3.3 Map coordinate systems, projections, and datums *(Jonathan W. Chipman, Dec. 2022)*

Geographic coordinates

Locations of points on or around the Earth are often specified using a spherical (or ellipsoidal) coordinate system. The Earth rotates around its **polar axis**, which connects the **north and south poles**, and corresponds to the minor (shorter) axis of the ellipsoid, due to polar flattening, as discussed in the previous section. The **equatorial plane** is perpendicular to the polar axis, and intersects it at the midpoint of that axis, the center of the ellipsoid. The **equator** is the circle or ellipse formed by the intersection of the equatorial plane with the ellipsoid. Finally, the **prime meridian** (an arbitrary reference meridian) is defined as the arc along the ellipsoid connecting the north and south poles and passing through a given reference point (Figure 3.3.1).

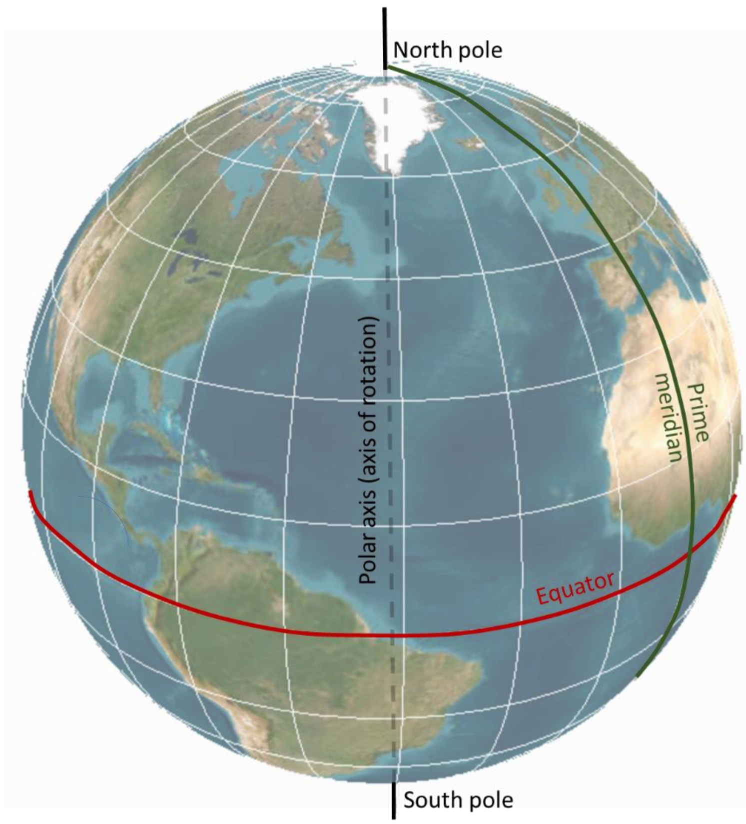


Figure 3.3.1 Basis for the spherical geometry of geographic coordinate systems. Key elements are the north and south poles, the polar axis (axis of rotation), the equator, and the prime meridian. The polar axis and equatorial plane are perpendicular and intersect at the center of the ellipsoid.

In this spherical geometry, the **geographic coordinates** of latitude and longitude are **angular coordinates**, meaning that they are angles rather than distances. Most commonly, these angles are specified in degrees, and that convention will be followed throughout this text, but note that for some purposes (e.g., computation), other units of measure such as radians are occasionally employed.

Latitude (denoted with the Greek letter **ϕ**, or phi) can be informally considered as an angle north or south of the equator, but more formal definitions are provided below. The latitude of the equator is 0°; the north and south poles are at 90° N and 90° S respectively. Lines of uniform latitude are called **parallels**; they circle the polar axis in planes parallel to the equatorial plane. The equator is the longest parallel of latitude; as latitude approaches 90° N or S, the length of each parallel decreases, approaching 0 at the poles. As their name suggests, parallels of latitude do not converge or intersect. The length of a degree of latitude is approximately 111 km.

As shown in Figure 3.3.2, there are two main definitions of latitude. **Geocentric latitude** is the angle between a point on the ellipsoid, the center of the Earth, and the equatorial plane. **Geodetic latitude** is the angle between a line perpendicular (or "normal") to the surface of the ellipsoid at a given point, and the equatorial plane. On a sphere, the two are identical, but on an ellipsoid, a line normal to the surface does not pass through the center of the ellipsoid, except for points at the poles or along the equator. Note that Figure 3.3.2 shows greatly exaggerated polar flattening, which in turn exaggerates the difference between geocentric and geodetic latitude.

For most purposes, coordinates are expressed using geodetic latitude, not geocentric. Among the various reasons for this are that the geodetic latitude of a point will be constant regardless of its elevation above (or depth below) the surface, while the geocentric latitude of a point will change with elevation or depth.

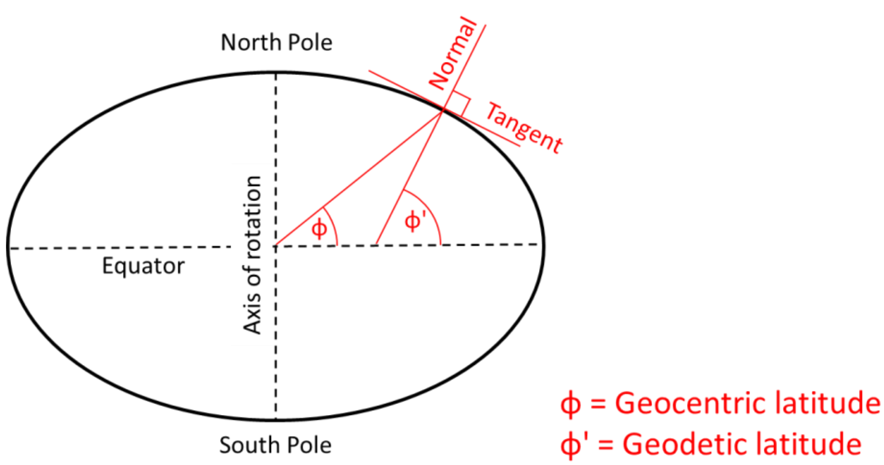


Figure 3.3.2 Geocentric vs geodetic latitude.

Longitude (the Greek letter **λ**, or lambda) is an angle east or west of the prime meridian. More formally, it is the angle between (a) the plane passing through a given point and the polar axis, and (b) the plane of the prime meridian. The longitude of the prime meridian is 0°; there are 360° of longitude, often divided into 180° E and W. Arcs of uniform longitude are called **meridians**; they run from one pole to the other in planes perpendicular to the equatorial plane. On the ellipsoid, all lines of longitude have equal length. They converge at the poles, so the distance between two meridians is greatest at the equator, and approaches 0 at the poles. The length of a degree of longitude is approximately 111 km at the equator, close to the length of a degree of latitude.

The set of parallels of latitude and meridians of longitude is referred to as the **graticule** (Figure 3.3.3). Traditionally, both latitude and longitude are commonly expressed in **degrees, minutes, and seconds (DMS)** format, with 60 minutes per degree and 60 seconds per minute, e.g., 37° 42′ 16″ E would be read as 37 degrees, 42 minutes, and 16 seconds east longitude. For use in GIS and geocomputation, they are typically expressed in **decimal degrees (DD)** format, with east and north coded as positive, and south and west as negative, e.g., 37.7044° (east) or -37.7044 (west). Other formats are less commonly used, such as degrees with decimal minutes, or using only positive degrees for longitude (0° to 360° instead of -180° to +180°).

On the curved surface of the ellipsoid, meridians and parallels intersect at right angles. When transferred to a flat map, some aspect of the graticule must be distorted – either the meridians do not converge, or the parallels are not parallel, or the two do not intersect at right angles.

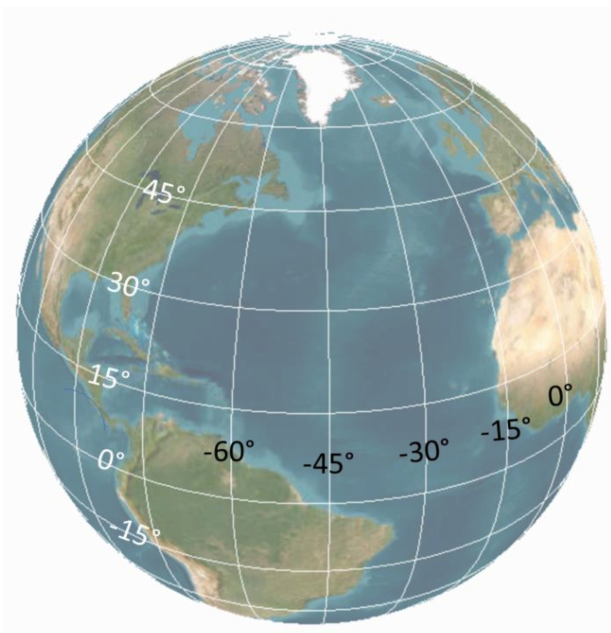
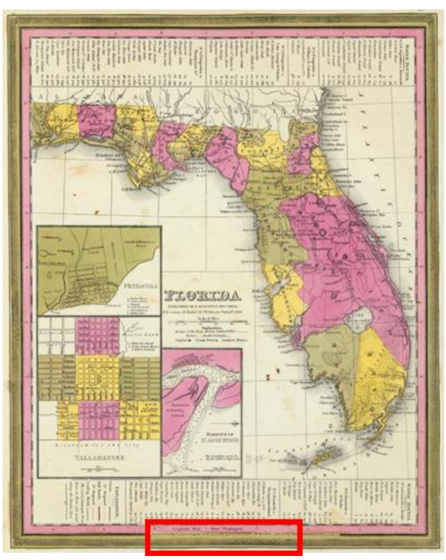


Figure 3.3.3 The graticule, consisting of parallels of latitude (labeled in white) and meridians of longitude (labeled in black). In this system, -60° longitude is 60° west longitude.

Because the length of a degree of longitude varies, it is almost always inappropriate to do calculations of distance or area in geographic (latitude/longitude) coordinates – the length of a 2.5° line will be different north-south vs east-west; the area of a 5° x 5° box will be different at 30° N vs at 60° N, and so on. Remember, latitude and longitude are *angles*, not *distances*.

Note that latitude is measured relative to the equator, an objectively-defined element of the Earth's ellipsoidal geometry. It can be found using traditional methods, such as measuring the elevation angle of the Sun at noon (taking into account seasonal variation), or of the star Polaris (in the northern hemisphere). In contrast, longitude is measured relative to an arbitrarily-chosen prime meridian, and is difficult to measure using pre-Space Age technology – the best solution (developed for navigation) required ships to carry a highly accurate clock, which gave the time of solar noon at a departure point of known longitude. The longitude at other points could then be determined by comparing the time of local solar noon to that on the clock – every hour difference in the timing of solar noon represented 15° of longitude. Today, latitude and longitude are easily and accurately measured using GPS receivers.

The dependence of longitude on an arbitrarily chosen prime meridian meant that, until the mid-19th century, maps were produced with differing longitude values based on different choices of prime meridian. Figure 3.3.4 shows an 1846 map of Florida, with longitudes based on the meridian of Washington DC.



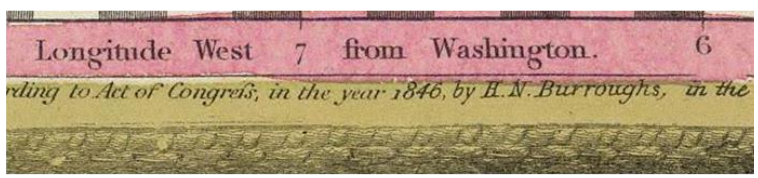


Figure 3.3.4 Map of Florida (1846) by S.A. Mitchell, from the Burroughs New Universal Atlas (Philadelphia, PA). Courtesy David Rumsey Map Collection, David Rumsey Map Center, Stanford Libraries.

As this example suggests, it was common for cartographers to use national capitals as the reference meridians for longitude values, such as Paris, London, and Washington DC. Other meridians were chosen for less subjective reasons; the Ferro (or Île de Fer) meridian was proposed in classical antiquity because it crossed the westernmost island in the Canaries archipelago, which represented the furthest edge of the Ptolemaic known world. Ferro was readopted by many cartographers in the 16th and subsequent centuries, particularly after its ratification by Louis XIII of France in 1634 (Withers 2017). Some French maps from this period show two different sets of longitude values along their upper and lower margins, one based on the Paris meridian and the other on Ferro. Ultimately, the International Meridian Conference of 1884, hosted at the US State Department in Washington DC, designated an international standard prime meridian passing through the grounds of the Greenwich Observatory in the UK, a historically significant facility for astronomical research. Conveniently, the antimeridian (on the opposite side of the planet from Greenwich) runs primarily through the western Pacific Ocean, avoiding bisecting major land masses.

Geodetic datums

Map coordinates, whether the angular coordinates of latitude and longitude, or the projected coordinates to be discussed later, depend on a model of shape of earth (an ellipsoid), which in turn is based on the geoid, as described in the previous section. More specifically, coordinates are based on a **horizontal datum**, which consists of an ellipsoid (or other mathematical model of the Earth, as discussed earlier), plus a reference point linking and orienting that model to a specific location on the Earth's surface.

These datums may be chosen to provide a good fit for maps of particular regions, or to provide an average fit for the globe as a whole. For example, the North American Datum of 1927 (NAD27), with its reference point at Meade's Ranch, Kansas, was chosen to provide an optimal fit of its ellipsoid to the continental United States. In contrast, the North American Datum of 1983 (NAD83) is an Earth-centered (geocentric) datum that provides a reasonably good fit between the ellipsoid and geoid over the Earth as a whole. Like NAD83, the widely-used World Geodetic System 1984 (WGS84) is an Earth-centered datum; it is the default datum used in most GPS surveying applications.

Because different datums can be built from different mathematical models of the Earth, and can be "attached" to the geoid at different places, the latitude and longitude coordinates for a given physical location on the surface will change when a different datum is used. In most cases, the shift in position associated with a change of datum will be on the order of meters to tens of meters, but in extreme cases, there can be hundreds of meters difference in position. Even meter-scale differences can be extraordinarily important in high-precision surveying applications, so it is important to record the horizontal datum used in a given map or data set.

Somewhat confusingly, the term “datum” is also used to refer to a **vertical datum**, or reference surface used to determine vertical coordinates (elevations or depths). The reference surface may be based on sea level (an indirect approximation of the gravity-defined geoid), or gravimetry (for a better approximation of the actual geoid), or an ellipsoid like those used in horizontal datums. Because the geoid-ellipsoid separation can be quite large in places due to irregularities in the geoid (Figure 3.2.2 earlier), it is common to record height measurements using a vertical datum that is based on the geoid rather than the ellipsoid. Thus, 3D GIS data and projects will typically contain coordinate information specifying different horizontal and vertical datums.

Projected coordinates

Since most maps are printed or displayed on flat surfaces, the map design process normally involves **projection** of spherical (latitude/longitude) coordinates into a **Cartesian** or **planar (XY)** coordinate system. Map projection can be visualized as an optical projection process, in which rays of light pass through a semitransparent globe to create a two-dimensional image of that three-dimensional globe on variously shaped flat surfaces (Figure 3.3.5) but fundamentally, the projection process is simply a mathematical transformation between two sets of coordinates, one spherical and one planar. The input to the transformation is a pair of latitude/longitude angular coordinates, and the output is a pair of Cartesian (planar) coordinates, typically labeled as X and Y, or Easting and Northing, with units of distance (meters, feet) rather than degrees.

Sticking with this optical projection analogy for the moment, there are three developable surfaces onto which the globe can be projected: a cylinder, a cone, and a plane. The plane can of course be used on a map directly – this is known as a planar or azimuthal projection – while the cylinder and cone can be cut along a line and “unwrapped” for use on a flat map (Figure 3.3.6, Figure 3.3.7). The developable surface can be oriented in various aspects, including normal (polar), transverse (equatorial), and oblique, as shown in Figure 3.3.6.

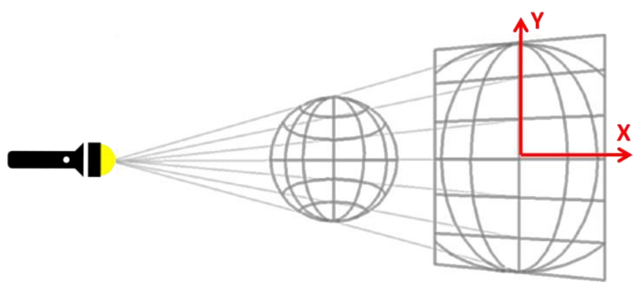


Figure 3.3.5 Analogy of the map projection process as optical projection, using rays of light from a source to produce a two-dimensional image of the three-dimensional globe.

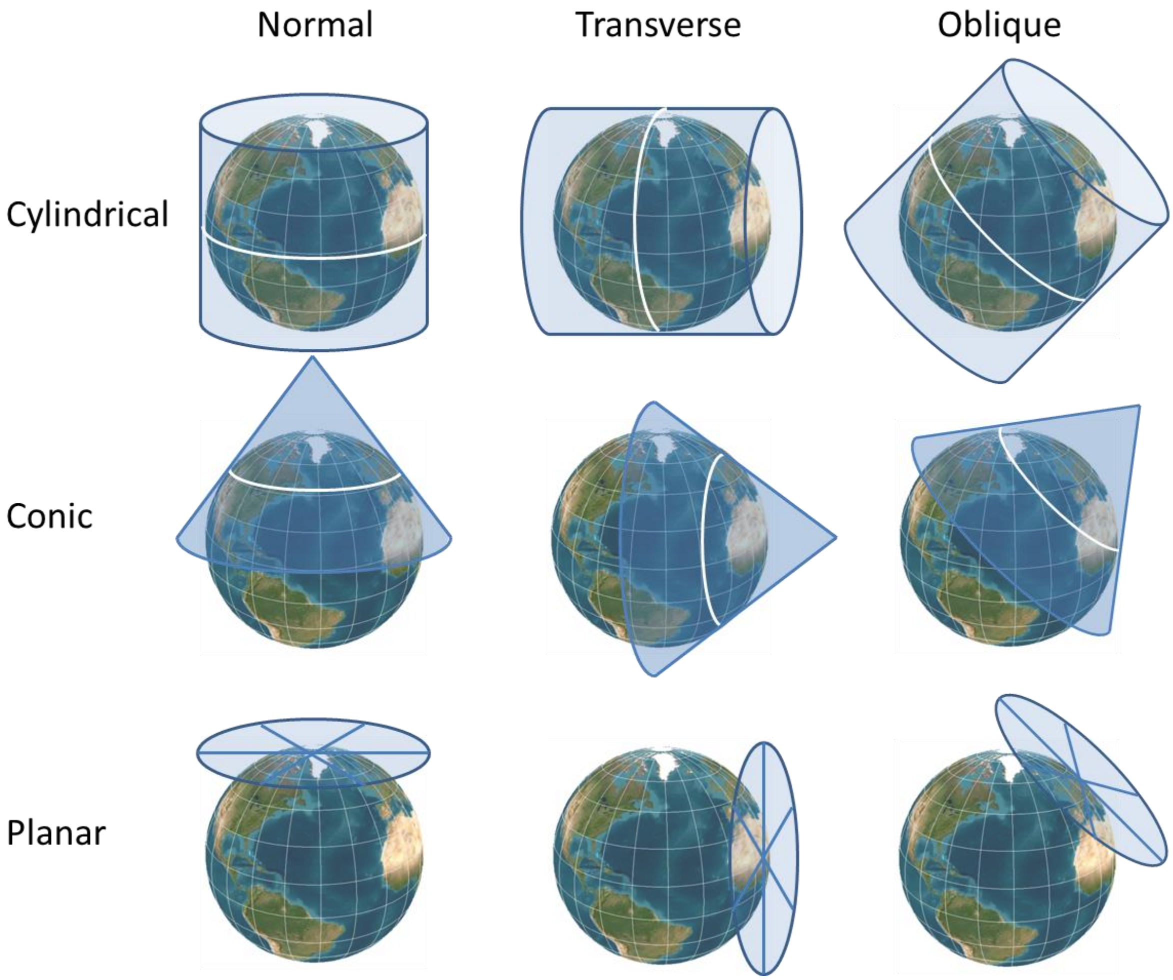


Figure 3.3.6 Developable surfaces (cylindrical, conic, and planar or azimuthal) and aspects (normal, transverse, oblique) for map projections.

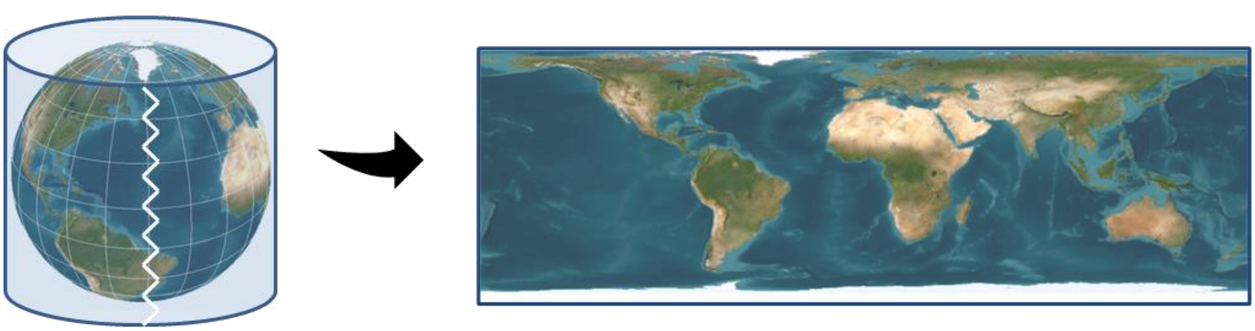


Figure 3.3.7 “Unwrapping” the cylindrical projection surface.

Projections such as those shown in Figure 3.3.6 – where the developable surface touches but does not cut through the surface of the generating globe – are known as **tangent projections**. The white lines shown for the conic and cylindrical projections in Figure 3.3.6 represent the places where the projection surface is in contact with (tangent to) the generating globe of the projection. These contacts – which are circles for the cylindrical and conic projections, and a single tangent point for the planar projection – are the only places where the scale of the projected map is identical to the scale of the generating globe, and thus where there is no distortion. In the normal (polar) aspect, the tangent line for a cylinder or cone will coincide with a parallel of latitude, and is thus called a **standard parallel**, while the tangent point for a planar projection will be the north or south pole.

As shown in Figure 3.3.8, it is also possible for the developable surface to cut through the generating globe. In this case, known as a **secant projection**, for a cylindrical or conic surface there will be two circles of contact, while for the secant planar projection, there will be one circle. As with the tangent case, these circles are the only areas with no distortion, and in the normal (polar) aspect are called standard parallels.

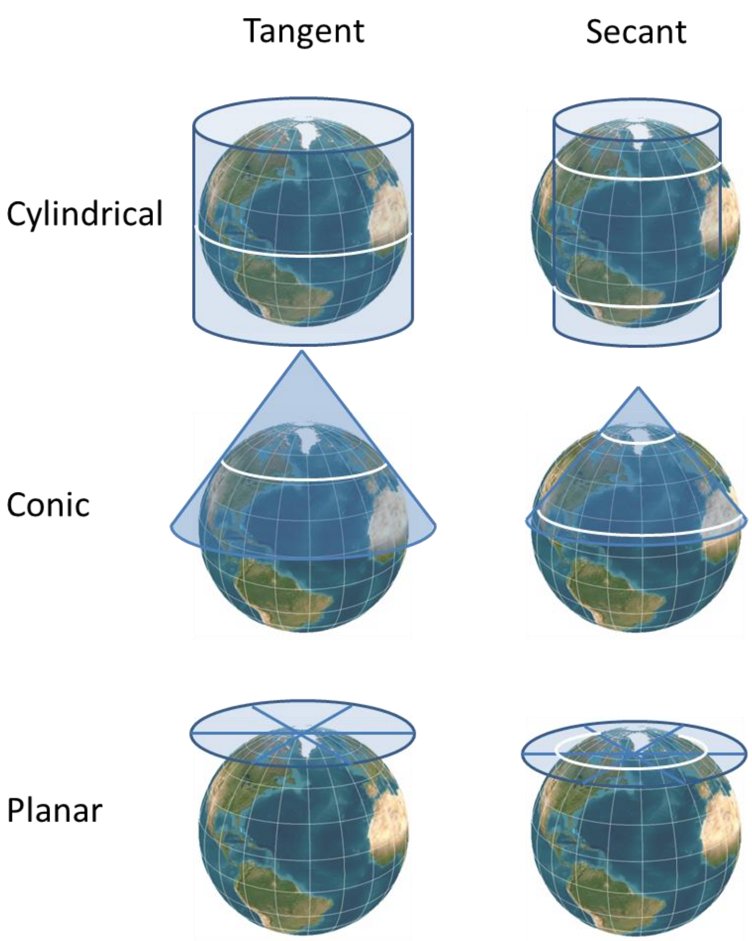


Figure 3.3.8 Tangent and secant projections, for three developable surfaces.

While these visualizations of the projection process can help explain the concepts involved, fundamentally projection is simply a mathematical transformation between two coordinate systems. For example, the equations for the cylindrical equal area projection (normal or polar aspect, tangent case), for a spherical globe, are as follows:

$$x = (\lambda - \lambda_0)\cos(\varphi_0)$$

$$y = \frac{\sin\varphi}{\cos\varphi_0}$$

where x = projected easting

y = projected northing

λ = longitude

λ_0 = longitude of central meridian

φ = latitude

φ_0 = latitude of standard parallel

with latitude and longitude expressed in radians (1 rad = 180°/π). Other projections, and those for ellipsoids rather than spheres, have more complicated equations for their transformation, but the principle is the same – projected x and y are functions of latitude and longitude, as well as various projection parameters.

The projection process inevitably introduces distortions – there is no non-distorting way to transform the curved surface of a sphere onto the flat surface of a plane. The properties of map projections that can be preserved or distorted include the following:

- Area: in an **equal-area (or equivalent) projection**, the size of a feature on a map is directly proportional to its size on the Earth (equal to its size on the generating globe).
- Shape: in a **conformal projection**, the shapes of features on the map are similar to those on the Earth. (Formally, angles on the map are identical to angles on the Earth; shapes approach true similarity as the diameter of the shape approaches 0).
- Direction: in an **azimuthal projection**, directions of all straight lines passing through a single, specified point match the corresponding directions on the Earth. No map has true direction everywhere.
- Distance: in an **equidistant projection**, scale is constant along all straight lines passing through a single, specified point (for an azimuthal equidistant projection), or along various other lines (e.g., all meridians and standard parallel(s), for a cylindrical or conic projection in normal or polar aspect).

Many map projections are compromises – they do not consistently preserve area, shape, direction, or distance, but are designed to minimize distortions in two or more of these properties across much or all of the globe.

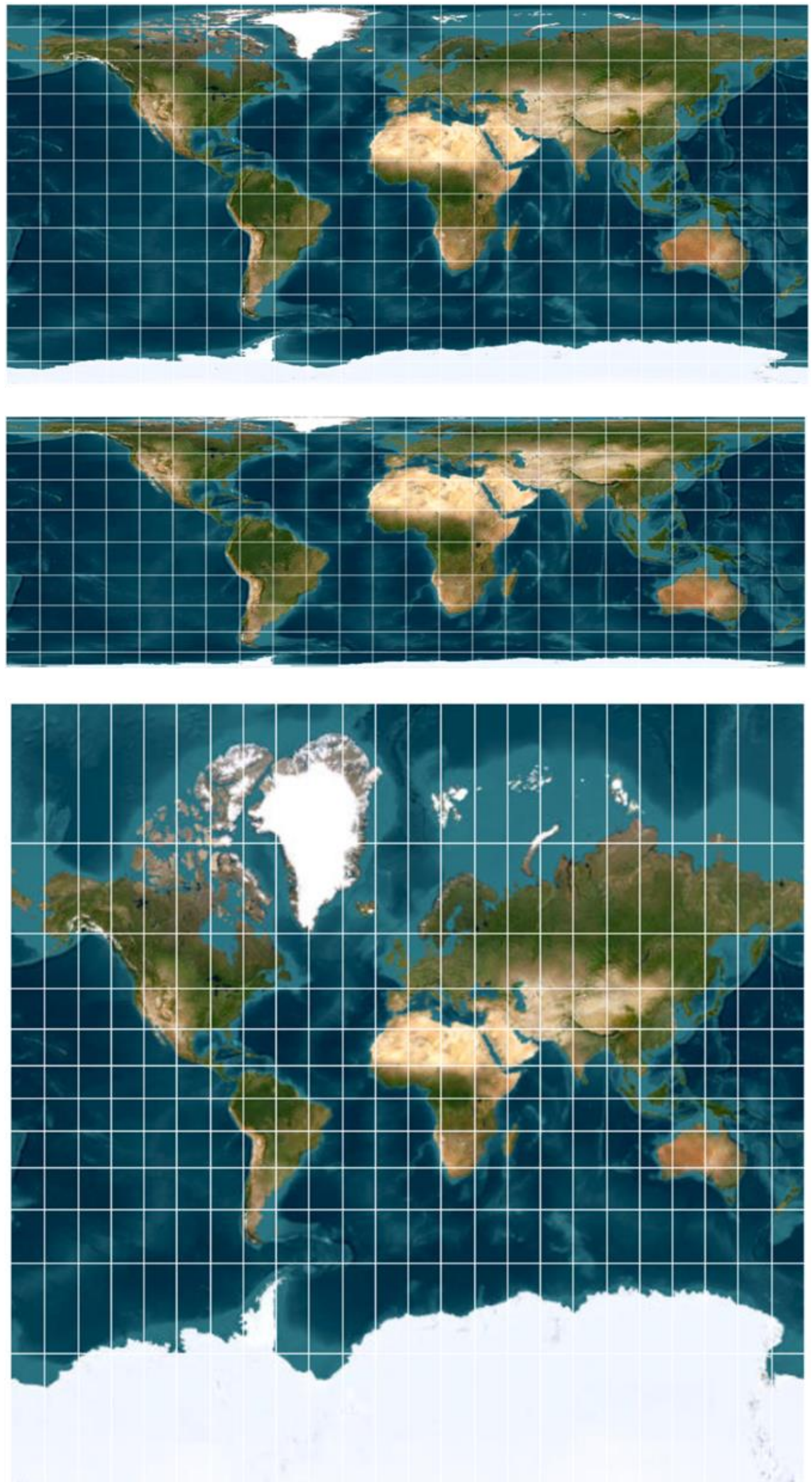


Figure 3.3.9 Examples of projections: (a) Plate Carrée, (b) Cylindrical Equal Area, (c) Mercator.

Figure 3.3.9 shows examples of three cylindrical projections. The **Plate Carrée** (also known as simple rectangular, equidistant cylindrical, or various other names) uses longitude and latitude as directly proportional to x and y, such that the graticule is an evenly-spaced grid. This projection distorts both area and shape, with distortion increasing away from the standard parallel (the equator, in this case), but it provides a compromise in that neither property is distorted as severely as in each of the other two projections in Figure 3.3.9. The **Cylindrical Equal Area** projection preserves area, at the cost of severe distortion in shape away from the standard parallel(s) – in the example shown in Figure 3.3.9, the equator is again the standard parallel, so distortion of shape becomes extreme near the poles. The opposite choice is adopted in the **Mercator** projection, a conformal projection that preserves angles and shapes, at the cost of severe distortion in area. Mercator maps were popular for navigational purposes in previous centuries, because a straight line on the map represents a line of constant compass bearing on the globe. This is not true of many other maps, and it is advantageous because while this line of constant bearing – known as a **rhumb line** – may not be the shortest line between two points, it can be followed using a single direction on the compass, without the need for repeated course changes.

Figure 3.1.10 shows a **Polar Azimuthal Equidistant** map projection, which is a central element in the logo of the United Nations. This planar projection is tangent at the north pole, the directions (azimuths) from the tangent point to all other points are correct, and the distances from the tangent point to all other points are proportional (i.e., scale is constant along all meridians).

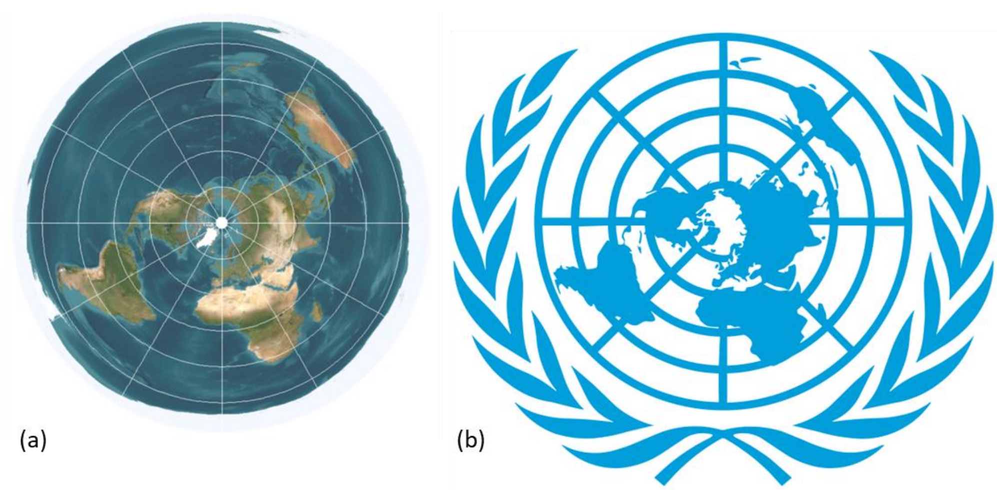
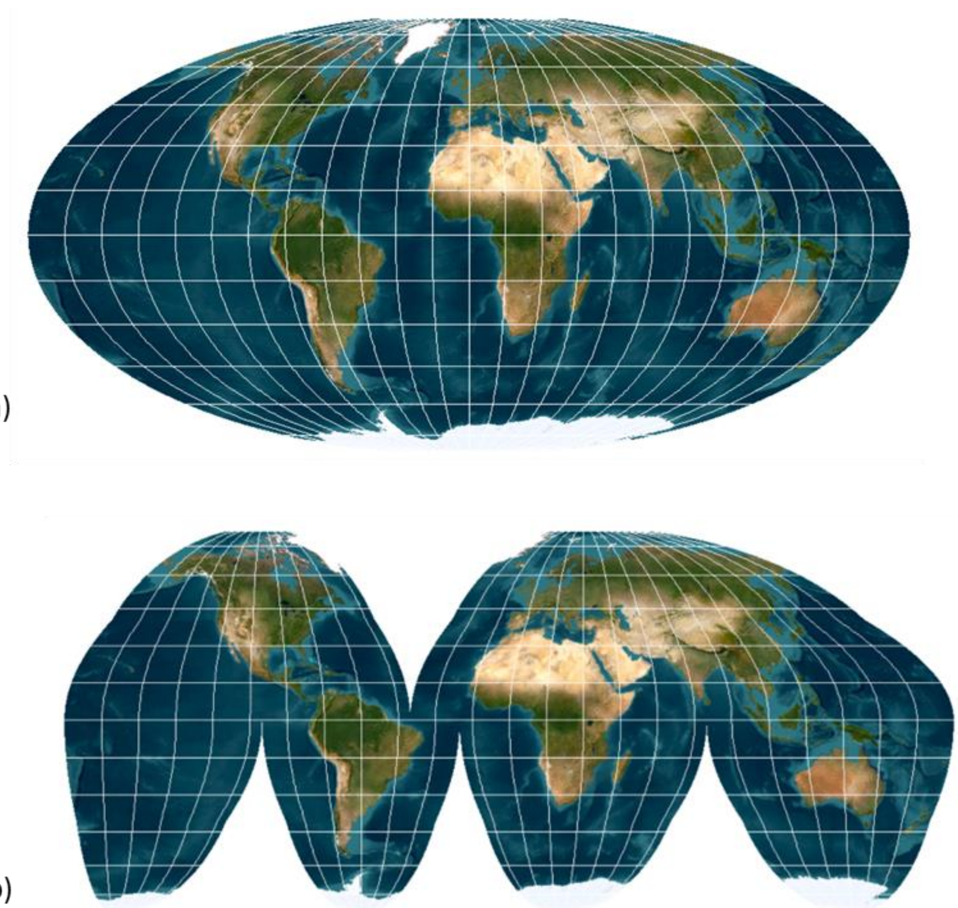


Figure 3.3.10 Examples of projections: (a) the Azimuthal Equidistant projection, (b) United Nations logo based on this projection.

Two more complex projections are shown in Figure 3.3.11. The **Mollweide** projection is an equal-area, pseudorectangular projection that distorts shape less than many other equal-area projections. Unlike in the three examples in Figure 3.3.9, north is not a consistent direction on the Mollweide projection. The **Goode's Homolosine** projection is a composite of the Mollweide and Sinuisoidal projections; it is equal-area, and preserves shape reasonably well, but has the geometry of two different projections, grafted together along latitude 40.7366° N. As shown in Figure 3.3.11(b) it is usually constructed as an **interrupted projection**, meaning that the globe is split along meridians (beyond the minimally required single split when "unwrapping" a rectangular projection from the globe). The interruptions on this example are the four approximately triangular notches, three in the southern hemisphere and one in the northern, which allow the continents to more closely approximate their true shapes. An alternative version – for oceanographic cartography – positions these interruptions on land, to optimize the representation of the oceans.



(a)

(b)

Figure 3.3.11 Examples of projections: (a) Mollweide, (b) Goode's Homolosine (interrupted).

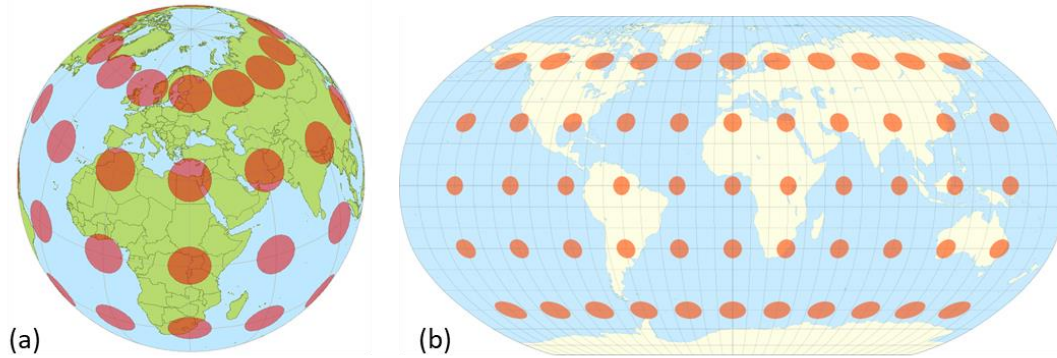


Figure 3.3.12 Tissot's Indicatrix, as displayed on a globe (a) and on the Robinson projection (b). Note the pattern of distortions of shape and area in (b). Courtesy S. Kühn and E. Gaba.

Because different map projections have different – and often complex – patterns of spatial distortion, various approaches have been developed to standardize the visual representation of distortion on a map. Among the most widely used is **Tissot's Indicatrix**, named for its 19th-century inventor, Nicolas Tissot. This consists of a grid of circles of uniform area, positioned on the globe and projected in the map projection. Distortions in the map projection become apparent through the distortion of the size, shape, and spacing of the circles (Figure 3.3.12).

While the examples in the preceding figures have focused on projections for world maps, designers of maps at the local, regional, and national scale also need to make appropriate choices of map projections. Many standard projections have been defined for specific geographic areas. Among the most widely used are the **Universal Transverse Mercator (UTM)** projections, defined for sixty north-south zones, numbered eastward around the globe in six-degree longitude bands (e.g., UTM zone 1 covers 180° to 174° W; zone 2 covers 174° to 168° W, and so on).

Many national and sub-national jurisdictions have one (or many) standard map projections. In the United States, each state has one or more State Plane coordinate systems, based on various types of projections, with different geodetic datums, and with map units in feet or meters. Cartographers can also define locally specific projections as desired, by adapting standard projection types (e.g., conformal conic) with user-selected standard parallels, central meridians, scale factors, and other parameters. For example, the Upper Valley region in the US states of Vermont and New Hampshire falls on the border between UTM zones 18 and 19, and between Vermont's and New Hampshire's State Plane zones. Thus, a local version of UTM has been defined, centered on 72 W longitude (Upper Valley Transverse Mercator or UVTM; sr-org projection 8853: https://spatialreference.org/ref/sr-org/uvtm/).

Many GIS novices will default to producing maps in which latitude/longitude angular coordinates are treated as Cartesian (planar) coordinates, along the lines of the previously mentioned Plate Carrée projection. While this can be a defensible choice for world maps, as a compromise between other projections that focus exclusively on equivalence or conformality (Figure 3.3.9), it is rarely a good choice for local to regional scale maps. Decisions about projections should be informed by an understanding of the projection process and of different projections' advantages and disadvantages, and should lead to the use of projections that minimize distortions of area and shape within the extent of the mapped region.

From a GIS logistics viewpoint, the term **Coordinate Reference System (CRS)** is often used to refer to the complete set of information describing the relationship between a geographic dataset and the corresponding location(s) on the Earth. This may include the horizontal and vertical geodetic datum; the type of coordinate system (geographic or projected); any projection parameters such as standard parallels, scale factors, etc.; and the units of measurement (degrees, radians, meters, feet). Software can usually deal with varying coordinate systems seamlessly, as long as each dataset's CRS is properly defined. Spatial calculations may be done in either a planar or geodetic mode; the latter takes into account the ellipsoidal geometry, and thus supports more accurate calculations of distances and areas. Finally, calculations involving both horizontal and vertical coordinates require either common units or careful unit conversion; e.g., avoid mixing units of degrees and meters, or meters and feet, when calculating topographic slope.