DATUMS: HOW TO WORK WITH THEM*

Robert Burtch
Surveying Engineering Department
Ferris State University
Big Rapids, MI 49307

Abstract

One of the least understood concepts of mapping is the role of datums in the mapping of spatial detail. This presentation will discuss the difference between a datum and a coordinate system. It will also identify the different types of datum that the GIS professional may encounter. Finally, the presentation will discuss the issues of transforming coordinates based on one datum to another datum, such as the conversion of NAD 27 to NAD 83. Pitfalls and accuracy will be presented.

Introduction

One of the most difficult technical issues facing the Geographic Information System (GIS) community is the concept of a datum and coordinate system. Frequently, these terms are used interchangeably, when in reality they are very different. Datums and coordinate systems are important because they form the basis from which features on the earth's surface will be depicted. They work together to provide the locational tools needed to depict the features and phenomenon of interest in GIS analysis.

Before beginning to look at the principles involved in working with a datum and coordinate system, it is important to review how the user requires data be located on a map. Position can be described as being either absolute or relative. An absolute position is one where the feature is located in its correct location on the surface of the earth. This location would be determined to some degree of accuracy with respect to all other points on the earth, even those that are outside of the area of interest. Relative location involves positioning features with respect to other features. For example, one might want to know very accurately the location of a water main with respect to the street that it may run parallel with. Here, the analyst is not interested in where the water main is located in the global scheme, just where does the repair crew need to dig to reach a particular water main.

A datum can be simply defined as a reference system. In the example above where the GIS analyst is concerned with location of the water main, the street becomes the datum from which the main can be located. Another simple example is construction of a house. Lets assume that the finished floor elevation needs to be located 2' above the road running in front of the house. The street then becomes the datum from which the height of the finished floor is established. Within the house, windows are to be set at a certain height above the floor. Here the floor becomes the datum. For mapping purposes, many

* Paper presented at the 2002 IMAGIN Conference, Traverse City, Mi

different datums exist and it is important when providing location that the basis of the datum be defined, just like the definition of the reference system is needed in the house construction example. It should be clear that a feature could have different values depending upon the datum being used. The finished floor elevation has one value when measured with respect to the road and another value if referred to "sea level", yet the location is identical.

A coordinate system is the realization of the datum. In other words, it is the means by which the feature is located on the accepted datum. Again, using the house construction example, the datum could be defined as the parcel and where the house is placed on the parcel is accomplished by using a coordinate system. Again, there are many different types of coordinate systems available. Additionally, there are different ways to express the location of a feature even if the datum does not change. As an example, assume that we are working on a plane. Then, point A can be defined using either polar or rectangular coordinates as shown in figure 1. They are related using simple trigonometric relationships.

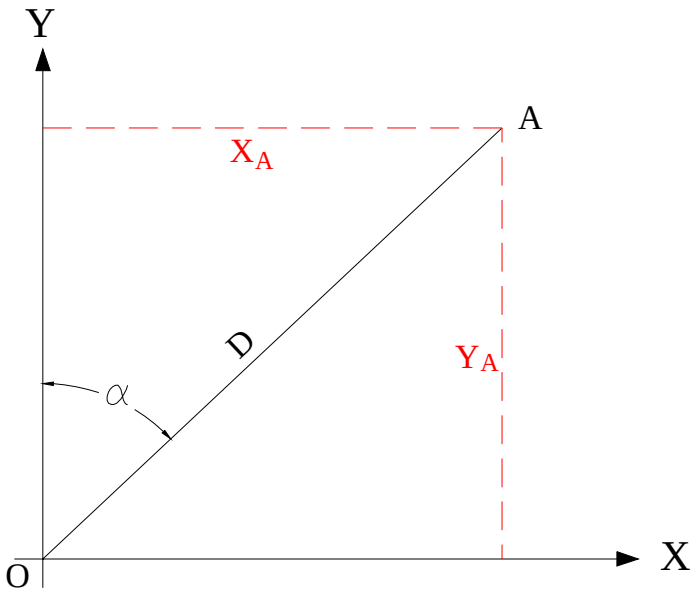

Figure 1. Rectangular and polar coordinates

$$X_A = D \sin \alpha$$

$$Y_A = D \cos \alpha$$

where X_A, Y_A are the coordinates of point A with respect to the origin, O, D is the distance from the origin to point A, and α is the direction of line OA.

There are different methods of expressing coordinates [Iliffee, 2000]:

- Projection coordinates – the representation of the curved earth's surface onto a plane
- Orthometric heights – the elevation of points above the geoid
- Geodetic coordinates – expressed in terms of latitude and longitude
- Spheroidal (or ellipsoidal) heights – the elevation of a feature above the spheroid
- Cartesian coordinates – a three-dimensional coordinate system referred to the center of the spheroid.

Coordinates

Three basic coordinate systems are utilized in mapping: spherical coordinates, spheroidal coordinates, or Cartesian coordinates. The first item that must be recognized is that the earth is not flat. This assumption may be acceptable for small areas, but once the map extends over any considerable distance, the curvature of the earth must be accounted for. The first approximation of the earth is as a sphere from which one can define spherical coordinates (figure 2). A feature is defined by a latitude (φ), longitude (λ), and height (h). The latitude is fixed physically with the equator. Longitude is another problem. Historically, different nations developed their own principal meridian from which longitude was referenced. Now, by convention, the meridian passing through Greenwich, England is accepted as the principal meridian. Latitudes are positive to the north and negative to the south and range from -90^o to $+ 90^o$. Longitude is positive to the east and negative to the west with a range from -180^o to $+ 180^o$.

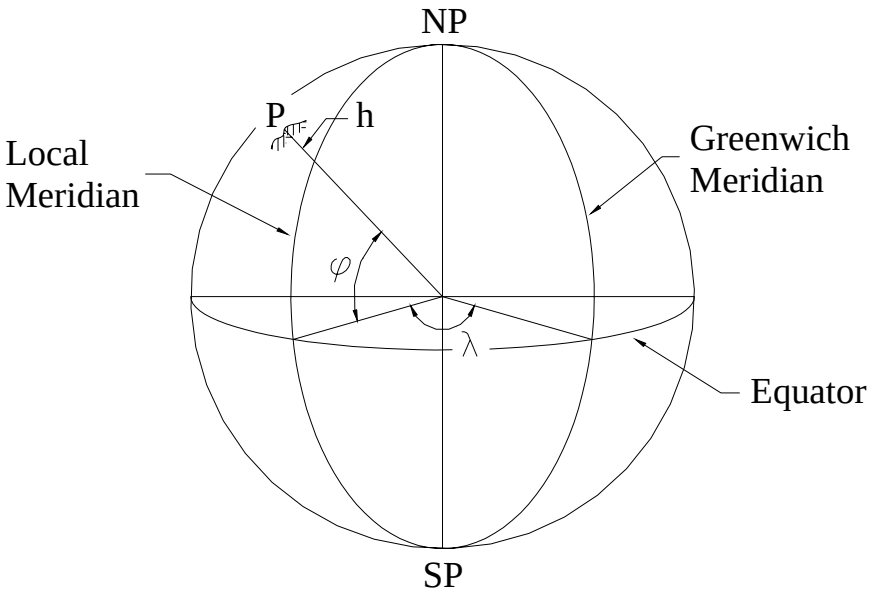

Figure 2. Spherical coordinates

The next approximation of the surface of the earth is the spheroid, also called the ellipsoid (figure 3). The spheroid is defined by rotating an ellipse around the semi-minor axis (line from the North to South poles and referred to as the rotational axis). The spheroid is defined by the semi-major axis (a), semi-minor axis (b), flattening (f), and eccentricity (e)[1]. Coordinates are defined by latitude, longitude, and height and are called geodetic coordinates. The reference direction is the normal to the spheroid. This defines a line that is perpendicular to the ellipsoid at a point. The significance of this is that, unlike the spherical coordinates, the lines do not intersect at the center of the ellipse. In fact, they do not intersect at a common point. The flattening is very small and appears negligible when looking at the earth from space. The known values for the ellipsoid parameters for the NAD 83 datum are given as:

- a = 6,378,137 m (exact)
- b = 6,356,752.3141 m
- 1/f = 298.257222101
- e^2 = 0.00669438002290

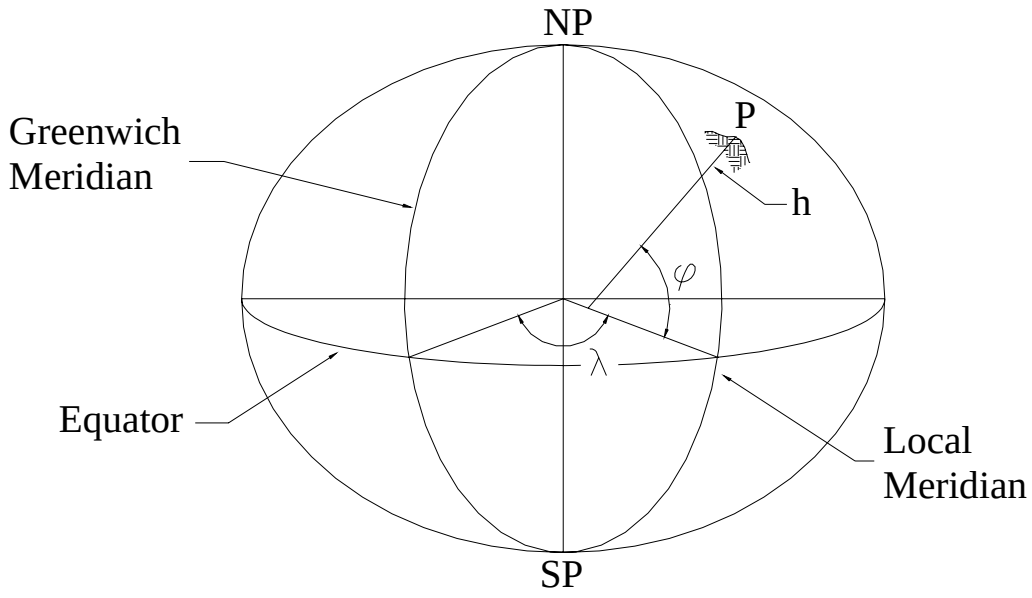

Figure 3. Spheroidal coordinates

The relationships between the flattening and eccentricity with respect to the semi-major and semi-minor axes are given as follows:

[1] This eccentricity term is often called the first eccentricity and is generally given in the square term.

$$f = \frac{a - b}{a}$$

$$e^2 = \frac{a^2 - b^2}{a^2} = 2f - f$$

The Cartesian coordinate system is a three-dimensional system with the origin at the center of the spheroid. The coordinates are defined as:

- Z-axis – axis aligned with the semi-major axis of the spheroid (polar axis)
- X-axis – axis in the plane of the equator passing through the Greenwich meridian
- Y-axis – axis in the plane of the equator 90^o counterclockwise from the +X-axis thus forming a right-handed coordinate system.

The Cartesian coordinates can be computes from the geodetic coordinates using the following relationships:

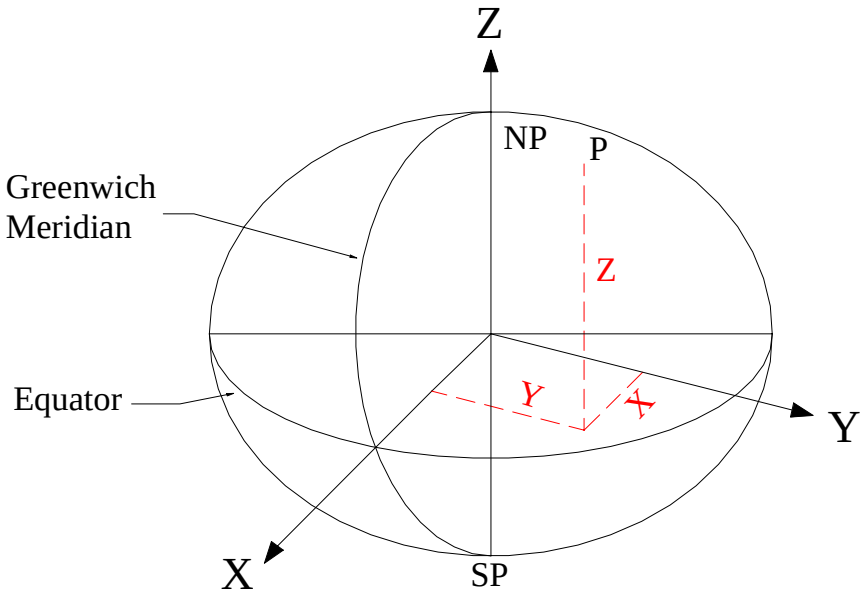

Figure 4. Cartesian coordinates from the center of the spheroid.

$$X = (v + h)\cos\varphi\cos\lambda$$

$$Y = (v + h)\cos\varphi\sin\lambda$$

$$Z = [(1 - e^2)v + h]\sin\varphi$$

where

$$v = \frac{a}{\left(1 - e^2 \sin^2 \varphi\right)^{1/2}}$$

A better shape of the earth is the geoid. The geoid is what is called an equipotential surface used as the reference to all other equipotential surfaces on the earth. What this means is that the distance between any two equipotential surfaces is the work required to move between these surfaces. An equipotential surface is a continuous surface that is everywhere perpendicular to the direction of gravity. As a good approximation, the geoid is often referred to as mean sea level. While this is not theoretically correct, the deviation is only about 1 meter worldwide.

The separation between the spheroid and the geoid is referred to as the geoid undulation (also called the geoid-spheroid separation). From figure 5, it is clearly evident that the geoid undulation is not constant over the earth. The line that is perpendicular to the geoid is called the plumb line. The angle between the plumb line and the normal to the ellipsoid is called the deflection of the vertical. The significance of this concept is that when an individual uses a level, the line that is being defined is perpendicular to the vertical line at that point whereas the coordinate system is generally related to the geoid.

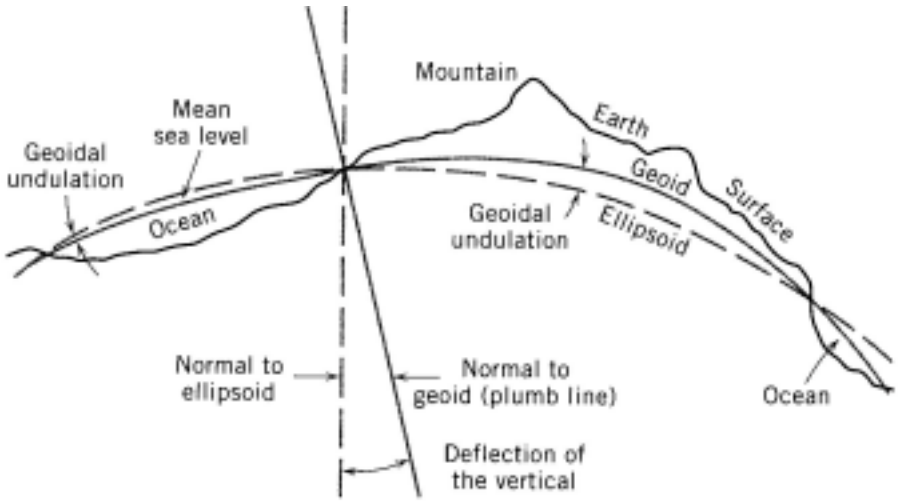

Figure 5. Relationship between the geoid and spheroid [from Smith, 1997, p.122].

North American Datum

Prior to the developments of satellite geodesy, many countries defined unique datums that could be used within their borders. These reference systems were refined as the knowledge of the size and shape of the earth became more accurately defined. The culmination of these refinement in the United States led to the definition of the North American Datum of 1927 (NAD 27). A point, called Meade's Ranch that was located in Kansas, the approximate geographical center of the U.S, defined this datum. The Clarke

1866 spheroid was used to define the NAD 27 because it was identified as the best overall fit for the United States.

Geodetic surveys were performed using triangulation (see figure 6). It was extremely difficult to measure distances using taping techniques so angle measurements were used. A triangulation network is one where angles within a quadrilateral are measured. With all of the angles within a triangle known along with the length of one side, all other elements of the triangle (the other two sides) can be determined. Thus, during much of the twentieth century, great triangulation arcs were run over the entire country.

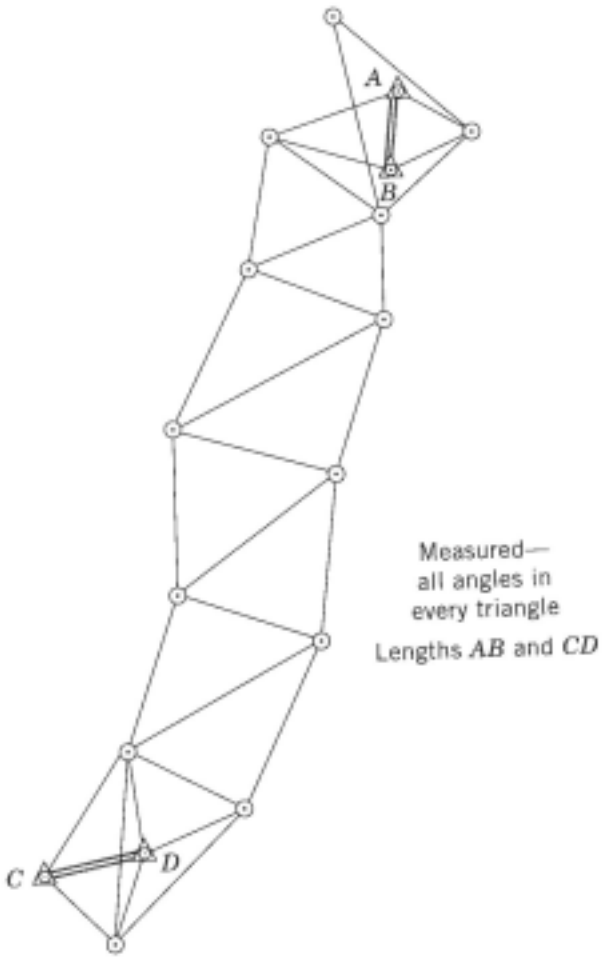

Figure 6. Example triangulation figure [from Smith, 1997, p.62].

But problems arose. One problem began almost immediately upon the adjustment of the NAD 27. A point in northern Michigan, on the U.S.-Canada border, was not held fixed during the adjustment as it was suppose to. This introduced a distortion within NAD 27. Even though a regional adjustment was performed of points within Michigan, Wisconsin, and Minnesota after the NAD 27 adjustment, small errors remained. In addition, during the 1960s and 1970s, electronic distance measurement (EDM) equipment found widespread use among surveyors. For the first time, surveyors began to utilize geodetic measurements to provide control for larger civil works projects. When they would start their survey from monuments in one triangulation arc and close the survey on a different arc, surveyors were finding large discrepancies in the results. Moreover, during this same period, geodesists began to use space-based techniques to measure long distances across the continent. Again, discrepancies occurred. Because of these problems, the National Geodetic Survey embarked on a new adjustment of the North American Datum.

One of the first considerations required in the readjustment was the selection of the datum. Because space-based measurement systems were being developed in both the military and civilian sectors, it was decided that the new datum should be a geocentric datum. This would provide for universal compatibility of survey control. Moreover, as the size and shape of the earth, through satellite observations, was known even better than

in previous years, the NGS decided to use the Geodetic Reference System 1980 (GRS 80) that was adopted by the International Association of Geodesy (IAG). The new datum was called the North American Datum 1983 (NAD 83). Thus, we see that the GRS 80 and NAD 83 datums are identical.

During this period, space-borne geodetic measurements were primarily being performed by the military and government. The Defense Mapping Agency (DMA), now called the National Imagery and Mapping Agency (NIMA), took the new GRS 80 ellipsoid parameters to define the World Geodetic System of 1984 (WGS 84). In doing this, the DMA used a truncated second zonal harmonic of the equipotential ellipsoid to derive a normalized form used in defining WGS 84. The effect is that the flattening of the spheroid between WGS 84 and NAD 83 is different at the eighth significant figure and the semi-minor axes differ at the 10^{th} significant figure. Thus, while WGS 84 and NAD 83 are not identical, for all practical purposes, they are the same. This means that results from GPS, which measures in the WGS 84 system, are in the NAD 83 datum.

The NAD 83 adjustment was based on geodetic measurements taken up until the adjustment began. Very Long Baseline Interferometry (VLBI), Doppler, and Satellite Laser Ranging (SLR) were used to help define the geodetic datum. Additionally, historical measurements based on triangulation, trilateration, traversing, and astronomic observations were also used within the adjustment. At about this time, the global positioning system (GPS) began to see more use in mapping applications. Despite the military's attempt to degrade the GPS survey results, those using this technology were finding survey accuracies that were not possible before.

The North American Datum 1983 has undergone changes over time, as had WGS 84. These changes are referred to as realizations – or how the reference frame was created[2]. The earth is a dynamic planet that is constantly changing. Moreover, each new realization of the datums employed different measurement tools. While the meter may be well defined theoretically, applying this definition to the measurement tool yields uncertainties in the length measurement.

As an analogy, assume that a manufacturer is charged with creating "meter sticks" made from different materials, which require different manufacturing processes. While all of these sticks are the same, namely one meter, they are not going to be exactly the same. Thus, a person using a wooden meter stick may find a different distance between two points from someone using a glass meter stick, etc. If the wooden meter sticks are used to form a network in an area and then the new glass meter sticks are used for measuring other points within the same network, the differences after all of the errors have been accounted for, will reflect a difference in scale between the two measurement systems. Thus, the meter used for EDM, GPS, VLBI, and other geodetic measurements are different by some scale factor.

[2] See the series of articles by R. Snay and T. Soler entitled "Modern Terrestrial Reference System" that appeared in four parts in the Professional Surveyor magazine beginning in December 1999.

Thus far, the National Geodetic Survey (NGS) has refined the NAD 83 reference frame a number of times. NAD 83 (1986) represents the original readjustment of the North American Datum. Soon after this adjustment, it was realized that the axes of the NAD 83 (1986) system were not correct and that the scale was different. This was determined through GPS, SLR, and VLBI measurements. This led individual states to establish High Precision Geodetic Networks, which are now called the High Accuracy Reference Network (HARN). As each state developed their own HARN, the NGS adjusted the NAD 83 reference system based on those HARN stations and this is referred to as the NAD 83 (HARN) realization. Three additional realizations have been developed. These relate the continuously operating reference stations (CORS) established by NGS. These are defined as NAD 83 (CORSXX), where XX is the year of the realization.

What Does This Mean for GIS?

The first question that often is asked is how do these different systems change the coordinates of my data. Assuming that the coordinates are in NAD 83, upwards to a meter difference may be found between NAD 83 (1986) and NAD 83 (HARN) control points. Since each realization is based on a better understanding of the earth, subsequent discrepancies are much smaller. Between NAD 83 (HARN) and NAD 83 (CORS93), most discrepancies will be below 10 cm. Finally, between any two NAD 83 (CORSXX) stations, the difference is generally less than 2 cm. Depending on the accuracy level of the GIS data, these discrepancies may be negligible.

Another consequence of these refinements is that the user needs to know what reference system is being used as the basis for the measurement. This issue may, in part, explain why data collected by GPS may not agree to existing data already residing within the database. There is a tendency to treat GPS data as absolute because of the capabilities that this tool gives us. Therefore, discrepancies may cause additional problems in the future as new realizations are developed.

State Plane Coordinates

An important part of any mapping project is the coordinate system utilized in depicting the spatial data. Converting data to a plane, or a surface that can be developed into a plane, is called a transformation. There is no map projection transformation that will allow the user to convert data from geographic coordinates to state plane coordinates without distortion. The transformation will distort shapes (or angles), distances, and areas. But, there are certain properties about projection transformations that can minimize these distortions. For example, the state plane coordinate system utilizes a conformal projection, which helps preserve the angles. Moreover, the system is usually designed to limit the size of the state plane coordinate zone to about 158 miles thereby minimizing the distance distortion as well. Theoretically, distance distortion should fall within 1:10,000 of the actual distance measured on the ground.

Two basic types of projections surfaces are used to create the plane surface: a cone and a cylinder. The cone, called the Lambert Conformal Projection, is oriented such that the

apex coincides with the axis of rotation of the earth. The cone intersects the earth along two lines, called the standard parallels – lines of latitude. A central meridian is designated as the line from which the X-coordinates are based. The Lambert projection is used for states with a predominant east-west extent.

The Transverse Mercator Projection is also a conformal projection but a cylinder is used instead. The axis of the cylinder is oriented perpendicular to the rotational axis of the earth. Like the Lambert, the cylinder intersects the earth along two lines, unfortunately, they are not related to the geographic coordinates of the earth. The Transverse Mercator projection is used for areas with a predominant north-south extent. When the state plane coordinate system was first established, Michigan was set up with this projection system. But, because development and growth at that time was predominantly in the southern part of the state, it was decided to change to the Lambert Conformal Projection.

Transformations

One of the most common programs used in coordinate transformations is CORPSCON. This program represents a global transformation model. The basis of CORPSCON is NADCON, developed by the NGS. Conceptually, the NADCON program has established a grid over the entire country and at each grid node transformation parameters have been developed. These parameters are based on known data from two systems. The CORPSCON menu is shown in figure 7.

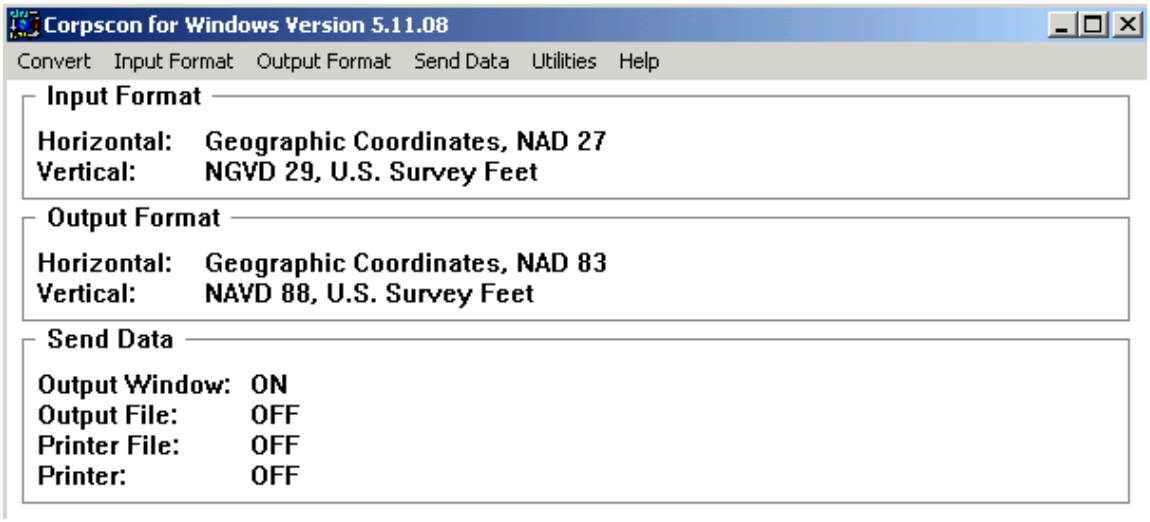

Figure 7. CORPSCON main menu.

The software will allow the user to convert between different coordinate systems. A printout of the input and output options are shown in figure 8. It is important to understand that any transformation is just a model and that the results may differ from the known values. For example, in figure 9, a couple of transformations using CORPSCON were performed on station SPICER, whose coordinates are known in NAD 27, NAD 83 (1986), and NAD 83 (HPGN). The NAD 83 (1994) data are in the HPGN/HARN

reference system. As a rule of thumb, 1" of arc in latitude and longitude represents about 100' on the ground. While the differences shown in figure 9 are not significant, they could be detrimental if not carefully evaluated. The only way that data from NAD 27, as an example, to NAD 83 can be done correctly is by adjusting the observations in the NAD 83 reference system.

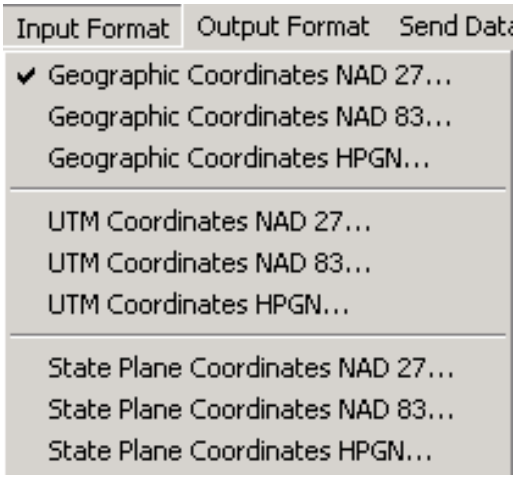

Figure 8. Transformation options within the CORPSCON software package.

NO VERTICAL DATUM USED

ADJUSTMENT DATE	LATITUDE	LONGITUDE
NAD 83 (1994) "True"	43° 40' 38.61563"	85° 36' 07.04729
NAD 83 (1994) "Calc"	43° 40' 38.61343"	85° 36' 07.05675
NAD 83 (1986) "True"	43° 40' 38.61471"	85° 36' 07.05917
NAD 83 (1986) "Calc"	43° 40' 38.60732"	85° 36' 07.06786
NAD 27 (Base Meas.)	43° 40' 38.53246"	85° 36' 06.89291

Figure 9. Differences in performing CORPSCON transformations.

To transform coordinates from one realization of NAD 83 to other realizations of NAD 83 or to different realizations of ITRF, NGS has a free software program available to users. The program is called Horizontal Time-Dependent Position (HTDP). The software can be used on-line or downloaded for use on a personal computer. The transformation parameters are based on a 14-parameter polynomial.

Conclusion

This paper provides a basic introduction to datums for GIS applications. The analyst needs to be aware that problems can occur when one accepts, carte blanche, the data from a data provider just because the measurement tool is considered "accurate". Coordinates and datum definitions will change in the future because our knowledge of the "true"

geometry of the earth will become better. As a GIS specialist, it is important to recognize this fact and to understand the changes that each definition and realization brings to the spatial data being depicted.

References

Iliffe, J.C., 2000. Datums and Map Projections, Boca Raton: CRC Press, 150 p.

Smith, J.R., 1997. Introduction to Geodesy: The History and Concepts of Modern Geodesy. New York: John Wiley & Sons, Inc., 224 p.