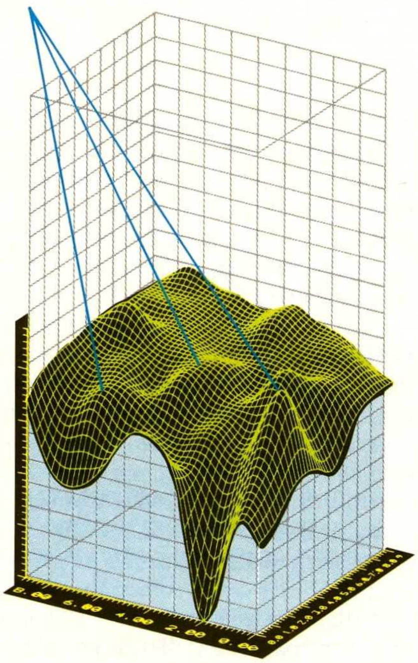

Basic Geodesy for GPS

Richard B. Langley

University of New Brunswick

"Innovation" is a regular column in GPS World *featuring discussions on recent advances in GPS technology and its applications as well as on the fundamentals of GPS positioning. This month we present another article in our tutorial series – this time on the basic principles of geodesy. These principles provide the scientific underpinning for all surveying and mapping activities that use the Global Positioning System.*

This column is coordinated by Richard Langley and Alfred Kleusberg of the Department of Surveying Engineering at the University of New Brunswick. We welcome your comments and suggestions of topics for future columns.

Just what is geodesy? The word itself comes from two Greek words: γῆ, meaning *earth*, and δαίω, meaning *I divide*. So clearly geodesy has something to do with dividing up or measuring the earth. A modern definition of geodesy is "the science of determining the size and shape of the earth, including its gravity field, in four-dimensional space-time." As such, its major subdisciplines are precise positioning in well-defined coordinate systems, description of the global gravity field, and the study of temporal variations in position due to natural and engineered causes.

HISTORICAL PERSPECTIVE

Geodesy is one of the oldest sciences. The first attempt to accurately measure the earth's size occurred in the third century B.C. It was commonly known by the intelligentsia of the day that the earth was round, or, more properly stated, spherical; but its size was unknown.

Eratosthenes, chief librarian and director of research at the Museum of Alexandria, which at the time was the preeminent center of learning in the Mediterranean world, measured the circumference of the earth using an extremely simple technique. He believed that at noon on the day of the summer solstice, the sun was directly overhead at Syene (now Aswan) so that the bottom of a well would be directly illuminated by the sun's rays. At the same time in Alexandria, the sun was a little lower in the sky. By measuring the length of the shadow of a small vertical rod at Alexandria, Eratosthenes concluded that the sun was one-fiftieth of a circle, or 7° 12′, below the zenith. He believed Syene was due south of Alexandria and that therefore the angle at the center of the earth formed by radii through the two cities also must be one-fiftieth of a circle. So Eratosthenes computed the earth's circumference by multiplying the distance between Alexandria and Syene by fifty. This distance was estimated to be 5,000 stadia, and hence the circumference of the earth worked out to be 250,000 stadia.

Eratosthenes, for some unknown reason, subsequently amended his estimate to 252,000 stadia. The precise length of the Egyptian stadium in terms of modern units is uncertain. If we assume the likely length of 158 meters, Eratosthenes' circumference works out to be about 39,820 kilometers, with the corresponding radius equal to 6,338 kilometers — very close to the actual values. However, the near agreement is a bit serendipitous considering the errors in some of Eratosthenes' assumptions. For example, Syene was not directly on the Tropic of Cancer but about 40 kilometers north, and the city was not due south of Alexandria but about 300 kilometers to the east. Nevertheless, Eratosthenes' determination of the earth's size was a remarkable achievement for its time and one that would not be improved upon until the 17th century A.D.

In the March 1991 Innovation column, we mentioned Sir Isaac Newton's derivation of Kepler's laws of orbital motion. Newton also made a significant contribution to geodesy. In the first edition of *Principia* published in 1687, Newton postulated that the earth was slightly ellipsoidal in shape rather than spherical as previously had been assumed. He developed this notion using his new theory of gravity and was able to confirm his prediction using accurate measurements of time kept by pendulum clocks. (Clocks set to give the correct time at, say, Paris were observed to run more slowly in places near the equator due to the slightly weaker pull of gravity there.) From both theory and observation, Newton concluded that "the earth is higher under the equator than at the poles, and that by an excess of about 17 miles" (*Principia*, Book III, Proposition XX). As a fraction of the earth's equatorial radius, this value worked out to be about 1/230. The ratio of the difference between the lengths of the equatorial and polar radii (the semimajor and semiminor axes, respectively, of the ellipsoid) to the length of the equatorial radius is called the *flattening*.

Not everyone believed Newton. The French astronomer Jacques Cassini, misled by somewhat inaccurate measurements, believed the earth was elongated at the poles, that its shape was that of a prolate, rather than an oblate, ellipsoid. A heated debate ensued and was not concluded until the Académie Royale des Sciences in Paris mounted expeditions to Lapland and Peru (now Ecuador) between 1736 and 1744 to determine, by precise angle and baseline measurements, the lengths of a pair of meridian arcs. With the latitudes of the endpoints of the arcs measured astronomically, the curvature of the arcs could be established and the equatorial radius of the earth and its flattening computed. The results confirmed once and for all that the earth had an equatorial bulge as Newton had predicted. Voltaire, alluding to Pierre Louis Moreau de Maupertuis, one of the leaders of the Lapland expedition, unkindly pointed out in the 1752 version of *Discours en Vers sur l'Homme*: *Vous avez confirmè dans ces lieux pleins d'ennui / Ce que Newton connut sans sortir de chez lui* (You have confirmed in these places full of difficulty / That which Newton knew without leaving home).

Over the next 200 years, the determinations of the radius and flattening of the earth became more accurate as the techniques of field geodesy were refined. However, it wasn't until the dawn of the space age that our knowledge of the earth's size and shape significantly improved. The orbital plane of a satellite precesses (rotates about the earth's polar axis, like an inclined spinning top) due to the gravitational force exerted by the earth's equatorial bulge, with the rate of precession dependent on the size of the bulge and hence on the value of the flattening, *f*. An analysis of the orbit of Sputnik 2, launched on November 3, 1957, showed that its orbital plane was precessing at a rate consistently 0.7 percent less than expected based on the then accepted value for *f* of 1/297.1. In order for the predicted rate to match the observed rate, the earth's flattening had to be 1/298.1. The new value for the flattening meant that the difference between the earth's equatorial and polar radii was about 85 meters less than had been thought. Although relatively small, this difference was important to geodesists, who were already able to make position measurements accurate to 10 meters.

This initial success at space geodesy was quickly followed by another. An analysis of the orbit of the Vanguard 1 satellite, launched on March 17, 1958, showed that the earth was slightly pear-shaped, with a slight hump around the North Pole, a slight depression around the South Pole, and a slight bulge just south of the equator. These deformations are small — on the order of 20 meters — but again, important to the work of geodesists.

THE GEOID

The scientists who analyzed the orbits of the Sputnik 2 and Vanguard 1 satellites were actually mapping the earth's gravity field. Because of latitudinal and longitudinal variations in the distribution of mass within the earth, the gravity field is quite complex. At each point above, on, and below the earth's surface, gravity has a certain magnitude and direction — it is a vector.

It is difficult to picture such a three-dimensional vector field, however, and even harder to mathematically manipulate it. As an alternative, geodesists have found it useful to represent the gravity field in terms of a scalar quantity called a *potential*. The gradient or spatial change in the potential at a certain point in space is equal to the gravity vector at that point.

The locus of all points with the same potential is a closed irregular but smooth surface surrounding the earth. An infinite number of such equipotential or geopotential surfaces exist nestled one inside the other, not unlike the layers of an onion. A characteristic of an equipotential surface is that the gravity vector at each point on the surface is perpendicular to it. In other words, an equipotential surface is a level surface. The undisturbed surface of any uniform body of water corresponds to a particular equipotential surface.

When geodesists talk about the shape of the earth, what they actually mean is the shape of its equipotential surfaces. As we mentioned, there are an infinite number of equipotential surfaces, but the one that most closely approximates mean sea level has special significance. The surface of the ocean is not quite level, even when the effects of waves and tides are averaged out. Prevailing winds, currents, and variations in salinity are responsible for the departures, called *sea surface topography*, which can be up to about a meter in size. Nevertheless, the equipotential surface best fitting the average sea surface over the whole globe can be determined. This surface is called the *geoid*.

One of the major tasks of geodesy is to map the geoid as accurately as possible (see "GPS and GEOID90 — the New Level Rod" by Dennis Milbert in this issue). The geoid usually is portrayed in terms of the height of a particular point on its surface above (or below) a corresponding point on a particular reference ellipsoid. This undulation, or *geoidal height*, *N*, can be positive or negative and range in value up to 100 meters or so.

The geoid is important both for long-term scientific research and for more mundane use. From maps of the geoid we can learn something about the structure of the earth's crust and upper mantle and its evolution through plate tectonics. However, the geoid also finds everyday use as the surface from which *orthometric heights*, the heights usually found on topographic maps, are measured.

GEODETIC COORDINATES

The three-dimensional position of a point on the surface of the earth is represented by a triplet of numbers, or coordinates, that refer to a particular coordinate system. For the coordinates to be meaningful, the system must be well defined; that is, the origin of the system (0, 0, 0) and the coordinate axes must be fixed with respect to the solid earth. The position of the origin and the direction of the axes can be chosen arbitrarily, though modern practice locates the origin at the earth's center of mass, the geocenter, and positions the *z*-axis so that it nearly coincides with the earth's axis of rotation.

The rotation axis actually moves slightly with respect to the solid earth as a result of the phenomenon known as *polar motion*, so an average pole position must be chosen in order to fix the *z*-axis. This was first done in the early 1900s, with the *z*-axis defined by the average position of the earth's rotation pole between the years 1900 and 1905. This position was known as the *Conventional International Origin*. In recent years, the concept of a reference pole has been refined and the adopted position is now referred to as the *Conventional Terrestrial Pole* (CTP). The *x*- and *y*-axes of this coordinate system are orthogonal to the *z*-axis, with the *x*-axis passing through the intersection of the Greenwich meridian and the earth's equatorial plane. The position of the Greenwich meridian used to be defined by a scribe mark at the Old Greenwich Observatory in London. Now it is defined implicitly by the International Earth Rotation Service in Paris via adopted coordinates for very long baseline interferometry (VLBI) and laser ranging stations. A coordinate system defined in this way is known as a *conventional terrestrial system* (CTS).

Whereas Cartesian coordinates (*x*,*y*,*z*) are very convenient for calculations, they are not the customary coordinates of cartographers and navigators. It was traditional for cartographers, going back to the third century B.C., to express positions on the surface of the earth using angular or spherical coordinates — latitude and longitude — rather than Cartesian coordinates. Up until the time of Newton, and even later for certain purposes, these coordinates were established by assuming the earth to be a sphere. But geodesists realized that for higher accuracies, the earth's ellipsoidal shape must be taken into account. Accordingly, spherical coordinates

Geodesists realized that for higher accuracies, the earth's ellipsoidal shape must be taken into account.

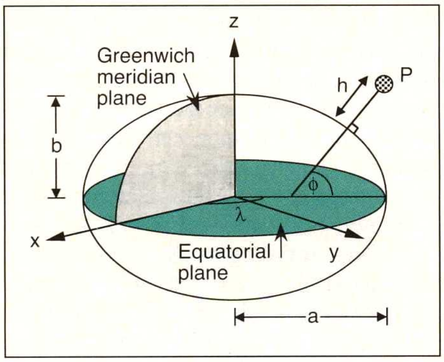

Figure 1. A geodetic coordinate system (the flattening of the ellipsoid has been greatly exaggerated), showing ellipsoid semimajor axis, a, and semiminor axis, b; flattening equals (a − b) / a

gave way to ellipsoidal coordinates.

The ellipsoidal coordinates of a point, P, are: the *geodetic latitude,* ϕ, the angle measured in the meridian plane through P between the equatorial (*x-y*) plane of the ellipsoid and a line perpendicular or normal to the surface of the ellipsoid at P; and the *geodetic longitude,* λ, the angle measured in the equatorial plane between the zero meridian (defined by the *x*-axis) and the meridian plane through P. In geodetic work, latitude conventionally is reckoned as positive toward the north and longitude as positive toward the east.

The coordinates (ϕ,λ) define a position on the surface of the ellipsoid. But to position a point on the physical surface of the earth, a third coordinate — the height above the ellipsoid — is required. This height, the *geodetic height, h*, is measured along the normal between the ellipsoid and the point. The position of a point in geodetic coordinates is then fully defined by the triplet, ϕ,λ,h (see Figure 1). It is a straightforward procedure to transform geodetic coordinates to Cartesian coordinates (x,y,z) or vice versa.

In general, the geodetic height of a point will not be the same as the orthometric height due to the noncoincidence of the ellipsoid and the geoid. However, the orthometric height can be computed by subtracting the geoidal undulation from the geodetic height (see the Milbert article for a more-detailed discussion of the relationship between geodetic and orthometric heights).

A specifically oriented reference ellipsoid constitutes a *geodetic datum*. Over the years, hundreds of *geodetic* or *horizontal datums* have been created by various agencies for surveying and mapping purposes within particular jurisdictions or regions (separate datums were established to provide orthometric heights). On each datum, a network of control stations was established to provide surveyors with access to accurate coordinates.

Eight parameters are required to define a geodetic datum: two to specify the dimensions of the ellipsoid, three to specify the location of its center, and three to specify the orientation of the ellipsoid. The definition was usually effected by adopting a reference ellipsoid of a particular shape; fixing the latitude, longitude, and geodetic height of an initial point or datum origin located near the center of the geodetic network being established; fixing the azimuth of a line from this point; and fixing the deflection of the vertical (the spatial angle between the gravity vector and a perpendicular to the ellipsoid) at the initial point.

The ellipsoids of these regional datums

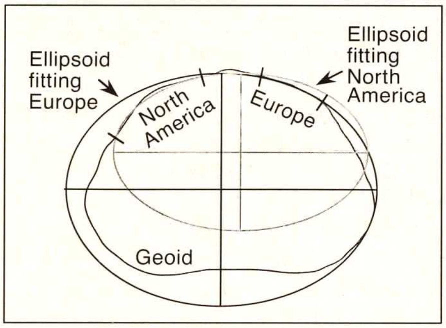

Figure 2. Traditionally, different regions have used different ellipsoids (the flattening of the ellipsoids and the undulations of the geoid are exaggerated)

were chosen to be nongeocentric so that the ellipsoid could conform as closely as possible to the geoid over the region (the area of interest for mapping) rather than the whole globe (see Figure 2). The result is that not only do the centers of different local reference ellipsoids not coincide, but the ellipsoids may have different semimajor axes and flattenings, and they are rotated slightly with respect to each other. Table 1 lists a few of the many such datums in use around the world today. Modern practice is to establish datums using geocentric ellipsoids with space geodesy techniques, as will be discussed later.

We should point out that the term datum can also refer to a description of the coordinate system and the set of all points and lines whose coordinates, lengths, and directions have been established by measurement or calculation together with the defining ellipsoid and its orientation.

WGS84

The struggle to tie different regional datums together for military and other purposes and the advent of satellite-based positioning systems asserted the need for a global geodetic reference system. One of the first such systems was the U.S. Department of Defense World Geodetic System (WGS) introduced in 1960. WGS60 was created from a global database of conventional geodetic measurements, satellite observations, and data from HIRAN (High-Precision Short-Range Navigation), an airborne trilateration, or range-measuring, system developed during and after World War II. In the years following the introduction of WGS60, the accuracy and number of satellite observations greatly increased and led to the development of WGS66 and subsequently WGS72.

WGS72 initially was adopted as the CTS for describing the orbits of the GPS satellites in their navigation messages. But as with WGS60 and WGS66, the accuracy of WGS72 eventually was found wanting. WGS72 was superseded by WGS84 and has been used for GPS navigation messages since January 1987.

The reference ellipsoid of WGS84 is essentially that of the International Union of Geodesy and Geophysics (IUGG) Geodetic Reference System 1980 (GRS80) with some minor changes. This ellipsoid was adopted by the international geodetic community at the IUGG's 17th quadrennial meeting, held in Canberra, Australia, in 1979, as best representing the size and shape of the earth. The WGS84 ellipsoid is specified by the value of 6,378.137 kilometers for its semimajor axis, a form factor describing the earth's equatorial bulge, from which a flattening of 1/

Table I. Relationships between selected geodetic datums and WGS84

Datum	Ellipsoid parameters a (m)	1/f	Datum shifts (m) Δx	Δy	Δz	Coordinate system rotations (") ϵ	Ψ	ω	Scale (ppm) ΔS
NAD83	6,378,137	298.257 222 101	0.42	0.95	−0.62	−0.012	−0.006	0.012	−0.1364
NAD27	6,378,206.4	294.978 698	−4	166	183	−0.257	0.341	−0.088	0.3723
European 1950	6,378,388	297.0	−102	−102	−129	0.413	−0.184	0.385	2.4664
South American 1969	6,378,160	298.25	−56	−3	−38	0.123	−0.569	−0.158	−0.6412
Australian Geodetic 1966	6,378,160	298.25	−127	−50	153	0.058	−0.018	−0.089	1.2065
Indian	6,377,276.345	300.8017	227	803	274	−0.444	−0.645	−0.353	6.5931

298.257 223 563 is derived, and values for the mean rotation rate of the earth and the product of the mass of the earth multiplied by the gravitational constant.

How well does the WGS84 ellipsoid represent the actual earth? The equatorial radius is probably in error by no more than 1 or 2 meters, and the value for the flattening is accurate to about 3 parts in a million.

In effect, WGS84's coordinate system was realized by adopting coordinates for more than 1,500 U.S. Navy Navigation Satellite System (Transit or Doppler) stations worldwide.

The coordinate system of WGS84 is a realization of the CTS as established by the Bureau International de l'Heure (BIH) on the basis of coordinates adopted for BIH stations (the BIH was a forerunner of the International Earth Rotation Service). The z-axis of the coordinate frame is parallel to the direction of the CTP; the x-axis lies at the intersection of the CTP's equatorial plane and the zero meridian; and the y-axis completes the system. The center of the coordinate frame coincides with the center of the WGS84 ellipsoid, and the coordinate axes coincide with the rotational axes of the ellipsoid.

In effect, WGS84's coordinate system was realized by adopting coordinates for more than 1,500 U.S. Navy Navigation Satellite System (Transit or Doppler) stations worldwide. However, the co-siting of Doppler stations with VLBI and satellite laser ranging (SLR) stations revealed that the Doppler coordinate system had meter-level errors. For example, the origin of the coordinate system was discovered to be about 4.5 meters above the earth's center of mass. So the Doppler station coordinate set was modified in origin, scale, and orientation to agree in the mean with the VLBI and SLR results.

WGS84 also includes a description of the earth's gravity field. Knowledge of the gravity field is needed, for example, for modeling the orbits of satellites. The field is described by a series of coefficients or values that account for smaller and smaller features with each additional term. These terms are called *spherical harmonics*. Spherical harmonics are identified by a pair of indices called the *degree, m*, and *order, n*. The harmonics with $m = 0$ are independent of longitude and are called *zonal harmonics*. The others are called *tesseral harmonics*. Those tesseral harmonics with $m = n$ are called *sectorial harmonics*. The WGS84 gravity field is complete up to degree and order 180, meaning that variations in the field over an area roughly 200 kilometers by 200 kilometers are described. However, because of the military importance of this gravity field model, only the coefficients up to degree and order 18 are unclassified. The field coefficients were obtained from the analysis of Doppler satellite tracking data, SLR data, surface gravity data, oceanic geoid heights from satellite altimetry, and GPS data as well as from analyses of data from a number of other satellites.

The coordinates directly computed by GPS receivers refer to the same coordinate system used to provide satellite coordinates to the receivers, usually WGS84. These coordinates may be expressed as Cartesian coordinates (x,y,z) or geodetic ellipsoidal coordinates (ϕ,λ,h). However, most GPS receivers also provide an option to transform the coordinates to one of a number of different regional datums, such as those listed in Table 1. The transformation parameter values in Table 1 were determined from investigations carried out by the Defense Mapping Agency (DMA). The datum shifts show the mean offsets of the reference ellipsoids from the center of the WGS84 ellipsoid; the coordinate system rotations represent the misalignment of the regional datum coordinate system axes with respect to those of WGS84; and the scale parameter accounts for differences in each datum's length scale with respect to WGS84's scale. These parameters can be used to transform coordinates given in a particular datum to WGS84. However, adjustments for variations in scale and distortions in the original datum are not included. For such transformations, DMA provides a set of multiterm polynomials that it has derived from extensive least squares analyses.

GPS receivers can also display orthometric heights rather than geodetic heights if the geoidal undulation is known.

NAD83

GPS users in North America currently have a slight advantage over users in other regions as a result of the recent introduction of a new datum, the North American Datum of 1983 (NAD83). NAD83 replaces the North American Datum of 1927 (NAD27), the coordinates of which had become inadequate for many purposes. Many of the published coordinates of survey control stations were unreliable due to errors and distortions in NAD27. In fact, relative coordinates were sometimes in error by as much as 1 part in 15,000. A further disadvantage of NAD27 was that its reference ellipsoid was nongeocentric and was not precisely oriented with respect to the CTS as established by the BIH.

The need for a readjustment of North American networks was realized in the late 1960s, and work on the new datum officially began in 1975. NAD83 was obtained by a least squares adjustment of more than 1.75 million geodetic observations at sites in the United States, Canada, Greenland, Mexico, Central America, and the Caribbean. Doppler and VLBI observations supplemented the observations obtained using traditional surveying techniques. Originally scheduled for completion in late 1982, the new datum was christened NAD83. But due to various delays and extensions, the National Geodetic Survey did not publish the first NAD83 coordinates until March 1987.

The reference ellipsoid and coordinate system of NAD83 are almost identical to those of WGS84; the two systems agree at about the 0.1-millimeter level. So, WGS84 coordinates provided by GPS receivers can be used as NAD83 coordinates. However, GPS surveyors should realize that if they occupy particular reference markers for which published NAD83 coordinates exist, the coordinates computed for the markers from GPS observations may differ from the published coordinates by a meter or more due to remaining errors and distortions in the datum.

To provide geodetic reference coordinates at a higher accuracy than afforded by standard NAD83 coordinates, many states have established special GPS-derived "NAD83 high-precision" networks.

UTM

Some GPS receivers can also project ellipsoidal coordinates onto a mapping plane, that is, a flat map. Projecting an ellipsoidal surface onto a flat surface causes some distortion. However, projections have been developed that minimize these distortions. One such projection is the Universal Transverse Mercator (UTM). This projection uses conformal mapping so that the magnitude and sense of angles measured on the ellipsoid and the shapes of small geographical features are preserved when coordinates are transformed to the mapping plane. The UTM projection, which can trace its lineage back to Karl Friedrich Gauss, has been adopted by the IUGG, NATO, and other military organizations and many civil administrations worldwide for various mapping needs.

The UTM projection divides the world into 60 zones, each with a width of 6 degrees of longitude, and superimposes a grid onto them. Each zone, which constitutes a segment of a reference ellipsoid, is projected onto a cylinder whose axis is parallel to the earth's equator and whose radius is chosen to keep the scale errors of the projection within acceptable limits. Coordinates of points on the ellipsoid within a particular zone can then be transformed to coordinates on the UTM grid.

UTM coordinates are generally referred to as *eastings* and *northings* and are expressed in meters. Eastings are reckoned from the central meridian of a zone and have 500,000 meters added to them — the so-called *false easting* — so that all coordinates remain positive. Northings are reckoned from the equator, which has a coordinate value of 0 meters for work in the northern hemisphere and a *false northing* of 10,000,000 meters for work in the southern hemisphere.

The U.S. State Plane Coordinate System uses the transverse Mercator projection and the Lambert conic map projection, another conformal projection, to map each state of the Union, Puerto Rico, and the U.S. Virgin Islands in one or more zones onto a plane rectangular coordinate system. (The panhandle of Alaska is a unique case with its own special projection.) The transformations from NAD83 geodetic coordinates to grid coordinates yield errors less than about 1 centimeter for points within the boundaries of a particular zone so that either geodetic coordinates or the corresponding grid coordinates of a point may be used depending on the application.

CONCLUSION

In this short article, we only scratch the surface of the subject of geodesy and its relationship to GPS. For a more in-depth look, the interested reader can consult one of several textbooks available on the subject. For a nonmathematical overview, see J.R. Smith, *Basic Geodesy* (Landmark Enterprises: Rancho Cordova, California, 1988). For those not put off by a little algebra or calculus, see P. Vaníček and E.J. Krakiwsky, *Geodesy: The Concepts*, 2nd ed. (North-Holland Publishing Company: Amsterdam, 1986; distributed by Elsevier Science Publishing Co., Inc., New York, New York). Both books are available in paperback. ■

From *GPS World's Big Book of GPS 2000*

UPDATE:

The Defense Mapping Agency and its successor, the National Imagery and Mapping Agency, have revised World Geodetic System 1984's (WGS 84's) coordinate frame twice since its introduction in an effort to bring it into better alignment with the International Terrestrial Reference Frame (ITRF) maintained by the International Earth Rotation Service. On June 29, 1994, the GPS Master Control Station implemented a new version of WGS 84, denoted WGS 84 (G730), by adopting improved coordinates for the GPS monitoring stations. A more refined version, WGS 84 (G873), was implemented on January 29, 1997.

In the designations G730 and G873, "G" indicates that the updates used GPS data; the numbers refer to the GPS week number when the new monitor station coordinates were implemented in the precise ephemeris estimation process. The latest incarnation of WGS 84 agrees with the ITRF at the 5-centimeter level. For further details on WGS 84 and its relationship to regional geodetic datums, see the latest version of *Department of Defense World Geodetic System 1984 — Its Definition and Relationships with Local Geodetic Systems,* 3rd edition, NIMA Technical Report TR8350.2, National Imagery and Mapping Agency, Washington, D.C., 1997; available as a PDF file from the NIMA Web site: <http://www.nima.mil/publications/pub.html>.

The ITRF is updated each year to take advantage of improvements in observational techniques and to account for tectonic plate motions. The latest version is ITRF97 and was derived from a combination of data from very long baseline interferometry, satellite laser ranging, GPS, and DORIS (for Doppler Orbitography and Radio-positioning Integrated by Satellite). For further details about the ITRF, see "International Terrestrial Reference Frame," in the September 1996 issue of *GPS World* (Vol. 7, No. 9, pp. 71–74).

The transformation parameter values in Table 1 illustrate the relationships between various datums and WGS 84, but NIMA does not recommend their use for practical transformations between datums. Coordinates should be converted using either the Standard Molodensky transformation formulas with appropriate datum shifts and ellipsoid parameter values or multiple regression equations that attempt to model distortions in large regional datums. For further information, see the NIMA report on WGS 84 and the article "Coordinates and Datums and Maps! Oh My!" in the January 1997 of *GPS World* (Vol. 8, No. 1, pp. 34–41).

With regard to the relationship between NAD 83 and WGS 84, while one can expect to see differences of as much as a meter or more between NAD 84 and WGS 84 coordinates of the same point because of inherent uncertainties of published NAD 83 coordinates and the limitations of GPS point positioning, the definitions of the NAD 83 and WGS 84 reference systems are virtually identical. For mapping, charting, and navigation, the two systems are indistinguishable at approximately the 2-meter accuracy level and map scales of 1:5,000 or smaller. The GPS-derived NAD 83 High Accuracy Reference Networks (HARNs) established by the U.S. National Geodetic Survey in many states have relative accuracies 1–2 orders of magnitude better than the original NAD 83 networks and these HARNs may be useful for applications demanding high positioning accuracies.

The Universal Transverse Mercator (UTM) map projection and grid system, which covers the globe between 80 degrees south and 84 degrees north, is discussed in more detail in "The UTM Grid System" in the February 1998 issue of *GPS World* (Vol. 9, No. 2, pp. 46–50).

Helpful letters to the editor concerning the initial publication of this article were received from Melvyn Grunthal and Muneendra Kumar and appeared in the April and October 1992 issues of *GPS World* respectively.

WGS84, NAD83 clarified

Dear Sir:
I would like to take this opportunity to thank Dr. Richard B. Langley for his article, "Basic Geodesy for GPS," which appeared in the February issue of *GPS World*. Such an article is long overdue and will be of great benefit to the GPS user community. There are, however, some inaccuracies in the article that may mislead GPS users.

First, the article incorrectly represents the relationship between WGS84 and NAD83 (Table I, page 48). These datums effectively are based on the same ellipsoid, and both adopted an identical geocentric center determined by Doppler point position observations. The datum shifts (e.g., Δx, Δy, and Δz) are essentially zero. The WGS84 to NAD83 transformation information provided by Dr. Langley in Table I was derived by the Defense Mapping Agency (DMA) in 1987 as part of an evaluation of transformation techniques. The relationship in the table was based on a sample of stations and was not intended to portray the definitive transformation. Thus, the inclusion of this WGS84 to NAD83 transformation information is out of context and should not be used. (See *NOAA Professional Paper NOS 2, North American Datum of 1983,* Chapter 22, by Charles R. Schwarz, for further details.)

Second, the relationship of WGS84 and NAD27 is oversimplified by the transformation parameters provided in Dr. Langley's Table I. NAD27 (as well as most other datums) contains significant regional distortions and cannot be transformed accurately to WGS84 (or NAD83) by a single set of transformation parameters. Dr. Langley references this problem with the statement, "However, adjustments for variations in scale and distortions in the original datum are not included," when referring to Table I. However, his comment is easily overlooked.

Finally, I have concerns regarding the description of the use of NAD83. The author states, "However, GPS surveyors should realize that if they occupy particular reference markers for which published NAD83 coordinates exist, the coordinates computed for the markers from GPS observations may differ from the published coordinates by a meter or more due to remaining errors and distortions in the datum."

This statement is highly misleading. NAD83 has local distortions, particularly between horizontal control points, that are not intervisible by classical surveying techniques. However, the vast majority of NAD83 first- and second-order coordinates are accurate relative to one another at the 3 to 4 parts-per-million level. (See "Bulletin Géodésique," Vol. 64, No. 1, 1990, by Dr. Richard A. Snay, for further information.) Therefore, errors at the 1-meter or greater level would not be expected unless the points were separated by a distance of 200 kilometers or more. Dr. Snay's analysis demonstrates that first-order control normally shows only 10-centimeter error at spacings of 50 kilometers.

GPS World provides a significant forum for the exchange of ideas and concepts related to GPS. Thank you for the opportunity to comment on Dr. Langley's article.

Sincerely,
Melvyn C. Grunthal
Captain, National Oceanic and Atmospheric Administration
Chief, National Geodetic Survey
Rockville, Maryland

Langley responds: *I would like to thank Capt. Grunthal for helping to clarify my discussion of the relationships among WGS84, NAD83, and NAD27. In a short, introductory article of this sort it is difficult to fully explain some of the more esoteric concepts involved in establishing and relating geodetic datums.*

As Capt. Grunthal points out, both the National Geodetic Survey (NGS) and DMA adopted the same set of transformation parameters to relate the coordinates of the Doppler stations in the NWSC 9Z-2 system to the Bureau International de l'Heure's Terrestrial System. By definition then, the NAD83 and WGS84 coordinate systems are identical. I attempted to get this across in the article when I said (taking into consideration the tiny difference in the size of the reference ellipsoids used): ". . . the two systems agree at about the 0.1-millimeter level." But both NAD83 and WGS84 involved the adjustment of different sets of data. Although both datums attempted to realize a geocentric system, neither is perfectly geocentric. As a result of the different adjustments, I believe the realizations of the geocenter are bound to be slightly different.

I included NAD83 in Table I in an attempt to illustrate the different coordinate system realizations, not intending that the values listed (for the conterminous states) were to be used for practical coordinate transformation purposes. As I said in the article, ". . . WGS84 coordinates provided by GPS receivers can be used as NAD83 coordinates." In any case, most GPS surveyors will not be in a position to determine point position coordinates at the decimeter level, and the effect of any decimeter-level difference in the actual origin of the datums on relative coordinates is insignificant for most GPS surveys.

My comment about possible differences between published NAD83 coordinates for a point and those obtained by a GPS surveyor in the field was prompted by a statement that Charles Schwarz made in the article that Capt. Grunthal references:

"NAD83 and WGS84 should be thought of as geographically overlapping datums (in the sense of datums as adopted coordinates). There will be points with coordinates in both datums. The action to take when confronted with two sets of coordinates for a single point is up to the user. If neither position determination contains a blunder, then the differences of the coordinates should be small . . . . 'Small' differences must be properly understood here. The actual difference between coordinates may quite possibly be a meter or more."

What I failed to stress in my article was that differences of a meter or more will be exceptional. What I should have said was, ". . . the coordinates computed for the markers from GPS observations may occasionally differ from the published coordinates by a meter or more. . . ."

Readers interested in learning more about the scientific niceties of the relationships between modern geodetic datums and their coordinate systems might wish to consult the article "Important Parameters Used in Geodetic Transformations," by Tomás Solder and Larry D. Hothem, in the Journal of Surveying Engineering, *Vol. 115, No. 4, pp. 414–417.*

Capt. Grunthal, I appreciate your interest in the article and I am sure your comments will help the magazine's readers to further their understanding of the complicated topic of geodetic datums. It is not my intention to be confrontational and I did want to help you to "clear the air," but I also want to point out that there are some scientific issues here besides the practical (and perhaps political) ones of having the typical GPS user believe that as far as coordinates are concerned NAD83 is exactly the same as WGS84. It might be a good idea for one of your staff to write an article for GPS World *specifically on NAD83 and its use in the area of GPS surveying. I would be very happy to help coordinate such an article.*

Geodesy revisited

Dear Sir:
I refer to Richard Langley's article entitled "Basic Geodesy for GPS," which was published in the February issue of *GPS World*. I want to commend the author on his excellent explanations to simplify basic geodesy ingredients used in everyday GPS and to make them easier for nontheoretical users. However, I want to comment on some portions of the article, especially those related to the WGS84 definition and its relationships with other geodetic datums.

First, the WGS84 coordinate system was realized by correcting the Doppler NSWC 9Z-2 system for identified biases and adopting the Conventional Terrestrial System for epoch 1984.0 (BTS84) as defined by the Bureau International del'Heure (BIH). The 1,500 or so stations, as quoted by the author, were used only to establish the datum transformations with the 83 local geodetic datums (DMA *TR 8350.2*, second printing).

Second, Table 1 of the article is an extracted summary of Table 7.13 from the Defense Mapping Agency's (DMA) publication *TR 8350.2-A*, published in December 1987. In that publication, these seven-parameter solutions were included for theoretical discussion/study, and should not be used for everyday mapping, charting, and survey applications.

DMA has since updated and revised its recommended (and DMA's official) set of datum transformations (previously available in DMA TR's *8350.2* and/or *8350.2-B*). The 8350.2, Second Edition, now lists transformations for 105 local geodetic datums and also for the rectangular Soviet Geodetic System 1985, which is used in GLONASS.

Third, a document jointly agreed on by DMA and the National Geodetic Survey explains similarities and dissimilarities between NAD83 and WGS84 for mapping, charting, and surveying. A few salient points of this explanatory write-up include:

- Both WGS84 and NAD83 are defined to be technically identical to BTS84 in their origin, orientation, and scale.
- WGS84 is primarily a three-dimensional system with emphasis on absolute point positioning of its control points, whereas NAD83 is primarily a two-dimensional horizontal system with its station coordinates linked to each other in an adjusted network.
- Though some geodetic definition differences exist, the systems are to be considered identical for mapping, charting, and surveying applications.
- The mean shifts (ΔX, ΔY, ΔZ) are zero at ± 2 m (1σ) level; however, at individual points the geoidal heights may show differences of 1–2 meters and the coordinates differences of up to 5 meters.

Fourth, DMA has recently reviewed the impact of general use of multiple regression equations (MREs) and their suitability for local geodetic datum transformations. The latest recommended set of MREs pertains to continental size local geodetic datums, and these equations are for use only over large contiguous land areas. Their extrapolation to any area for which the MRE's were not developed can result in large errors and/or blunders and thus is not allowed.

Finally, the author mentioned the adoption of UTM by the North Atlantic Treaty Organization and other military organizations. He has not included its area of coverage, i.e., from latitude 80°S to 84°N, and also its exceptions in the longitudinal width of six UTM zones west and north of Norway.

Sincerely,
Muneendra Kumar
Defense Mapping Agency
Fairfax, Virginia