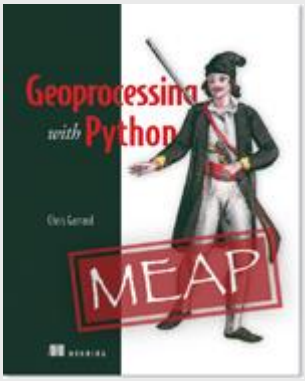

Introduction to spatial reference systems

By Chris Garrard

In order to effectively work with coordinate systems you need to understand why there are so many of them in the first place and how to select an appropriate one for your purposes. In this article, based on my book Geoprocessing with Python, I give you some background information about spatial reference systems..

A spatial reference system is made up of two components, a datum and a projection, both of which affect where on the earth a set of coordinates refers. Briefly, datums are used to represent the curvature of the earth, and projections transform coordinates from a three-dimensional globe to a two-dimensional map. Different projections are appropriate for different purposes, such as web mapping, accurately measuring distances, or calculating areas.

There is a lot more to it than that, however, and it is important to understand the role that both datums and projections play. In order to do that, let's back up and review how coordinates are represented on a globe. Latitude and longitude are the distance, in degrees, from the equator and the prime meridian, respectively. Latitude values range from -90 to 90, with positive values north of the equator. Longitudes range from -180 to 180, with positive values east of the Greenwich prime meridian (figure 1). Using degrees makes perfect sense on a spherical surface, and although the earth is not a perfect sphere, it's close enough for this to be a very convenient way to specify a precise location on the planet.

There is a complication, however, and it arises from the fact that the earth is not a perfect sphere, or even a perfect ellipsoid. As you probably learned in geometry class, but then promptly forgot, there are simple equations to model the shape of ellipsoids, including spheres. But these equations assume a perfect geometry with a nice smooth surface and no protrusions and dips.

It would be quite something if a planet were to form that perfectly, and ours certainly did not. Have you ever seen a worn out ball, like a volleyball, that has developed a weak spot and has a bulge that wasn't there when the ball was new? Not only does the earth have mountains and valleys, but it is a little lopsided like that volleyball, which definitely makes describing its surface with a simple set of equations more complicated.

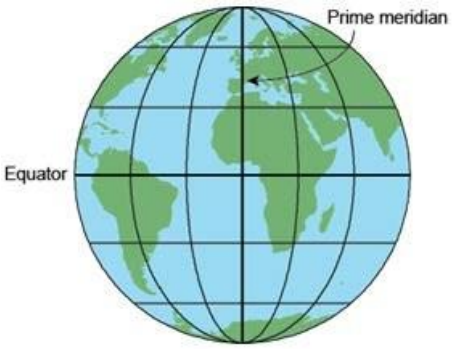

Figure 1 Latitude and longitude lines at 30° intervals. Positive latitude values are north of the equator, and positive longitudes are east of the prime meridian.

DEFINITION The prime meridian is the line of longitude that passes through the Royal Observatory, Greenwich, in London. This has been recognized by much of the world as the reference meridian since 1884.

Methods of specifying latitude and longitude

There are multiple methods for specifying latitude and longitude coordinates. For example, these are all equivalent:

Decimal degrees (DD): 37.8197° N, 122.4786° W
Degrees decimal minutes (DM): 37° 49.182' N, 122° 28.716' W
Degrees minutes seconds (DMS): 37° 49' 11" N, 122° 28' 43" W

These different notations are based on the fact that angles are divided up into minutes, where one degree in an angle is made up of 60 minutes, and each minute is made up of 60 seconds. Because latitude and longitude are degree measurements, they are also divided up into minutes and seconds. To get decimal minutes from decimal degrees, multiply the fractional part of the DD value by 60, so for example, 60 * 0.8197 = 49.182. Therefore, 37.8197 degrees equals 37 degrees and 49.182 minutes. Similarly, you can multiply the fractional part of the minutes value by 60 in order to get seconds. Since 60 * 0.182 = 10.92, now you have 37 degrees, 49 minutes, and about 11 seconds.

Additionally, south and west values are represented as negative numbers if the directions are not specified. For example, -122.4786° is the same as 122.4786° W.

In order to use latitude and longitude values in your Python code, you will need to make sure that they use the decimal degrees format and specify directions using positive and negative values instead of N, S, E, or W.

Because of these anomalies in the planet's surface and also because measurement accuracies vary, there are multiple models of the earth's ellipsoid. These models are called datums, and every spatial reference system is based on one of them. There is one widely used global datum, called the World Geodetic System, which was last revised in 1984. This datum, called WGS84 for short, is the one used for data with a global coverage, including the Global Positioning System (GPS).

Most datums are designed to model the curvature of the earth in a more localized area, such as a continent or even a smaller area. A datum designed for one area will not work well elsewhere. For example, the North American Datum of 1983 (NAD83) should not be used in Europe.

Depending on which datum is being used, the same set of latitude and longitude coordinates can refer to slightly different locations, because the underlying ellipsoids are different shapes. Sometimes the difference between coordinates using two different datums is negligible, but other times it can be hundreds of meters. Because of this, you always need to know which datum your geographic data is based on.

So far we've only talked about three-dimensional ellipsoids, but what you really want in most cases is a two-dimensional map because they tend to be more convenient for most purposes. After all, it's hard to fold up a globe and put it in your pocket, or embed one inside of a book!

So how do mapmakers go from three to two dimensions? One way to solve the problem is with what is called an interrupted map, like that shown in figure 2. You have probably seen something like this before, and perhaps you've even cut one out and bent the paper to make a globe. That's kind of cool, but in its two-dimensional form the map would be much easier to use if land masses weren't split up into chunks and separated by wasted space. This is where projections come in.

As their name implies, projections are used to project, or transform, locational data into a different coordinate system. These map projections use Cartesian coordinate systems, so there are two perpendicular axes and locations are specified with *x,y* coordinate pairs, just like scatterplots or line graphs. The tricky part is converting coordinates on a sphere to a two dimensional plane.

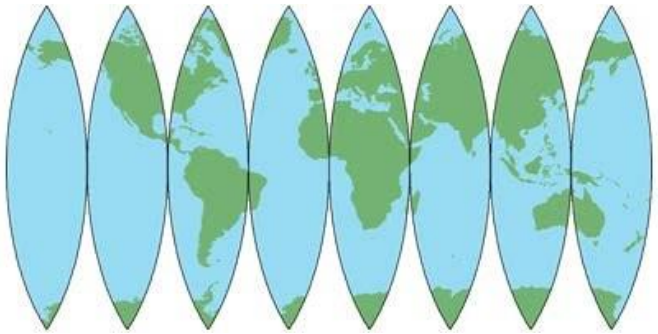

Figure 2 An interrupted map.

In fact, there are a lot of ways to accomplish this, and they all have their own strengths and weaknesses. Think about stretching the different parts of the interrupted map shown in figure 2 so that the map is a single rectangle with no cutouts. Geographic features would obviously get warped, especially near the poles where you had to stretch farther. No matter how you project geographic data to two dimensions, you will get distortion, but the type of distortion depends on how you do the conversion.

Depending on what you plan to use the data for, some types of distortion may be acceptable while others would not. Figure 3 shows a couple of ways a piece of paper could be

wrapped around a globe and used to convert the geographic data to 2D, but there are others. Even with those shown here, the angle of the paper could be changed to get a different effect.

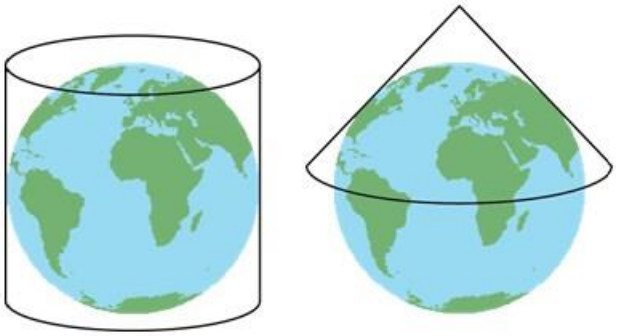

Figure 3 Two different ways that a piece of paper could be wrapped around a globe and used to project geographic data onto a two-dimensional surface. The example on the left is cylindrical, and the one on the right is conical.

Some projections, called conformal, preserve local shapes. For example, the shape of Lake Titicaca on the border of Bolivia and Peru would not change between the globe and the 2D map. No mathematical trickery can preserve the shape of a large area, such as all of Eurasia, however. Mercator projections, including the Transverse Mercator (UTM), are examples of this type of projection. Others, called equal-area projections, keep the amount of area the same, so the measured area of Greenland would not change, although the shape might. The Lambert equal-area and Gall-Peters projections are two examples. Equidistant projections, such as the Azimuthal equidistant, keep distances and scales the same, but only for a certain part of the map such as the equator. The farther you get from this true line, the greater the distortion. Figure 4 shows some examples of different projections.

TIP There are a few different terms for data that use latitude and longitude coordinate values. You might see them referred to as having a geographic projection, or see them simply called unprojected or geographic.

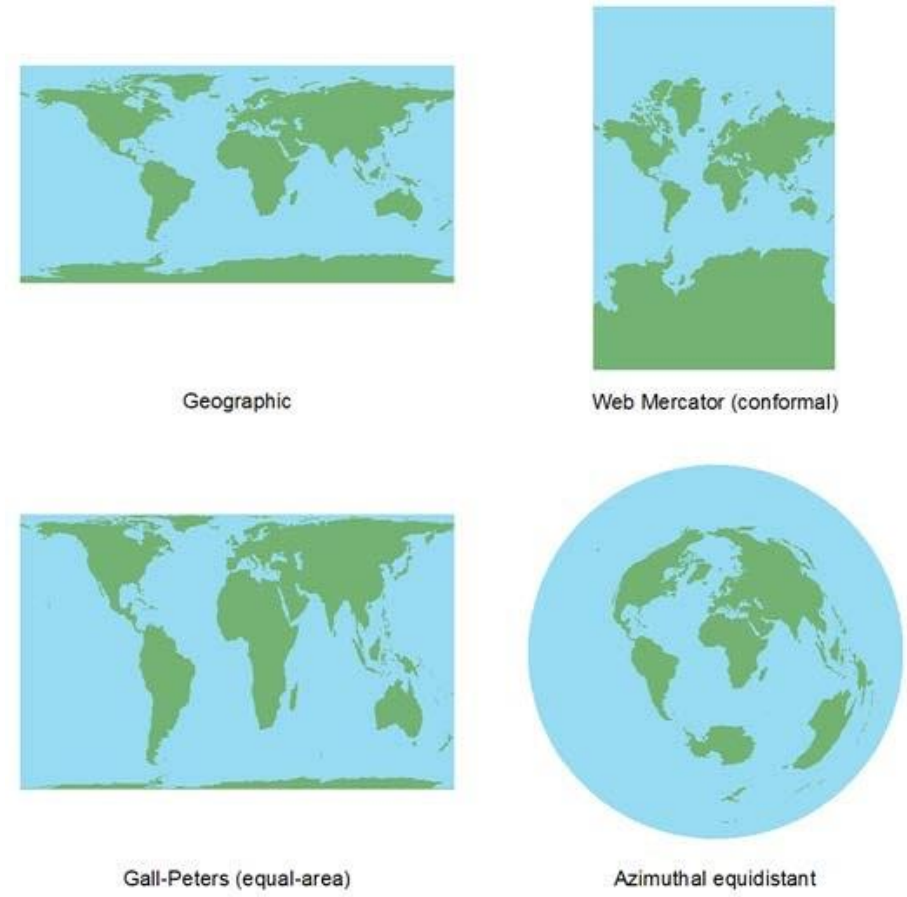

Figure 4 Examples of different types of projections.

So why should you care about all of these differences? Depending on your purposes, maybe you don't. I doubt I would be very worried about it if I was making a map of the small town I live in. But if I was making a map of the state I live in, I might care if it looked short and fat or a little taller and skinnier, as shown in figure 5.

What if you cared more about measurements and less about appearances? Let's consider a dramatic example and think about what would happen if you needed to compare the amount of forested area in Columbia and Chile. Sticking with latitude and longitude wouldn't work,

because the lines of longitude converge at the poles, so one degree of longitude does not represent a constant distance. In fact, one degree of longitude is equal to approximately 111 kilometers at the equator, but only about 79 km at a latitude of 45 degrees.

Although latitude distance can vary slightly because the earth is not a perfect sphere, it is generally around 111 kilometers per degree. This means that a square 100 km on a side would measure about 0.8 square degrees in Columbia, but closer to 0.5 degrees^2 at the southern tip of Chile. Using latitude and longitude to compare the amount of forested area in the two countries would obviously give inaccurate results. Instead, you would want to choose an appropriate equal-area projection for this purpose.

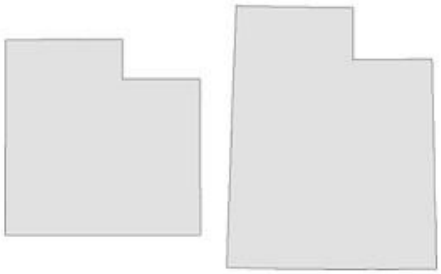

Figure 5 The state of Utah shown using geographic (lat/lon) coordinates on the left, and UTM Zone 12N on the right. Both examples use the NAD83 datum.

Projections are not tied to specific datums, so knowing the projection of your data is not enough. You also need to know the datum, and it is the combination of the two that defines the spatial reference system. For example, most of the data I get for Utah uses a UTM projection and the NAD83 datum, but I cannot safely assume that all UTM data I receive uses NAD83. It could easily be NAD27 or WGS84 instead, so I don't have a complete spatial reference system unless I know both the projection and the datum.

If you do not know both components, you might map your data in the wrong location. I have known people who unknowingly set their GPS to display coordinates in an unusual spatial reference system and then collected data by writing down the coordinates shown on their screen. Unfortunately, their data was then unusable because they didn't know what spatial reference system the GPS had been set to display at the time.

On the other hand, I have also known people who lacked spatial reference information for their data, but fortunately the data was in a common system and we were able to figure it out. If you are collecting data, please simplify your life by paying attention to this crucial information at the beginning of the process, no matter how boring it might seem.