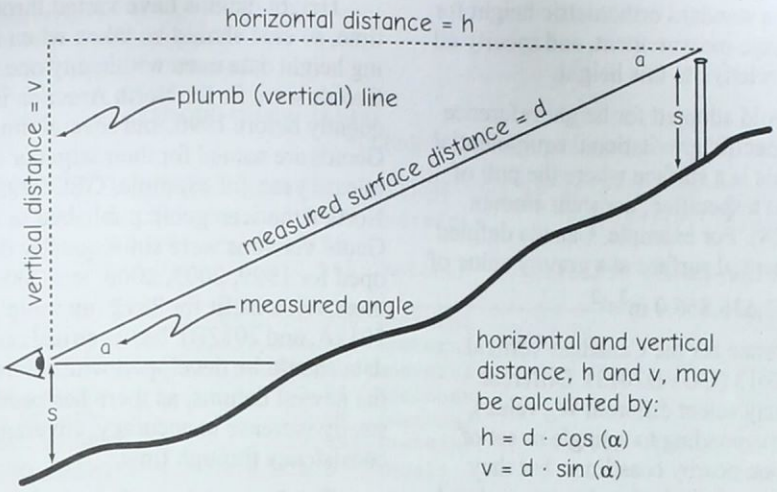


Figure 3-25:Leveling surveys often employ optical measurements of vertical angle (α) with measurements of surface distance (d) and knowledge of trigonometric relationships to calculate horizontal distance (h) and vertical distance (v).

method uses optical instruments and trigonometry to measure changes in height, as shown in Figure 3-25. Surface distance along the slope was measured to avoid the tedious process of establishing vertical posts and leveling rods.The vertical angle was also measured from a known station, typically by a small telescope fitted with a precisely scribed angle gage. Surface distance would then be combined with the measured vertical angle to calculate the horizontal and vertical distances. Early surveys measured surface distance along the slope with ropes, metal chains, and steel tapes. Modern height measurements have largely replaced trigonometric measurements, and primarily use a variety of laser and satellite-based methods.

In North America, we no longer use mean sea-level as a base for orthometric heights (elevations); except for specific projects near the seashore, orthometric heights above a specific vertical datum are now our standard elevation specification. We stopped using mean sea level because it varies too much in time and space. The mean varies in time because of daily through multi-decadal solar and lunar cycles, and across the globe because of persistent differences in water density with temperature, salinity, and ocean currents. Global sea level has been rising over the past century, so the mean at any one seaside station will depend on the length of measurement, even for stations collecting for longer than the 19-year lunar/solar cycles. The mean sea level will differ from Miami to New York, or Amsterdam to Genoa. We weren't able to address this variation until the past few decades, after which methods improved to where the discrepancies in sea levels across the globe became apparent.

Since we want a surface that is consistent in time and space, most countries have picked one or a set of tidal stations, and based orthometric heights relative to a geoidal height passing through or near the station height(s). North American orthometric heights are based on a height specified relative to a long-term tidal gage in Quebec. In mainland Australia, heights are relative to measurements averaged over 30 tidal gages spread along the coast, because they have an approximately 1 meter decline in the geoid height relative to tidal gage measurements from the northeast to the southwestern part of the country. Various European countries adopt base points near different long-term tidal gages, or if landlocked, for points related to gages in adjacent countries. Most countries then adopt an appropriate geoid

and assign a standard orthometric height for the mean gage measurement, and specify all elevations relative to this height.

The geoid adopted for height reference is often a specific gravitational equipotential surface. This is a surface where the pull of gravity is at a specified, constant amount (Figure 3-24). For example, Canada defined the equipotential surface at a gravity value of

$$W_0 = 62{,}636{,}856.0\ m^2 s^{-2}$$

as the reference for the Canadian Vertical Datum of 2013 (CGVD2013). Different countries may select different W_0 values, usually corresponding to a single or set of tidal gages on nearby coastlines, but they don't all assign the calculated mean sea level a height of zero. Nonzero heights may be assigned to best match historical data, or when a mean of several stations is used.

Orthometric heights (elevations) in North America are defined as the vertical distance measured from our adopted reference geoid to the ground surface height, along a line that is always at right angles to all intervening equipotential surfaces (Figure 3-24). This height line may bend, as there are often small undulations in the successive equipotential surfaces. The height paths are not the same as a straight line normal to the ellipsoid and up to the surface, and not the same as a straight line that is normal to the geoid surface at the starting point.

Because the zero height may differ among countries, you must be careful when mixing heights across countries. Orthometric heights referenced to one geoid and set of bench marks in Poland may differ from heights referenced to another set in the Netherlands, or ones in Jamaica different from Florida. Unless heights are adjusted, they may be inconsistent when combined across vertical datums. Cooperation among governments is common; for example, datums are compatible across the United States, Canada, and Mexico in North America, and there is a European Vertical Reference System to unify European height datums.

Height datums have varied through time, so care should be taken when combining height data even within any one country. Geoids were fit for North America infrequently before 1990, and several times since. Geoids are named for their target or effective release year, for example, GEOID96 for the North American geoid published in 1996. Geoid versions were subsequently developed for 1999, 2003, 2006, and 2009, with three versions fit for 2012 (an initial, a 2012A, and 2012B). New vertical coordinate data should be developed with reference to the newest datums, as there has been a steady increase in accuracy, coverage, and consistency through time.

The first continental vertical datum in North America was the *National Geodetic Vertical Datum* of 1929, also referred to as NGVD29. Vertical leveling was adjusted to 26 tidal gages, including 5 in Canada, to match measured local mean sea level. Geodesists realized that mean sea level varied across the continent, but assumed these differences would be similar or smaller than measurement errors. They wanted to avoid confusion caused by seaside bench marks having heights that differed from mean sea level.

The latest North American datum is labeled NAVD88. This datum is based on over 600,000 kilometers (373,000 miles) of control leveling performed since 1929, and also reflects geologic crustal movements or subsidence that may have changed bench mark elevation. NAVD88 was fixed relative to only one tidal station because improved measurements yielded errors much smaller than among-station differences in mean sea level, as noted before.

Improved surface, aerial, and satellite gravity measurements, particularly the NASA GRACE and ESA GOCE satellite missions, have led to a dense network of gravity measurements, including regions far from coastal tidal stations. These measurements are combined with previous surveys to update geoid models and allow calculation of the geoidal height at any point on the ellipsoid. Now we most often combine mod-

```
DESIGNATION -  E 58
PID         -  FB1004
STATE/COUNTY-  NC/MADISON
COUNTRY     -  US
USGS QUAD   -  SPRING CREEK (1946)
FB1004                        *CURRENT SURVEY CONTROL

______________________________________________________________________
NAD 83(2011) POSITION- 35 47 30.13346(N) 082 51 55.76123(W)  ADJUSTED
NAD 83(2011) ELLIP HT-   623.632 (meters)        (06/27/12)  ADJUSTED
NAD 83(2011) EPOCH   -  2010.00
NAVD 88 ORTHO HEIGHT -   653.568 (meters)     2144.25  (feet) ADJUSTED

______________________________________________________________________
NAD 83(2011) X  -    643,358.550 (meters)                    COMP
NAD 83(2011) Y  - -5,139,947.911 (meters)                    COMP
NAD 83(2011) Z  -  3,709,833.794 (meters)                    COMP
LAPLACE CORR    -         -2.95  (seconds)                   DEFLEC12A
GEOID HEIGHT    -        -29.93  (meters)                    GEOID12A
DYNAMIC HEIGHT  -        652.892 (meters)     2142.03  (feet) COMP
MODELED GRAVITY -    979,578.6   (mgal)                      NAVD 88
```

Figure 3-26: A portion of a data sheet for a vertical control bench mark.

els of geoidal height with measurements of ellipsoidal height (easily given by GNSS systems, described in Chapter 5) to establish orthometric heights.

At this writing, the most current model for North America, GEOID12B, incorporates the best available gravity data with bench marks, leveling, and GPS/GNSS surveys. It has integrated nearly 23,000 vertical bench marks to estimate geoidal and orthometric heights. These heights are known across the continent, and reported on NGS data sheets for vertical bench marks (Figure 3-26). The bench mark sheets also note the vertical datum (here NAVD88), the geoid model (GEOID12), the orthometric height (here 653.568 meters), and the ellipsoidal and geoidal heights. Hybrid vertical datums we use are not entirely independent of horizontal datums, so we should pair our horizontal/vertical datums when combining/converting coordinate data (Figure 3-27).

There is currently an effort to modernize the North American vertical datum, in concert with the horizontal NATRF2022 datum. This will integrate airborne gravity surveys of the entire U.S. and its holdings, to yield a geoid surface estimate accurate to within 1 cm. It will also result in vertical height shifts of up to 1.3 m (4 feet) from NAVD88 to the new datum.

NGVD29, no geoid
with
NAD27, NAD83(1986)

NAVD88, GEOID03
with
NAD83(1996)

NAVD88, GEOID09
with
NAD83(NSRS2007)

NAVD88, GEOID12B
with
NAD83(2011)

Figure 3-27: Recommended pairing for horizontal and vertical datums in North America.

Because vertical datums differ among regions, and have changed through time within most regions, datum confusion often reigns. Failure to adjust for height differences between vertical datums has caused many errors in height reference, both for older and modern height measurements.

These errors are becoming more commonplace with the widespread use of inexpensive, precise satellite positioning, and with high accuracy laser positioning. Knowledge of the sequence of vertical datums, associated geoid evolution, and vertical datum conversion tools are needed to avoid vertical measurement errors.

Vdatum

Given that vertical datums and associated geoids change through time, the United States National Geodetic Survey (NGS) has created a tool, VDatum, to estimate conversions among vertical datums in the U.S. (Figure 3-28). VDatum calculates the vertical difference from one datum to another at any given horizontal coordinate location and height. Conversions are provided between the 1929 and modern datums, between WGS84/ITRF and NAVD datums, and between various ellipsoid versions within the NAVD88 datum.

Because the vertical datum shift will vary as a function of position, a latitude and longitude must be provided, and because the shift may also depend somewhat on elevation, a vertical height entered. As shown in the example in Figure 3-28, the shifts can be quite large, particularly when converting between NAVD and WGS84/ITRF, and also from NGVD1929 to NAVD88 datums. The vertical datum shift typically changes slowly with distance, so one offset may be suitable for all height shifts over a few to tens of square kilometers. The amount of error and "safe" distance to span varies by region, so the magnitude of the transformation should be verified at several points across any new study area to see how broadly an offset may be used.

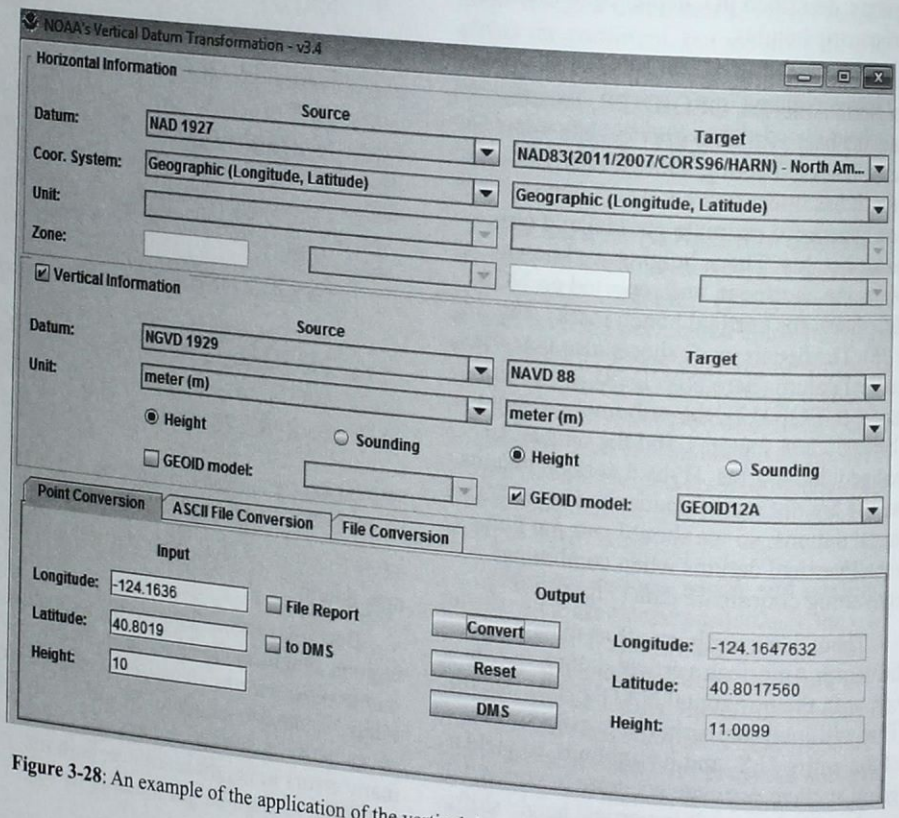


Figure 3-28: An example of the application of the vertical datum transformation software VDatum.

VDatum may also be used to estimate shifts in height among geoid versions. New geoid surfaces have been estimated approximately every three years since 1996 for North America, and heights at any given point will change between geoids. If heights relative to different geoids are to be combined, one set of heights must be adjusted to match the geoid of the other. This is typically achieved by adding an offset calculated from the models included in VDatum.

As an example, I may have two elevation data sets, both in Eureka, California, near a point with latitude 40.8019, longitude -124.1636, and approximately a 10-meter height. One elevation is measured relative to the GEOID96 version of the NAVD88, and the other using the GEOID12A version. I can use VDatum to calculate the vertical height shift due to this difference in geoids; at that coordinate and height, it estimates a 31 cm, or approximately 1 foot, increase in height between these two geoids. This means I would have to add 31 cm to all my 96 heights before combining them with my 12A heights.

Dynamic Heights

We must discuss another kind of height, called a *dynamic height*, because it is important for certain applications. Dynamic heights measure the change in gravitational pull from a given equipotential surface. Dynamic heights are important when interested in water levels and flows across elevations. Points that have the same dynamic heights can be thought of as being at the same water level. Surprisingly, points with the same dynamic heights often have different orthometric heights (Figure 3-29). To be clear, two distinct points at water's edge on a large lake often do not have the same elevations; often, they are different orthometric heights above our reference geoid. Since orthometric heights are our standard for specifying elevation, this means water may indeed flow uphill relative to our standard height measurement, or as confusingly, a lake may have a different elevation on one shore than on the opposite shore.

To understand why water may flow uphill (from lower to higher orthometric heights), it is important to remember how

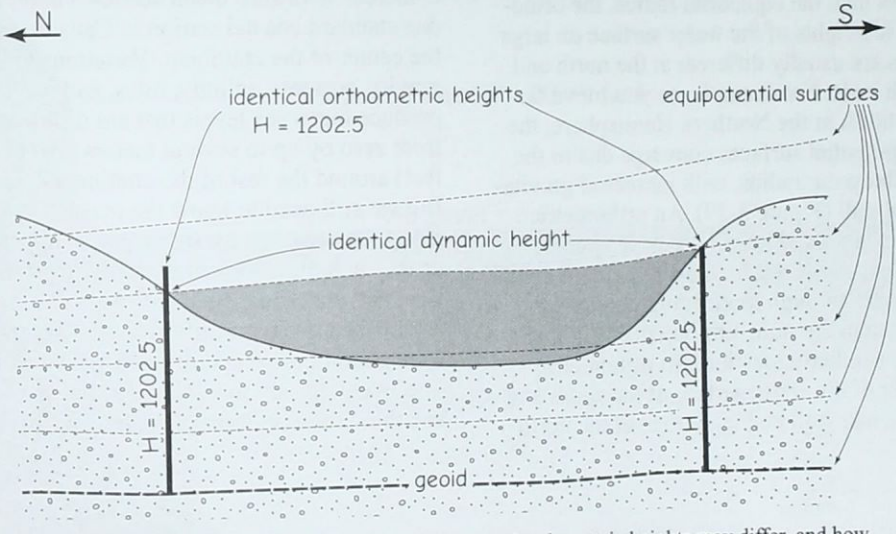


Figure 3-29: An illustration of how dynamic heights and orthometric heights may differ, and how equal orthometric heights may correspond to different heights above the water level on a large lake. Because equipotential surfaces converge, the water level at the northern and southern extremes of a lake will have different orthometric heights. Dynamic heights and water levels are equal across an equipotential surface.

orthometric heights are defined. An orthometric height is the distance, in the direction of gravitational pull, from the geoid up to a point. But remember, the geoid is a specified gravity value, an "equipotential" surface, where the pull of gravity is at some specified level. As we move up from the geoid toward the surface, we pass through other equipotential surfaces, each at a slightly weaker gravitational force, until we arrive at the surface point. But these gravity surfaces are not always parallel, and may be more closely packed in one portion of the globe than another.

There are two key points. First, water spreads out to level across an equipotential surface, absent wind, waves, and other factors. The water level in a still bathtub, pond, or lake has the same equipotential surface at one end as another. Gravity ensures this. Second, the equipotential surfaces are closer together when nearer the mass center of Earth. As the equipotential surfaces converge, or become "denser," the water surface seems to dip below our fixed orthometric height.

Because water follows an equipotential surface, and because the Earth's polar radius is less than the equatorial radius, the orthometric heights of the water surface on large lakes are usually different at the north and south ends. For example, as you move farther north in the Northern Hemisphere, the equipotential surfaces converge due to the smaller polar radius, with increased gravitational pull (Figure 3-29). An orthometric height is a fixed height above the geoidal surface, so the northern orthometric height will pass through more equipotential surfaces than the same orthometric height at a more southerly location. An orthometric height of the water surface at the south end of the lake will be higher than at the north end. For example, in Lake Michigan, a large lake in North America, the elevation of the the water surface at the south end is approximately 15 cm higher than the elevation of the water surface at the north end.

Dynamic heights are most often used when we're interested in relative heights for water levels, particularly over large lakes or connected water bodies. Because equal dynamic heights are at the same water level, we can use them when interested in accurately representing hydrologic drop, head, pressure, and other variables related to water levels across distances. But these differences could be confusing when observing bench mark or sea level heights, and underscore again that our height reference is not mean sea level, but rather an estimated geoidal surface.

Local Sea Level Datums

Water height measurements along the U.S. coast are typically reference to local sea level datums. As noted earlier, mean sea level is not zero for almost all points along North America's coastline. Elevations are measured relative to a geoid. Zero elevation coincides with zero mean sea level at only one standard coastal station in Canada, near the center of the continent. Variations in gravity, currents, salinity, tides, and wind produce mean sea levels that are different from zero by up to several meters (10s of feet) around the rest of the continental rim. But we still need to know the ocean level along the coastline for many practical purposes, including construction, flood protection, and water management. We have established a network of long-term, reference measurement stations along the coastline. We precisely measure both sea level and the station orthometric height, so that we

can tie our standard elevation to local water heights.

Data for measured tidal stations are available from the NOAA web page:

tidesandcurrents.noaa.gov/stations.html

These sites report mean sea level, as well as mean high, low, and extreme water levels (Figure 3-30). Most importantly, they also report the NAVD88 orthometric heights for each tidal station, allowing a conversion from local sea level heights to measured surface elevation.

Figure 3-30 shows data for a station in Seattle measured since 1899. Mean sea level has a local reference height of 6.64 feet, meaning the sea level averages that height above the long-term measurement of a given low water height. The NAVD88 height at the same point is 2.34 feet, which yields a sea level height of 6.64 - 2.34, or 4.3 feet. As strange as it may seem at first, the mean sea level at this Seattle station has an elevation of 4.3 feet. Any point nearby that has an elevation less than 4.3 feet will be below sea level, and will likely flood frequently if there is access to the sea. Local construction, water level measurements, or other activities dependent on sea level will reference this station measurements, and the 4.3 foot offset between mean sea level and seaside orthometric heights.

This mean sea level offset varies by location, for example, Port San Luis, CA, has a vertical offset of 2.7 feet, and Vaca Key, FL, has an offset of -0.8 feet. When heights of sea level are important for an analysis, projects should reference the nearest local datum.

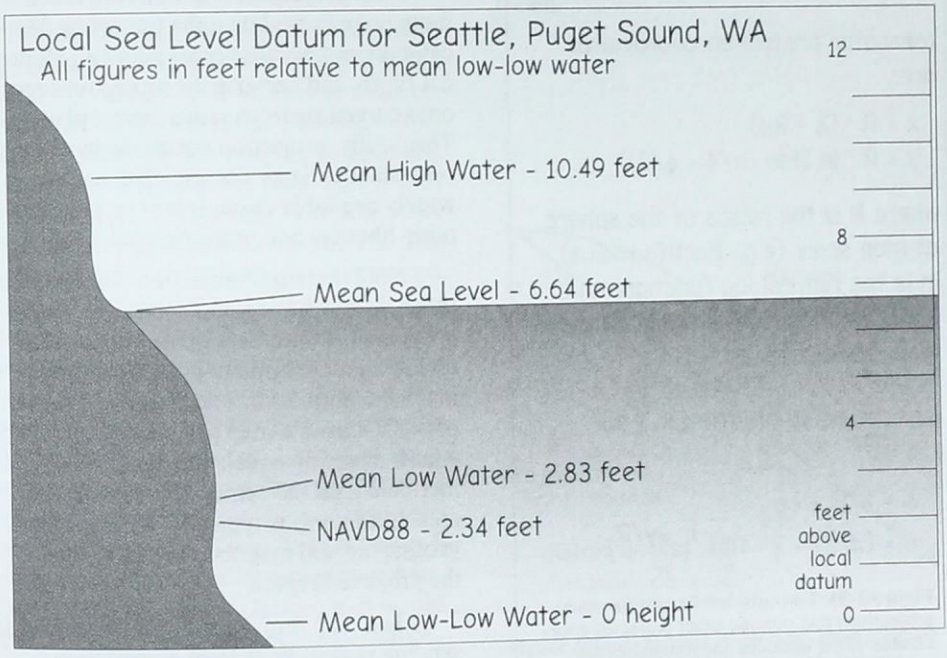


Figure 3-30:An illustration of mean sea level and other measures at a NOAA long-term tidal gage. Note that the mean sea level at this station is 6.64 - 2.34, or 4.3 feet above the NAVD88 zero height.

Map Projections and Coordinate Systems

Datums tell us the latitudes and longitudes of features on an ellipsoid. We need to transfer these from the curved ellipsoid to a flat map. A *map projection* is a systematic rendering of locations from the curved Earth surface onto a flat map surface.

Nearly all projections are applied via exact or iterated mathematical formulas that convert between geographic latitude/longitude pairs and projected X/Y (easting and northing) coordinates. Figure 3-31 shows one of the simpler projection equations, between Mercator and geographic coordinates, assuming a spherical Earth. These equations would be applied for every point, vertex, node, or grid cell in a data set, converting the vector or raster data feature by feature from geographic to Mercator coordinates.

Conversion from geographic (lon, lat) to projected coordinates

Given longitude = λ, latitude = ϕ (all angles in radians)

Mercator projection coordinates are:

$$x = R \cdot (\lambda - \lambda_0)$$
$$y = R \cdot \ln(\tan(\pi/4 + \phi/2))$$

where R is the radius of the sphere at map scale (e.g., Earth's radius), ln is the natural log function, and λ_0 is the longitudinal origin (Greenwich meridian)

Inverse equation, from x, y to λ, ϕ:

$$\lambda = x/R + \lambda_0$$
$$\phi = (\pi/2) - 2 \cdot \tan^{-1}[e^{-y/R}]$$

Figure 3-31: Formulas are known for most projections that provide exact projected coordinates, if the latitudes and longitudes are known. This example shows the formulas defining the Mercator projection for a sphere.

Notice that there are parameters we must specify for this projection – here R, the Earth's radius, and λ_0, the longitudinal origin. Different values for these parameters give different values for the coordinates, so even though we may have the same kind of projection (transverse Mercator), we have different versions each time we specify different parameters.

Projection equations must also be specified in the "backward" direction, from projected coordinates to geographic coordinates, if they are to be useful. The projection coordinates in this backward, or "inverse," direction are often much more complicated than the forward direction, but are specified for every commonly used projection.

Most projection equations are much more complicated than the transverse Mercator, in part because most adopt an ellipsoidal Earth, and because the projections are onto curved surfaces rather than a plane. Thankfully, projection equations have long been standardized, documented, and made widely available through proven programming libraries and projection calculators.

Note that each projection defines a Cartesian coordinate system and hence creates *grid north*, a third version of the northern direction, in addition to geographic and magnetic norths. Grid north is the direction of the Y axis in a map projection, and often equals or nearly equals the direction of a meridian near the center of the projected area. Grid north is typically different from geographic and magnetic north for most of the projected region.

Most map projections may be viewed as sending rays of light from a projection source through the ellipsoid and onto a map surface (Figure 3-32). In some projections,

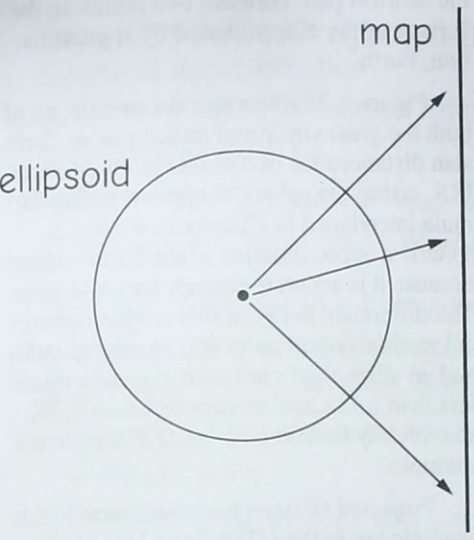


Figure 3-32: A conceptual view of a map projection.

the source is not a single point; however, the basic process involves the systematic transfer of points from the curved ellipsoidal surface to a flat map surface.

Distortions are unavoidable when making flat maps because of the transition from a complexly curved Earth surface to a flat or simply curved map surface. Portions of the rendered Earth surface must be compressed or stretched to fit onto the map. This is illustrated in Figure 3-33, a side view of a projection from an ellipsoid onto a plane.The map surface intersects the Earth at two locations, I_1 and I_2. Points toward the edge of the map surface, such as D and E, are stretched apart. The scaled map distance between d and e is greater than the distance from D to E measured on the surface of the Earth. More simply put, the distance along the map plane is greater than the corresponding distance along the curved Earth surface. Conversely, points such as A and B that lie in between I_1 and I_2 would appear compressed together. The scaled map distance from a to b would be less than the surface measured distance from A to B. Distortions at I_1 and I_2 are zero.

Figure 3-33 demonstrates a few important facts. First, distortion may differ in sense across the map. Parts of the map may have compressed areas or distances relative to the scaled Earth's surface measurements, while other parts may have expanded areas or distances. Second, there are often a few points or lines where distortions are zero and where length, direction, or some other geometric property is preserved. Finally, distortion is usually small near the points or lines of intersection, and increases with increasing distance from the points or lines of intersection.

Different map projections may distort the globe in different ways. The projection source, represented by the point at the middle of the circle in Figure 3-33, may change locations. We may project on to different shapes, and we may place the projection surface at different locations at or near the globe. If we change any of these three factors, we will change how or where our map is distorted. The type and amount of projection distortion may guide the selection of the appropriate projection or limit the area projected.

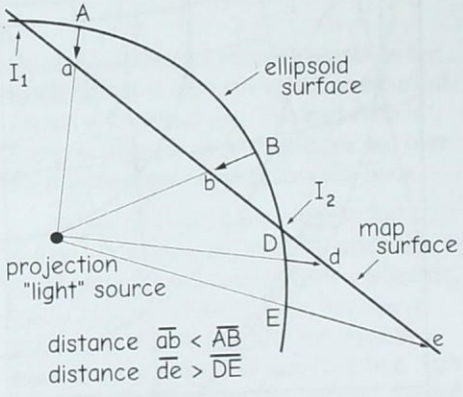


Figure 3-33: Distortion during map projection. This side view shows both expansion and compression of areas on a planar map.

Figure 3-34 shows an example of distortion with a projection onto a planar surface, but from above rather than the side view in Figure 3-33. This planar surface intersects the globe at a line of true scale, the solid circle shown in Figure 3-34. Distortion increases away from the line of true scale, with features inside the circle compressed or reduced in size,while features outside the standard circle are expanded. Calculations show a scale error of -1% near the center of the circle, and increasing scale error in concentric bands outside the circle to over 2% near the outer edges of the projected area.

An approximation of the distance distortion may be obtained for any projection by comparing grid coordinate distances to *great circle distances*. A great circle distance is defined on the surface of the spheroid or ellipsoid (Figure 3-35). The circle distance is the shortest path between two points on the surface of the ellipsoid, and by approximation, Earth.

Figure 3-35 illustrates the calculation of both the great circle and projection, or Cartesian distances for two points in the southern U.S., using the spherical approximation formula introduced in Chapter 2. We use a spherical approximation of the Earth's shape because it is accurate enough for illustration. The difference between this simpler spheroidal method (equal polar and equatorial radii) and an ellipsoidal method is typically much less than 0.1%, and always less than 0.3%, so typically less than 50 cm (1.5 feet) in our example.

Projected (Cartesian) coordinates in this example are in the UTM Zone 15N coordinate system, and derived from the appropriate coordinate transformation equations.

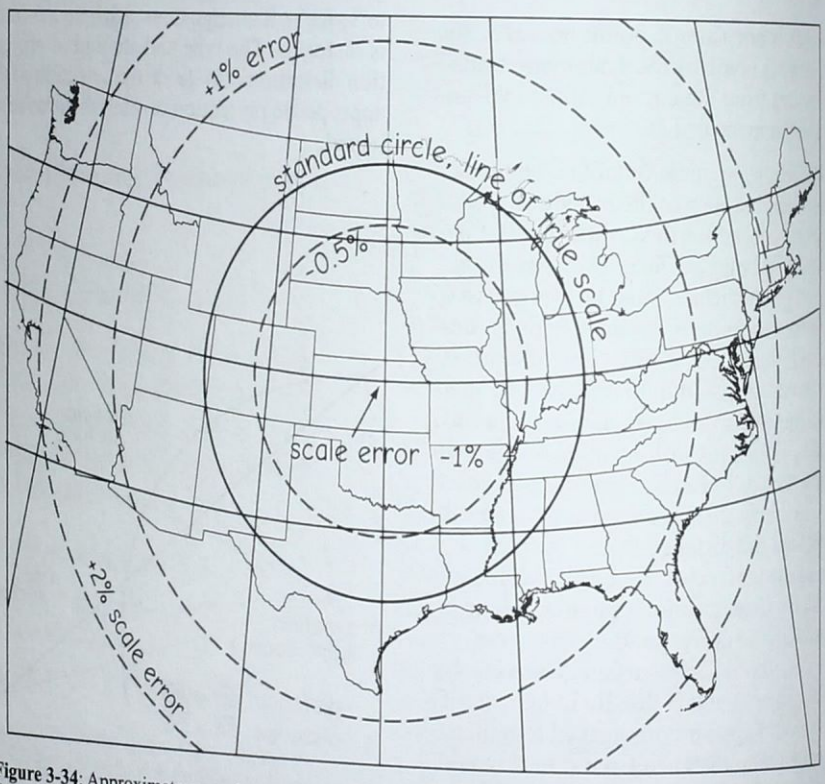


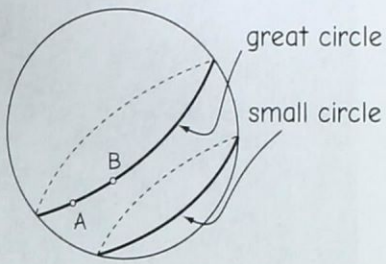
Figure 3-34: Approximate error due to projection distortion for a specific oblique stereographic projection. A plane intersects the globe at a standard circle. This standard circle defines a line of true scale, where there is no distance distortion. Distortion increases away from this line, and varies from -1% to over 2% in this example (adapted from Snyder, 1987).

Great Circle vs. Projected Distance

Spherical Approximation

Using the great circle formula from our example in Chapter 2,

A with latitude, longitude of (ϕ_A, λ_A), and
B, with latitude, longitude of (ϕ_B, λ_B)



The great circle distance from point A to point B is given by the formula:

A corresponding to Baton Rouge, LA = 30.4877456°, -91.1693348°
B corresponding to Houston, Texas = 29.7507171°, -95.370003°

$$d = 6378 \cdot 2\sqrt{\sin^{-1}[(\sin^2(0.368514)) + \cos(30.4877456) \cdot \cos(29.7507171) \cdot \sin^2(2.1003341)]}$$
$$= 412.681 \text{ km}$$

Grid distance (UTM Zone 15N coordinates):
Grid coordinates of Baton Rouge, LA = 675,708.2, 3,374,258.0
Grid coordinates of Houston, Texas = 270,816.1, 3,293,516.3

$$dg = [(X_A - X_B)^2 + (Y_A - Y_B)^2]^{0.5}$$
$$= [(675{,}708.2 - 270{,}816.1)^2 + (3{,}374{,}258.0 - 3{,}293{,}516.3)^2]^{0.5}$$
$$= 412.864 \text{ km}$$

distortion is 412.681 - 412.864 = -0.183 km, or a 183 meter lengthening

Figure 3-35: Example calculation of the distance distortion due to a map projection. The great circle and grid distances are compared for two points on the Earth's surface, the first measuring along the curved surface, the second on the projected surface. The difference in these two measures is the distance distortion due to the map projection. Calculations of the great circle distances are approximate, due to the assumption of a spheroidal rather than ellipsoidal Earth, but are at worst within 0.3% of the true value along the ellipsoid. Note that various great circle distance calculators are available via the World Wide Web, and these often don't specify the formula or Earth radius values used, so different great circle distances may be provided.

Armed with the coordinates for both pairs of points in both the geographic and projected coordinates, we can calculate the distance in the two systems, and subtract to find the length distortion due to projecting from the spherical surface to a flat surface.

Note that web-based or other software may use the ellipsoidal approximation, and may not specify the Earth radii used, so it is best to calculate the values from the original formulas when answers differ substantially.

A straight line between two points shown on a projected map is usually not a straight line nor the shortest path when traveling on the surface of the Earth. Conversely, the shortest distance between points on the Earth surface is likely to appear as a curved line on a projected map. The distortion is imperceptible for large scale maps and over short distances, but exists for most lines.

Figure 3-36 illustrates straight line distortion. This figure shows the shortest distance path (the great circle) between Seattle, USA, and Paris, France. Paris lies almost due east of Seattle, but the shortest path traces a route north of an east-west line. This shortest path is distorted and appears curved by the Plate Carrée projection commonly used for global maps.

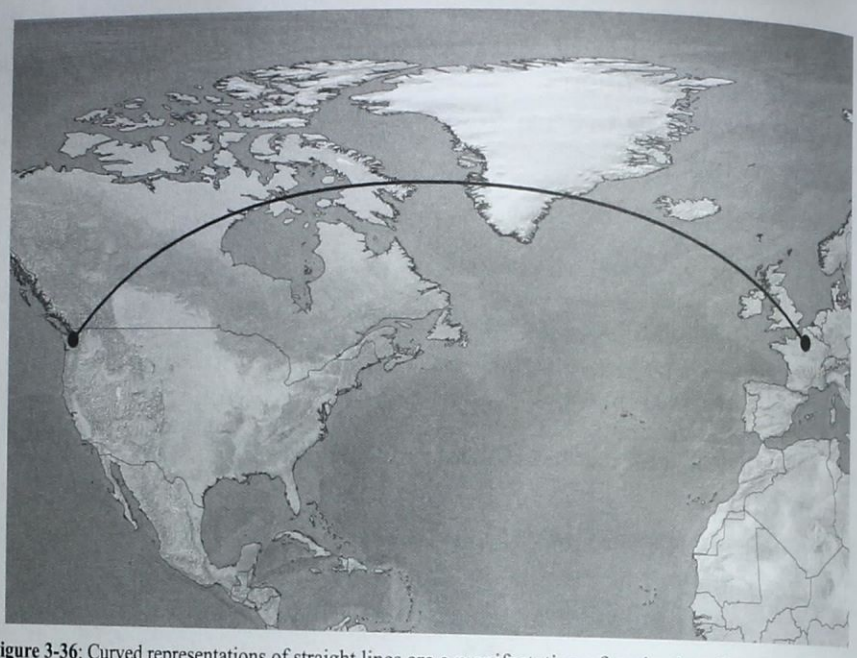


Figure 3-36: Curved representations of straight lines are a manifestation of projection distortion. A great circle path, shown above, is the shortest route when flying from Paris to Seattle, and commonly appears distorted when displayed.

Projections may also substantially distort the shape and area of polygons. Figure 3-37 shows various projections for Greenland, from an approximately "unprojected" view from space through geographic coordinates cast on a plane, to Mercator and transverse Mercator projections. Note the changes in size and shape of the polygon depicting Greenland.

Most map projections are based on a *developable surface*, a geometric shape onto which the Earth's surface is projected. Cones, cylinders, and planes are the most common developable surfaces. A plane is already flat, and cones and cylinders may be mathematically "cut" and "unrolled" to develop a flat surface (Figure 3-38). Projections may be characterized according to the shape of the developable surface, as *conic* (cone), *cylindrical* (cylinder), and *azimuthal* (plane). The orientation of the developable surface may also change among projections; for example, the axis of a cylinder may coincide with the poles (equatorial) or the axis may pass through the Equator (transverse).

Note that while the most common map projections used for spatial data in a GIS are based on a developable surface, many map projections are not. Projections with names such as pseudocylindrical, Mollweide, sinusoidal, and Goode homolosine are examples. These projections often specify a direct mathematical projection from an ellipsoid onto a flat surface. They use mathematical forms not related to cones, cylinders, planes, or other three-dimensional figures, and may change the projection surface for different parts of the globe, but generally are used only for display, and not for spatial analysis, because the coordinate systems are not strictly Cartesian.

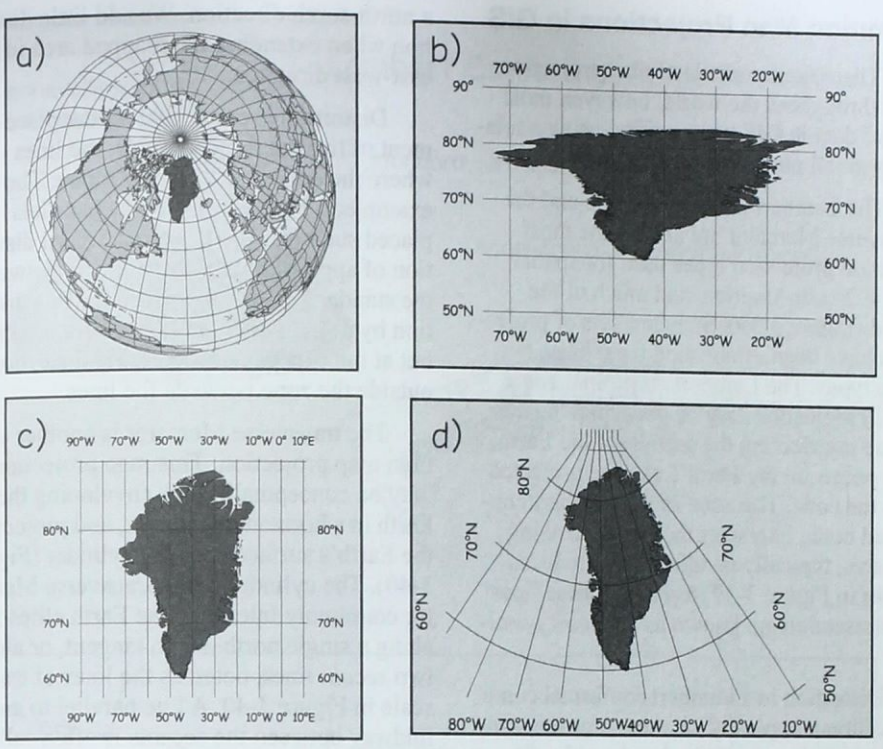


Figure 3-37: Map projections can distort the shape and area of features, as illustrated with these various projections of Greenland, from a) approximately unprojected, b) geographic coordinates on a plane, c) a Mercator projection, and d) a transverse Mercator projection.

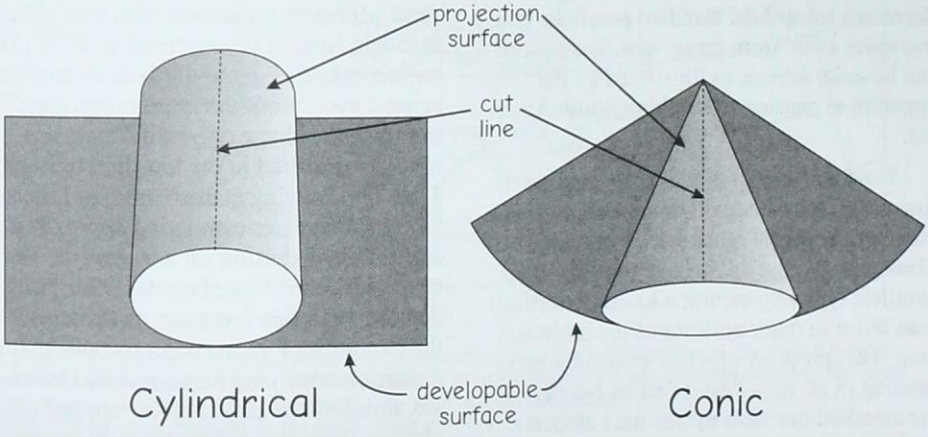


Figure 3-38: Projection surfaces are derived from curved "developable" surfaces that may be mathematically "unrolled" to a flat surface.

Common Map Projections in GIS

There are hundreds of map projections used throughout the world; however, most spatial data in GIS are specified using a relatively small number of projection types.

The Lambert conformal conic and the transverse Mercator are among the most common projection types used for spatial data in North America, and much of the world (Figure 3-39). Standard sets of projections have been established from these two basic types. The Lambert conformal conic (LCC) projection may be conceptualized as a cone intersecting the surface of the Earth, with points on the Earth's surface projected onto the cone. The cone in the Lambert conformal conic intersects the ellipsoid along two arcs, typically parallels of latitude, as shown in Figure 3-39 (top left). These lines of intersection are known as *standard parallels*.

Distortion in a Lambert conformal conic projection is typically smallest near the standard parallels, where the developable surface intersects Earth. Distortion increases in a complex fashion as distance from these parallels increases. This characteristic is illustrated at the top right and bottom of Figure 3-39. Circles of a constant 5-degree radius are drawn on the projected surface at the top right, and approximate lines of constant distortion and a line of true scale are shown in Figure 3-39, bottom. Distortion decreases toward the standard parallels, and increases away from these lines. Distortions can be quite severe, as illustrated by the apparent expansion of southern South America.

Note that sets of circles in an east-west row are distorted in the Lambert conformal conic projection (Figure 3-39, top right). Those circles that fall between the standard parallels typically exhibit a lower distortion than those in other portions of the projected map. This property of a low-distortion band running in an east-west direction between the standard parallels makes the Lambert conformal conic projection popular for mapping areas that are larger in an east-west than a north-south direction. We add little distortion when extending the mapped area in the east-west direction.

Distortion is controlled by the placement of the standard parallels, the lines where the cone intersects the globe. The example in Figure 3-39 shows parallels placed such that there is a maximum distortion of approximately 1% midway between the standard parallels. We reduce this distortion by moving the parallels closer together, but at the expense of increasing distortion outside the zone between the lines.

The transverse Mercator is another common map projection. This map projection may be conceptualized as enveloping the Earth in a horizontal cylinder, and projecting the Earth's surface onto the cylinder (Figure 3-40). The cylinder in the transverse Mercator commonly intersects the Earth ellipsoid along a single north–south tangent, or along two *secant* lines, noted as the lines of true scale in Figure 3-40. A line parallel to and midway between the secants is often called the *central meridian*. The central meridian extends north and south through transverse Mercator projections.

As with the Lambert conformal conic, the transverse Mercator projection has a band of low distortion, but this band runs in a north-south direction. Distortion is least near the line(s) of intersection. The graph at the top right of Figure 3-40 shows a transverse Mercator projection with the central meridian (line of intersection) at W96^0. Distortion increases markedly with distance east or west away from the intersection line; for example, the shape of South America is severely distorted in the top right of Figure 3-40. The drawing at the bottom of this same figure shows lines estimating approximately equal scale distortion for a transverse Mercator projection centered on the USA. Notice that the distortion increases as distance from the two lines of intersection increases. Scale distortion error may be maintained below any threshold by ensuring the mapped area is close to these two secant lines intersecting the globe. Transverse Mercator projections are often used for areas that extend in a

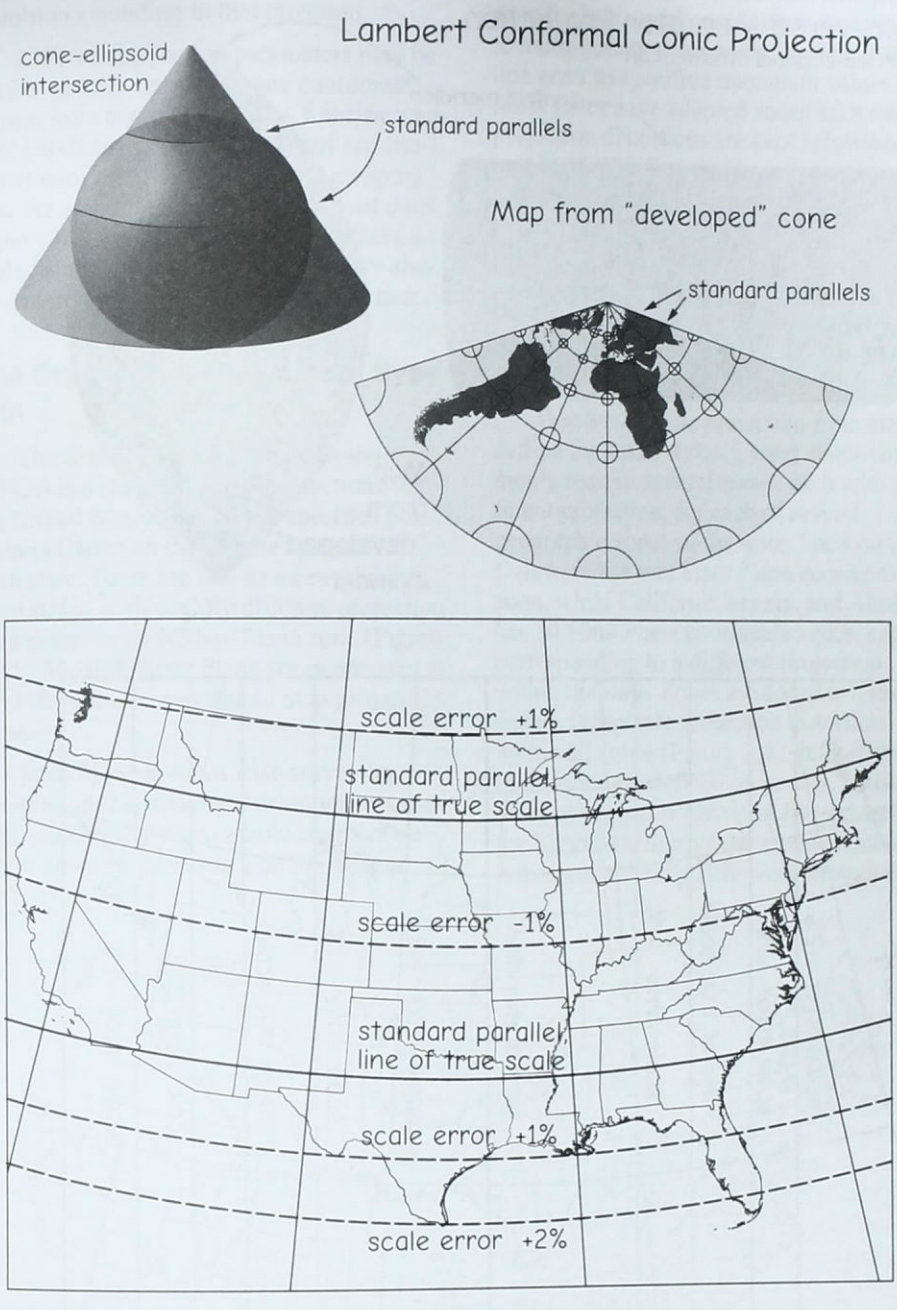


Figure 3-39: Lambert conformal conic (LCC) projection (top) and an illustration of the scale distortion associated with the projection. The LCC is derived from a cone intersecting the ellipsoid along two standard parallels (top left). The "developed" map surface is mathematically unrolled from the cone (top right). Distortion is primarily in the north-south direction, and is illustrated in the developed surfaces by the deformation of the 5-degree diameter geographic circles (top) and by the lines of approximately equal distortion (bottom). Note that there is no scale distortion where the standard parallels intersect the globe, at the lines of true scale (bottom, adapted from Snyder, 1987).

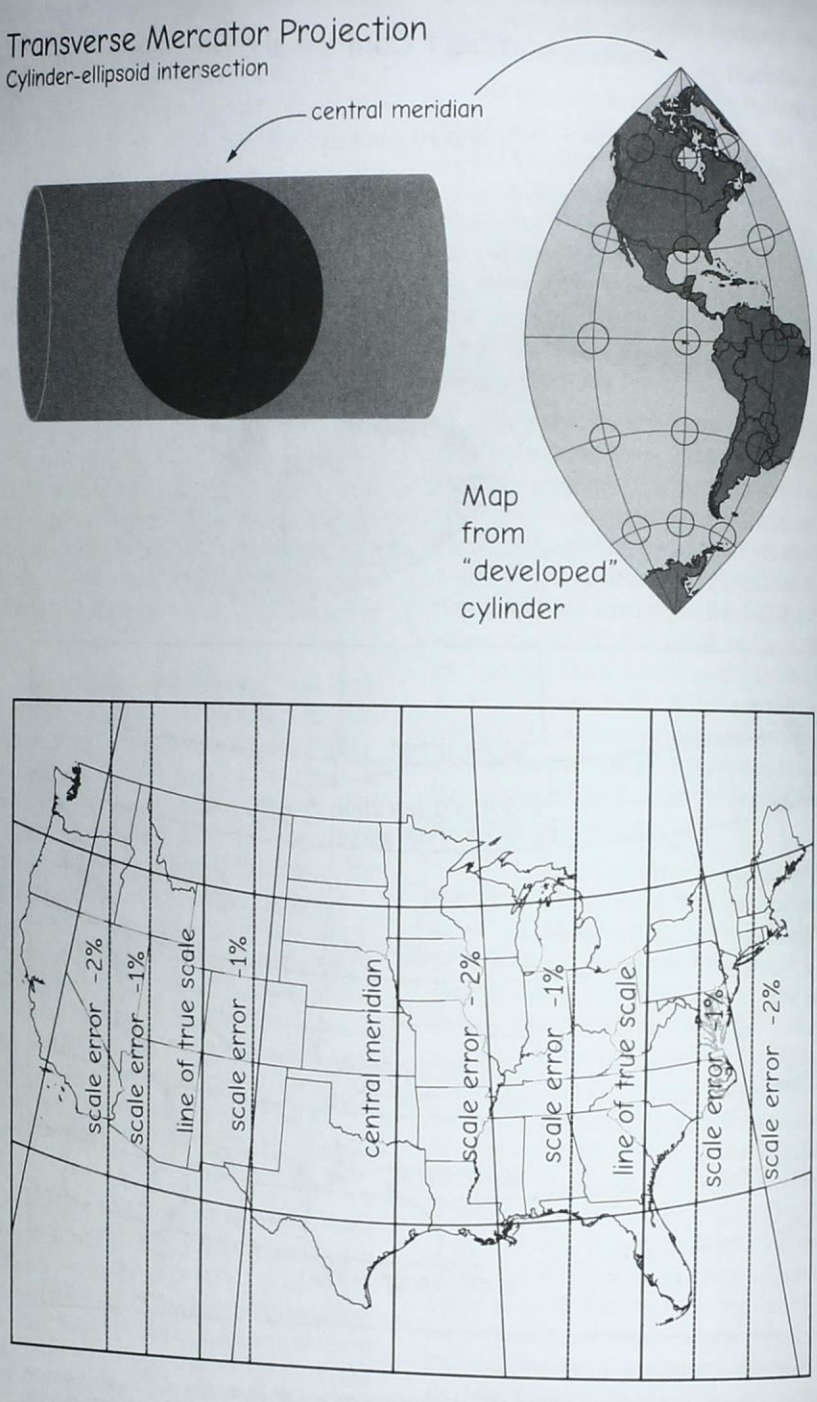


Figure 3-40: Transverse Mercator (TM) projection (top), and an illustration of the scale distortion associated with the projection (bottom). The TM projection distorts distances in an east-west direction, but has relatively little distortion in a north-south direction. This TM intersects the sphere along two lines, and distortion increases with distance from these lines (bottom, adapted from Snyder, 1987).

north-south direction, as there is little added distortion extending in that direction.

Different projection parameters may be used to specify an appropriate coordinate system for a region of interest. Specific standard parallels or central meridians are chosen to minimize distortion over a mapping area. An origin location, measurement units, x and y (or northing and easting) offsets, a scale factor, and other parameters may also be required to define a specific projection.

The State Plane Coordinate System

The State Plane Coordinate System (SPCS) is a standard set of projections for the United States. The SPCS specifies positions in Cartesian coordinate systems for each state. There are one or more zones in most states, with slightly different projection parameters in each State Plane zone (Figure 3-41). Multiple State Plane zones are used to limit distortion errors due to map projections.

State Plane systems ease surveying, mapping, and spatial data development in a GIS, particularly when whole counties or larger areas are covered. The State Plane system provides a common coordinate reference for horizontal coordinates over county to multi-county areas while limiting distortion error to specified maximum values. Most states have adopted zones such that projection distortions are kept below one part in 10,000. Some states allow larger distortions (e.g., Montana, Nebraska) for the sake of having only one state plane zone. SPCSs are used in many types of work, including property surveys, property subdivisions, large-scale construction projects, and photogrammetric mapping, and the zones and SPCSs are often adopted for GIS.

One State Plane projection zone may suffice for small states. Larger states commonly require several zones, each with a different projection, for each of several geographic zones of the state. For example, Delaware has one State Plane coordinate zone, while California has six, and Alaska has 10 State Plane coordinate zones, each corresponding to a different projection within the state. Zones are added to a state to ensure acceptable projection distortion within all zones (Figure 3-42, left). Zone boundaries are defined by county, parish, or other municipal boundaries. For example, the Minnesota south/central zone boundary runs approximately east–west through the

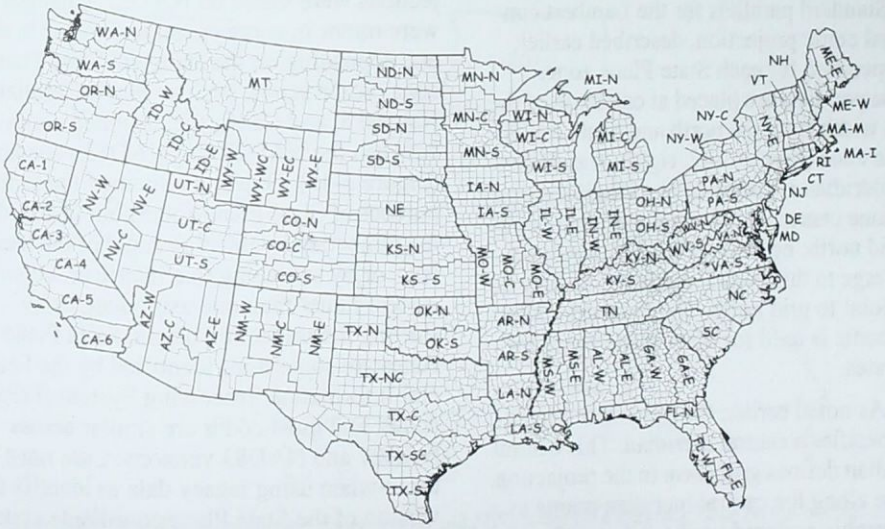


Figure 3-41: State plane zone boundaries, NAD83.

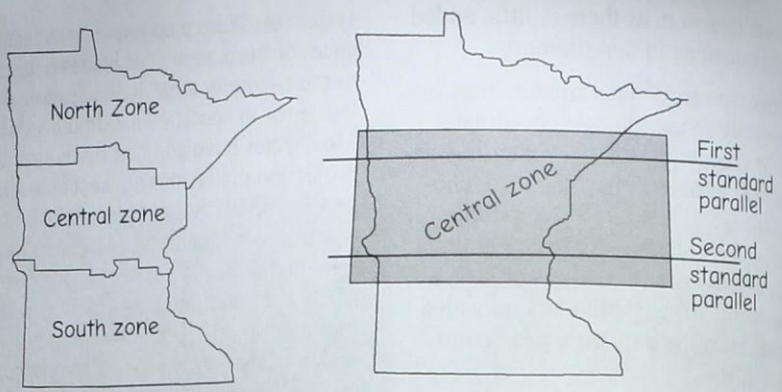


Figure 3-42: The State Plane zones of Minnesota, and details of the standard parallel placement for the Minnesota central State Plane zone.

state along defined county boundaries (Figure 3-42, left).

Most State Plane coordinate systems are based on one of two types of map projections: the Lambert conformal conic or the transverse Mercator projections. Because distortion in a transverse Mercator increases with distance from the central meridian, this projection type is most often used with states or zones that have a long north-south axis (e.g., Illinois or New Hampshire). Conversely, a Lambert conformal conic projection is most often used when the long axis of a state or zone is in the east-west direction (examples are North Carolina and Virginia).

Standard parallels for the Lambert conformal conic projection, described earlier, are specified for each State Plane zone. These parallels are placed at one-sixth of the zone width from the north and south limits of the zone (Figure 3-42, right). A zone central meridian is specified at a longitude near the zone center. This central meridian points at grid north; however, all other meridians converge to this central meridian, so they do not point to grid north. The Lambert conformal conic is used for State Plane zones for 31 states.

As noted earlier, the transverse Mercator specifies a central meridian. This central meridian defines grid north in the projection. A line along the central meridian points to geographic and grid north, and specifies the Cartesian grid direction for the map projection. All parallels of latitude and all meridians except the central meridian are curved for a transverse Mercator projection, and hence these lines do not parallel the grid x or y directions. The transverse Mercator is used for 22 State Plane systems (the sum of states is greater than 50 because both the transverse Mercator and Lambert conformal conic are used in some states, e.g., Florida).

Finally, note that more than one version of the State Plane coordinate system has been defined. Changes were introduced with the adoption of the North American Datum of 1983. Prior to 1983, the State Plane projections were based on NAD27. Changes were minor in some cases, and major in others, depending on the state and State Plane zone. Some states, such as South Carolina, Nebraska, and California, dropped zones between the NAD27 and NAD83 versions (Figure 3-43). Others maintained the same number of State Plane zones, but changed the projection by the placement of the meridians, or by switching to a metric coordinate system rather than one using feet, or by shifting the projection origin. State Plane zones are sometimes identified by the Federal Information Processing System (FIPS) codes, and most codes are similar across NAD27 and NAD83 versions. Care must be taken when using legacy data to identify the version of the State Plane coordinate system used because the FIPS and State Plane zone

designators may be the same, but the projection parameters may have changed from NAD27 to NAD83.

Conversion among State Plane projections may be further confused by the various definitions used to translate from feet to meters. The metric system was first developed during the French Revolution in the late 1700s, and it was adopted as the official unit of distance in the United States, by the initiative of Thomas Jefferson. President Jefferson was a proponent of the metric system because it improved scientific measurements, was based on well-defined, integrated units, reduced commercial fraud, and improved trade within the new nation. The conversion was defined in the United States as one meter equal to exactly 39.97 inches. This yields a conversion for a *U.S. survey foot* of:

1 foot = 0.3048006096012 meters

Unfortunately, revolutionary tumult, national competition, and scientific differences led to the eventual adoption of a different conversion factor in Europe and most of the rest of the world. They adopted an *international foot* of:

1 foot = 0.3048 meters

The U.S. definition of a foot is slightly longer than the European definition, by about one part in five million. Both conversions are used in the U.S., and the international conversion elsewhere. The European conversion was adopted as the standard for all measures under an international agreement in the 1950s. However, there was a long history of the use of the U.S. conversion in U.S. geodetic and land surveys. Therefore, the U.S. conversion was called the U.S. survey foot. This slightly longer metric-to-foot conversion factor should be used as the default for conversions among geodetic coordinate systems within the United States, for example, when converting from a State Plane coordinate system specified in feet to one specified in meters.

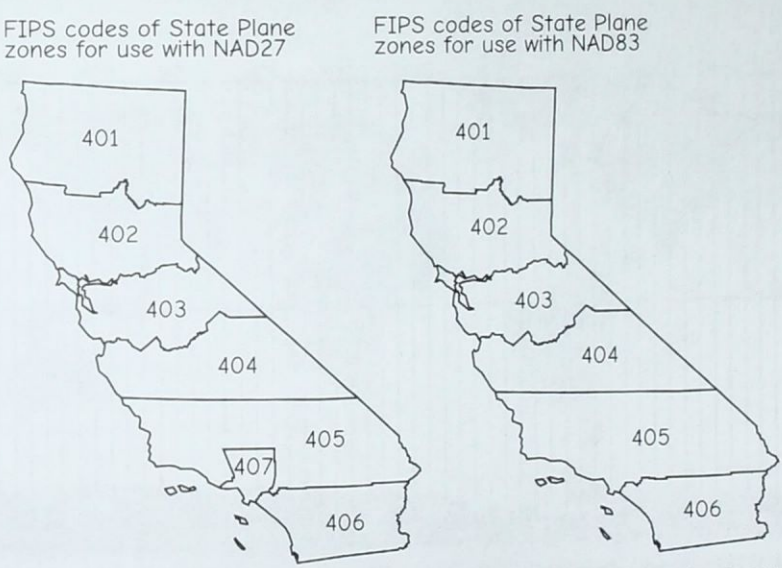


Figure 3-43: State Plane coordinate system zones and FIPS codes for California based on the NAD27 and NAD83 datums. Note that zone 407 from NAD27 is incorporated into zone 405 in NAD83.

Universal Transverse Mercator Coordinate System

The Universal Transverse Mercator (UTM) coordinate system is another standard, distinct from the State Plane system. The UTM is a global coordinate system, based on the transverse Mercator projection. It is widely used in the United States and other parts of North America, and is also used in many other countries.

The UTM system divides the Earth into zones that are 6 degrees wide in longitude and extend from 80 degrees south latitude to 84 degrees north latitude. UTM zones are numbered from 1 to 60 in an easterly direction, starting at longitude 180 degrees West (Figure 3-44). Zones are further split north and south of the Equator. Therefore, the zone containing most of England is identified as UTM Zone 30 North, while the zones containing most of New Zealand are designated UTM Zones 59 South and 60 South. Directional designations are here abbreviated, for example, 30N in place of 30 North.

Distances in the UTM system are specified in meters north and east of a zone origin (Figure 3-45). The y values are known as *UTM northings*, and increase in a northerly direction. The x values are referred to as *UTM eastings* and increase in an easterly direction.

The origins of the UTM coordinate system are defined differently depending on whether the zone is north or south of the Equator. In either case, the UTM coordinate system is defined so that all coordinates are positive within the zone. Zone easting coordinates are all greater than zero because the central meridian for each zone is assigned an easting value of 500,000 meters. This effectively places the origin (E = 0) at a point 500,000 meters west of the central meridian. All zones are less than 1,000,000 meters wide, ensuring that all eastings will be positive.

The Equator is used as the northing origin for all north zones. Thus, the Equator is assigned a northing value of zero for north zones. This avoids negative coordinates, because all of the UTM north zones are defined to be north of the Equator.

Universal Transverse Mercator zones south of the Equator are slightly different than those north of the Equator (Figure 3-46). South zones have a *false northing* value added to ensure all coordinates within a zone

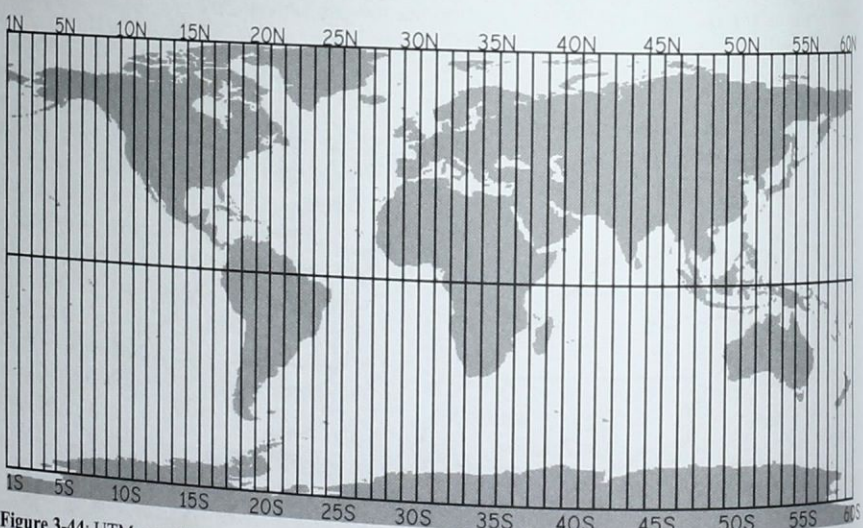


Figure 3-44: UTM zone boundaries and zone designators. Zones are six degrees wide and numbered from 1 to 60 from the International Date Line, 180°W. Zones are also identified by their position north and south of the Equator, e.g., Zone 7 North, Zone 16 South.

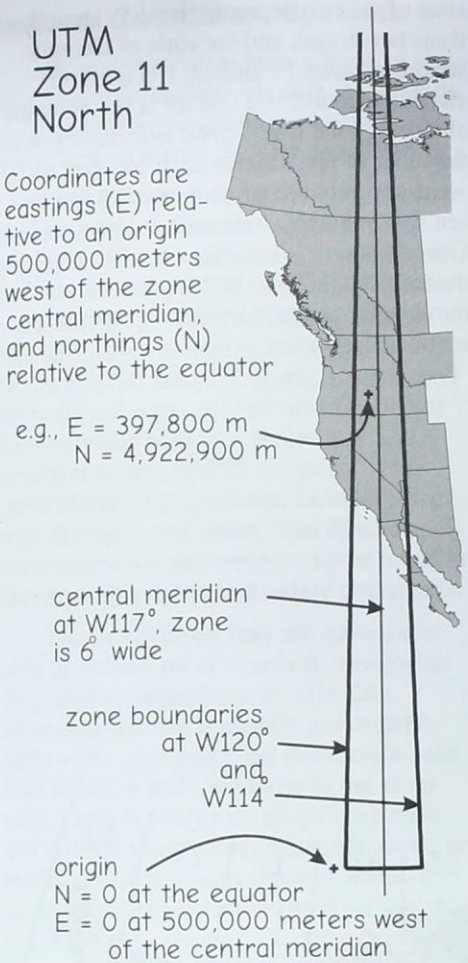


Figure 3-45: UTM zone 11N. The zone origin is on the Equator, with a false easting of 500,000 to ensure positive coordinates throughout the zone.

are positive. UTM coordinate values increase as one moves from south to north in a projection area. If the origin were placed at the Equator with a value of zero for south zone coordinate systems, then all the northing values would be negative. An offset is applied by assigning a false northing, a non-zero value, to an origin or other appropriate location. For UTM south zones, the northing values at the Equator are set to equal 10,000,000 meters. Because the distance from the Equator to the most southerly point in a UTM south zone is less than 10,000,000 meters, this assures that all northing coordinate values will be positive within each UTM south zone (Figure 3-46).

The UTM coordinate system is common for data and study areas spanning large regions, for example, several State Plane zones. Many data from U.S. federal government sources are in a UTM coordinate system because many agencies manage large areas. Many state government agencies in the United States distribute data in UTM coordinate systems because the entire state

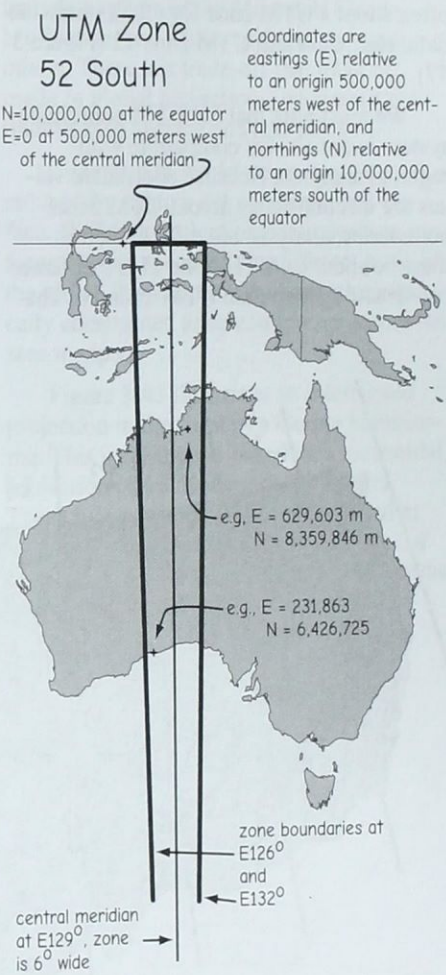


Figure 3-46: UTM south zones are defined to maintain positive northing and easting values within the zone. To that end, a false northing of 10,000,000 is applied to the Equator, and a false easting of 500,000 is applied to the central meridian.

fits predominantly or entirely into one UTM zone.

As noted before, all data for an analysis area must be in the same coordinate system if they are to be analyzed together. If not, the data will not co-occur as they should. The large width of the UTM zones accommodates many large-area analyses, and many states, national forests, or multicounty agencies have adopted the dominant UTM coordinate system as a standard. States that fall predominantly or entirely within a zone often adopt a UTM zone for much statewide data, e.g., Utah and UTM zone 12 (Figure 3-47).

We must note that the UTM coordinate system is not always compatible with regional analyses. Because coordinate values are discontinuous across UTM zone boundaries, analyses are difficult across these boundaries. UTM zone 15 is a different coordinate system than UTM zone 16. The state of Wisconsin approximately straddles these two zones, and the state of Georgia straddles zones 16 and 17. If a uniform, statewide coordinate system is required, the choice of zone is not clear, and either one or the other of these zones must be used, or some compromise projection must be chosen. For example, statewide analyses in Georgia and in Wisconsin are often conducted using UTM-like systems that involve moving the central meridian to near the center of each state.

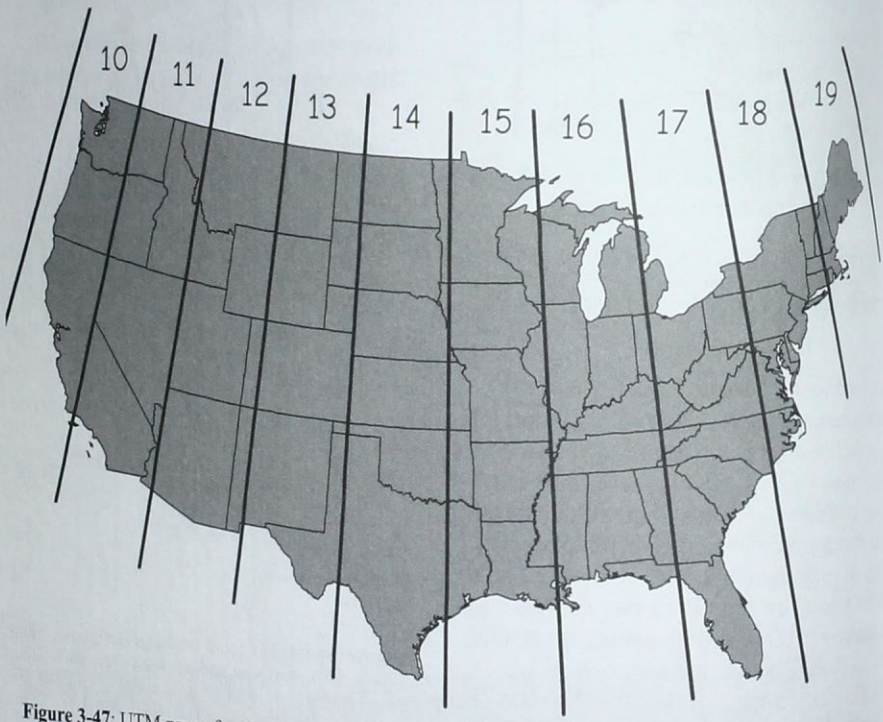


Figure 3-47: UTM zones for the lower 48 contiguous states of the United States of America. Each UTM zone is 6 degrees wide. All zones in the Northern Hemisphere are north zones, e.g., Zone 10 North.

National Coordinate Systems

Many governments have adopted a standard project for nationwide data, particularly small and midsized countries where distortion is limited across the spanned distances.

Many European countries have standard map projections covering a national extent; for example, Belgium, Estonia, and France each have different Lambert Conformal Conic projections defined for use on standard nation-spanning maps and data sets, while Germany, Bulgaria, Croatia, and Slovenia use a specialized modification of the transverse Mercator projection. Some countries adopt specific Universal Transverse Mercator projections, including Norway, Portugal, and Spain. Specifications of these projection parameters may be found in the respective national standard documents.

Larger countries may not have a specific or unified set of standard, nationwide projections, particularly for GIS data, because distortion is usually unavoidably large when spanning great distances across both latitudes and longitudes in the same map. There is simply no single projection that faithfully represents distances, areas, or angles across the entire country, so more constrained projections are used for analysis, and the results aggregated to larger areas.

Continental and Global Projections

There are map projections that are commonly used when depicting maps of the world. Directions, distances, and areas are typically not measured or computed on them, as distortions are too great. Most worldwide projections are used for visualization, but not quantitative analysis.

There are a number of projections that have been widely used for the world. These include variants of the Mercator, Goode, Mollweide, and Miller projections, among others. There is a trade-off that must be made in global projections, between a continuous map surface and distortion.

Distortion in world maps may be reduced by using a cut or interrupted surface. Different projection parameters or surfaces may be specified for different parts of the globe. Projections may be mathematically constrained to be continuous across the area mapped.

Figure 3-48 illustrates an interrupted projection in the form of a Goode homolosine. This projection is based on a sinusoidal projection and a Mollweide projection. These two projection types are merged at parallels of identical scale. The parallel of identical scale in this example is set near the midnorthern latitude of $44^{o}\ 40'$ N.

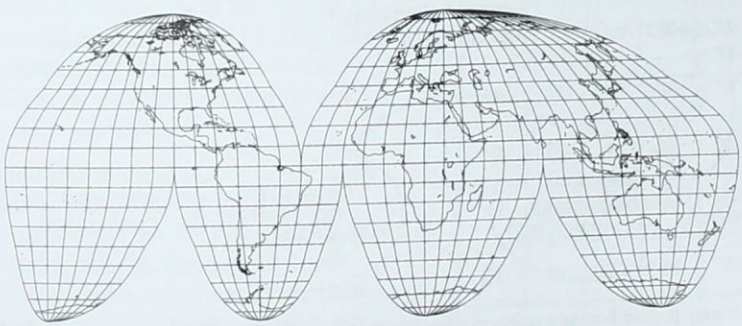


Figure 3-48: A Goode homolosine projection. This is an example of an interrupted projection, often used to reduce some forms of distortion when displaying the entire Earth surface (from Snyder and Voxland, 1989).

Conversion Among Coordinate Systems

Conversion from one projected coordinate system to another requires using the inverse and forward projection equations, described in an earlier section, passing through the geographic coordinate set. This allows a flexible conversion between any two projections, given our requirement that both the forward and inverse, or "backward" projection equations are specified for any map projection. For example, given a coordinate pair in the State Plane system, you may calculate the corresponding geographic coordinates. You may then apply a formula that converts geographic coordinates to UTM coordinates for a specific zone using another set of equations. Since the backward and forward projections from geographic to projected coordinate systems are known, we may convert among most coordinate systems by passing through a geographic system (Figure 3-49, a).

Care must be taken when converting among projections that use different datums. If appropriate, we must insert a datum transformation when converting from one projected coordinate system to another (Figure 3-49, b). A datum transformation, described earlier in this chapter, is a calculation of the change in geographic coordinates when moving from one datum to another.

Users of GIS software should be careful when applying coordinate projection tools

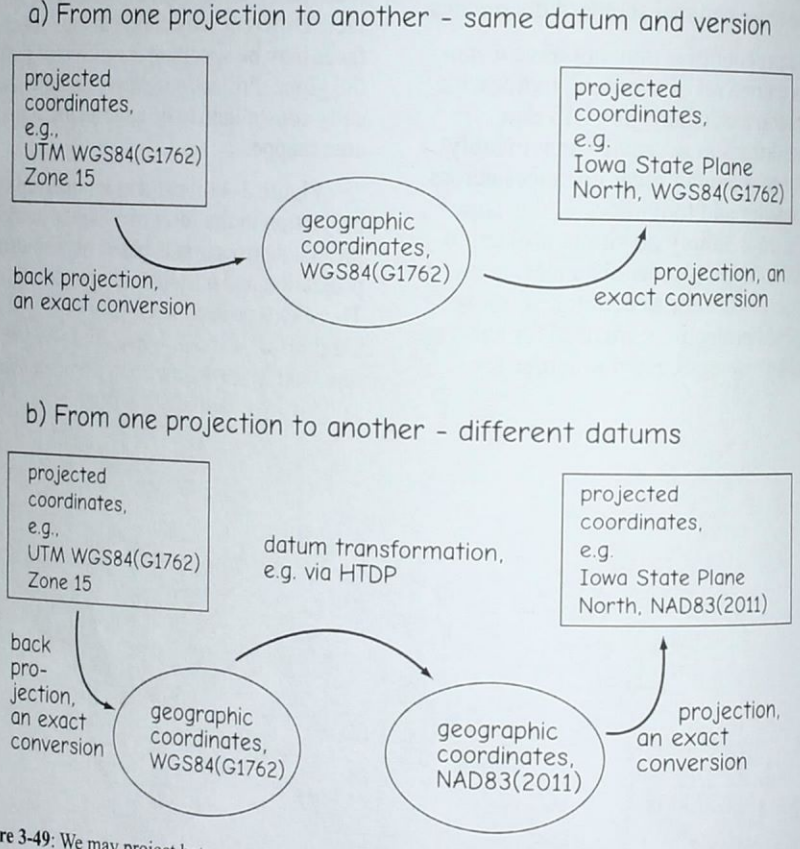


Figure 3-49: We may project between most coordinate systems via the back (or inverse) and forward projection equations. These calculate exact geographic coordinates from projected coordinates (a), and then new projected coordinates from the geographic coordinates. We must insert an extra step when a projection conversion includes a datum change. A datum transformation must be used to convert from one geodetic datum to another (b).

because the datum transformation may be omitted, or an inappropriate datum manually or automatically selected. For some software, the projection tool does not check or maintain information on the datum of the input spatial layer. This will often lead to an inappropriate or no datum transformation, and the output from the projection will be in error. Often these errors are small relative to other errors, for example, spatial imprecision in the collection of the line or point features. As shown in Figure 3-23, errors between NAD83(1986) and NAD83(CORS96) may be less than 10 cm (4 inches) in some regions, often much less than the average spatial error of the data themselves. However, errors due to ignoring the datum transformation may be quite large, for example, tens to hundreds of meters between NAD27 and most versions of NAD83, and errors of up to a meter are common between recent versions of WGS84/ITRF and NAD83. Given the sub-meter accuracy of many new GPS and other GNSS receivers used in data collection, datum transformation error of one meter is significant. As data collection accuracy improves, users develop applications based on those accuracies, so datum transformation errors should be avoided in all cases.

The Public Land Survey System

For the benefit of GIS practitioners in the United States, we must cover one final land designation system, known as the *Public Land Survey System*, or PLSS. The PLSS is not a coordinate system, but PLSS points are often used as reference points in the United States, so the PLSS should be well understood for work there.

The PLSS divided lands by north-south lines, 6 miles apart, running parallel to a principal meridian. East-west lines were surveyed perpendicular to these north-south lines, also at six mile intervals. These lines form square townships. Each township was further subdivided into 36 sections, each section approximately a mile on a side. Each section was subdivided further, to quarter-sections (one-half mile on a side), or sixteenth sections (one-quarter mile on a side). Sections were numbered in a zigzag pattern from one to 36, beginning in the northeast corner (Figure 3-50).

The PLSS is a standardized method for designating and describing the location of land parcels. It was used for the initial surveys over most of the United States after the early 1800s; therefore, nearly all land outside the original thirteen colonies uses the PLSS. An approximately uniform grid system was established across the landscape, with periodic adjustments incorporated to account for the anticipated error. Parcels were designated by their location within this grid system.

The PLSS was developed for a number of reasons. First, it was seen as a method to remedy many of the shortcomings of *metes and bounds* surveying, the most common method for surveying prior to the adoption of the PLSS. Metes and bounds describe a parcel relative to features on the landscape, sometimes supplemented with angle or distance measurements. Metes and bounds was used in colonial times, but parcel descriptions were often ambiguous. Subdivided parcels were often poorly described, and hence the source of much litigation, ill will, and many questionable real estate transactions.

6	5	4	3	2	1
7	8	9	10	11	12
18	17	16	15	14	13
19	20	21	22	23	24
30	29	28	27	26	25
31	32	33	34	35	36

Figure 3-50: Typical layout and section numbering of a PLSS township.

The U.S. government needed a system that would provide unambiguous descriptions of parcels in unsettled territories west and south of the original colonies. The federal government saw public land sales as a way to generate revenue, to pay revolutionary war veterans, to expand the country, and to protect against encroachment by European powers. Parcels could not be sold until they were surveyed, so the PLSS was created. Land surveyed under the PLSS can be found in 30 states, including Alaska and most of the midwestern and western United States. Lands in the original 13 colonies, as well as West Virginia, Tennessee, Texas, and Kentucky were not surveyed under the PLSS system.

Surveyors typically marked the section corners and quarter-corners while running survey lines. Points were marked by a number of methods, including stone piles, pits, blaze marks chiseled in trees, and pipes or posts sunk in the ground.

Because the primary purpose of the PLSS survey was to identify parcels, lines and corner locations were considered static on completion of the survey, even if the corners were far from their intended location. Survey errors were inevitable given the large areas and number of different survey parties involved. Rather than invite endless dispute and readjustment, the PLSS specifies that boundaries established by the appointed PLSS surveyors are unchangeable, and that township and section corners must be accepted as true. The typical section contains approximately 640 acres, but due in part to errors in surveying, sections larger than 1200 acres and smaller than 20 acres were also established (Figure 3-51).

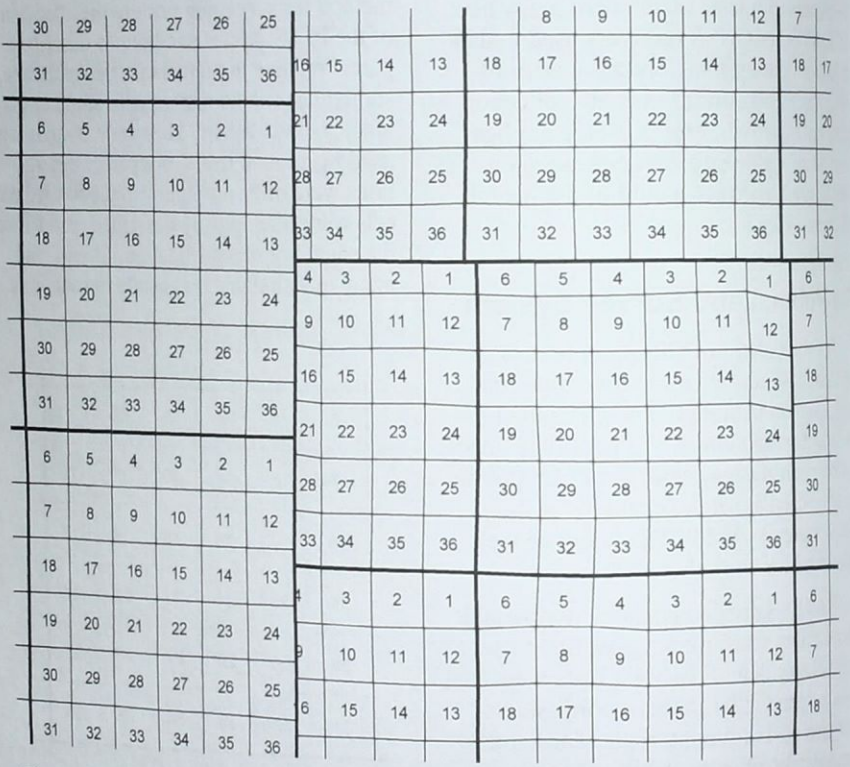


Figure 3-51: Example of variation in the size and shape of PLSS sections. Most sections are approximately one mile square with section lines parallel or perpendicular to the primary meridian, as illustrated by the township in the upper left of this figure. However, adjustments due to different primary meridians, different survey parties, and errors result in irregular section sizes and shapes.

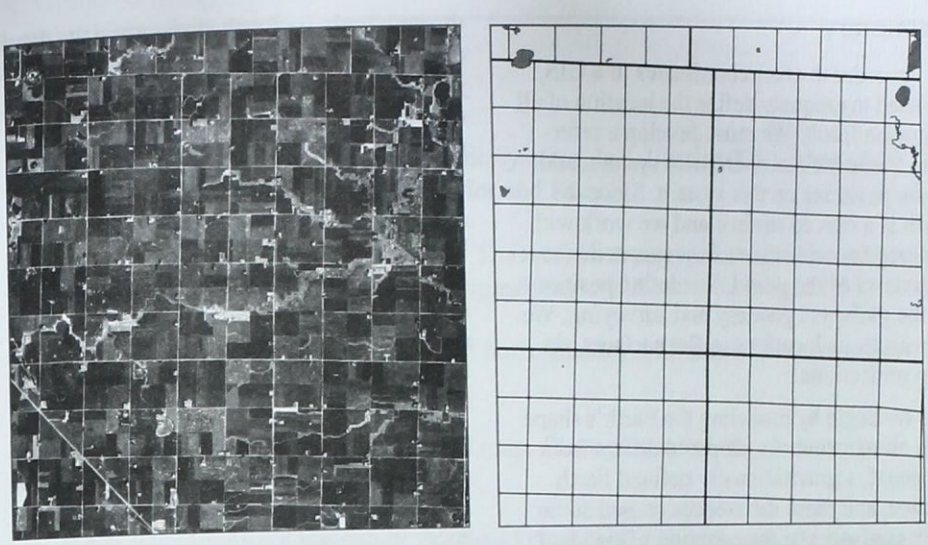


Figure 3-52: PLSS lines are often visible on the landscape. Roads (light lines on the image, above left) often follow the section and township lines (above right).

The PLSS is important today for several reasons. First, since PLSS lines are often property boundaries, they form natural corridors in which to place roads, powerlines, and other public services; they are often evident on the landscape (Figure 3-52). Many road intersections occur at PLSS corner points, and these can be viewed and referenced on many maps or imagery used for GIS database development efforts. Thus, the PLSS often forms a convenient system to coregister GIS data layers. PLSS corners and lines are often plotted on government maps (e.g., 1:24,000 quads) or available as digital data (e.g., National Cartographic Information Center Digital Line Graphs). Further, PLSS corners are sometimes resurveyed using high precision methods to provide property line control, particularly when a GIS is to be developed (Figure 3-53). These points may be useful to properly locate and orient spatial data layers on the Earth's surface.

Figure 3-53: A PLSS corner that has been surveyed and marked with a monument. This monument shows the physical location of a section corner. These points are often used as control points for further spatial data development.

Summary

In order to enter coordinates in a GIS, we need to uniquely define the location of all points on Earth. We must develop a reference frame for our coordinate system, and locate positions on this system. Since the Earth is a curved surface and we work with flat maps, we must somehow reconcile these two views of the world. We define positions on the globe via geodesy and surveying. We convert these locations to flat surfaces via map projections.

We begin by modeling the Earth's shape with an ellipsoid. An ellipsoid differs from the geoid, a gravitationally defined Earth surface, and these differences caused some early confusion in the adoption of standard global ellipsoids. There is a long history of ellipsoidal measurement, and we have arrived at our best estimates of global and regional ellipsoids after collecting large, painstakingly developed sets of precise surface and astronomical measurements. These measurements are combined into datums, and these datums are used to specify the coordinate locations of points on the surface of the Earth.

Map projections are a systematic rendering of points from the curved Earth surface onto a flat map surface. While there are many purely mathematical or purely empirical map projections, the most common map projections used in GIS are based on developable surfaces. Cones, cylinders, and planes are the most common developable surfaces. A map projection is constructed by passing rays from a projection center through both the Earth surface and the developable surface. Points on the Earth are projected along the rays and onto the developable surface. This surface is then mathematically unrolled to form a flat map.

Standard sets of projections are commonly used for spatial data in a GIS. In the United States, the UTM and State Plane coordinate systems define a standard set of map projections that are widely used. Other map projections are commonly used for continental or global maps, and for smaller maps in other regions of the world.

A datum transformation is often required when performing map projections. Datum transformations account for differences in geographic coordinates due to changes in the shape or origin of the spheroid, and in some cases to datum adjustments. Datum transformation should be applied as a step in the map projection process when input and output datums differ.

A system of land division known as the Public Land Survey System (PLSS) was established in the United States. This is not a coordinate system, but rather a method for unambiguously and systematically defining parcels of land based on regularly spaced survey lines in approximately north-south and east-west directions. Intersection coordinates have been precisely measured for many of these survey lines, and are often used as a reference grid for further surveys or land subdivision.

Suggested Reading

Bossler, J.D. (2002). Datums and geodetic systems. In J. Bossler (Ed.), *Manual of Geospatial Technology.* London: Taylor and Francis.

Brandenburger, A.J., Gosh, S K. (1985). The world's topographic and cadastral mapping operations. *Photogrammetric Engineering and Remote Sensing*, 51:437-444.

Burkholder, E.F. (1993). Computation of horizontal/level distances. *Journal of Surveying Engineering*, 117:104-119.

Colvocoresses, A.P. (1997). The gridded map. *Photogrammetric Engineering and Remote Sensing*, 63:371-376.

Doyle, F.J. (1997). Map conversion and the UTM Grid. *Photogrammetric Engineering and Remote Sensing*, 63:367-370.

Elithorpe, J.A.Jr., Findorff, D.D. (2009). *Geodesy for Geomatics and GIS Professionals*. Acton: Copley Custom Textbooks.

Featherstone, W.E., Kuhn, M. (2006). Height systems and vertical datums: a review in the Australian context. *Journal of Spatial Science*, 51:21-41.

Flacke, W., Kraus, B. (2005). *Working with Projections and Datum Transformations in ArcGIS: Theory and Practical Examples*. Norden: Points Verlag.

Habib, A. (2002). Coordinate transformation. In J. Bossler (Ed.), *Manual of Geospatial Technology*. London: Taylor and Francis.

Iliffe, J.C., Lott, R. (2008). *Datums and Map Projections for Remote Sensing, GIS, and Surveying*. 2nd ed. Boca Raton: CRC Press.

International Association of Oil and Gas Producers (2016). *Coordinate Conversion and Transformations including Formulas. Geomatics Guidance Note Number 7, Part 12*. www.epsg.org.

Janssen, V. (2009). Understanding coordinate reference systems, datums, and transformations. *International Journal of Geoinformatics*, 5:41-53.

Keay, J. (2000). *The Great Arc*. New York: Harper Collins.

Leick, A. (1993). Accuracy standards for modern three-dimensional geodetic networks. *Surveying and Land Information Systems,* 53:111-127.

Maling, D.H. (1992). *Coordinate Systems and Map Projections*. London: George Phillip.