3 Geodesy, Datums, Map Projections, and Coordinate Systems

Introduction

Geographic information systems are different from other information systems because they contain spatial data. These spatial data include coordinates that define the location, shape, and extent of geographic objects. To effectively use GIS, we must develop a clear understanding of how coordinate systems are established and how coordinates are measured. 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 two main factors. First, most people best understand geography in a Cartesian coordinate system on a flat surface. Humans naturally perceive the Earth's surface as flat, because at human scales the Earth's curvature is barely perceptible. Humans have been using flat maps for more than 40 centuries, and although globes are quite useful for perception and visualization at extremely small scales, they are not practical 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 distortion is quite apparent on large-area maps. Because measurements on maps are affected by the distortion, we must somehow reconcile the portrayal of the Earth's truly 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. These deformations affect how we best map the surface of the Earth, and how we define Cartesian coordinate systems for mapping and GIS.

Early Measurements

Humans have long speculated on the shape and size of the Earth. Babylonians believed the Earth was a flat disk floating in an endless ocean, a notion adopted by Homer, one of the more widely known Greek writers. The Greeks were early champions of geometry, and they had many competing views of the shape of the Earth. One early Greek, Anaximenes, believed the Earth was a rectangular box, while Pythagoras and later Aristotle reasoned that the Earth must be a sphere. Their deduction was based on many lines of evidence, and also on a belief of divine direction by the gods. Pythagoras believed the sphere was

the most perfect geometric figure, and that the gods would use this perfect shape for what he considered their greatest creation. Aristotle also observed that ships disappeared over the horizon, the moon appeared to be a sphere, and that the stars moved in circular patterns, and he noted reports from wandering fishermen on shifts in the constellations. These observations were all consistent with a spherical earth.

After scientific support for a spherical earth became entrenched, Greek scientists turned toward estimating the size of the sphere. Eratosthenes, a Greek scholar in Egypt, performed one of the earliest well-founded measurements of the Earth's circumference. He noticed that on the summer solstice the Sun at noon shone to the bottom of a deep well in Syene. He believed that the well was located on the Tropic of Cancer, so that the Sun would be exactly overhead during the summer solstice. He also observed that 705 km north in Alexandria, at exactly the same date and time, a vertical post cast a shadow. The shadow/post combination defined an angle which was about $7^o12\tilde{0}$, or about 1/50th of a circle (Figure 3-1).

Eratosthenes deduced that the Earth must be 805 multiplied by 50, or about 40,250 kilometers in circumference. His calculations were all in stadia, the unit of measure of the time, and have been converted here to the metric equivalent, using our best idea of the length of a stadia. Eratosthenes' estimate differs from our modern measurements of the Earth's circumference by less than 4%.

The accuracy of Eratosthenes' estimate is quite remarkable, given the equipment for measuring distance and angles at that time, and because a number of his assumptions were incorrect. The well at Syene was located about 60 kilometers off the Tropic of

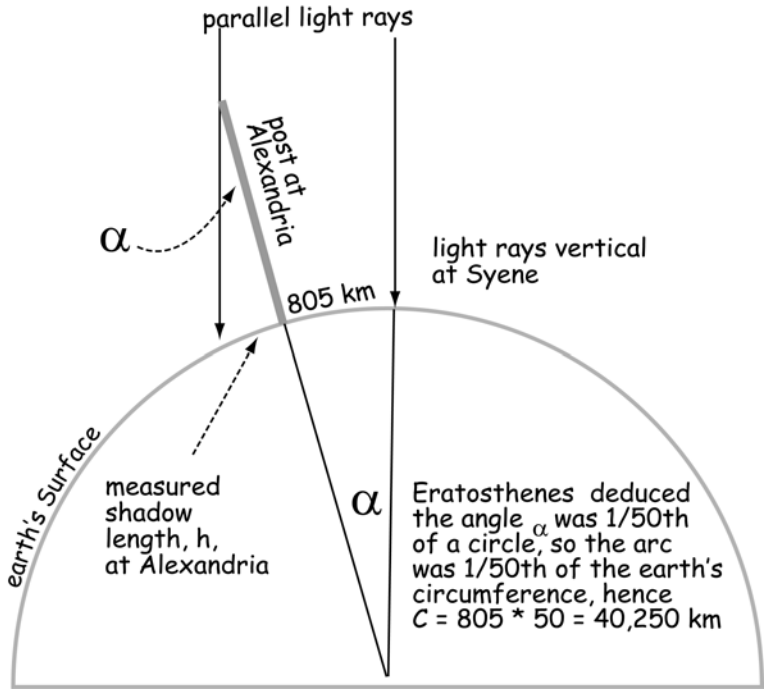


Figure 3-1: Measurements made by Eratosthenes to determine the circumference of the Earth.

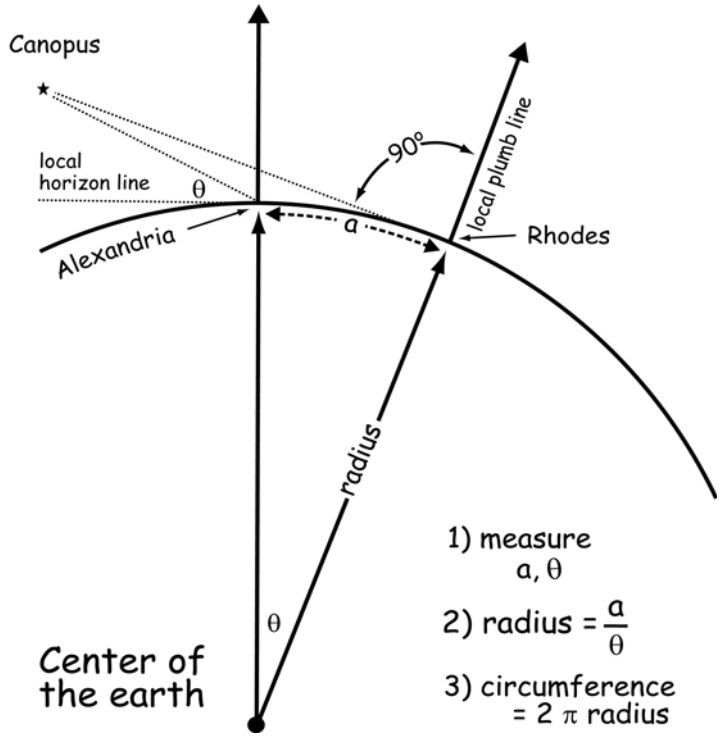


Figure 3-2: Poseidonius approximated the Earth's radius by simultaneous measurement of zenith angles at two points. Two points are separated by an arc distance a measured on the Earth surface. These points also span an angle θ defined at the Earth center. The Earth radius is related to a and θ. Once the radius is calculated, the Earth circumference may be determined. Note this is an approximation, not an exact estimate, but was appropriate for the measurements available at the time (adapted from Smith, 1997).

Cancer, so the Sun was not directly overhead. The true distance between the well location and Alexandria was about 729 kilometers, not 805, and the well was $3^{o}3'$ east of the meridian of Alexandria, and not due north. However these errors either compensated for or were offset by measurement errors to end up with an amazingly accurate estimate.

Posidonius made an independent estimate of the size of the Earth by measuring angles from local vertical (plumb) lines to a star near the horizon (Figure 3-2). Stars visible in the night sky define a uniform reference. The angle between a plumb line and a star location is called a *zenith angle*. The zenith angle can be measured simultaneously at two locations on Earth, and the difference between the two zenith angles can be used to calculate the circumference of the Earth. Figure 3-2 illustrates the observation by Posidonius at Rhodes. The star named Canopus was on the horizon at Rhodes, meaning the zenith angle at Rhodes was 90 degrees. He also noticed Canopus was above the horizon at Alexandria, meaning the zenith angle was less than 90 degrees. The surface distance between these two locations was also measured, and the measurements combined with an approximate geometric relationships to calculate the Earth's circumference. Posidonius calculated the difference in the zenith angles at Canopus as about 1/48th of a circle between Rhodes and Alexandria. By estimating these two towns to be about 800 kilometers apart, he calculated the circumference of the Earth to be 38,600 kilometers. Again there were compensating

errors, resulting in an accurate value. Another Greek scientist determined the circumference to be 28,960 kilometers, and unfortunately this shorter measurement was adopted by Ptomely for his world maps. This estimate was widely accepted until the 1500s, when Gerardus Mercator revised the figure upward.

During the 17th and 18th centuries two developments led to intense activity directed at measuring the size and shape of the Earth. Sir Isaac Newton and others reasoned the Earth must be flattened somewhat due to rotational forces. They argued that centrifugal forces cause the equatorial regions of the Earth to bulge as it spins on its axis. They proposed the Earth would be better modeled by an *ellipsoid*, a sphere that was slightly flattened at the poles. Measurements by their French contemporaries taken north and south of Paris suggested the Earth was flattened in an equatorial direction and not in a polar direction. The controversy persisted until expeditions by the French Royal Academy of Sciences between 1730 and 1745 measured the shape of the Earth near the equator in South America and in the high northern latitudes of Europe. Complex, repeated, and highly accurate measurements established that the curvature of the Earth was greater at the equator than the poles, and that an ellipsoid flattened at the poles was indeed the best geometric model of the Earth's surface.

Note that the words spheroid and ellipsoid are often used interchangeably. For example, the Clarke 1880 ellipsoid is often referred to as the Clarke 1880 spheroid, even though Clarke provided parameters for an ellipsoidal model of the Earth's shape. GIS software often prompts the user for a spheroid when defining a coordinate projection, and then lists a set of ellipsoids for choices.

An ellipsoid is sometimes referred to as a special class of spheroid known as an "oblate" spheroid. Thus, it is less precise but still correct to refer to an ellipsoid more generally as a spheroid. It would perhaps cause less confusion if the terms were used more consistently, but the usage is widespread.

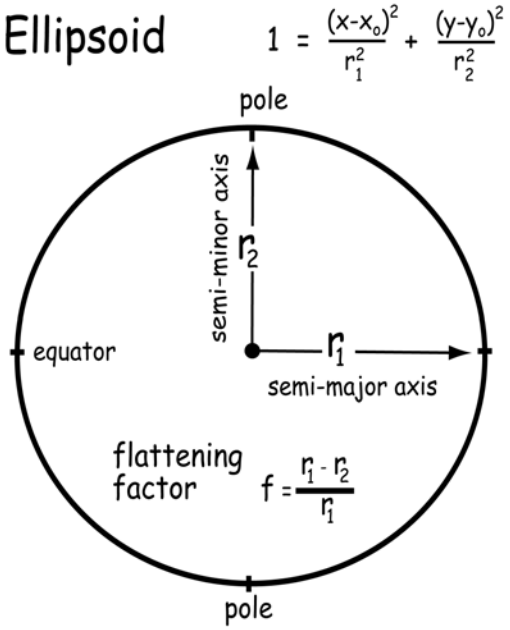


Figure 3-3: An ellipsoidal model of the Earth's shape.

Specifying the Ellipsoid

Once the general shape of the Earth was determined, geodesists focused on precisely measuring the size of the ellipsoid. The ellipsoid has two characteristic dimensions (Figure 3-3). These are the *semi-major axis*, the radius r_1 in the equatorial direction, and the *semi-minor axis*, the radius r_2 in the polar direction. The equatorial radius is always greater than the polar radius for the Earth ellipsoid. This difference in polar and equatorial radii can also be described by the flattening factor, as shown in Figure 3-3.

Earth radii have been determined since the 18th century using a number of methods. The most common methods until recently have involved astronomical observations similar to the those performed by Posidonius. These astronomical observations, also called celestial observations, are combined with long-distance surveys over large areas. Star and sun locations have been observed and cataloged for centuries, and combined

with accurate clocks, the positions of these celestial bodies may be measured to precisely establish the latitudes and longitudes of points on the surface of the Earth. Measurements during the 18th, 19th and early 20th centuries used optical instruments for celestial observations (Figure 3-4).

Measurement efforts through the 19th and 20th centuries led to the establishment of a set of official ellipsoids (Table 3-1). Why not use the same ellipsoid everywhere on Earth, instead of the different ellipsoids listed in Table 3-1? Different ellipsoids were adopted in various parts of the world, primarily because there were different sets of measurements used in each region or continent, and these measurements often could not be

Table 3-1: Official ellipsoids. Radii may be specified more precisely than the 0.1 meter shown here (from Snyder, 1987 and other sources).

Name	Year	Equatorial Radius, r_1 meters	Polar Radius, r_2 meters	Flattening Factor, f	Users
Airy	1830	6,377,563.4	6,356,256.9	1/ 299.32	Great Britain
Bessel	1841	6,377,397.2	6,356,079.0	1/ 299.15	Central Europe, Chile, Indonesia
Clarke	1866	6,378,206.4	6,356,583.8	1/ 294.98	North America; Philippines
Clarke	1880	6,378,249.1	6,356,514.9	1/ 293.46	Most of Africa; France
International	1924	6,378,388.0	6,356,911.9	1/ 297.00	Much of the world
Australian	1965	6,378,160.0	6,356,774.7	1/ 298.25	Australia
WGS72	1972	6,378,135.0	6,356,750.5	1/ 298.26	NASA, US Defense Dept.
GRS80	1980	6,378,137.0	6,356,752.3	1/ 298.26	Worldwide
WGS84	1984 - current	6,378,137.0	6,356,752.3	1/ 298.26	Worldwide

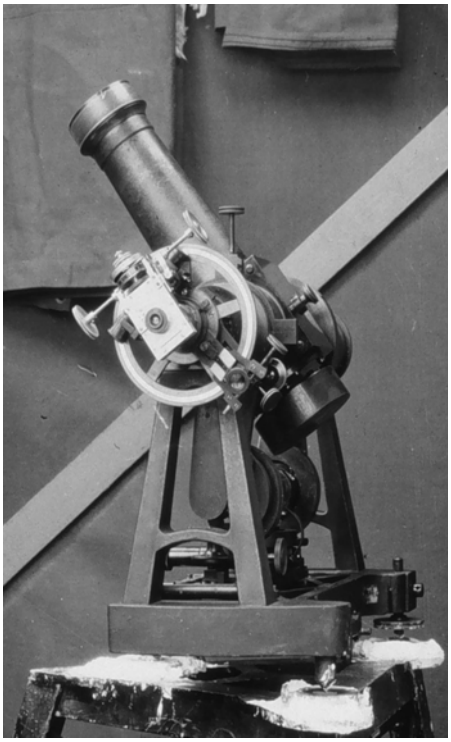


Figure 3-4: An instrument used in the early 1900s for measuring the position of celestial bodies.

tied together or combined in a unified analysis.

Historically, geodetic surveys were isolated by large water bodies. For example, surveys in Australia did not span the Pacific Ocean to reach Asia. Geodetic surveys relied primarily on optical instruments prior to the early 20th century. These instruments were essentially precise telescopes, and thus sighting distances were limited by the Earth's curvature. Individual survey legs greater than 50 kilometers (30 miles) were rare, so during this period there were no good ways to connect surveys between continents.

Because continental surveys were isolated, ellipsoidal parameters were fit for each country, continent, or comparably large survey area. These ellipsoids represented continental measurements and conditions. Because the Earth's shape is not a perfect ellipsoid (described in the next section), surveys of one portion of the Earth will produce different ellipsoidal parameters than surveys of any other portion of the Earth. Measurements based on Australian 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 18th and 19th centuries there was no clear way to combine a global set of measurements.

Differences in the ellipsoids were also due to differences in survey methods and data analyses. Computational resources, the sheer number of survey points, and the scarcity of survey points for many areas were barriers to the development of global ellipsoids. Methods for computing positions, removing errors, or adjusting point locations were not the same worldwide, and led to differences in ellipsoidal estimates. It took time for the best methods to be developed, widely recognized, and adopted.

More recently, data derived from satellites, lasers, and broadcast timing signals have been used for extremely precise measurements of relative positions across continents and oceans. A global set of measurements became possible, and with it an increase in computing capability that allows us to combine and compare measurements on all continents simultaneously. This has led to the calculation of globally applicable ellipsoids. Global ellipsoids such as WGS72, GRS80, or WGS84 have become widely used as more precise measurements have been developed over a larger portion of the globe.

The Geoid

As noted in the previous section, the true shape of the Earth varies slightly from the mathematically smooth surface of an ellipsoid. Differences in the density of the Earth cause variation in the strength of the gravitational pull, in turn causing regions to dip or bulge above or below a reference ellipsoid (Figure 3-5).This undulating shape is called a *geoid*.

Geodesists have defined the geoid as the three-dimensional 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 is in this way related to mean sea level, or other references against which heights are measured. Geodesists often measure surface heights relative to the geoid, and at any point on Earth there are three important surfaces, the ellipsoid, the geoid, and the Earth surface (Figure 3-6).

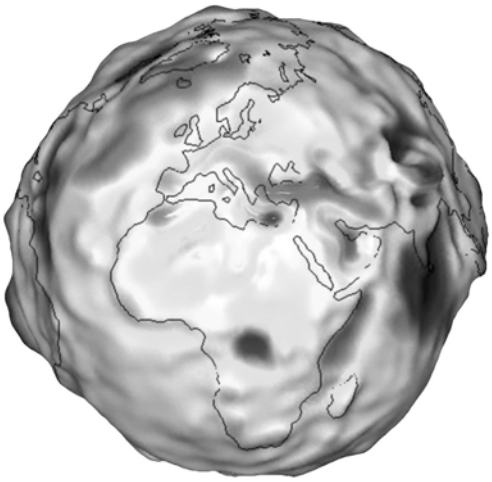


Figure 3-5: Depictions of the Earth's gravity field, as estimated from satellite measurements. These show the undulations, greatly exaggerated, in the Earth's gravity, and hence the geoid (courtesy University of Texas Center for Space Research, and NASA).

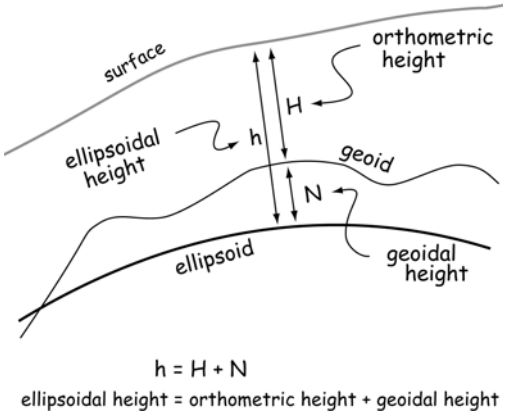


Figure 3-6: Ellipsoidal, orthometric, and geoidal height are interrelated. Note that values for N are highly exaggerated in this figure - values for N ar are typically much less than H.

Because we have two reference surfaces, a geiod and an ellipsoid, against which we measure the Earth's surface, we also have two bases from which to measure height. Elevation is typically defined as the vertical distance above a geoid. This height above a geoid is also called the *orthometric height*. Heights above the ellipsoid are often referred to as *ellipsoidal height*. These are illustrated in Figure 3-6, with the ellipsoidal height labeled h, and orthometric height labeled H. The difference between the ellipsoidal height and geoidal height at any location, shown in Figure 3-6 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

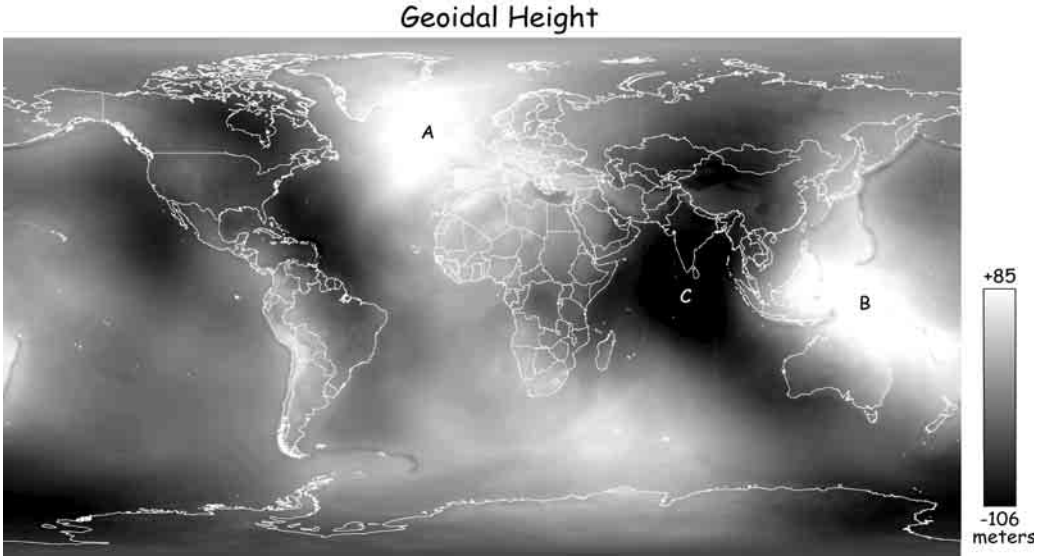


Figure 3-7: Geoidal heights vary across the globe. This figure depicts positive geoidal heights in lighter tones (geoid above the ellipsoid) and negative geoidal heights in darker tones. Note that geoidal heights are positive for large areas near Iceland and the Philippines (A and B, respectively), while large negative values are found south of India (C). Continental and country borders are shown in white.

Earth (Figure 3-7), 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 permanent bulges and dips in the mean ocean surface. 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. As noted in Table 3-1, 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 would be imperceptible in an object at human scales. For example, the largest geoidal height is less than the relative thickness of a coat of paint on a ball three meters in diameter. However, while relatively small, the geoidal variations in shape must still be considered for accurate vertical and horizontal mapping over continental or global scales.

The geoid is a measured and interpolated surface, and not a mathematically defined surface. The geoid's surface is measured using a number of methods, initially by a combination of *plumb bob*, a weighted line that indicates the direction of gravity, and horizontal and vertical distance measurements, and later with various types of *gravimeters*. These instruments measure the gravitational pull as they are towed, flown, or driven on or above the Earth's surface. Variation in the orbits of artificial satellites may also be used to estimate the shape of the geoid. The height from the satellite to the geoid is estimated by measuring the change in gravity with elevation. Gravimetric surveys have been conducted for the entire globe, although measurements are more frequent over the continents and in developed countries.

Satellite-based measurements in the late 20th century substantially improved the coverage, quality, and density of geoidal height measurements across the globe. 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. The satellites are pulled closer or drift farther from the Earth due to variation in the gravity field. Satellites speed up as they are pulled lower, so the leading satellite increases its separation as it speeds up ahead of the following satellite, thereby increasing distance. The lead satellite slows down as it rises higher in regions of weaker gravitational pull. 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. GRACE observations have substantially improved our estimates of the gravitational field and geoidal shape (Figure 3-5).

GRACE and other observations are used by geodesists to develop geoidal models. Models are needed because we measured 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 predict the geoidal height. These models provide an accurate estimation of the geoidal height at unmeasured heights for the entire globe.

Geoidal variation in the Earth's shape is the main cause for different ellipsoids being employed in different parts of the world. The best local fit of an ellipsoid to the geoidal surface in one portion of the globe may not be the best fit in another portion. This is illustrated in Figure 3-8. Ellipsoid A fits well over one portion of the geoid, ellipsoid B in another, but both provide a poor fit in many other areas of the Earth.

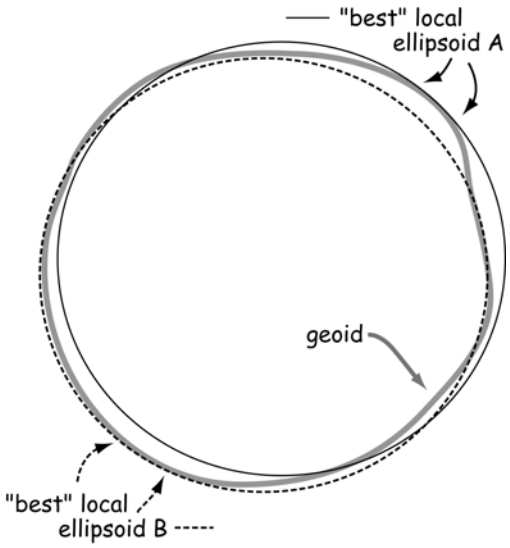


Figure 3-8: An ellipsoid that fits well in one portion of the Earth may fit poorly in another.

Geographic Coordinates, Latitude, and Longitude

Once a size and shape of the reference ellipsoid has been determined, the Earth poles and equator are also defined. The poles are defined by the axis of revolution of the ellipsoid, and the equator is defined as the circle mid-way between the two poles, at a right angle to the polar axis, and spanning the widest dimension of the ellipsoid. We estimate these locations from precise surface and astronomical measurements. Once the locations of the polar axis and equator have been estimated, we can define a set of geographic coordinates. This creates a reference system by which we may specify the position of features on the ellipsoidal surface.

Geographic coordinate systems consist of latitude, which varies from north to south, and longitude, which varies from east to west (Figure 3-9). Lines of constant longitude are called meridians, and lines of constant latitude are called parallels. Parallels run parallel to each other in an east-west direction around the Earth. The

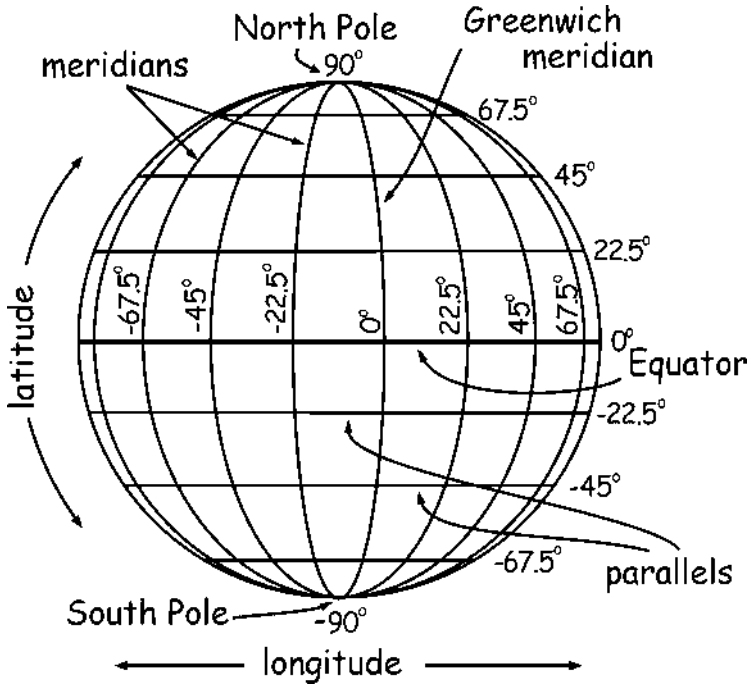


Figure 3-9: The geographic coordinate system.

meridians are geographic north/south lines that converge at the poles.

By convention, the equator is taken as zero degrees latitude, and latitudes increase in absolute value to the north and south (Figure 3-9). Latitudes are thus designated by their magnitude and direction, for example 35^{o}N or 72^{o}S. When signed values are required, northern latitudes are designated positive and southern latitudes designated negative. An international meeting in 1884 established a longitudinal origin intersecting the Royal Greenwich Observatory in England. Known as the *prime* or *Greenwich meridian*, this north-to-south line is the origin, or zero value, for longitudes. East or west longitudes are specified as angles of rotation away from the Prime Meridian. When required, west is considered negative and east positive.

There is often confusion between magnetic north and geographic north. Magnetic north and the geographic north do not coincide (Figure 3-10). Magnetic north is the

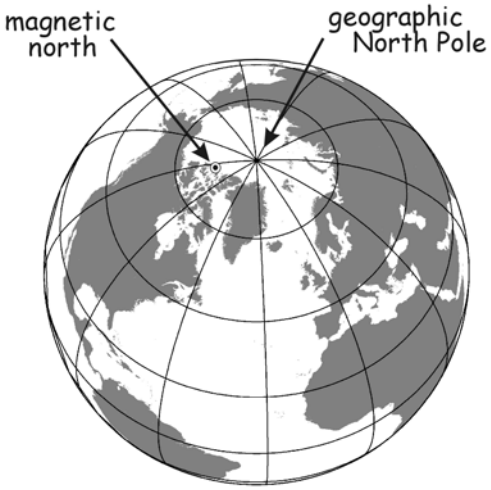


Figure 3-10: Magnetic north and the geographic North Pole.

location towards which a compass points. The geographic North Pole is the northern pole of the Earth's axis of rotation. If you were standing on the geographic North Pole with a compass, it would point approximately in the direction of northern Canada, towards magnetic north some 600 kilometers away.

Because magnetic north and the geographic North Pole are not in the same place, a compass does not point at geographic north when observed from most places on Earth. The compass will usually point east or west of geographic north, defining an angular difference in direction to the poles. This angular difference is called the magnetic *declination* and varies across the globe. The specification of map projections and coordinate systems is always in reference to the geographic North Pole, not magnetic north.

Geographic coordinates do not form a Cartesian system (Figure 3-11). As previously described in Chapter 2, a Cartesian system defines lines of equal value to form a right-angle grid. Geographic coordinates are defined on a curved surface, and the longitudinal lines converge at the poles. Because lines of equal longitude converge at the poles, the distance spanned by a degree of longitude varies from south to north. A degree of longitude spans approximately 111.3 kilometers at the equator, but 0 kilometers at the poles. This means figures specified in geographic coordinates appear distorted when displayed in Cartesian coordinates.

This distortion is seen on the left in Figure 3-11. Circles with a fixed 5 degree radius appear distorted near the poles when drawn on a globe. The circles become flattened in the east-west direction. In contrast, circles appear as circles when the geographic coordinates are plotted in a Cartesian system, as at the right of Figure 3-11, but the underlying geography is distorted; note the erroneous size and shape of Antarctica.While the ground distance spanned by a degree of longitude changes markedly across the globe, the ground distance for a degree of latitude varies only slightly, from 110.6 kilometers at the equator to 111.7 kilometers at the poles.

Spherical coordinates are most often recorded in a degrees-minutes-seconds (DMS) notation, for example N43^{o} 35′ 20″, signifying 43 degrees, 35 minutes, and 20 seconds of latitude. Minutes and seconds range from 0 to 60. Alternatively, spherical coordinates may be expressed as decimal degrees (DD). Examples for conversion DMS to DD and back are shown in Figure 2-7.

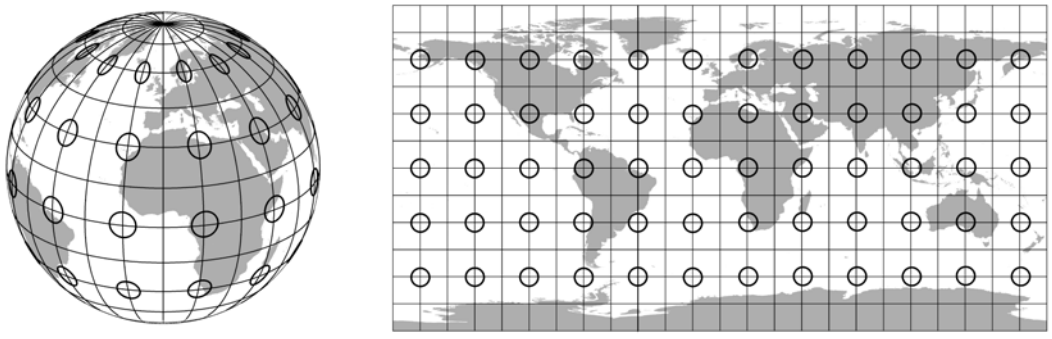


Figure 3-11: Geographic coordinates on a spherical (left) and Cartesian (right) representation. Notice the circles with a 5 degree radius appear distorted on the spherical representation, illustrating the change in surface distance represented by a degree of longitude from the equator to near the poles.

Horizontal Datums

The geographic coordinate system we have just described provides for specifying locations on the Earth. However, this gives us the exact longitude of only one location, the zero line of longitude. By definition the longitude was agreed to be through an observatory in Greenwich, England. We must estimate the longitudes and latitudes of all other locations through surveying measurements, primarily by observing stars and by measuring distances and directions between points. Through these methods we establish a set of points for which the horizontal and vertical positions have been accurately determined. All other coordinate locations we use are measured with reference to this set of precisely surveyed points, including the coordinates we enter in our GIS to represent spatial features.

Many countries have a government body charged with making precise geodetic surveys. For example, nearly all horizontal locations in the United States are traceable to a national network of highly accurate survey points established or maintained by the National Geodetic Survey (NGS). This unit of the U.S. federal government establishes geodetic latitudes and longitudes of known, specifically marked (monumented) points. These points, taken together, form the basis for a *geodetic datum*, upon which most subsequent surveys and positional measurements are based.

A *datum* is a reference surface. A geodetic datum consists of two major components. The first component is the previously described specification of an ellipsoid with a spherical or three-dimensional Cartesian coordinate system and an origin. The second part of a datum consists of a set of points and lines that have been painstakingly surveyed using the best methods and equipment, and an estimate of the coordinate location of each point in the datum. Some authors define the datum as a specified reference surface, and a *realization of a datum* as that surface plus a physical network of precisely measured points. In this nomenclature, the measured points describe a *Terrestrial Reference Frame*. This clearly separates the theoretical surface, the reference system or datum, from the terrestrial reference frame, a specific set of measurement points that help fix the datum. While this more precise language may avoid some confusion, datum will continue to refer to both the defined surface and the various realizations of each datum.

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 a sufficiently large number of new survey points has been measured. We do this by re-estimating the coordinates of our datum points after including these newer measurements, thereby improving our estimate of the position of each point.

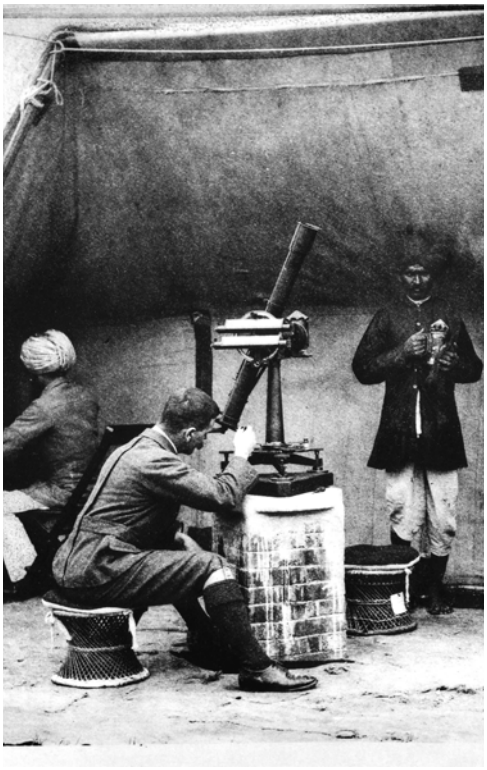


Figure 3-12: Astronomical observations were used in early geodetic surveys to measure datum locations (courtesy NMSI)

Historically, the relative positions of a set of datum points were determined using celestial measurements in combination with high-accuracy ground measurements. Most early measurements involved precise field surveys with optical instruments (Figure 3-12). These methods have been replaced in recent years by sophisticated electronic and satellite-based surveying systems.

Precisely surveyed points used in developing a datum are often monumented, and these monumented points are known as *benchmarks*. Benchmarks usually consist of a brass 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 benchmark, they are often redundantly monumented, and their distance and direction from specific local features are recorded. For example, in areas of deep soil, the NGS may bore a hole one to three meters deep at the point location. Successive concrete posts may be placed one atop another in the hole, with a brass monument affixed to each post, with the top of the final post resting flush with the ground surface. If the surface monument is lost, buried, or destroyed, for example, covered during road improvements, the point may still be recovered, albeit with some effort. Control survey points are often identified with a number of nearby signs to aid in recovery (Figure 3-14).

Figure 3-13: A brass disk used to monument a survey benchmark.

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

Geodetic surveys in the 18th and 19th centuries combined horizontal measurements with repeated, excruciatingly precise astronomical observations to determine latitude and longitude of a small set of points. Only a few datum points were determined using astronomical observations. Astronomical observations were typically used at the starting point, a few intermediate points, and near the end of geodetic surveys. This is because star positions 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 which bends light and changes the apparent position of stars. Refraction depends on how high the star is in the sky at the time of measurement, as well as temperature, atmospheric humidity, and other factors.

Horizontal measurements were as precise and much faster than astronomical measurements when surveys originated at known

locations. These horizontal surface measurements were then used to connect these astronomically surveyed points and thereby create an expanded, well-distributed set of known datum points. Figure 3-15 shows an example survey, where open circles signify points established by astronomical measurements and filled circles denote points established by surface measurements.

Figure 3-15 shows an *triangulation survey*, until recently the method commonly used to establish datum points via horizontal surface measurements. Triangulation surveys utilize a network of interlocking triangles to determine positions at survey stations. Because there are multiple measurements to each survey station, the location at each station may be computed by various paths. The survey accuracy can be field checked, because large differences in a calculated station location via different paths indicate a survey error. There are always some differences in the measured locations when traversing different paths. An acceptable error limit is often set, usually as a proportion of the distance surveyed, for example, differences in the measured location of more than 1 part in 100,000 would be considered unacceptable. When unacceptable errors were found, survey lines were remeasured.

Figure 3-16: A metal survey chain, an early device used to measure distances (courtesy NMSI).

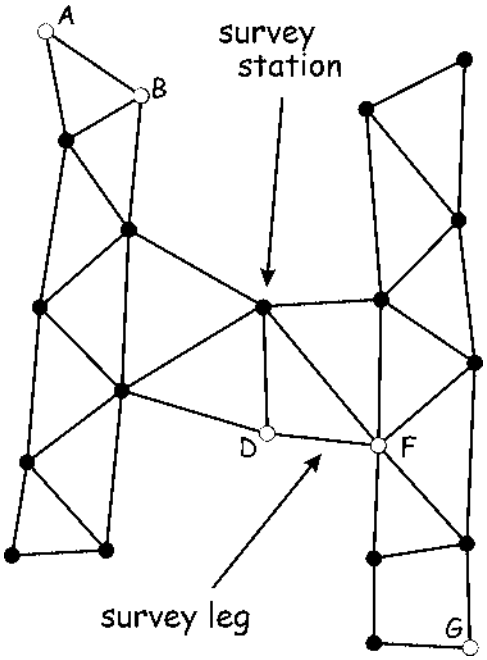


Figure 3-15: A triangulation survey network. Stations may be measured using astronomical (open circles) or surface surveys (filled circles).

Triangulation surveys were also adopted because they required few surface distance measurements. This was an advantage between the late 18th and early 20th century, the period during which the widespread networks of datum points were established. During that time it was much easier to accurately measure horizontal and vertical angles than to measure horizontal distances. Direct measurements of surface distances were less certain because they relied on metal tapes, chains, or rods (Figure 3-16). Tapes and chains sagged by varying amounts. Metal changed length with temperature and stretched with use. Ingenious bi-metallic compensation bars were developed n the 1700s that combined two metals along their length. These metals changed length at dif-

ferent rates with temperature. A scale engraved in the bars indicated the amount of expansion and facilitated a compensation to offset the temperature-caused error. Nonetheless, laying the bars end-to-end over great distances was a painstaking, slow process, and so surface measurements were to be avoided.

Traditional triangulation surveys started with a precisely known point, and then surface measurement of an initial baseline, giving a second known point. From these, angle measurements are combined with geometric relationships between angles and distances in triangles to calculate all subsequent distances. This minimizes errors. Figure 3-17 shows a sequence of measurements for a typical triangulation survey conducted in the 19th or early 20th century. The positions of stations S1 and S2 (Figure 3-17, top) were determined by celestial observations, and the length A was measured using compensation bars or equally precise methods. A surveying instrument was placed at stations S1, S2, and S3 and used to measure angles a, b, and c. The *law of sines* allows the calculation of lengths B and C, and combined with the angle measurements allows precise estimates of the location of station S3. Angles e, f, and g (Figure 3-17, bottom) were then measured, and the precise location of station S4 estimated. The geodetic survey could then be extended indefinitely by repeated application of angle measurements and the law of sines, with only occasional surface distance measurements as a check on positions.

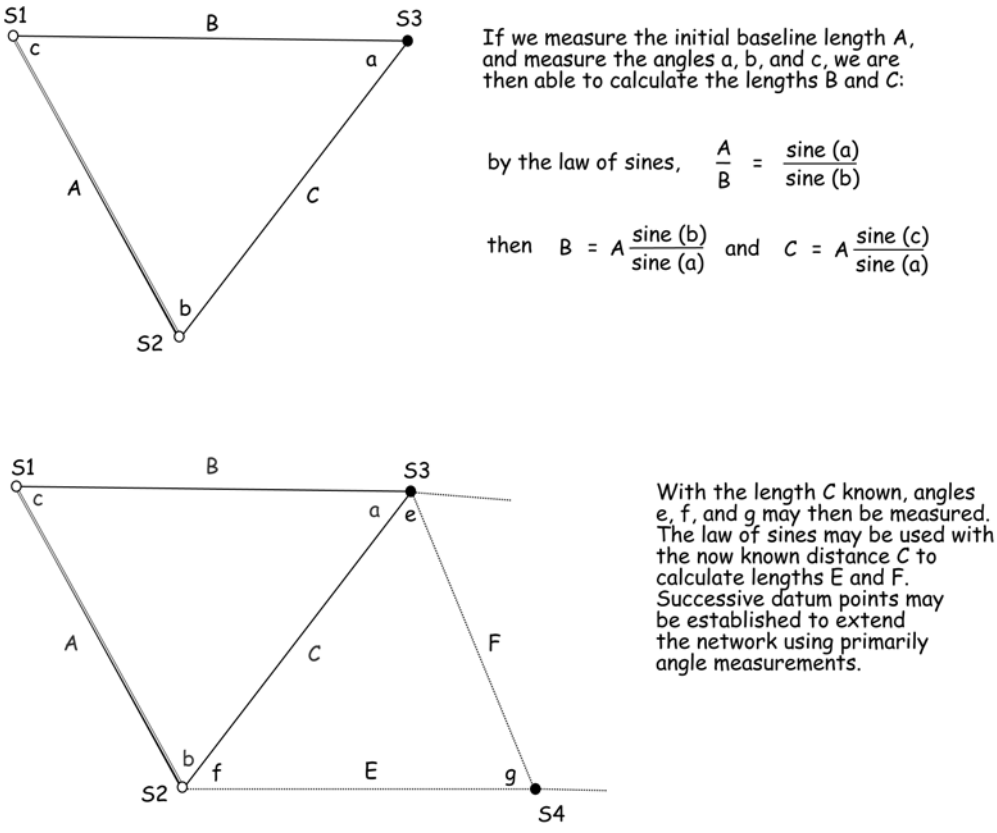


Figure 3-17: Early geodetic surveys used triangulation networks to reduce the number of surface distance measurements.

Survey stations were typically placed on high vantage points to lengthen the distance between triangulation stations. Longer distances meant fewer stations were required to cover a given area. Since the triangulation networks spanned continents, there was a strong incentive to keep costs down. The survey stations were quite far apart, up to tens of kilometers, and measurements were often made from tall objects such as mountaintops, church steeples, or specially constructed Bilby towers to give long sight lines and hence long survey legs (Figure 3-18).

Triangulation networks spanned long distances, from countries to continents (Figure 3-19). Individual measurements of these triangulation surveys were rarely longer than a few kilometers, however triangulations were nested, in that triangulation legs were combined to form larger triangles spanning hundreds of kilometers. These are demonstrated in Figure 3-19 where the sides of each large triangle are made up themselves of smaller triangulation traverses.

Although triangulation surveys and angle measurements were effective ways to control measurement errors, surveys still included errors, and these errors accumulated over long surveys. Despite repeated measurements, periodic checks and astronomical observations, and careful calibration of instruments and calibration for changes in temperature and humidity, inconsistencies remained when data were combined across large areas. For example, the length of a 30 kilometer (20 mile) survey leg may differ by 30 centimeters (1 foot) when measured using two independent methods. These errors had to be accounted for when calculating survey positions.

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 mis-measurements or other mistakes, and also by mathematically minimizing errors by combining repeat measurements and statistically assigning higher influence to consistent or more precise measurements. Errors are reduce and reconciled across the network through these datum adjustments.

Periodic datum adjustments result in series of regional or global reference datums. Each datum is succeeded by an improved, more accurate datum. For example, there are several reference datums for North America. New datums are periodically established as measurement networks expand and as survey instruments and com-

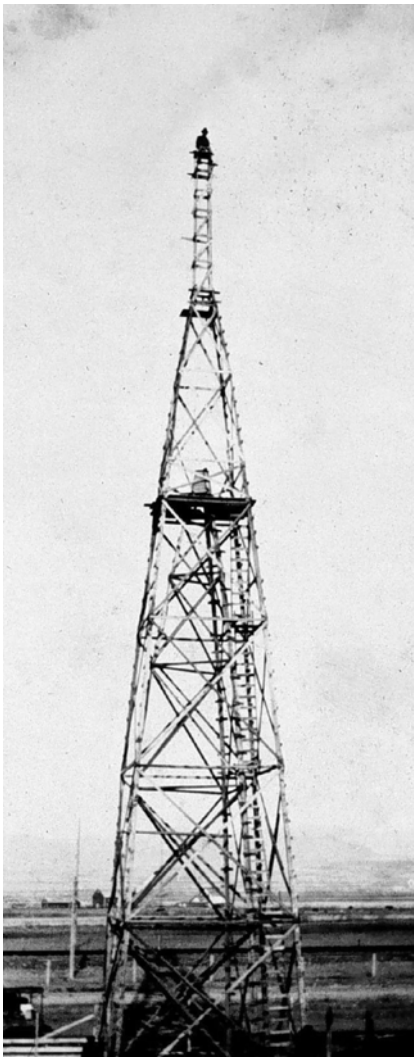


Figure 3-18: A Bilby tower near Bozeman, Montana, USA, used as a platform for a triangulation survey. Towers or other high vantage points increased the distance between survey stations. (courtesy NCGS)

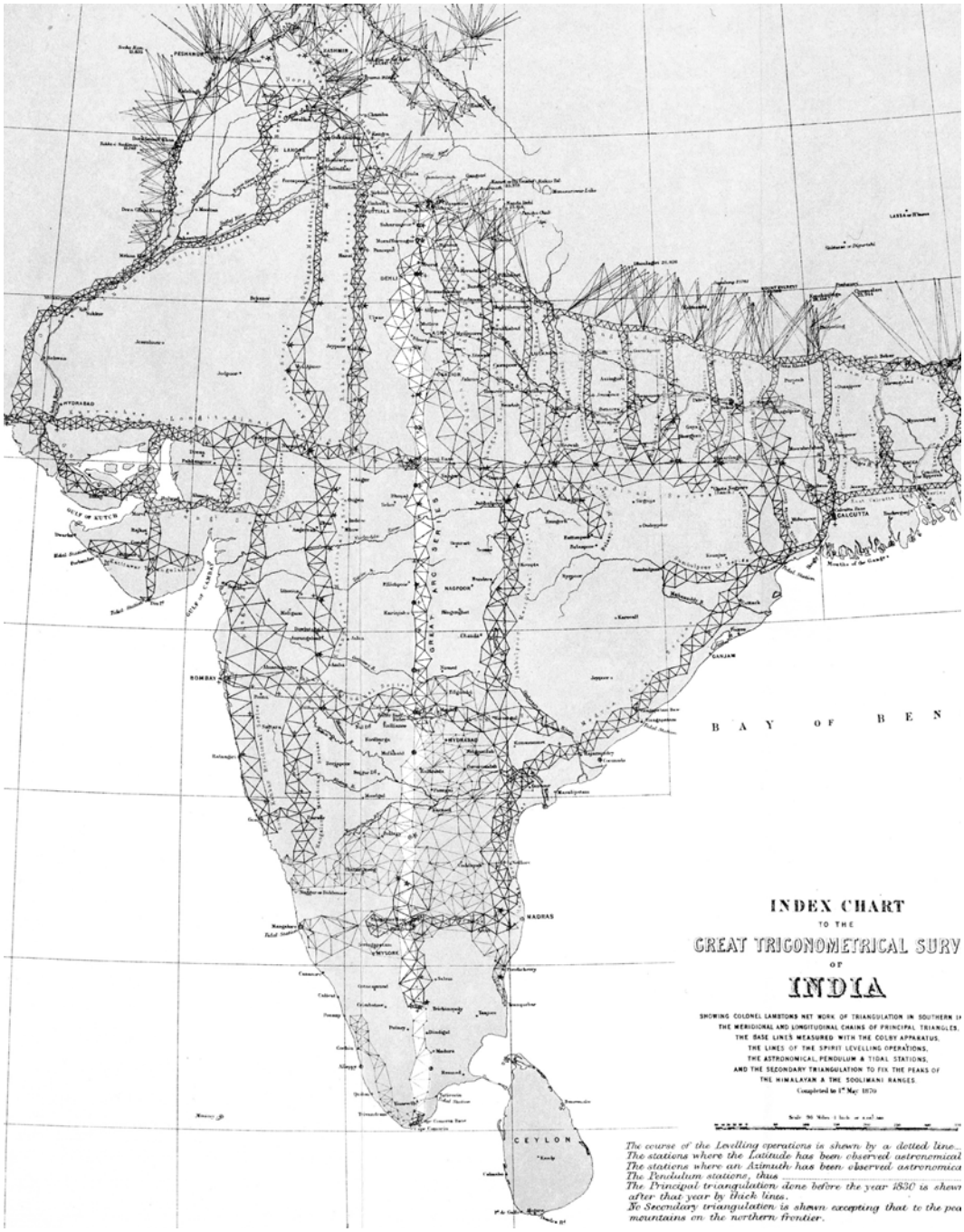


Figure 3-19: A map of the triangulation survey network established across India in the 1800s. Each leg of the triangles, shown here as a single line, is in turn a triangulation survey. This nested triangulation provides reinforcing measurements, thereby increasing the accuracy of the surveyed positions (courtesy NMSI).

putational methods improve. When enough new survey points have been collected, a new datum is estimated. This means new coordinate locations are estimated for all the datum points. The datum points do not move, but our best estimates of the datum point coordinates will change. Differences between the datums reflect differences in the control points, survey methods, and mathematical models and assumptions used in the datum adjustment.

The calculation of a new datum requires that all surveys must be simultaneously adjusted to reflect our current "best" estimate of the true positions of each datum point. Generally a statistical least-squares adjustment is performed, but 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 have historically been quite infrequent. Computational barriers to datum adjustments have diminished in the past few decades, and so datum adjustments and new versions of datums are becoming more common.

Commonly Used Datums

Three main series of horizontal datums have been used widely in North America. The first of these is the *North American Datum of 1927* (NAD27). NAD27 is a general least-squares adjustment that included all horizontal geodetic surveys completed at that time. The geodesists used the Clarke Ellipsoid of 1866 and held fixed the latitude and longitude of a survey station in Kansas. NAD27 yielded adjusted latitudes and longitudes for approximately 26,000 survey stations in the United States and Canada.

The *North American Datum of 1983* (NAD83) is the immediate successor datum to NAD27. It was undertaken by the National Coast and Geodetic Survey to include the large number of geodetic survey points established between the mid-1920s and the early 1980s. Approximately 250,000 stations and 2,000,000 distance measurements were included in the adjustment. The GRS80 ellipsoid was used as reference. NAD83(1986) uses an Earth-centered reference, rather than the fixed station selected for NAD27. We place the (1986) after and NAD83 designator to indicate the year, or version, of the datum adjustment.

The *World Geodetic System of 1984* (WGS84) is also commonly used, and was developed by the U.S. Department of Defense (DOD). It was developed in 1987 based on Doppler satellite measurements of the Earth, and is the base system used in most DOD maps and positional data. WGS84 has been updated based on more recent satellite measurements and is specified using a version designator, e.g., the update based on data collected up to January 1994 is designated as WGS84 (G730). The WGS84 ellipsoid is very similar to the GRS80 ellipsoid.

You should note that there are several versions of both NAD83 and of WGS84, and not acknowledging the version may cause significant confusion. Geodesists at the U.S. National Geodetic Survey have adjusted the NAD83 datum at least four times since the initial 1986 estimation. Precise GPS data became widely available soon after the initial NAD83(1986) adjustment, and these were often more accurate than NAD83(1986) position estimates. State governments and the National Geodetic Survey collaborated in creating *High Accuracy Reference Networks* (HARNs), also known as *High Precision Geodetic Networks* (HPGN) in each state. New GPS points were established, adjustments were made within each state, and then combined to produce an updated datum known as NAD83(HARN), also known as NAD83(HPGN).

Subsequent NAD83 adjustments have been based on the Continuously Operating Reference Station (CORS) network (Figure 3-20). This growing network of satellite observation stations is the basis for newer datums, including NAD83(CORS93), NAD83(CORS94), and NAD83(CORS96). These are all datum adjustments, or new

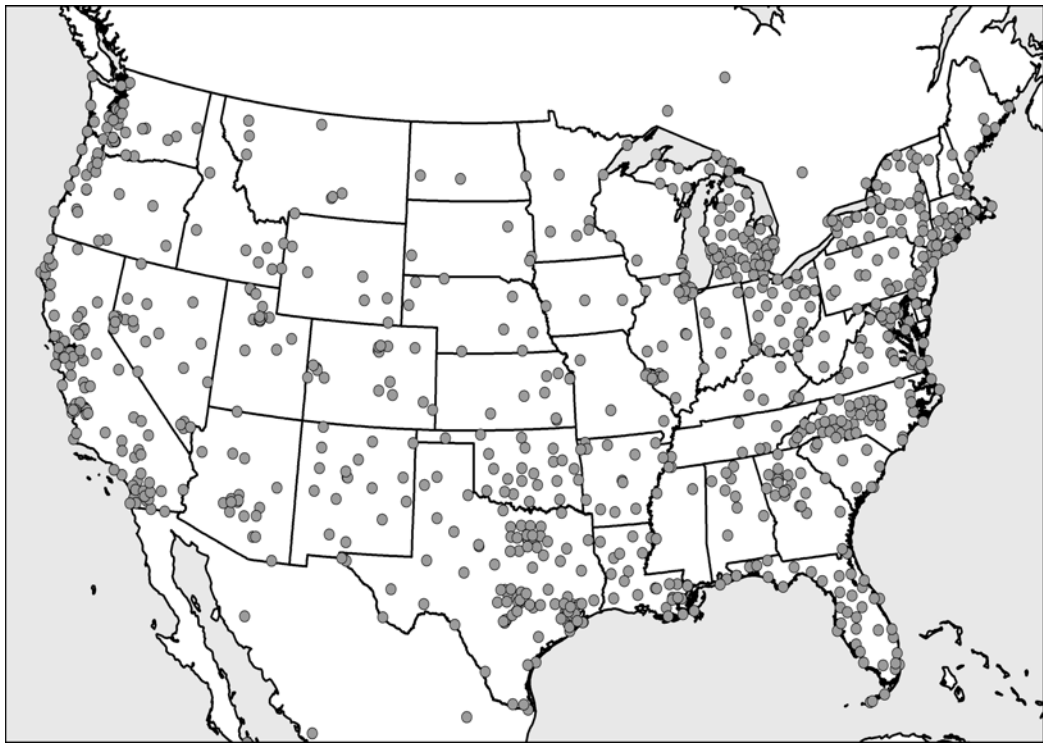


Figure 3-20: Partial distribution of the CORS network, as of 2000. This evolving network is the basis for the NAD83 (1993), NAD83(1994), NAD83(1996), and subsequent U.S. datum adjustments.

datum realizations, based on improved sets of survey and satellite measurements. Position estimates for locations vary by less then 2 cm when compared among the NAD83(CORSxx) datums, important improvements for geodesists and extremely precise surveying, but small relative to spatial error budgets for many GIS projects. Differences among current and future NAD83(CORSxx) datums are likely to remain small.

The WGS84 datum also has several versions. WGS84 was originally based on a set of observations from the TRANSIT satellite system. These observations resulted in a positional accuracy of key datum parameters to within between one and two meters. Subsequent satellite observations improved accuracies. A re-analysis was conducted on data collected through week 730 of the GPS satellite schedule, resulting in the more accurate WGS84(G730). Successive readjustments in weeks 873 and 1150 are known as WGS84(G873) and WGS(G1150), respectively. There will likely be more adjustments in the future.

Another set of datums used worldwide, known *International Terrestrial Reference Frames*, (ITRF), and are annual realizations of the International Terrestrial 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 annual datum is noted by the year, e.g., ITRF89, ITRF90, ITRF91, and so forth, and includes the X, Y, and Z location of each point and the velocity of each point in three dimensions. The European Terrestrial Reference System (ETRS89 and annual updates thereafter) is based on ITRF measurements.

We must underscore that while many spatial data are collected using the WGS84 series of datums due to their use in GPS satellite positioning systems, these WGS84 datums are not often used as a base coordinate system for GIS. For a datum to be practically useful in a a GIS, we typically need the datum coordinates for a widely distributed and uniformly documented set of monumented benchmarks. The development of new data through local surveys and image interpretation requires that we tie our new data to this existing network of surveyed points. In the U.S., the widely distributed set of benchmarked points are reported in the NAD83 (CORSxx) datums, and state, county, and local surveys referenced to these points. In the U.S., we generally must convert any data, such as GPS or other GNSS data, from WGS84 to NAD83 datums before using them. The error introduced in ignoring the differences between versions of WGS84 and the NAD83 datums can be quite large, generally up to 1 meter or more (Figure 3-21). Errors in ignoring differences among older datums are larger still, up to 100's of meters. We must use a technique called a datum transformation to correctly convert data among datums.

Datum Transformations

Since different datums are based on different sets of measurements and ellipsoids, the coordinates for benchmark datum points typically differ between datums. For example, the latitude and longitude location of a given benchmark in the NAD27 datum will likely be different from the latitude and longitude of that same benchmark in NAD83 or WGS84 datums. This is described as a *datum shift*. Figure 3-21 indicates the relative size of datum shifts at an NGS benchmark between NAD27 and NAD83(86), based on estimates provided by the National

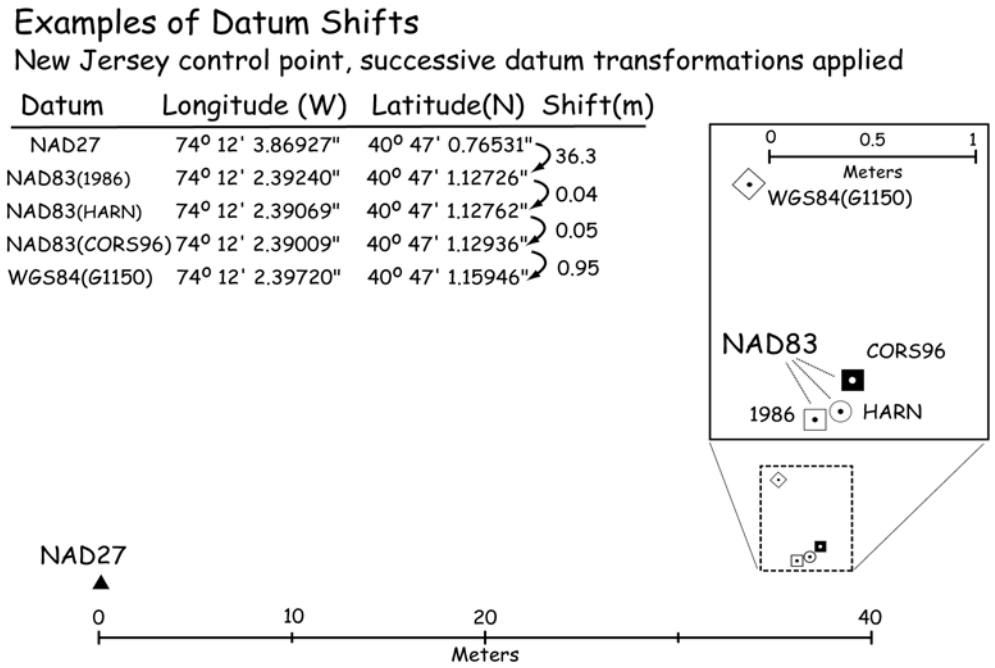


Figure 3-21: Datum shifts in the coordinates of a point for some common datums. Note that the estimate of coordinate position shifts approximately 36 meters from the NAD27 to the NAD83(1986) datum, while the shift from NAD83(1983) to NAD83(HARN) then to NAD83(CORS96) are approximately 0.05 meters. The shift to WGS84(G1150) is also shown, here approximately 0.95 m. Note that the point may not be moving, only our datum estimate of the point's coordinates. Calculations are based on NGS NADCON and HTDP software.

Geodetic Survey. Notice that the datum shift between NAD27 and NAD83(86) is quite large, approximately 40 meters (140 feet).

Note that a datum shift does not imply that points have shifted physical location. Most monumented points have not moved relative to their immediate surroundings. Physically, the points remain in the same location, 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.

Generally, the differences among later versions of modern datums are small when compared to shifts from "legacy" datums such as NAD27. While the datum shifts from NAD27 to any of the NAD83-based datums are often in the tens to hundreds of meters, shifts among recent versions of NAD83 are typically less than a few centimeters (Figure 3-21). Datum shifts from NAD83 to WGS84 are different yet, and differences among WGS versions are often from one to a few meters. Note that the original versions of WGS84 and NAD83 were quite similar, but they have diverged over successive adjustments, such that the differences between WGS84(1150) and NAD83(CORS96) are often over a meter (3 feet).

Estimating the shift and converting geographic 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 benchmark in NAD83(HARN) when we know these geographic coordinates in NAD83(CORS96) (Figure 3-22).

Datum transformations are often more complicated when they involve older datums. As described earlier in this chapter, datums are defined by both an ellipsoid and a set of measured points on the Earth, and many older datums were created piecemeal to optimize fit for a country or continent. The amount of shift between one datum and another often varies across the globe because the errors in measurements may be distributed idiosyncratically. Measurements in one area or period may have been particularly accurate, while another area or time they may exhibit particularly large errors. These errors are an historical artifact, contingent on the unique methods and events associated with the geodetic measurements. When they are included in the datum adjustment, they affect the local and global differences among datums in their own unique way. 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, using a number of different methods. As an example, in the United States the National Geodetic Survey has published a number of papers on datum transformations and provided datum transformation software tools, including NADCON to convert between NAD27 and NAD83 datums, and HTDP for conversion among a subset of NAD and ITRF datums.

A common, alternative approach relies on a mathematical datum transformation between three-dimensional, Cartesian coordinate systems (Figure 3-22). These Earth-centered (geocentric) coordinate systems have been adopted for most modern global datums, including the WGS84 and ITRF systems, and are supported in large part by improved global measurements from artificial satellites, as described in the previous few pages. This three-dimensional approach typically allows for a shift in the origin, a rotation, and a change in scale from one datum to another.

A mathematical geocentric datum transformation is typically a multi-step process. These datum transformations are based on one of a few methods, for example, a *Molodenski transformation* using a system of equations with three or five parameters, or more commonly, a *Helmert transformation* using seven parameters (Figure 3-22). First,

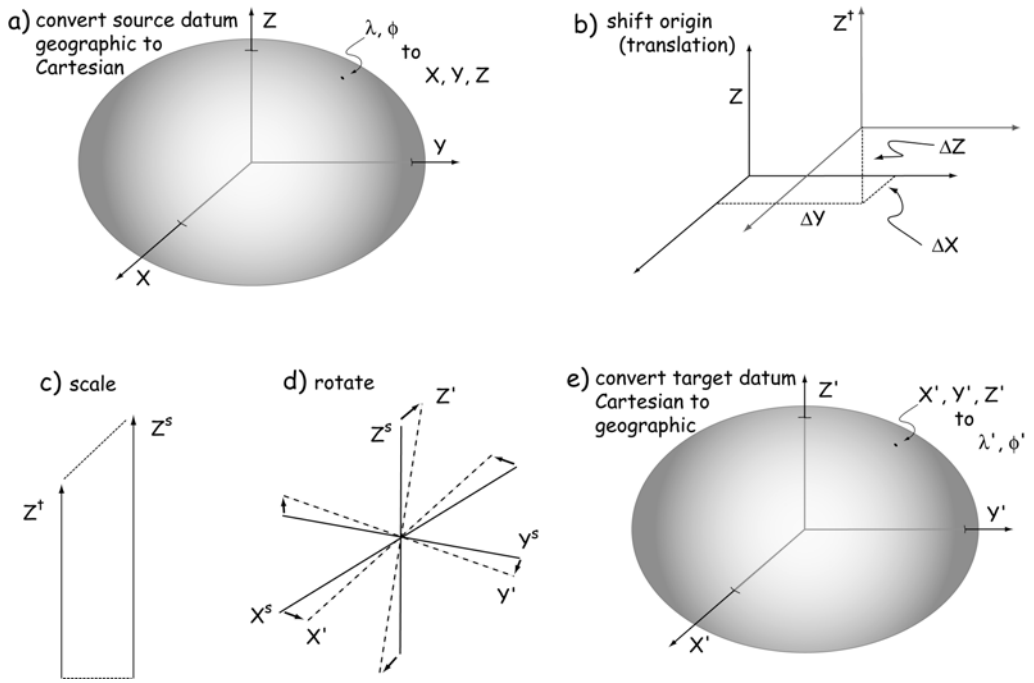


Figure 3-22: Application of a modern datum transformation. Geographic coordinates (longitude, λ, and latitude, ϕ), are transformed to a new datum by a) conversion from geographic to Cartesian coordinates in the old datum (through a set of equations that are not shown), b) applying an origin shift, c)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

geographic coordinates on the source datum ellipsoid 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 geographic coordinates, longitudes and latitudes (λ' and ϕ'), in the target datum.

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. However, many datum shifts are quite large, from meters to tens of meters. One should assume that coordinate shifts between datums are significant until specifically proven otherwise.

Second, datum transformations are estimated, or empirical relationships which are developed with a specific data set and for a specific area and time. There is typically some spatial error or uncertainty in the transformations, and are specific to the input and output datums and versions. 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 special circumstances. All data should be converted to the same coordinate system, based on the same datum. If not, data may mis-align.

Much caution is required when converting among datums because the best datum transformation to use is often not well documented in software, nor always clearly identified in the relevant literature. Until quite recently, spatial error due to improper datum

transformation has been below a detectable threshold in many analyses, so it caused few problems. As data collection accuracies improve, datum transformation errors become more apparent. The datum transformation method within any software package should be documented and the accuracy of the method known before it is adopted. Unfortunately, both of these recommendations are too often ignored or only partially adopted by software vendors and users.

The NGS maintains and disseminates a list of control points in the United States, including those points used in datum definitions and adjustment. Point descriptions are provided in hardcopy and digital forms, including 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 or accuracy, date, or coordinate location.

Figure 3-23 is a partial description of a control point data sheet for a station named Guernsey, located in Kings County, California. This station was first surveyed in 1944, and is a first-order horizontal and vertical control point. Only a portion of the multipage description is shown here. Datasheets for each point include the name, history, physical description, directions for location, coordinates, relevant datums, and adjustment history.

A description of precisely surveyed points may also be obtained from state, county, city, or other surveyors. The accuracy of these points is usually available, and point description and location may be obtained from the appropriate surveying authority. These control points may then be

```
GT0651 DESIGNATION - GUERNSEY
GT0651 PID       - GT0651
GT0651 STATE/COUNTY- CA/KINGS
GT0651 USGS QUAD   - GUERNSEY (1954)
GT0651                 *CURRENT SURVEY CONTROL
GT0651
GT0651* NAD 83(1992)-  36 12 41.51519(N)   119 38 24.98749(W)    ADJUSTED
GT0651* NAVD 88    -      65.61  (+/-2cm)    215.3   (feet)  VERTCON
GT0651
GT0651 EPOCH DATE  -      1991.35
GT0651 LAPLACE CORR-        4.10  (seconds)              DEFLEC99
GT0651 GEOID HEIGHT-       -33.14  (meters)               GEOID03
GT0651
GT0651 HORZ ORDER  - FIRST
GT0651 VERT ORDER  - FIRST    CLASS II (See Below)
GT0651
.....
GT0651
GT0651                  STATION DESCRIPTION
GT0651
GT0651'DESCRIBED BY COAST AND GEODETIC SURVEY 1944 (JCS)
GT0651'STATION IS LOCATED ABOUT 8 MILES S OF HANFORD, AND 0.1 MILE
GT0651'SE OF GUERNSEY RAILROAD STATION, IN THE SE CORNER OF
GT0651'SECTION 1, T. 20 S., R. 21 E., ON LAND OWNED BY THE ATCHISON,
GT0651'TOPEKA AND SANTA FE RAILWAY.
GT0651'
GT0651'UNDERGROUND MARK IS A BRONZE STATION DISK SET IN THE TOP OF
GT0651'A CONCRETE POST.
```

Figure 3-23 A portion of a National Geodetic Survey control point data sheet.

used as starting locations for local surveys to develop a denser network of control points, and as a basis for the development of spatial data.

Vertical Datums

Just as there are networks of well-measured points to define horizontal position, there are networks of points to define vertical position and vertical datums. *Vertical datums* are used as a reference for specifying heights. Much like horizontal datums, they are established through a set of painstakingly surveyed control points. These point elevations are precisely measured, initially through a set of optical surface measurements, but more recently using GPS, laser, satellite, and other measurement systems. Establishing vertical datums also requires estimating the strength and direction of the gravitational force near the surface of the Earth.

Leveling surveys, or leveling, involve the precise measurement of horizontal and vertical distance, and are among the oldest surveying techniques.(Figure 3-24). Distances and elevation differences are precisely measured from an initial point. Early leveling surveys were performed with the simplest of instruments, including a plumb bob to establish leveling posts, and a simple liquid level to establish horizontal lines. Early surveys used an approach known as *spirit leveling*, illustrated in Figure 3-24. Horizontal rods were placed between succeeding leveling posts across the landscape to physically measure height differences.

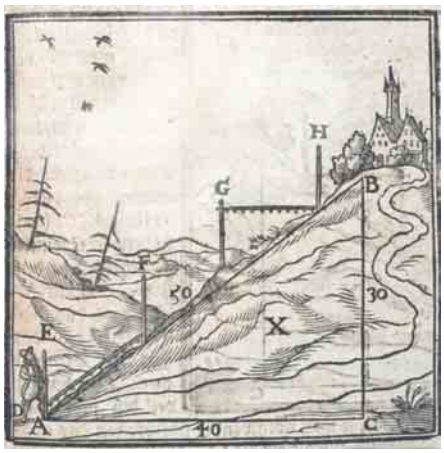


Figure 3-24: Ancient surveys often used "level" bars placed on vertical posts, hence the name leveling surveys.

The number, accuracy, and extent of leveling surveys increased substantially in the 18th and 19th centuries. Grand and 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 due to illness and mishaps over a six year period. Surface leveling provided most height measurements for vertical datums until the mid to late 20th century, when a variety of satellite-based methods were introduced.

Most leveling surveys from the late 1700s through the mid 20th century employed *trigonometric leveling*. This 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 to an unknown station. The angle was typically measured with a small telescope fitted with a precisely scribed angle gauge. The gauge could be referenced to zero at a horizontal position, usually with an integrated bubble level, or later, with an electronic level. 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, but these physical devices have largely been replaced by improved optical methods, or by laser-based methods.

Leveling surveys for geodetic reference may be thought of as beginning at a station near the ocean. This ocean station has been vertically referenced to a mean ocean height nearby. Leveling surveys may then be used

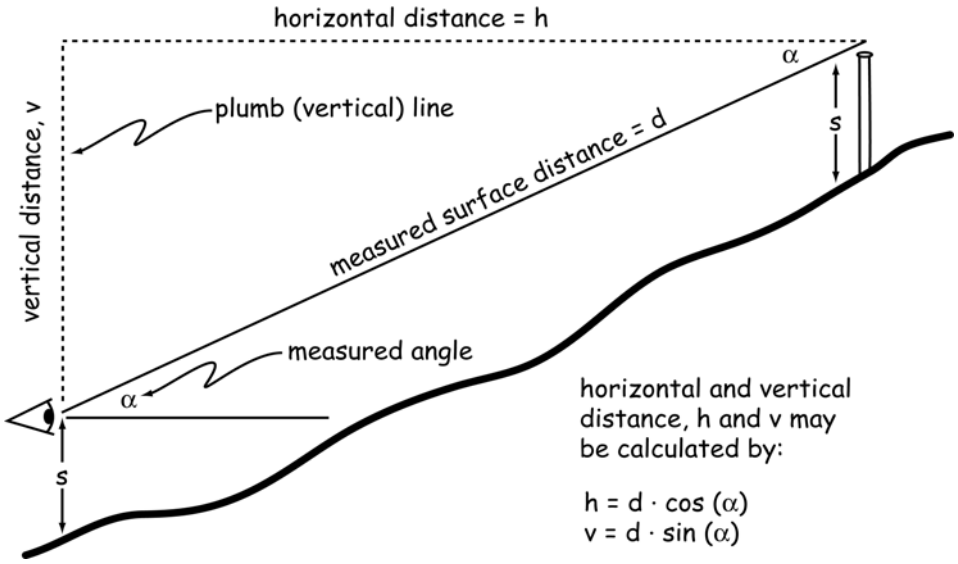


Figure 3-25:Leveling surveys often employ optical measurements of vertical angle (α) with measurements of surface distance (d) and knowledge of trigonometric relationships to calculate horizontal distance (h) and vertical distance (v).

to establish the heights of points that are far from the coast, but relative to the standard coastal benchmark or station. Leveling surveys often share stations with horizontal datum surveys because the horizontal positions are precisely known. Independent leveling surveys may be conducted, but they must also include precise horizontal distance and direction measurements to ensure high mapping accuracy.

As with horizontal datums, the primary vertical datums in use have changed through time as the number, distribution, and accuracy of vertical survey points have increased. The two most common vertical datums in North America are the *National Geodetic Vertical Datum* of 1929 (NGVD29) and the *North American Vertical Datum of 1988* (NAVD88). The NGVD29 was derived from a best fit of the average elevation of 26 stations in North America. The 1988 datum is based on over 600,000 kilometers (360,000 miles) of control leveling performed since 1929, and also reflects geologic crustal movements or subsidence that may have changed benchmark elevation. NAVD88 adjusted the estimate of mean sea level slightly.

Control Accuracy

In most cases the horizontal datum control points are too sparse to be sufficient for all needs in GIS data development. For example, precise point locations may be required when setting up a GPS receiving station, to georegister a scanned photograph or other imagery, or as the basis for a detailed subdivision or highway survey. It is unlikely there will be more than one or two datum points within any given work area. Because a denser network of known points is required for many projects, datum points are often used as a starting locations for additional surveying. These smaller area surveys increase the density of precisely known points. The quality of the point locations depends on the quality of the intervening survey.

The Federal Geodetic Control Committee of the United States (FGCC) has published a detailed set of survey accuracy specifications. These specifications set a

minimum acceptable accuracy for surveys and establish procedures and protocols to ensure the advertised accuracy has been obtained. The FGCC specifications establish a hierarchy of accuracy. First order survey measurements are accurate to within 1 part in 100,000. This means the error of the survey is no larger than one unit of measure for each 100,000 units of distance surveyed. The maximum horizontal measurement error of a 5,000 meter baseline (about 3 miles) would be no larger than 5 centimeters (about 2 inches). Accuracies are specified by Class and Orders, down to a Class III, 2nd order point with an error of no more than 1 part in 5,000.

Map Projections and Coordinate Systems

Datums tell us the latitudes and longitudes of a set of points on an ellipsoid. We need to transfer the locations of features measured with reference to these datum points 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. Points are "projected" from the Earth surface and onto the map surface.

Most map projections may be viewed as sending rays of light from a projection source (Figure 3-26). Rays radiate from a source to intersect both the ellipsoid surface and the map surface. The rays specify where each point from the ellipsoid surface is placed on the map surface. In some projections 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.

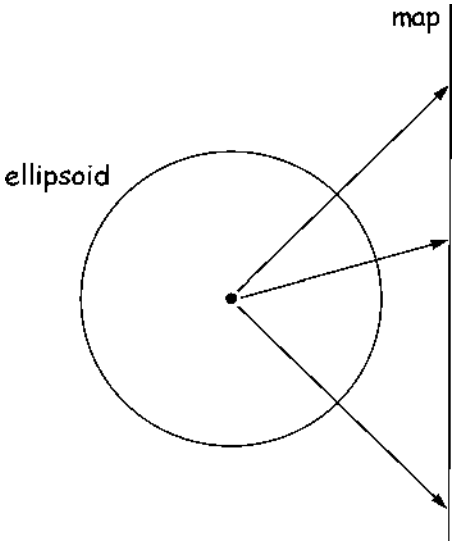


Figure 3-26: A conceptual view of a map projection.

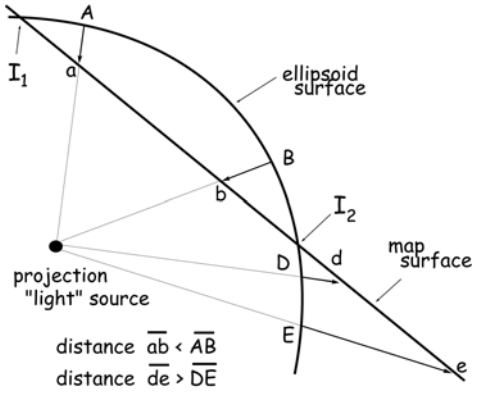


Figure 3-27: Distortion during map projection.

Distortions are unavoidable when making flat maps, because as we've said, locations are projected 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 3-27, 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 dis-

tance 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 3-27 demonstrates a few important facts. First, distortion may take different forms in different portions of the map. In one portion of the map features may be compressed and exhibit reduced areas or distances relative to the Earth's surface measurements, while in another portion of the map areas or distances may be expanded. 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 less near the points or lines of intersection, where the map surface intersects the imaginary globe. Distortion usually increases with increasing distance from the intersection points or lines.

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 3-27, may change locations. The surface onto which we are projecting may change in shape, 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 selection of the appropriate projection or limit the area projected.

Figure 3-28 shows an example of distortion with a projection onto a planar surface. This planar surface intersects the globe line of true scale, the solid line labeled as the standard circle shown in Figure 3-28. Distortion increases away from the line of true scale, with features inside the circle compressed or reduced in size, for a negative

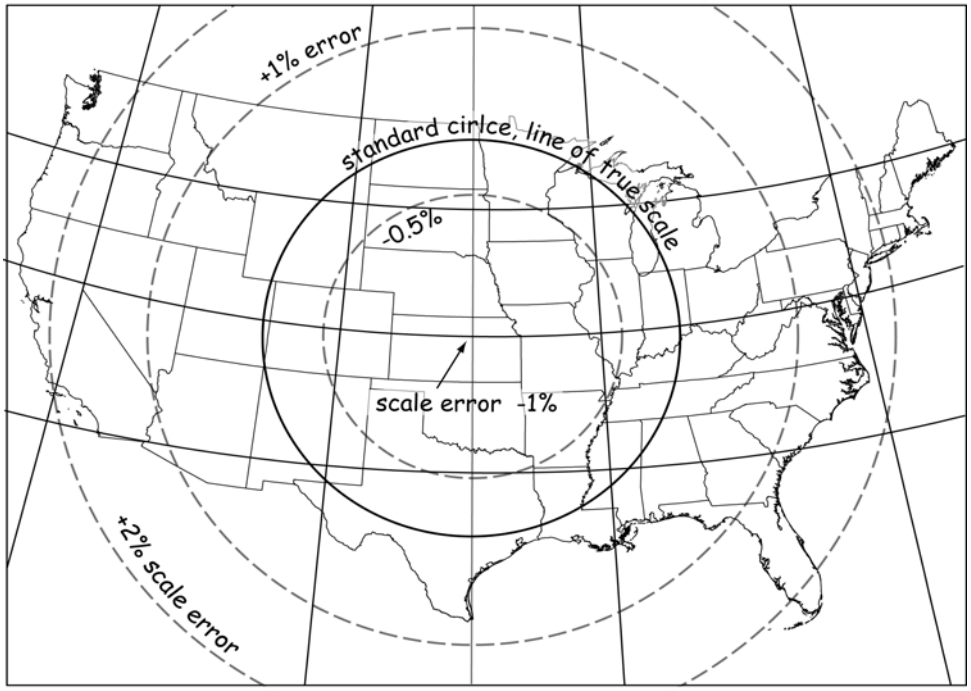
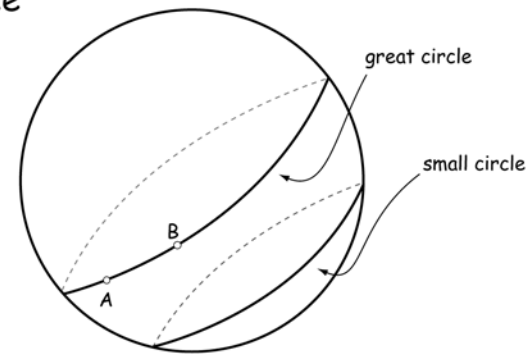


Figure 3-28: Approximate error due to projection distortion for a specific oblique stereographic projection. A plane intersects the globe at a standard circle. This standard circle defines a line of true scale, where there is no distance distortion. Distortion increases away from this line, and varies from -1% to over 2% in this example (adapted from Snyder, 1987).

Great Circle Distance

Consder two points on the Earth's surface, A with geographic coordinates (lat.,lon.) (ϕ_A, λ_A), and

B, with geographic coordinates (ϕ_B, λ_B)



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

$$d = r \cdot \cos^{-1}[(\cos(\phi_A)\cos(\phi_B)\cos(\lambda_A-\lambda_B)\} + \sin(\phi_A)\sin(\phi_B)],$$

where d is the shortest distance on the surface of the Earth from A to B, and r is the Earth's radius, approximately 6378 km.

This formula may be used to find the distance distortion caused by a projection between two points, for example, between Ursine and Moab, Utah, when using UTM Zone 12N coordinates, NAD83?

Great circle distance:

Latitude, longitude of Ursine, Utah = 37.98481°, -114.216944°
Latitude, longitude of Moab, Utah = 38.57361°, -109.551111°

$$d = 6378 \cdot \cos^{-1}[(\cos(37.98481)\cos(38.57361)\cos(-114.216944 - 109.551111\} + \sin(37.98481)\sin(38.57361)\ ]$$
$$= 412.906 \text{ km}$$

Grid distance (UTM Zone 12N coordinates):

Grid coordinates of Ursine, Utah = 217,529.8, 4,208,972.8
Grid coordinates of Moab, Utah = 626,239.2, 4,270,405.9

$$dg = [(X_A - X_B)^2 + (Y_A - Y_B)^2]^{0.5}$$
$$= [(217{,}529.8 - 626{,}239.2)^2 + (4{,}208{,}972.8 - 4{,}270{,}405.9)^2]^{0.5}$$
$$= 413.300 \text{ km}$$

distortion is 412.906 - 413.300 = -0.394 km, or a 394 meter lengthening

Figure 3-29: 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 very close.

scale distortion. Conversely, features outside the standard circle are expanded, for a positive scale distortion. 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 a distance measured on the ellipsoid and in a plane through the Earth's center. This planar surface intersects the two points on the Earth's surface and also splits the spheroid into two equal halves (Figure 3-29). The smallest great circle distance is the shortest path between two points on the surface of the ellipsoid, and by approximation, Earth.

As noted earlier, a straight line between two points on the projected map is likely not to be a straight line on the surface of the Earth, and is not the shortest distance between two points when traveling on the surface of the Earth. Conversely, the shortest distance between points when traveling on the surface of the Earth 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 3-30 illustrates straight line distortion. This figure shows the shortest distance path between Adelaide, Australia, and Tokyo, Japan. Tokyo lies almost due north of Adelaide, and the shortest path approximates a line of longitude, by definition a great circle path. This shortest path is distorted and appears curved by the projection used for this map.

The magnitude of this distortion may be approximated by simple formulas (Figure 3-29). Coordinates may be identified for any two points in the grid system, and the

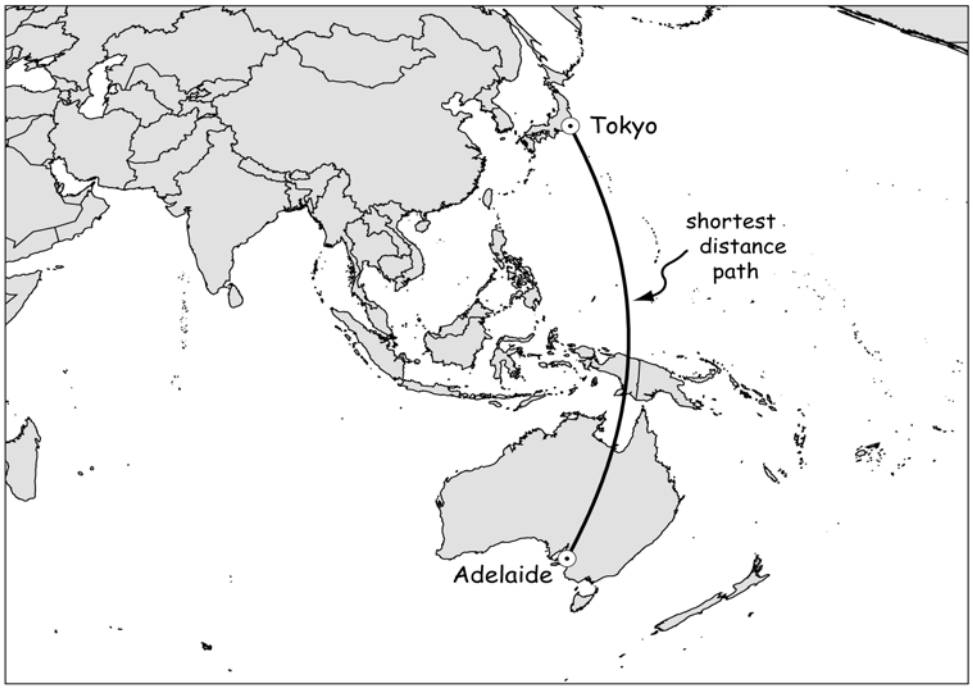


Figure 3-30: Curved representations of straight lines are a manifestation of projection distortion. A great circle path, shown above, is the shortest route when traveling on the surface of the Earth. This path appears curved when plotted on this sinusoidal projection.

Pythagorean formula used to calculate distance between the two points. The resulting distance will be expressed in the grid coordinate system, and therefore will include the projection distortion. The distance may also be calculated for a great circle route along the spheroid surface. This calculation will approximate the unprojected distance, measured on the surface of the Earth. This is only an approximation, as we know from the previous section, because the Earth is shaped more like an ellipsoid, and has geoidal undulations. However, the approximation is quite accurate, generally off by less than a few parts per thousand over several hundred kilometers. The great circle and grid coordinate distance may then be compared to estimate the distance distortion (Figure 3-29).

Projections may also substantially distort the shape and area of polygons. Figure 3-31 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 surface locations are projected. Cones, cylinders, and planes are the most common types of developable surfaces. A plane is already flat, and cones and cylinders may be mathematically "cut" and "unrolled" to develop a flat surface (Figure 3-32). Projections may be characterized according to the developable surface, for

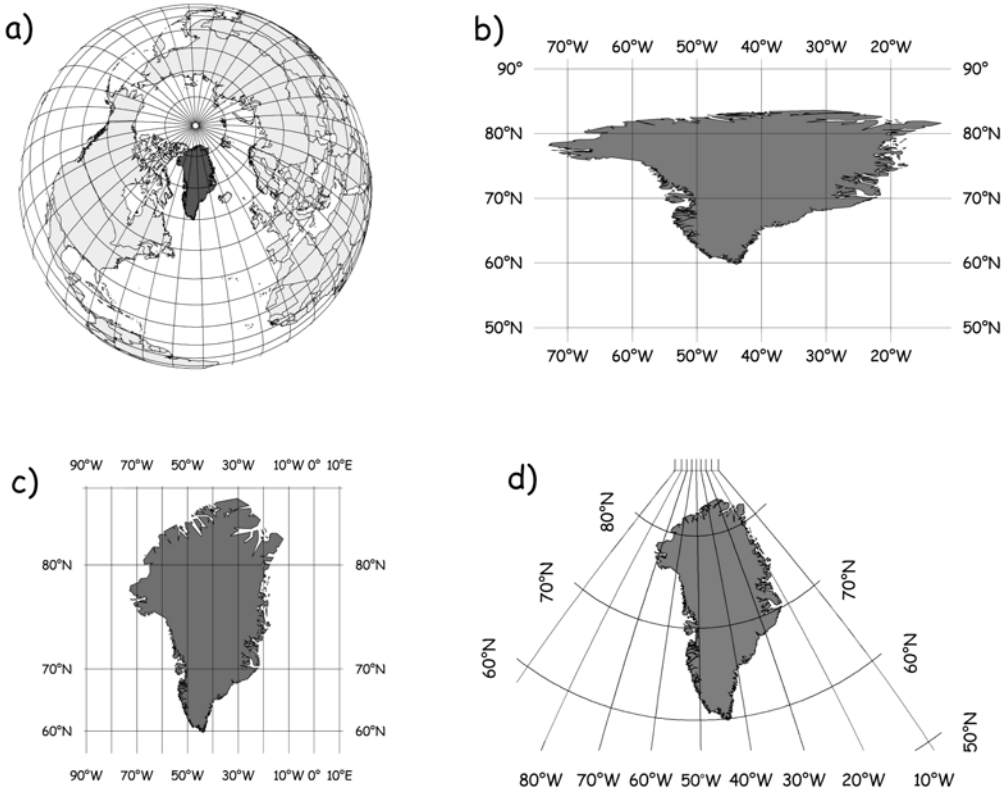


Figure 3-31: Map projections can distort the shape and area of features, as illustrated with these various projections of Greenland, from a) approximately unprojected, b) geographic coordinates on a plane, c) a Mercator projection, and d) a transverse Mercator projection.

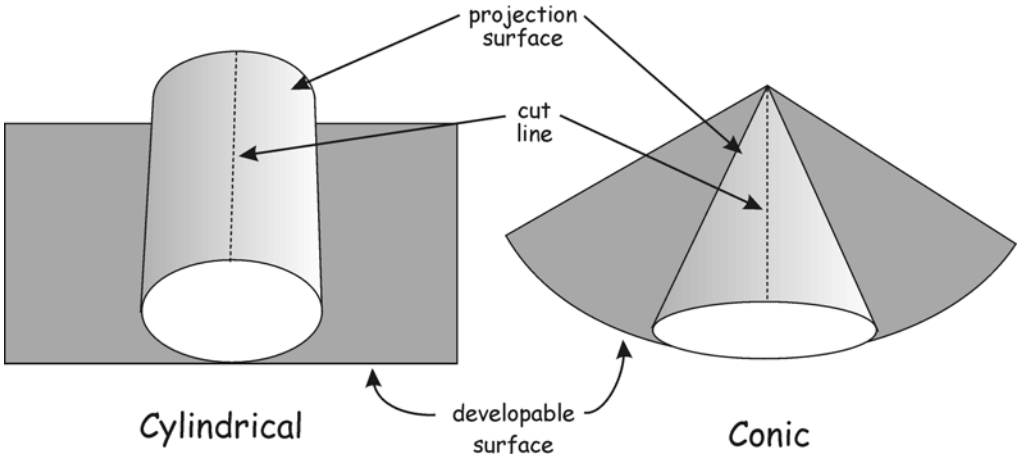


Figure 3-32: Projection surfaces are derived from curved "developable" surfaces that may be mathematically "unrolled" to a flat surface.

example, 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. For example, projections such as the Goode homolosine projection are formed by fusing two or more projections along specified line segments. These projections use complex rules and breaks to reduce distortion for many continents.

We typically have to specify several characteristics when we specify a map projection. For example, for an azimuthal projection we must specify the location of the projection center (Figure 3-33) and the location and orientation of the plane onto which the globe is projected. Azimuthal projections are often tangent to (just touch) the ellipsoid at one point, and we must specify the location of this point. A projection center ("light" source location) must also be specified, most often placed at one of three locations. The projection center may be at the center of the ellipsoid (a *gnomonic* projection), at the antipodal surface of the ellipsoid (diametrically opposite the tangent point, a *stereographic* projection), or at infinity (an *orthographic* projection). Scale factors, the location of the origin of the coordinate sys-

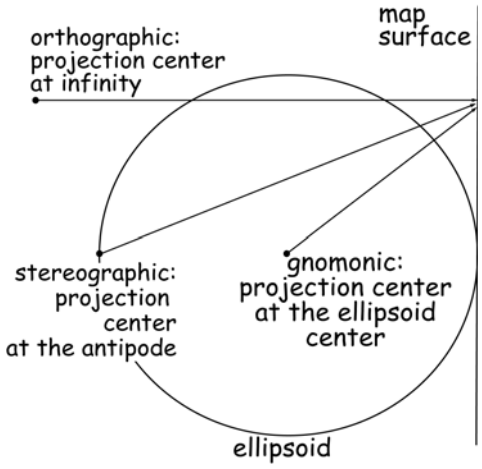


Figure 3-33: The projection center of map projections is most often placed at the center of the ellipsoid, or the antipode, or at infinity.

tem, and other projection parameters may be required. Defining characteristics must be specified for all projections, such as the size and orientation of a cone in a conic projection, or the size, intersection properties, and orientation of a cylinder in a cylindrical projection

Note that the use of a projection defines a projected coordinate system and hence typically adds a third version of North to our description of geography. We have already described magnetic north, towards which a compass points, and geographic north, the pole around which the globe revolves (Figure 3-10). We must add *grid north* to these, defined as the direction of the Y axis in the projection. Grid north is typically defined by some meridian in the projection, often known as the central meridian. Grid north is typically not the same as geographic or magnetic north over most of the projected area.

Common Map Projections in GIS

There are hundreds of map projections used throughout the world, however most spatial data in GIS are specified using projections from 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 3-34). 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 3-34 (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 the Earth. Distortion increases in a complex fashion as distance from these intersection lines increases. This characteristic is illustrated at the top right and bottom of Figure 3-34. 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 3-34, bottom. Distortion decreases towards the standard parallels, and increases away from these lines. Those farther away tend to be more distorted. 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 show are distorted in the Lambert conformal conic projection (Figure 3-34, top right). Those circles that fall between the standard parallels exhibit a uniformly lower distortion than those in other portions of the projected map. One property of the Lambert conformal conic projection is a low-distortion band running in an east-west direction between the standard parallels. Thus, the Lambert conformal conic projection may be used for areas that are larger in an east-west than a north-south direction, as there is little added distortion when extending the mapped area in the east-west direction.

Distortion is controlled by the placement and spacing of the standard parallels, the lines where the cone intersects the globe. The example in Figure 3-34 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 reducing the area mapped at this lower distortotion level.

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 3-35). 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 3-35. A line parallel to and midway between the secants is often called the central meridian. The central meridian

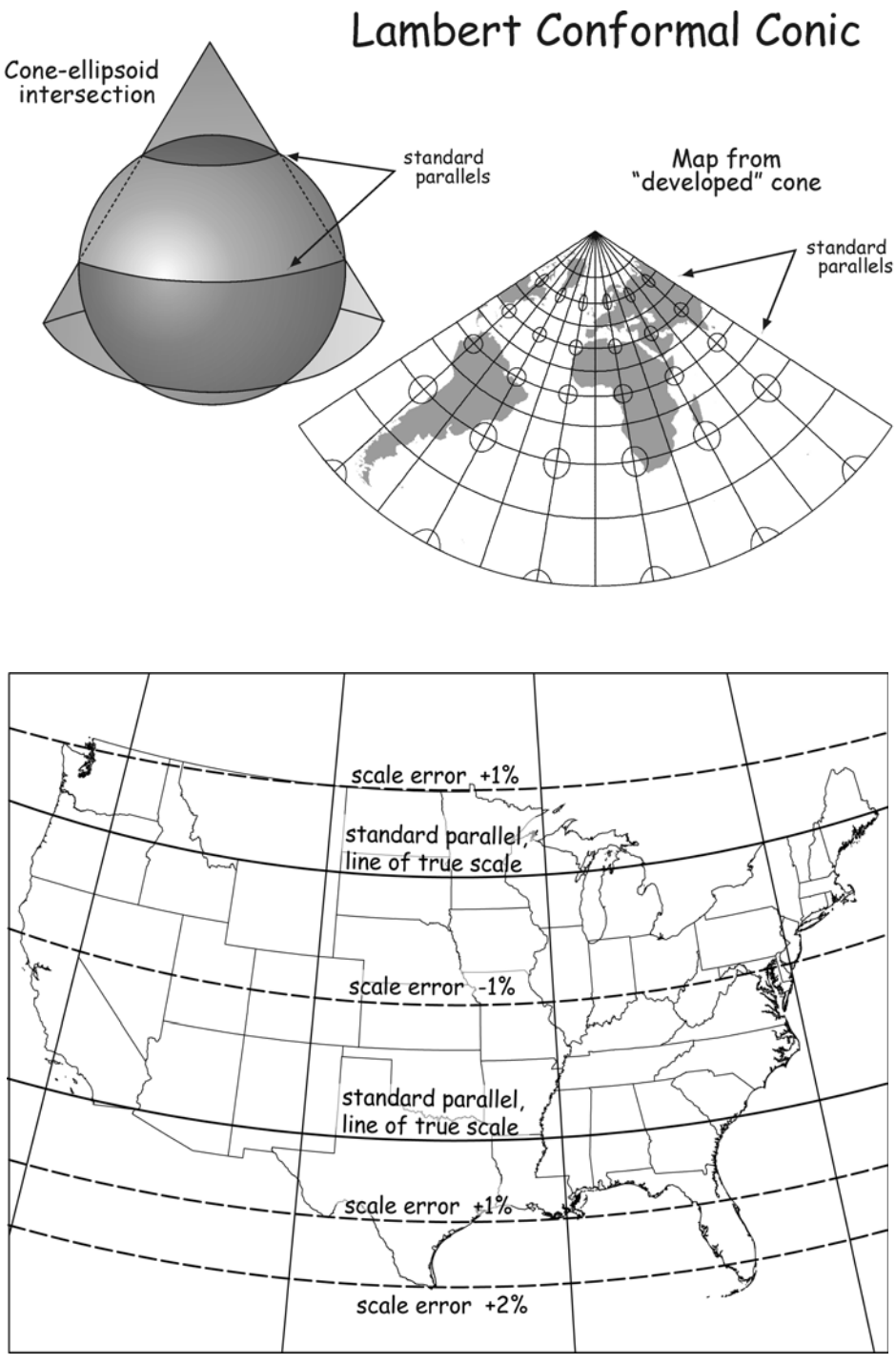


Figure 3-34: Lambert conformal conic (LCC) projection (top) and an illustration of the scale distortion associated with the projection. The LCC is derived from a cone intersecting the ellipsoid along two standard parallels (top left). The "developed" map surface is mathematically unrolled from the cone (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).

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 3-35 shows a transverse Mercator projection with the central meridian (line of intersection) at 0 degrees longitude, traversing western Africa, eastern Spain, and England. 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 3-35. The drawing at the bottom of Figure 3-35 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 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. Once a projection is defined, the coordinates of every point on the surface of the Earth may be determined, usually by a closed-form or approximate mathematical formula.

The State Plane Coordinate System

The State Plane Coordinate System is a standard set of projections for the United States. The State Plane coordinate system specifies positions in Cartesian coordinate systems for each state. There are one or more zones in each state, with slightly different projections in each State Plane zone (Figure 3-36). Multiple State Plane zones are used to limit distortion errors due to map projections.

State Plane systems greatly facilitate surveying, mapping, and spatial data development in a GIS, particularly when whole county or larger areas are involved. Over relatively small areas, the surface of the Earth can be assumed to be flat without introducing much distortion. As noted earlier, distortion typically increases over larger distances. The State Plane system provides a common coordinate reference for horizontal coordinates over county to multi-county areas while limiting error to specified maximum values. Zones are specified in each state such that projection distortions are kept below one part in 10,000. State Plane coordinate systems are used in many types of work, including property surveys, property subdivisions, large-scale construction projects, and photogrammetric mapping, and the zones and state plane coordinate system 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 6, 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 3-37, left). Within most zones, the distance on the curving. Zone boundaries are defined by county, parish, or other municipal boundaries. For example, the Minnesota south/central zone boundary runs

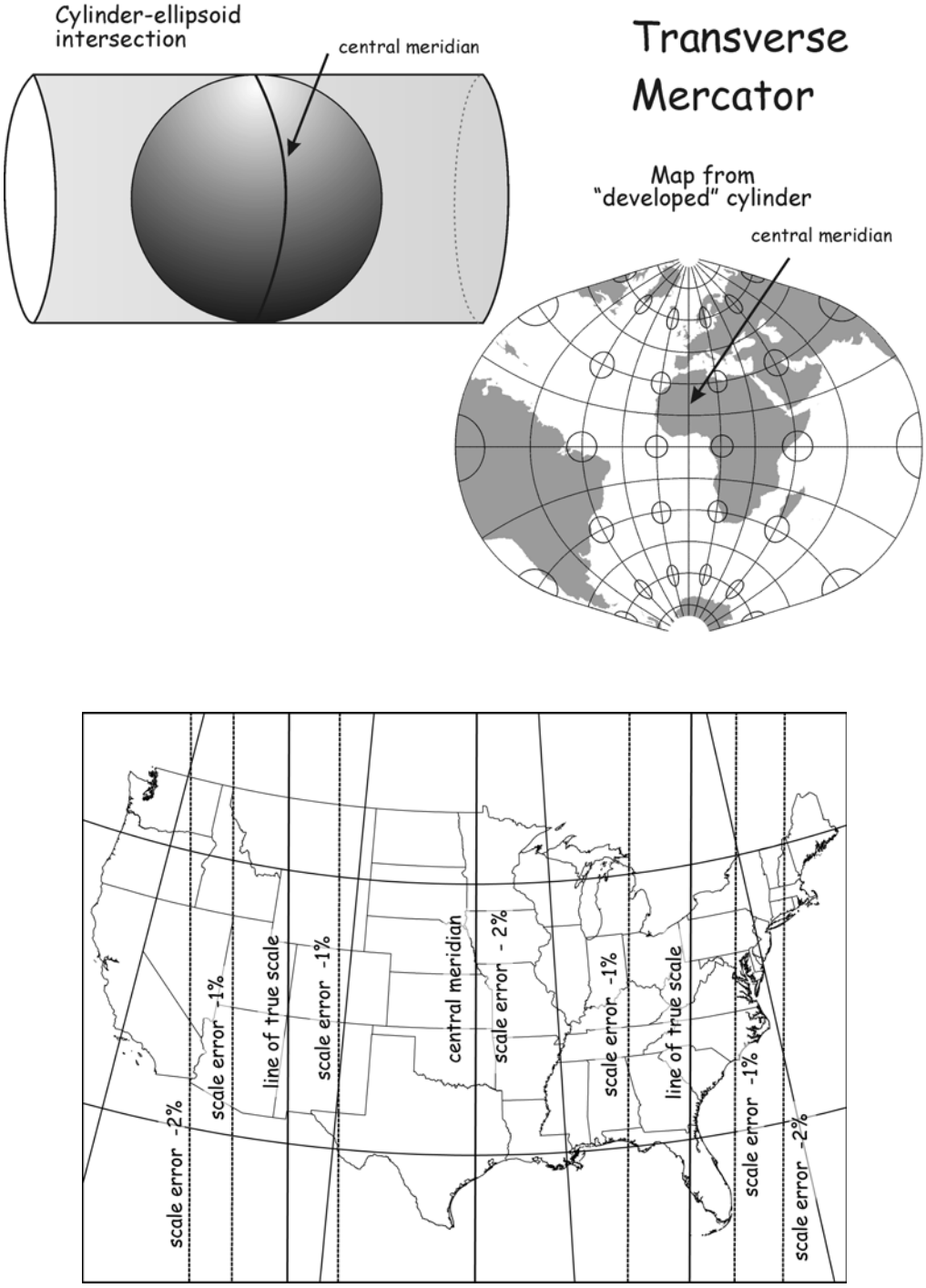


Figure 3-35: Transverse Mercator (TM) projection (top), and an illustration of the scale distortion associated with the projection (bottom). The TM projection distorts distances in an east-west direction, but has relatively little distortion in a north-south direction. This TM intersects the sphere along two lines, and distortion increases with distance from these lines (bottom, adapted from Snyder, 1987).

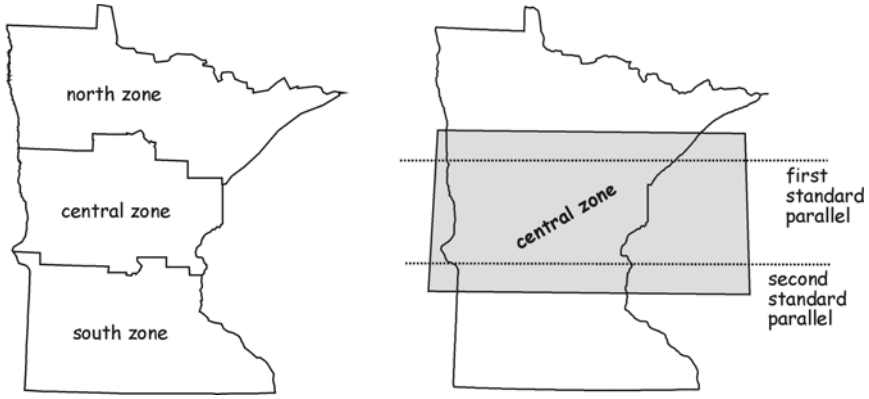


Figure 3-37: The State Plane zones of Minnesota, and details of the standard parallel placement for the Minnesota central State Plane zone.

approximately east-west through the state along defined county boundaries (Figure 3-37, left).

The State Plane coordinate system is based on two basic types of map projections: the Lambert conformal conic and 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 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 is in the east-west direction (e.g. North Carolina and Virginia). When computing the State Plane coordinates, points are projected from their geodetic latitudes and longitudes to x and y coordinates in the State Plane systems.

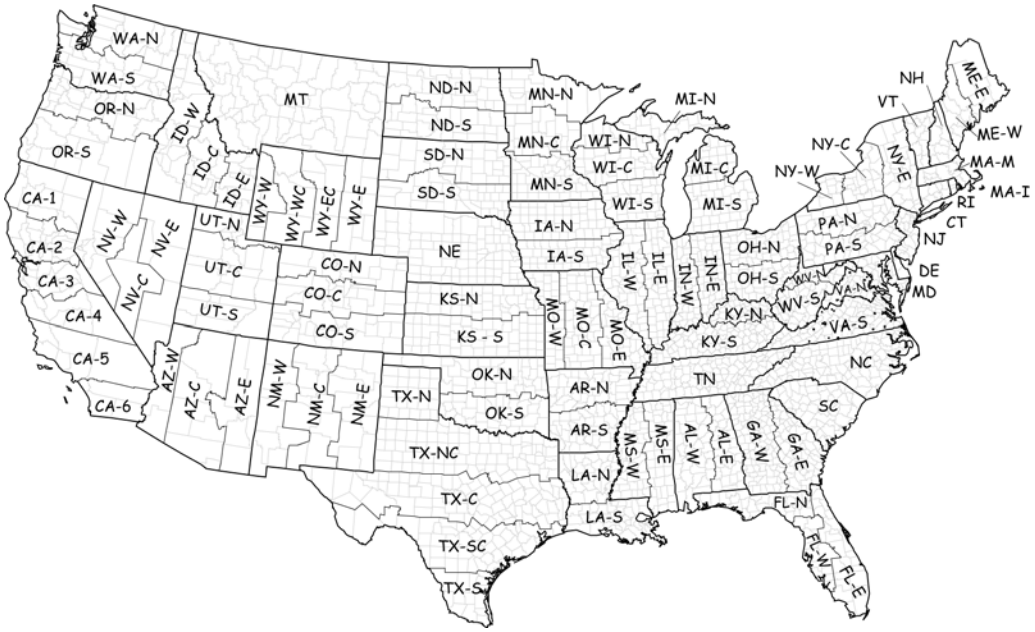


Figure 3-36: State plane zone boundaries, NAD83.

The Lambert conformal conic projection is specified in part by two standard parallels that run in an east-west direction. A different set of standard parallels is defined 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 3-37, right). The zone projection is defined by specifying the standard parallels and a central meridian that has a longitude near the center of the zone. This central meridian points in the direction of geographic north, however all other meridians converge to this central meridian, so they do not point to geographic north. The Lambert conformal conic is used to specify projections 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 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 the NAD27 datum. 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 3-38). 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

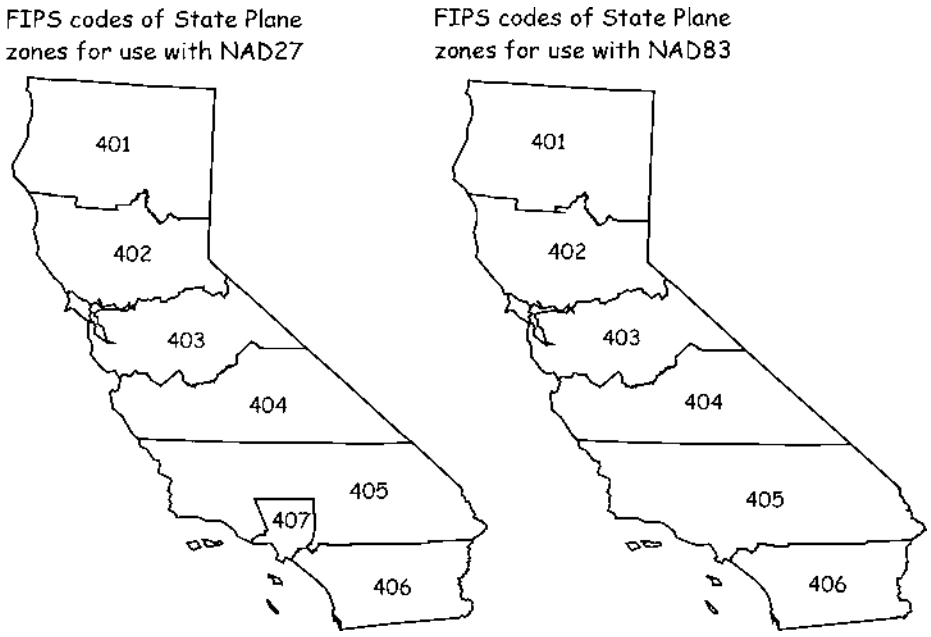


Figure 3-38: State Plane coordinate system zones and FIPS codes for California based on the NAD27 and NAD83 datums. Note that zone 407 from NAD27 is incorporated into zone 405 in NAD83.

across NAD27 and NAD83 versions. Care must be taken when using older 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 additionally 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 based on well-defined, integrated units and thus reduced commercial fraud and improve 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 United States definition of a foot is slightly longer than the European definition, by about one part in five million. Both systems were used in measuring distance, the U.S. conversion for surveys in the U.S., and the international conversion of surveys or measurements 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 for all 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.

Universal Transverse Mercator Coordinate System

Another standard coordinate system has been based on the transverse Mercator projection, distinct from the State Plane system. This system is known as the Universal Transverse Mercator (UTM) coordinate system. The UTM is a global coordinate system. 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 3-39). 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. The UTM system also defines a subzone numbering system called the Military Grid Reference System (MGRS). The MGRS is rarely used in civilian settings.

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

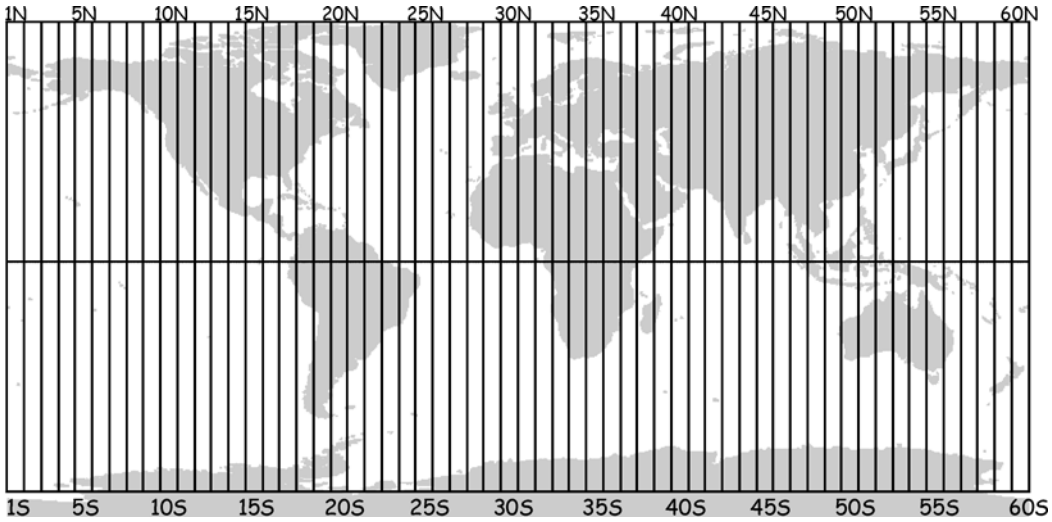


Figure 3-39: UTM zone boundaries and zone designators. Zones are six degrees wide and numbered from 1 to 60 from the International Date Line, 180^{o}W. Zones are also identified by their position north and south of the equator, e.g., Zone 7 North, Zone 16 South.

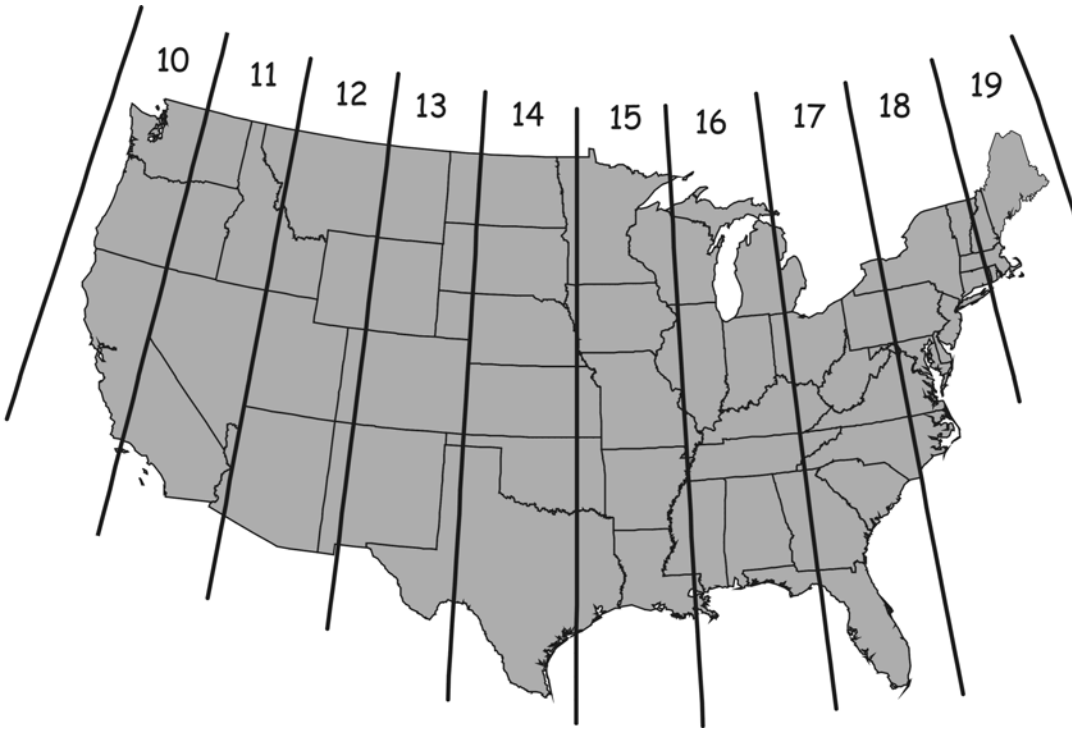


Figure 3-40: UTM zones for the lower 48 contiguous states of the United States of America. Each UTM zone is 6 degrees wide. All zones in the Northern Hemisphere are north zones, e.g., Zone 10 North, 11 North,...19 North.

fits predominantly or entirely into one UTM zone (Figure 3-40).

As indicated before, all regions 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 multi-county agencies have adopted the dominant UTM coordinate system as a standard.

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 zones must be chosen, 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.

Distances in the UTM system are specified in meters north and east of a zone origin (Figure 3-41). The y values are known as *northings*, and increase in a northerly direction. The x values are referred to as *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.

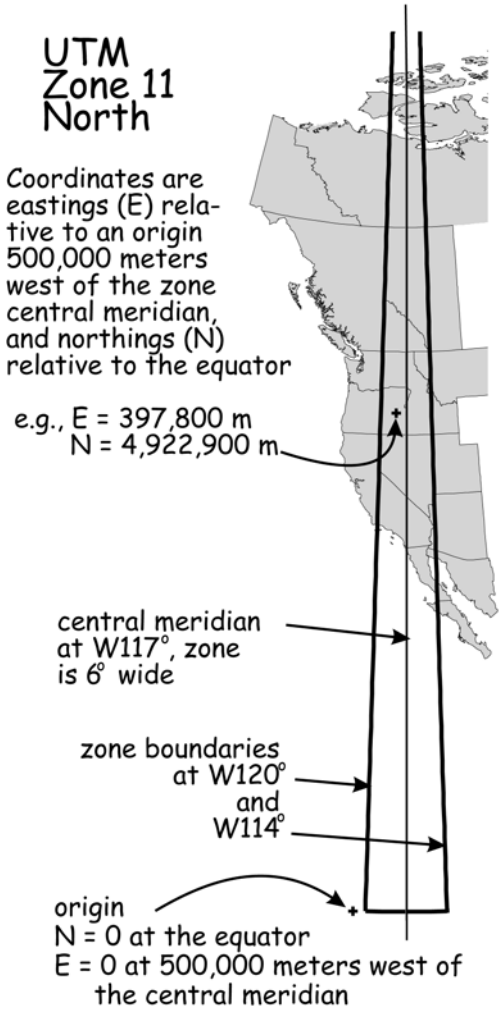


Figure 3-41: UTM zone 11N.

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.

University Transverse Mercator zones south of the equator are slightly different than those north of the equator (Figure 3-42). South zones have a *false northing* value added to ensure all coordinates within a 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, assuring that all northing coordinate values will be positive within each UTM south zone (Figure 3-42).

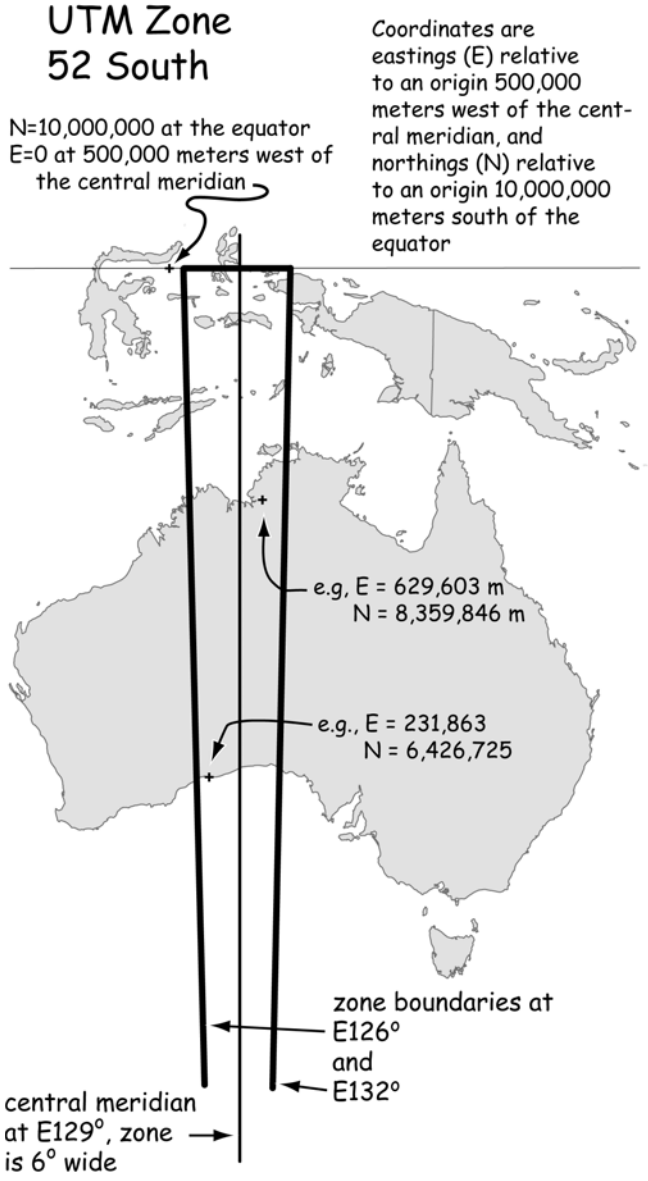


Figure 3-42: UTM south zones, such as Zone 52S shown here, are defined such that all the northing and easting values within the zone are positive. A false northing of 10,000,000 is applied to the equator, and a false easting of 500,000 is applied to the central meridian to ensure positive coordinate values throughout each zone.

Continental and Global Projections

There are map projections that are commonly used when depicting maps of continents, hemispheres, or regions. Just as with smaller areas, map projections for continental or larger areas may be selected based on the distortion properties of the resultant map. Sizeable projection distortion in area, distance, and angle are observed in most large-area projections. Angles, distances, and areas are typically not measured or computed from these projections, as the differences between the map-derived and surface-measured values are too great for most uses. Large-area maps are most often used to display or communicate data for continental or global areas.

There are a number of projections that have been or are 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. If a single, uncut surface is mapped, then there is severe distortion in some portion of the map. Figure 3-43 shows a Miller cylindrical projection, often used in maps of the world. This projection is similar to a Mercator projection, and is based on a cylinder that intersects the Earth at the equator. Distortion increases towards the poles, although not as much as with the Mercator.

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 3-44 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 the parallels of identical scale. The parallel of

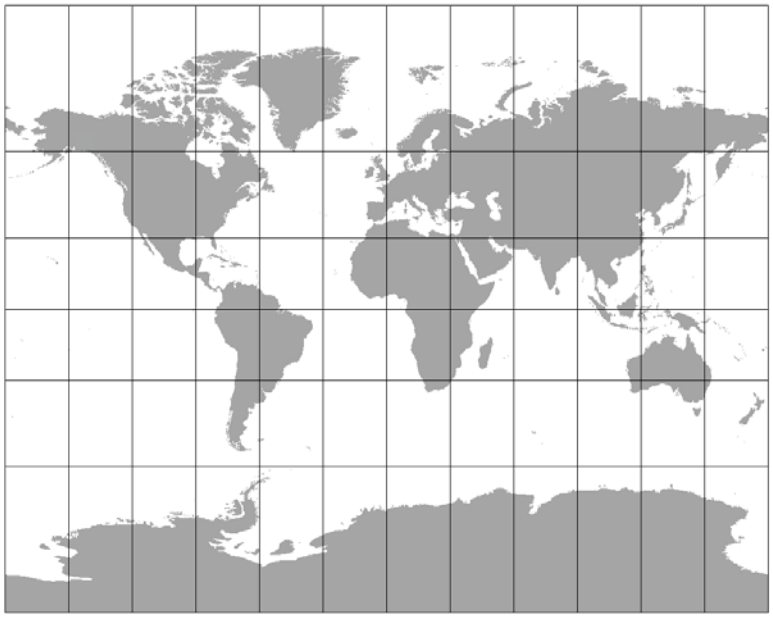


Figure 3-43: A Miller cylindrical projection, commonly used for maps of the world. This is an example of an uninterrupted map surface.

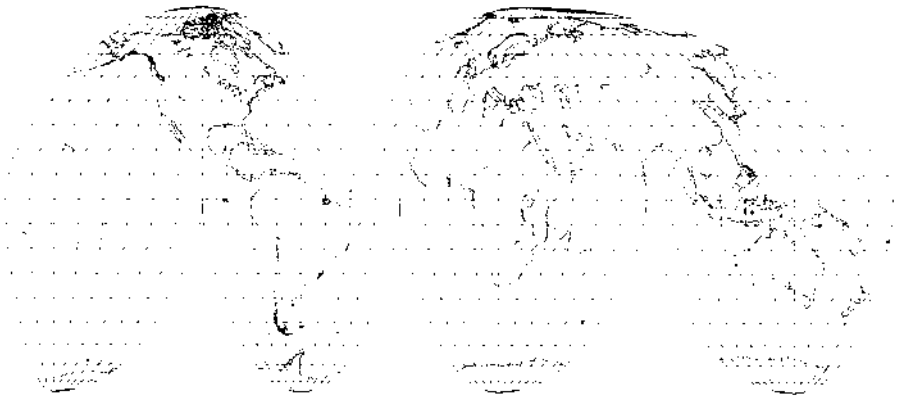


Figure 3-44: A Goode homolosine projection. This is an example of an interrupted projection, often used to reduce some forms of distortion when displaying the entire Earth surface. (from Snyder and Voxland, 1989)

identical scale in this example is set near the mid-northern latitude of 44^{o} 40' N.

Continental projections may also be established. Generally, the projections are chosen to minimize area or shape distortion for the region to be mapped. Lambert conformal conic or other conic projections are often chosen for areas with a long east-west dimension, for example when mapping the contiguous 48 United States of America, or North America (Figure 3-45). Standard parallels are placed near the top and bottom of the continental area to reduce distortion across the region mapped. Transverse cylindrical projections are often used for large north-south continents.

None of these worldwide or continental projections are commonly used in a GIS for data storage or analysis. Uninterrupted coordinate systems show too much distortion to be of use in measuring most spatial quantities, and interrupted projections do not specify a Cartesian coordinate system that defines positions for all points on the Earth surface. Worldwide data are typically stored in geographic coordinates (latitudes and longitudes). These data may then be projected to a specific coordinate system for display or document preparation.

Figure 3-45: A Lambert conformal conic projection of North America. Standard parallels are placed to reduce distortion within the projected continent.

Conversion Among Coordinate Systems

You might ask, how do I convert between geographic and projected coordinate systems? Exact or approximate mathematical formulas have been developed to convert to and from geographic (latitude and longitude) to all commonly used coordinate projections (Figure 3-46). These formulas are incorporated into "coordinate calculator" software packages, and are also integrated into most GIS software. 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 3-47, a).

Conversion from geographic (lon, lat) to projected coordinates

Given longitude = λ, latitude = ϕ

Mercator projection coordinates are:

$x = R \cdot (\lambda - \lambda_0)$
$y = R \cdot \ln (\tan (90^\circ + \phi/2)$

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)

Figure 3-46: Formulas are known for most projections that provide exact projected coordinates, if the latitude and longitudes are known. This example shows the formulas defining the Mercator projection.

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 3-47, b). A datum transformation 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 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 3-21, 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, 10s to 100s of meters between NAD27 and most versions of NAD83, and errors of up to a meter are common between recent versions of WGS84 and NAD83. Given the submeter 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.

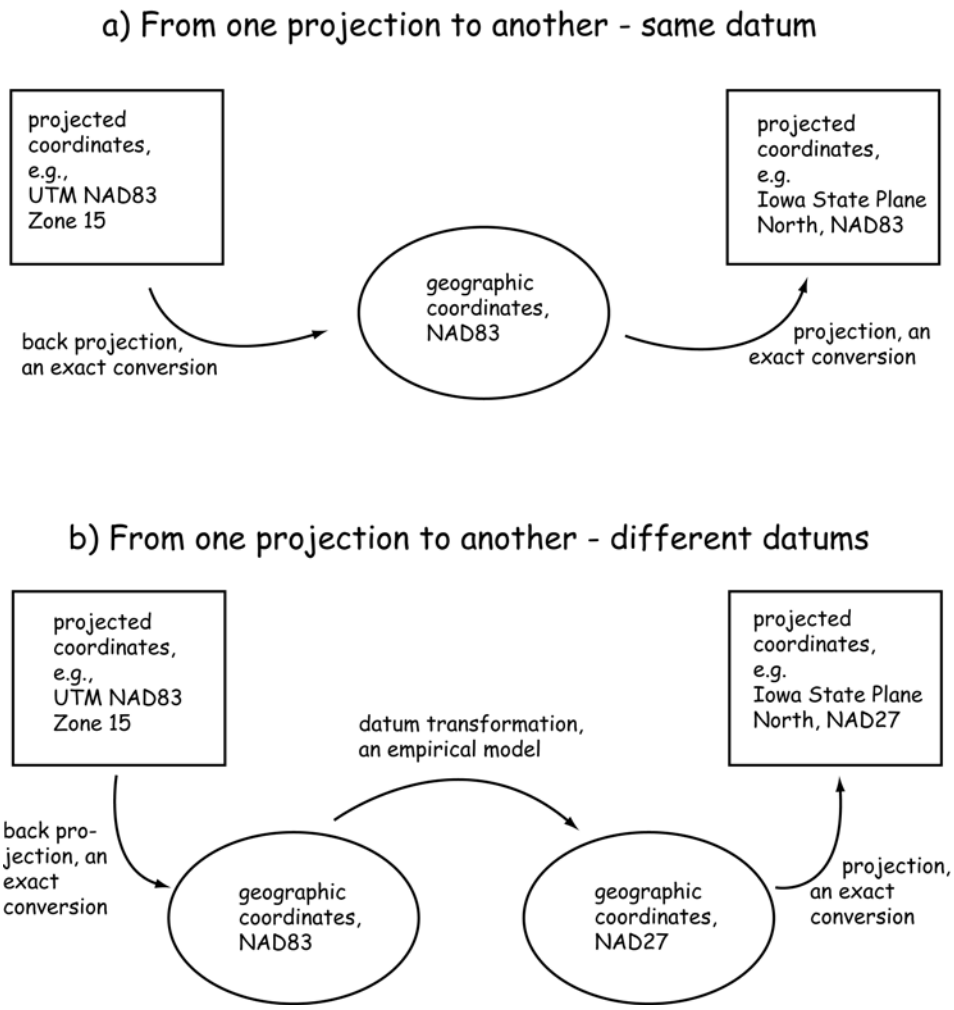


Figure 3-47: We may project between most coordinate systems via the back (or inverse) and forward projection equations. These calculate exact geographic coordinates from projected coordinates (a), and then new projected coordinates from the geographic coordinates. We must insert an extra step when a projection conversion includes a datum change. A datum transformation must be used to convert from one geodetic datum to another (b).

The Public Land Survey System

For the benefit 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 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. In colonial times a parcel description might state "beginning at the joining of Shope Fork and Coweeta Creek, downstream along the creek approximately 280 feet to a large rock on the right bank, thence approximately northwest 420 feet to a large chestnut oak blazed with an S, thence west 800 feet to Mill Creek, thence down Mill Creek to Shope Fork Road, thence east on Shope Fork Road to the confluence of Shope Fork and Coweeta Creek."

Metes and bounds descriptions require a minimum of surveying measurements, but it was a less than ideal system for describing locations or parcels. As you might surmise, these metes and bounds descriptions could be vague, the features in the landscape might be moved or change, and it was difficult to describe parcels when there were few readily distinguished landscape features. Subdivided parcels were often poorly described, and hence the source of much litigation, ill will, and many questionable real estate transactions.

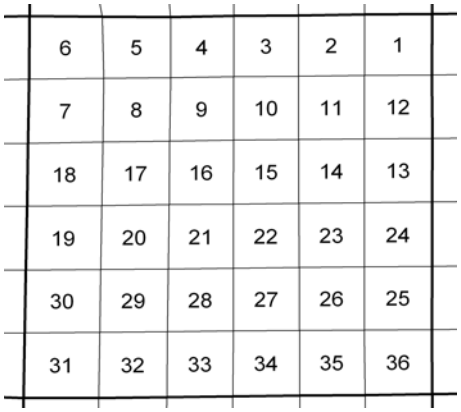


Figure 3-48: Typical layout and section numbering of a PLSS township

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, therefore the PLSS was created. Land surveyed under the PLSS can be found in thirty 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.

The PLSS divided lands by north-south lines running parallel to a principal meridian. New north-south lines were surveyed at six mile intervals. Additional lines were surveyed that were perpendicular to these north-south lines, in approximately east-west directions, and crossing meridian lines. These east-west lines were also run at six-mile intervals. These lines form townships that were six miles square. 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 six-teenth sections, (one-quarter mile on a side, commonly referred to as quarter-quarter sections). Sections were numbered in a zig-zag pattern from one to 36, beginning in the northeast corner (Figure 3-48).

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 re-adjustment, 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 3-49).

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, so they are often evident on the landscape. 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 co-register GIS data

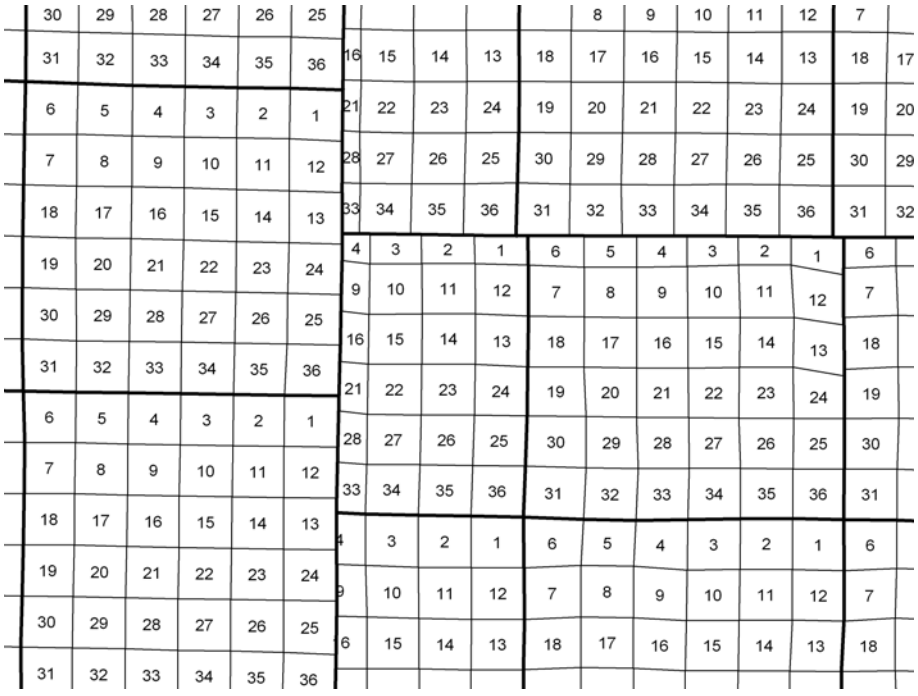


Figure 3-49: Example of variation in the size and shape of PLSS sections. Most sections are approximately one mile square with section lines parallel or perpendicular to the primary meridian, as illustrated by the township in the upper left of this figure. However, adjustments due to different primary meridians, different survey parties, and errors result in irregular section sizes and shapes.

Figure 3-50: 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 as control point for further spatial data development.

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 re-surveyed using high precision methods to provide property line control, particularly when a GIS is to be developed (Figure 3-50). Thus these points may be useful to properly locate and orient spatial data layers on the Earth's surface.

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 was established in the United States known as the

Public Land Survey System (PLSS). 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.

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

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

Iliffe, J.C. (2000). *Datums and Map Projections for Remote Sensing, GIS, and Surveying*. Boca Raton: CRC Press.

Keay, J. (2000). *The Great Arc*. London: 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.

Schwartz, C.R. (1989). *North American Datum of 1983, NOAA Professional Paper NOS 2*, Rockville: National Geodetic Survey.

Smith, J. (1997). *Introduction to Geodesy: The History and Concepts of Modern Geodesy*, New York: Wiley.

Sobel, D. (1995). *Longitude*. New York: Penguin Books.

Soler, T. & Snay, R.A.(2004). Transforming positions and velocities between the International Terrestrial Reference Frame of 2000 and the North American Datum of 1983. *Journal of Surveying Engineering*, 130:49-55.

Snay, R.A. & Soler, T. (1999). Modern terrestrial reference systems, part 1. *Professional Surveyor*, 19:32-33.

Snay, R.A. & Soler, T. (2000). Modern terrestrial reference systems, part 2. the evolution of NAD83, *Professional Surveyor*, 20:16-18.

Snay, R.A., & Soler, T. (2000). Modern terrestrial reference systems, part 3. WGS84 and ITRS, *Professional Surveyor*, 20:24-28.

Snay, R.A., & Soler, T. (2000). Modern terrestrial reference systems, part 4, practical considerations for accurate positioning. *Professional Surveyor*, 20:32-34.

Snyder, J. (1993). *Flattening the Earth: Two Thousand Years of Map Projections*. Chicago: University of Chicago Press, Chicago.

Snyder, J. P. (1987). *Map Projections, A Working Manual, USGS Professional Paper No. 1396*. Washington D.C.: United States Government Printing Office.

Snyder, J.P., & Voxland, P.M. (1989). *An Album of Map Projections, USGS Professional Paper No. 1453*. Washington D.C.: United States Government Printing Office.

Tobler, W.R. (1962). A classification of map projections. *Annals of the Association of American Geographers*, 52:167-175.

Van Sickle, J. (2004). Basic GIS Coordinates. Boca Raton: CRC Press.

Welch, R., & Homsey, A. (1997). Datum shifts for UTM coordinates. *Photogrammetric Engineering and Remote Sensing*, 63:371-376.

Wolf, P. R., & Ghilani, C.D. (2002). Elementary Surveying (10th ed.). Upper Saddle River: Prentice-Hall.

Yang, Q., Snyder, J.P. & Tobler, W.R. (2000). *Map Projection Transformation: Principles and Applications.* London: Taylor & Francis.

Zilkoski, D., Richards, J. & Young, G. (1992). Results of the general adjustment of the North American Vertical Datum of 1988. *Surveying and Land Information Systems*, Prentice-Hall 53:133-149

Study Questions

3.1 - Can you describe how Eratosthenes estimated the circumference of the Earth? What value did he obtain?

3.2 - What is an ellipsoid? How does an ellipse differ from a sphere? What is the equation for the flattening factor?

3.3 - Why do different ellipsoids have different radii? Can you provide three reasons?

3.4 - Can you define the geoid? How does it differ from the ellipsoid, or the surface of the Earth? How do we measure the position of the geoid?

3.5 - Can you define a parallel or meridian in a geographic coordinate system? Where do the “horizontal” and “vertical” zero lines occur?

3.6 - How does magnetic north differ from the geographic North Pole?

3.7 - Can you define a datum? Can you describe how datums are developed?

3.8 - Why are there multiple datums, even for the same place on Earth? Can you define what we mean when we say there is a datum shift?

3.9 - What is a triangulation survey, a Bilby tower, and a benchmark?

3.10 - Use the NADCON software available from the U.S. NOAA/NGS website (http://www.ngs.noaa.gov/TOOLS/program_descriptions.html) to fill the following table. Note that all of these points are in CONUS, and longitudes are west, but entered as positive numbers.

	NAD27		NAD83(86)		HPGN	
Pnt	latitude	longitude	latitude	longitude	latitude	longitude
1	32°44'15"	117°09'42"	32°44'15.1827"	117°09'45.1202"	32°44'15.1820"	117°09'45.1200"
2	47°27'55"	122°18'06"	47°27'54.3574"	122°18'10.4453"	47°27'54.3642"	122°18'10.4366"
3	43°07'59"	89°20'11"	43°07'58.9806"	89°20'11.4226"		
4	29°58'07"	95°21'31"	29°58'07.7975"	95°21'31.7705"		
5	40°00'00"	105°16'01"			39°59'59.9552"	105°16'02.9712"
6	24°33'30"	81°45'19"			24°33'31.5216"	81°45'18.3362"
7			38°51'10.4052"	77°02'19.9165"	38°51'10.4063"	77°02'19.9041"
8			46°52'0.1524"	68°00'59.0974"	46°52'0.1580"	68°00'59.0995"

3.11 - Use the web version or download and start the HTDP software from the U.S. NOAA/NGS site listed above, and complete the following table. Enter epoch start and stop dates of 1, 1, 1986 and 1,1, 2005, respectively, and specify a zero height or z.

	NAD83(CORS96)		WGS(G1150)		ITRF2005	
Pnt	latitude	longitude	latitude	longitude	latitude	longitude
1	32°44'15"	117°09'42"	32°44'15.0321"	117°09'42.0662"	32°44'15.0325"	117°09'45.0663"
2	47°27'55"	122°18'06"	47°27'55.0183"	122°18'06.0583"	47°27'55.0186"	122°18'06.0583"
3	43°07'59"	89°20'11"	43°07'59.0283"	89°20'11.0293"		
4	29°58'07"	95°21'31"	29°58'07.0177"	95°21'31.0293"		
5	40°00'00"	105°16'01"			40°00'00.2143"	105°16'01.422"
6	24°33'30"	81°45'19"			24°33'30.0164"	81°45'19.0145"
7			38°51'1.0288"	77°02'21.0137"	38°51'1.0293"	77°02'21.0136"
8			46°52'0.0363"	68°01'00.0061"	46°52'0.0367"	68°01'01.0061"

3.12 - What is a developable surface? What are the most common shapes for a developable surface?

3.13 - Can you describe the State Plane coordinate system? What type of projections are used in a State Plane coordinate system?

3.14 - Can you define and describe the Universal Transverse Mercator coordinate system? What type of developable surface is used with a UTM projection? What are UTM zones, where is the origin of a zone, and how are negative coordinates avoided?

3.15 - What is a datum transformation? How does it differ from a map projection?

3.16 - Can you describe the Public Land Survey System? Is it a coordinate system? What is its main purpose?