Chapter 3: Geodesy, Datums, Map Projections, and Coordinate Systems

Meng Lu

Juniorprofessor of Geoinformatics, Bayreuth University

1 Introduction

2 Map Projections and Coordinate Systems

Introduction

To use GIS, we must be clear about how coordinate systems are established for the Earth, how these coordinates are measured on the Earths curving surface, and how these coordinates are converted for use in flat maps, either digital or paper. Here we will introduce two important concepts:

- Geodesy: Is the science of measuring the shape of the Earth
- Map projections: Transformation of coordinate locations from the Earths curved surface onto flat maps.

To use GIS, we must be clear about how coordinate systems are established for the Earth, how these coordinates are measured on the Earths curving surface, and how these coordinates are converted for use in flat maps, either digital or paper. Here we will introduce two important concepts:

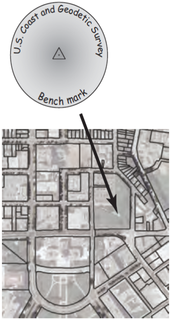
- Geodesy: Is the science of measuring the shape of the Earth
- Map projections: Transformation of coordinate locations from the Earths curved surface onto flat maps.

Defining the coordinates for the Earths surface is complicated by four main factors:

1. The Earths curvature: a flat map distort geometry
2. The Earths deviation from our idealized shape: the irregular shape of the Earth
3. Inevitable inaccuracies in measurement: our measurements are rarely perfect, and this applies when measuring both the shape of the Earth, and the exact position of features on it.
4. Physical shifts: the physical locations of points on the Earth change through time, due to plate tectonics and vertical crustal movements.

We often have several different sets of coordinates to define the same location on the surface of the Earth (Fig.1).

Coordinates for a Point Location



From Surveyor Data:

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

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

From Data Layers:

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

Figure 1: An example of different coordinate values for the same point (Bolstad (2016)).

Early Measurements

In specifying a coordinate system, we must first define the size and shape of the Earth (Fig. 2).

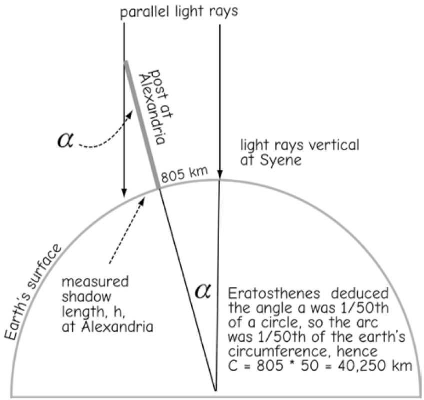


Figure 2: Measurements made by Eratosthenes to determine the circumference of the Earth (Bolstad (2016)).

Specifying the Ellipsoid

Earth radii have been determined by a number of methods. Measurements during the 18th, 19th and early 20th centuries used optical instruments for celestial observations (Fig 3).

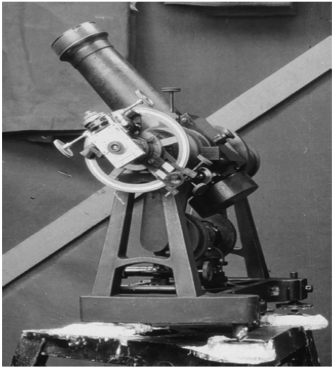


Figure 3: An instrument used in the early 1900s for measuring the position of celestial bodies (Bolstad (2016)).

Celestial observations of the stars are combined with long-distance surface measurements over large areas (Fig 4).

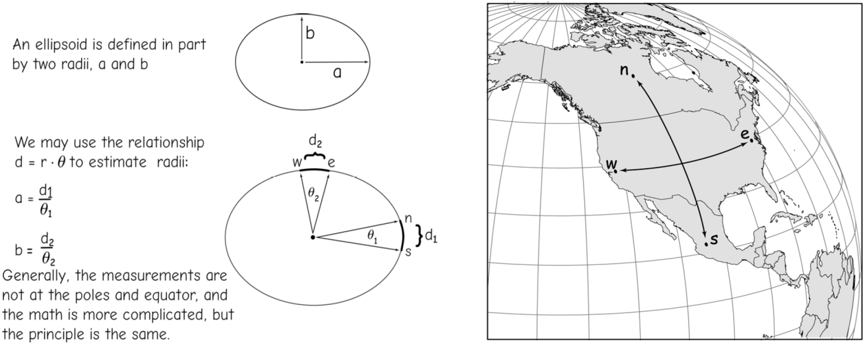


Figure 4: Two arcs illustrate the surface measurements and calculations used to estimate the semi-major and semi-minor axes, here for North America (Bolstad (2016)).

Since the 1980s, data derived from satellites, lasers, and broadcast timing signals have been used for much more precise measurements of relative positions across continents and oceans (E.g. GRS80).

Name	Year	equatorial Radius, a meters	Polar Radius, b meters	Flatten-ing Factor	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; Philip-pines
Clarke	1880	6,378,249.1	6,356,514.9	1/293.46	Most of Africa; France
Interna-tional	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, U.S. DOD
GRS80	1980	6,378,137.0	6,356,752.3	1/298.26	Worldwide
WGS84	1987 – current	6,378,137.0	6,356,752.3	1/298.26	U.S. DOD, Worldwide

The Geoid

Geoid is the **shape that the ocean surface would take under the influence of the gravity of Earth.**
Variation's in the strength of the gravitational pull are caused by the differences in the density of the Earth (Fig. 6).

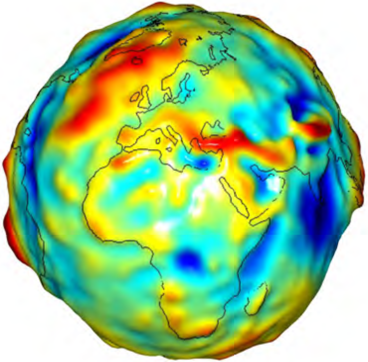


Figure 6: Depictions of the Earths gravity field, as estimated from satellite measurements (Bolstad (2016)).

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 Earths gravity.

The next figure shows the differences in the Earths shape due to geoidal deviations will produce different best local ellipsoids.

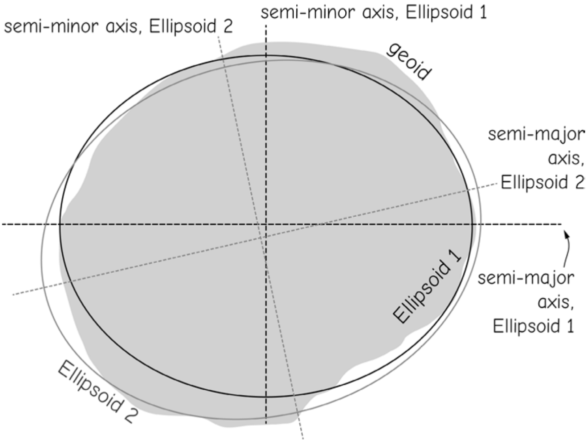


Figure 7: Different ellipsoids were estimated due to local irregularities in the Earths shape (Bolstad (2016)).

Because we have two reference surfaces, a geoid and an ellipsoid, we also have two bases from which to measure height.

- Elevation is defined as the distance above a geoid and is called the *orthometric height*.
- Heights above an ellipsoid are often referred to as *ellipsoidal height*.

Orthometric height labeled H and ellipsoidal height labeled h (Figure 8).

The difference between the ellipsoidal height and geoidal height at any location (N), has various names, including geoidal height and geoidal separation.

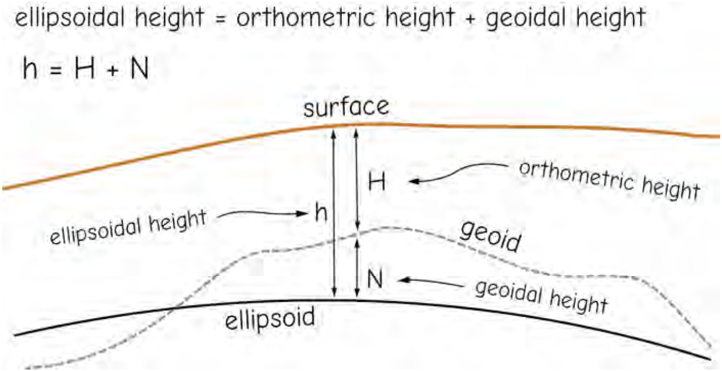


Figure 8: Ellipsoidal, orthometric, and geoidal height are interrelated. This formula, e.g., to calculate orthometric height (elevation) if when we know the ellipsoidal height (say, from GPS), and geoidal height (from national models) (Bolstad (2016)).

The absolute value of the geoidal height is less than 100 meters over most of the Earth (Fig 9)

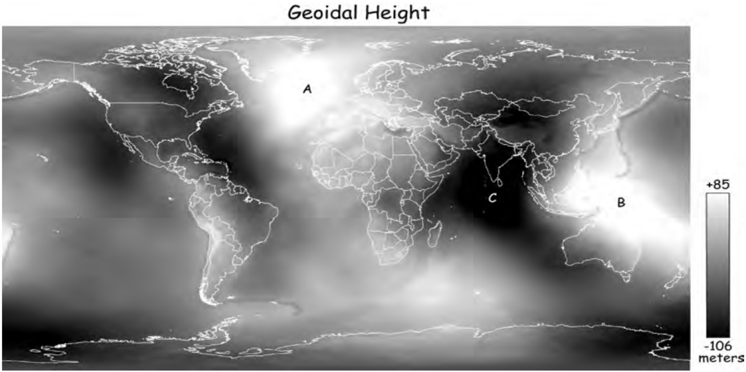


Figure 9: Geoidal heights vary across the globe. This figure depicts positive geoidal heights in lighter tones (geoid above the ellipsoid) and negative geoidal heights in darker tones (Bolstad (2016)).

Datums

Datum

A geodetic datum is a specific, defined coordinate system. With a datum we could describe the location of the geographic measurements. It consists of well-surveyed (reference) points that allow us to specify a *reference frame*.

Geographic coordinate systems typically use an ellipsoid to calculate positions on the earth. A datum defines the position of the ellipsoid relative to the center of the earth.

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

Different datums are specified through time because our estimates of the datum, change through time: New points are added and survey methods improve.

There are two main eras of datums, those created before satellites geodesy (astronomical observations), and those after (much more precise with larger number of datum points).

Datums and coordinates found today are a mix of those developed under pre-satellite datums, and those referenced to post-satellite datums, so the GIS user should be familiar with both.

Datum Adjustment

The positions of all points in a reference datum are estimated in a network-wide *datum adjustment*.

Figure 10 illustrates how ellipsoids might change over time, even for the same survey region. Ellipsoid A is estimated with the datum coordinates for pt1 and pt2, with the shown corresponding coordinate axes, origin, and orientation. Ellipsoid B is subsequently fit, after pts 3 through 7 have been collected.

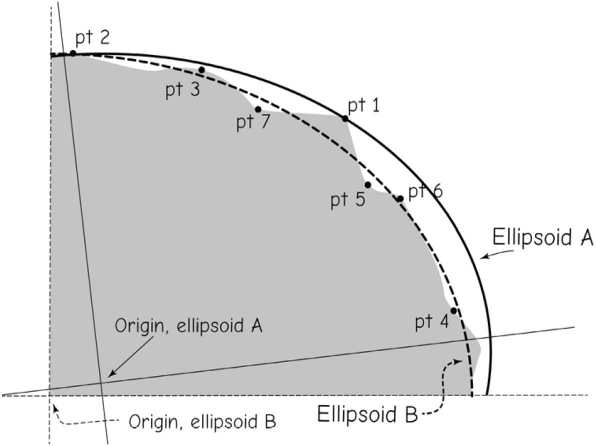


Figure 10: An illustration of two datums, one corresponding to Ellipsoid A and based on the fit to pt1 and pt2, and a subsequent datum resulting in Ellipsoid B, and based on a fit of pt1 through pt7 (Bolstad (2016)).

Commonly Used Datums

Three main series of horizontal datums have been used widely in North America.

- North American Datum of 1983: NAD83 (GRS80 ellipsoid: the best overall fit to observed measurements across the globe), NAD83 is the successor to NAD27 (Clarke ellipsoid of 1986)
- The World Geodetic System of 1984: WGS84 (WGS84 ellipsoid, similar to GRS80)
- International Terrestrial Reference Frames: ITRF (do not directly use an ellipsoid, if needed can be transferred to GRS80). It aligns with WGS84 but not NAD83. A primary purpose is to estimate continental drift and crustal deformation.

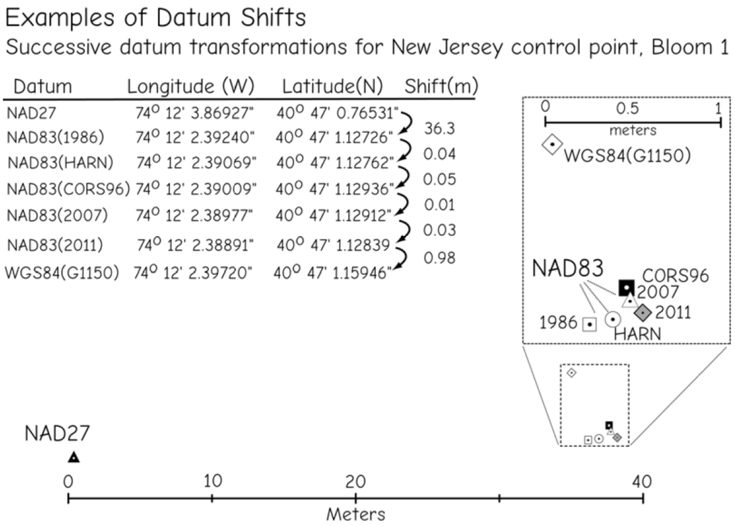


Figure 11: Datum shifts in the coordinates of a point for some common datums (Bolstad (2016)).

There are a few points about datums that must be emphasized.

- Different datums specify different coordinate systems
- The version of the datum is important, for example, NAD83(1996) is a different realization than NAD83(2011). The datum is incompletely specified unless the version is noted.
- Differences between families of datums change through time. This means you should assume all data should be converted to the same datum, via a datum transformation, before combination in a GIS.

Datum Transformations

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(2011) when we know these geographic coordinates in NAD83(1986) (Fig 12)

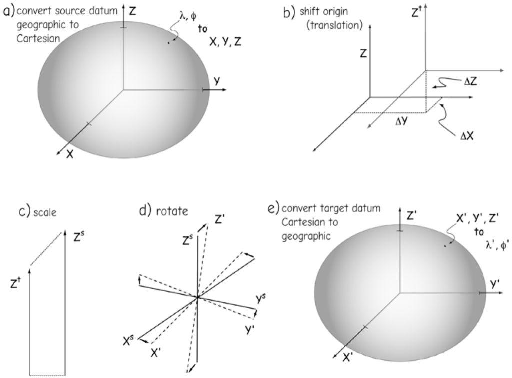


Figure 12: Application of a modern datum transformation. Geographic coordinates (longitude, λ, and latitude, ϕ, are transformed to a new datum (Bolstad (2016)).

A mathematical geocentric datum transformation is typically a multistep process. These datum transformations are based on one of a few methods.

- *Molodenski transformation* was common, using a system of equations with three or five parameters.
- *Helmert transformation* is employed using seven or 14 parameters (in the previous Fig 12)

First, geographic coordinates on the source datum are converged from longitude (λ) and latitude (ϕ) to X, Y, and Z Cartesian coordinates. An origin shift (translation), rotation, and scale are applied. This system produces new X', Y', and Z' coordinates in the target datum.

Datums shifts associated with datum transformations have changed with each successive datum realization, as summarized in Figure 13.

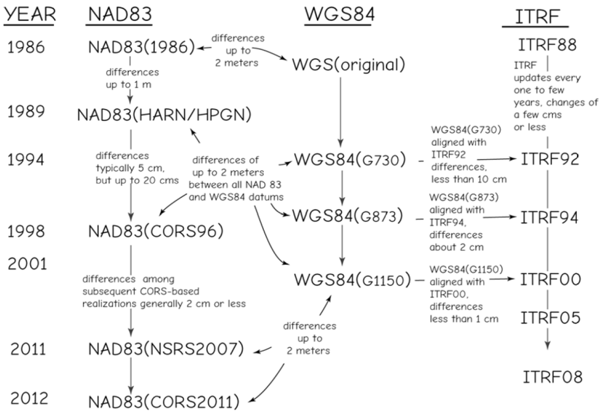


Figure 13: This graphic summarizes the evolution of the three main families of datums used in North America (Bolstad (2016)).

The datum transformation method within any hardware or software system should be documented and the accuracy of the method known before it is adopted. Unfortunately, much data are now degraded because of improper datum transformations.

There are a number of factors that we should keep in mind when applying datum transformations.

1. Changing a datum changes our best estimate of the coordinate locations of most points.
2. Datum transformations are estimated relationships that are developed with a specific data set and for a specific area and time.
3. GIS projects should not mix datums except under circumstances when the datum shift is small relative to the requirements of the analysis.

Vertical Heights and Datums

A *vertical datum* is:

A reference that we use for measuring heights. A vertical datum is found by measuring gravity and identifying variations in constant gravity (equipotential) surfaces, and then combining these with carefully measuring control heights above a specific equipotential surface.

Most government or other organizations use a specific geoid as a reference surface for height, although not everyone adopts the same geoid. For example, governments adopt hybrid geoids, that combine their own precise vertical surveys with gravity measurements and models.

- Because the word elevation is used for many other things, we call these elevations *orthometric heights*.
- *Leveling surveys* were used for establishing heights prior to satellites.
- *Spirit leveling* using plumb bobs and bubble or tube levels.

All the points on an equipotential surface have the same gravitational pull, and they may be envisioned as layers, like an onion, with decreasing strength the further they are away from the mass center of the Earth (Fig 14).

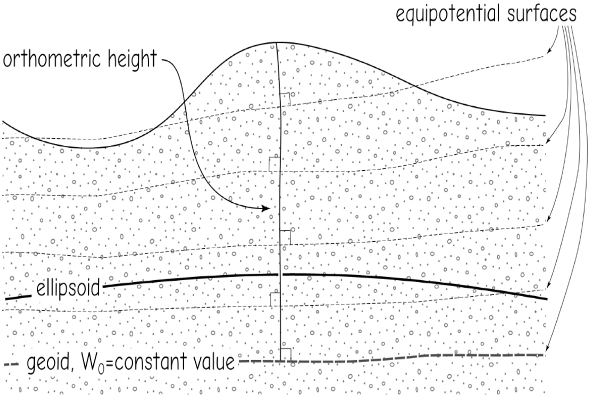


Figure 14: Heights are referenced to a geoid, corresponding to a given equipotential surface (Bolstad (2016)).

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 (Fig 15).

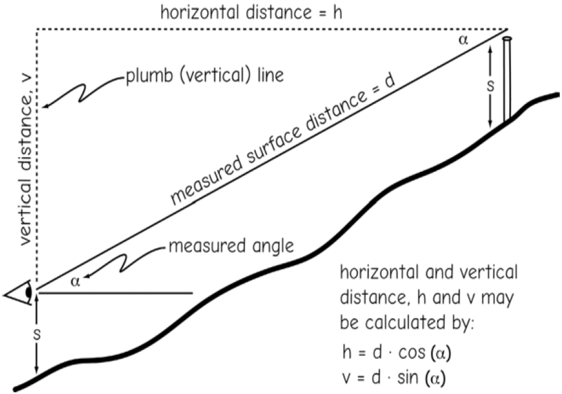


Figure 15: Leveling surveys often employ optical measurements of vertical angle (α) with measurements of surface distance (d) and knowledge of trigonometric relationships to calculate horizontal

Surface distance along the slope was measured to avoid the tedious process of establishing vertical posts and leveling rods.The vertical angle was also measured from a known station, typically by a small telescope fitted with a precisely scribed angle gauge. Surface distance would then be combined with the measured vertical angle to calculate the horizontal and vertical distances.

Vdatum

Given that vertical datums and associated geoids change through time, the United States National Geodetic Survey (NGS) has created a tool, VDatum, to estimate conversions among vertical datums in the U.S (Fig 16).

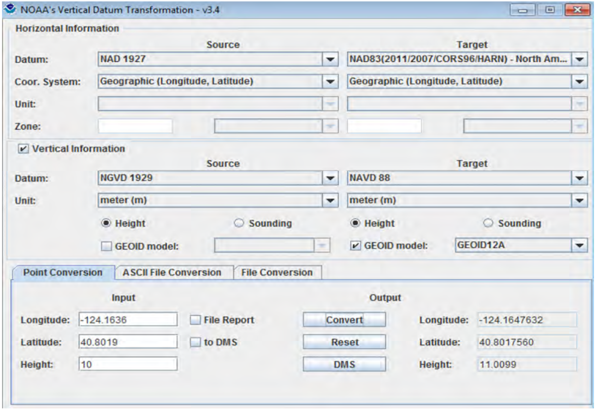


Figure 16: An example of the application of the vertical datum transformation software VDatum (Bolstad (2016)).

VDatum calculates the vertical difference from one datum to another at any given horizontal coordinate location and height.

Dynamic Heights

We must discuss one final kind of height, called a dynamic height, because they are important for certain applications.

Dynamic heights measure the change in gravitational pull from a given equipotential surface. They are important when we are interested in water levels and flows across elevations.

Sometimes the points with the same dynamic heights often have different orthometric heights (Fig 17).

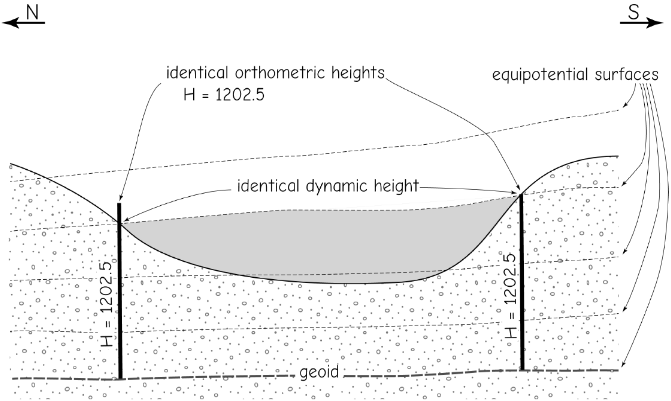


Figure 17: An illustration of how dynamic heights and orthometric heights may differ, and how equal orthometric heights may correspond to different heights above the water level on a large lake (Bolstad (2016)).

To remember:

An orthometric height is the distance, in the direction of gravitational pull, from the geoid up to a point. The geoid is a specified gravity value, an equipotential surface, where the pull of gravity is at some specified level.

Because water follows an equipotential surface, and because the Earths polar radius is less than the equatorial radius, the orthometric heights of the water surface on large lakes are usually different at the north and south ends.

Dynamic heights are most often used when were interested in relative heights for water levels, particularly over large lakes or connected water bodies.

Because water follows an equipotential surface, and because the Earths polar radius is less than the equatorial radius, the orthometric heights of the water surface on large lakes are usually different at the north and south ends.

Dynamic heights are most often used when were interested in relative heights for water levels, particularly over large lakes or connected water bodies.

Map Projections and Coordinate Systems

Datums tell us the latitudes and longitudes of features on an ellipsoid, thus we need to transfer these from the curved ellipsoid to a flat map.

A map projection is a systematic rendering of locations from the curved Earth surface onto a flat map surface.

Datums tell us the latitudes and longitudes of features on an ellipsoid, thus we need to transfer these from the curved ellipsoid to a flat map.

A map projection is a systematic rendering of locations from the curved Earth surface onto a flat map surface.

Conversion from geographic (longitude, latitude) to projected coordinates. Here longitude $= \lambda$, latitude $= \phi$ (all the angles in radians)

Mercator projection coordinates are:

$$
\begin{aligned}
x =& R * (\lambda - \lambda_0) \\
y =& R * \ln(\tan(\pi/4 + \phi/2))
\end{aligned}
$$

where R is the radius of the sphere at map scale and λ_0 is the longitudinal origin.

From x, y to λ and ϕ

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

Most map projections may be viewed as sending rays of light from a projection source through the ellipsoid and onto a map surface (Fig 18)

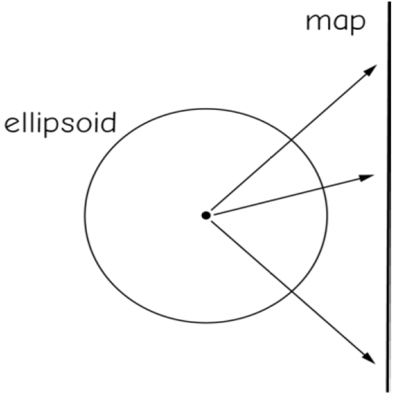


Figure 18: conceptual view of a map projection (Bolstad (2016)).

Distortions are unavoidable when making flat maps because of the transition from a complexly curved Earth surface to a flat or simply curved map surface (Fig 19).

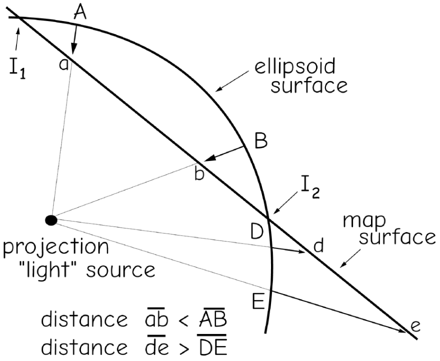


Figure 19: Distortion during map projection (Bolstad (2016)).

In the Figure 19 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 Earths surface measurements
    - 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

Figure 20 shows an example of distortion with a projection onto a planar surface.

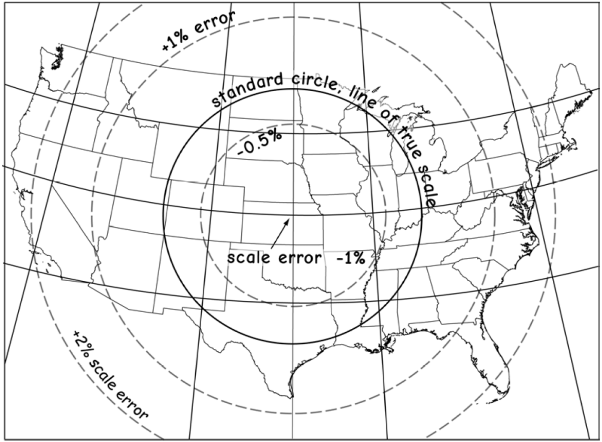


Figure 20: Approximate error due to projection distortion for a specific oblique stereographic projection (Bolstad (2016)).

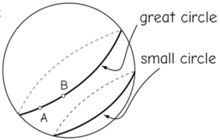
An approximation of the distance distortion may be obtained for any projection by comparing grid coordinate distances to *great circle distances*. A great circle distance is a distance measured on the spheroid or ellipsoid and in a plane through the Earths center.

This planar surface intersects the two points on the Earths surface and also splits the spheroid into two equal halves (Fig 21).

Great Circle vs. Projected Distance
Spherical Approximation

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

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



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

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

$$d = 6378 \cdot \cos^{-1}[(\sin(30.4877456) \cdot \sin(29.7507171) + \cos(30.4877456)\cos(29.7507171)\cdot\cos(-91.1693348 - \{-95.370003\})]$$
$$= 412.681 \text{ km}$$

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

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

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

Figure 21: Example calculation of the distance distortion due to a map projection (Bolstad (2016)).

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

Figure 22: Curved representations of straight lines are a manifestation of projection distortion (Bolstad (2016)).

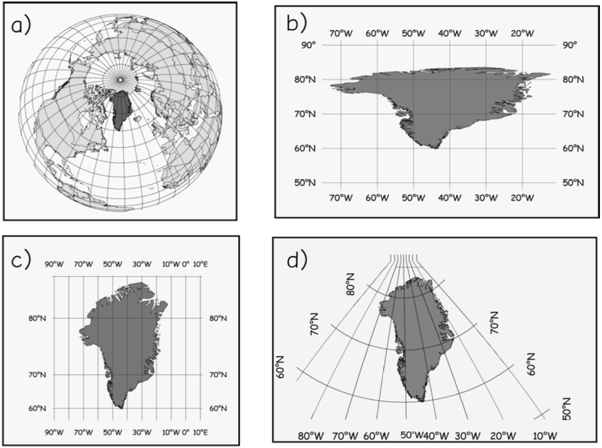


Figure 23: Map projections can distort the shape and area of features, as illustrated with these various projections of Greenland, from a) approximately unprojected, b) geographic coordinates on a plane, c) a Mercator projection, and d) a transverse Mercator projection. (Bolstad (2016)).

Most map projections are based on a *developable surface*, a geometric shape onto which the Earth surface is projected. Cones, cylinders, and planes are the most common developable surfaces (Fig 24).

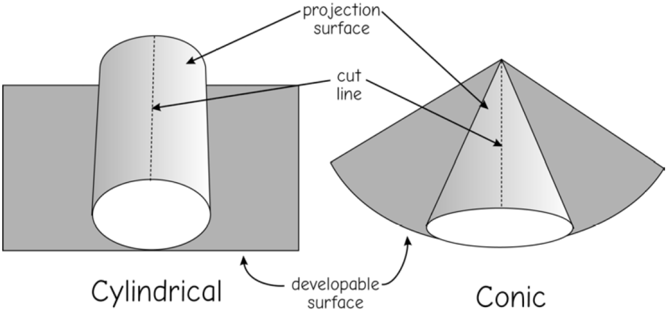


Figure 24: Projection surfaces are derived from curved developable surfaces that may be mathematically unrolled to a flat surface (Bolstad (2016)).

Common Map Projections in GIS

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

The Lambert conformal conic and the transverse Mercator are among the most common projection types used for spatial data in North America, and much of the world.

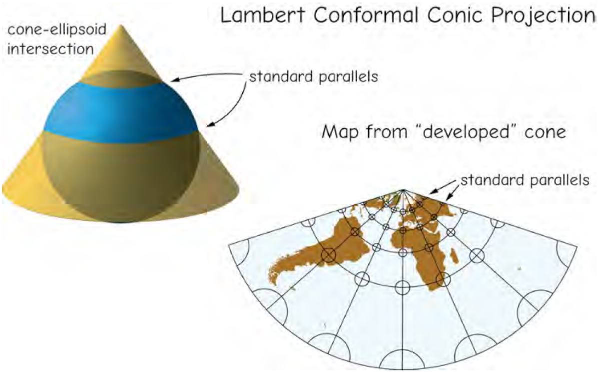


Figure 25: Lambert conformal conic (LCC) projection and an illustration of the scale distortion associated with the projection (Bolstad (2016)).

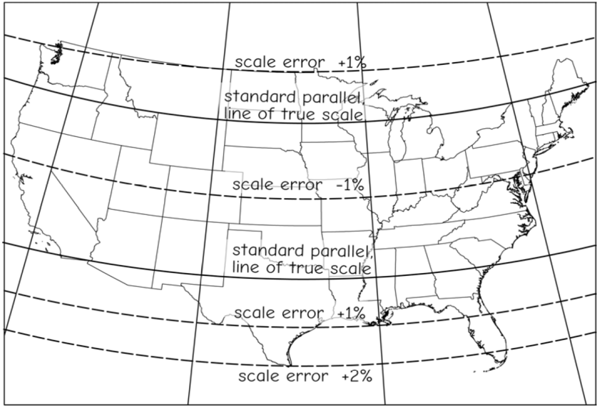


Figure 26: Distortion is primarily in the northsouth direction, and is illustrated in the developed surfaces by the deformation of the 5-degree diameter geographic circles (top) and by the lines of approximately equal distortion (Bolstad (2016)).

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 Earths surface onto the cylinder (Figure 27).

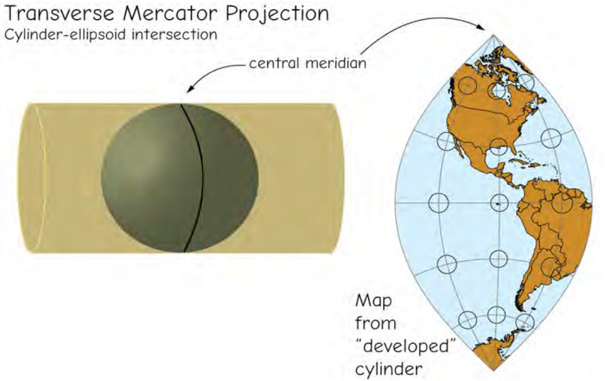


Figure 27: Transverse Mercator (TM) projection (Bolstad (2016)).

Figure 28: Scale distortion associated with the projection (Bolstad (2016)).

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 projection parameters in each State Plane zone.

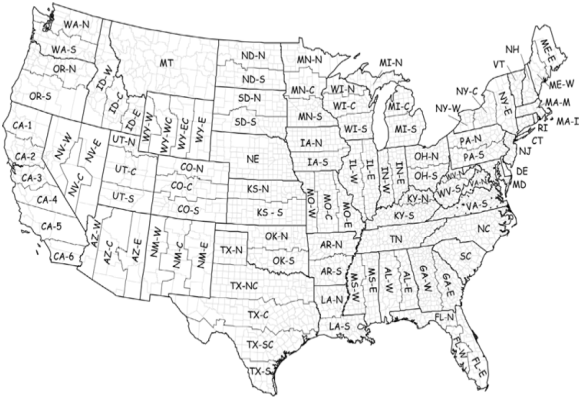


Figure 29: Scale distortion associated with the projection (Bolstad (2016)).

The State Plane system provides a common coordinate reference for horizontal coordinates over county to multicounty areas while limiting distortion error to specified maximum values.

The State Plane coordinate system is based on two types of map projections:

- The Lambert conformal conic: is most often used when the long axis of a state or zone is in the eastwest direction
- The transverse Mercator projections: this projection type is most often used with states or zones that have a long northsouth axis.

Universal Transverse Mercator Coordinate System

The UTM is a global coordinate system, based on the transverse Mercator projection. It is widely used in many 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 (Fig 30).

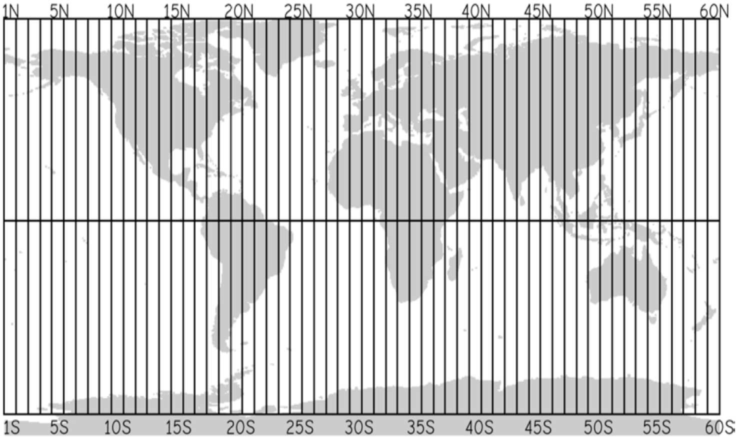


Figure 30: UTM zones for the lower 48 contiguous states of the United States of America (Bolstad (2016)).

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.

Distances in the UTM system are specified in meters north and east of a zone origin (Fig 31). 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.

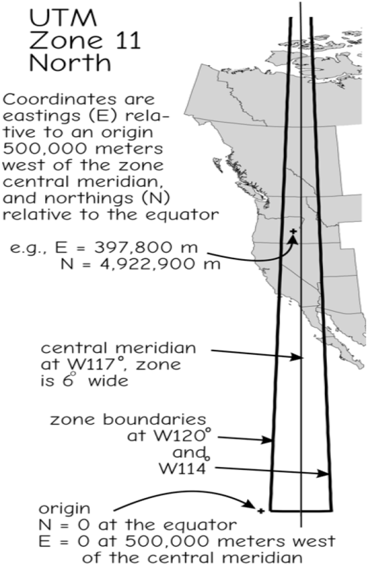


Figure 31: UTM zone 11N (Bolstad (2016)).

University Transverse Mercator zones south of the equator are slightly different than those north of the equator (Fig 32). 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).

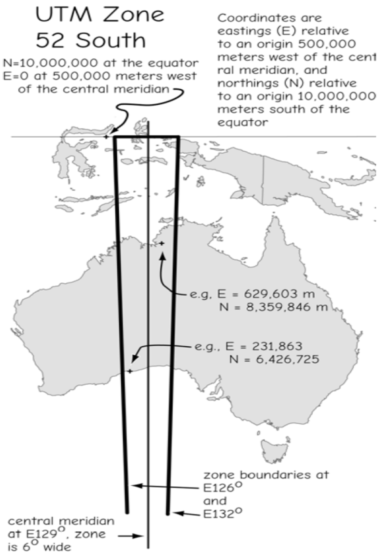


Figure 32: UTM south zones(Bolstad (2016)).

National Coordinate Systems

Many governments have adopted a standard project for nationwide data. European countries have standard map projections covering a national extent; for example:

- Belgium, Estonia, and France (each have different Lambert Conformal Conic projections)
- Germany, Bulgaria, Croatia, and Slovenia (use a specialized modification of the transverse Mercator projection)
- Norway, Portugal, and Spain (Universal Transverse Mercator projections).

Larger countries may not have a specific or unified set of standard, nationwide projections, particularly for GIS data, because distortion is usually unavoidably large when spanning great distances across both latitudes and longitudes in the same map.

Continental and Global Projections

Just as with smaller areas, map projections for continental areas may be selected based on the distortion properties of the resultant map. Sizeable projection distortion in area, distance, and angle are common 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.

Most worldwide projections are used for visualization, but not quantitative analysis.

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.

Most worldwide projections are used for visualization, but not quantitative analysis.

The Figure 33 is based in the projection of a sinusoidal projection and a Mollweide projection. These two projection types are merged at parallels of identical scale. The parallel of identical scale in this example is set near the midnorthern latitude of $44^\circ\ 40'$ N.

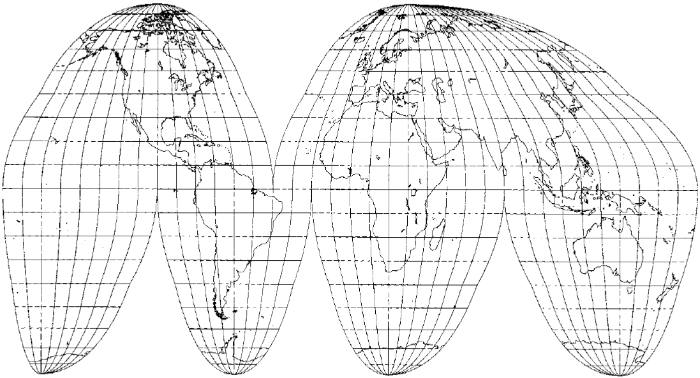


Figure 33: UTM south zones(Bolstad (2016)).

Conversion Among Coordinate Systems

Conversion from one projected coordinate system to another requires using the inverse and forward projection equations, described in an earlier section, passing through the geographic coordinate set.

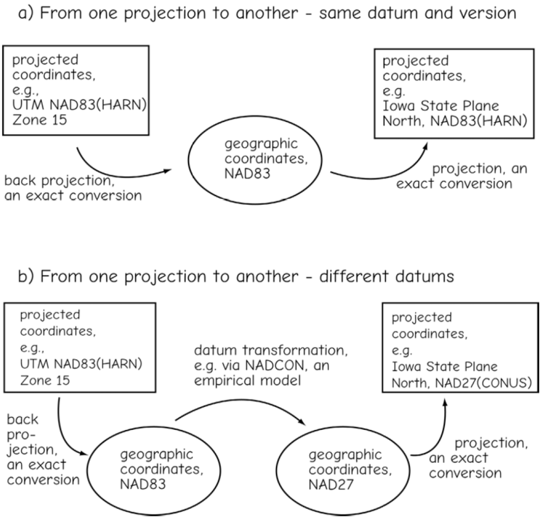


Figure 34: We may project between most coordinate systems via the back (or inverse) and forward projection equations (Bolstad (2016)).

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.

Thank You

References I

Bolstad, P. (2016). *GIS fundamentals: A first text on geographic information systems*. Eider (PressMinnesota).