

Datums - Map Coordinate Reference Frames:

Part 1 - Basics

James R. Clynych, February 2006

Part 1

- I. What are Datums - in Geodesy and Mapping
- II. Origin of Datums
- III. Datums and Ellipsoids
- IV. Modern Datums and Reference Frames
- V. Realizations
 - A. How Datums Really Work
 - B. Pre-satellite Era Realizations
 - C. Satellite Era Realizations
 - D. Datums as Rubber Sheets
- VI. What Datum Am I On
 - A. Map Datums - Paper and Electronic
 - B. Navigation and Survey Equipment Datums
 - 1. Standalone GPS Users
 - 2. Differential GPS Users
 - 3. Other Electronic Navigation Systems
 - 4. Surveying

Part 2

- VII. Datum Transformations
 - A. Basic Methods
 - B. Vector Method - The 7 Parameter Transform
 - C. Molodensky Transform - The Historically Common Method
 - 1. Vector Viewpoint of Molodensky Transformation
 - 2. The Usual Statement of Molodensky Transformation
 - 3. Modern Uses of Molodensky Transformation
 - D. Local Fit Equations - Multiple Regression Method

I. What Are Datums – in Geodesy and Mapping?

The basic fact is: the **coordinates** of all locations on the earth are on **datums**. At high accuracy the coordinates of a point, the latitude, longitude and height or Cartesian coordinates, may exist in several self-consistent sets. These will be in different datums.

A datum is the answer to the practical problem of making an accurate map. If you wish to determine the relative location of a pair of points a few meters apart, the solution is obvious. Just measure the difference with a tape measure. The issue of orientation still exists though, but this can be solved using two "known" points to measure a third. Or observations of the stars can be used to define north.

In effect this defines a **local datum**. The known point, together with some method for determining the direction of north, defines the location of points measured from it. If the reference point is in error by 100 m north, so will all the points using it. They move together. This of course assumes these errors are small, at least as compared to the radius of the earth.

If you look at the legend of a topographic map, you will find that it lists the "datum" that is used. In fact there may be several datums, one for horizontal, one for vertical etc. These are important because they define the reference system that is used for the coordinates.

If you use a GPS navigation system not set to the map datum, you can be off by 100 m (usually) to a kilometer (sometimes). Navigation systems produce coordinates on datums. If this is not the same datum as the map being used, an error occurs. Several ship accidents have happened because the GPS navigation system and the navigation chart were on different datums.

The practical way to define a datum is with a whole set of reference markers and their associated coordinates. They should be carefully surveyed together. This gives a network that serves as a "**realization** of the datum". This provides a practical set of points spread out over the region covered. Surveyors use the closest survey marker that meets the accuracy needs. In practice almost all surveying is relative, from one point to the next.

This means that datums are the reference frames used in the construction of maps. Realization, the disks in the ground and the catalogue of their coordinates, are the practical way datums are used.

Things are complicated in practice as the same area may have two datums giving each point two different coordinates. In addition datums were often generated by individual countries, and did not match at the boundaries. The 1900 era German maps of France and the English maps of Germany did not match, even in the areas of overlap.

In general if a datum covers areas not directly connected by a survey, such as over a body of water, there are really different versions of the datum. This is not an academic

point. At least one recent ship grounding in the Caribbean was caused by using the wrong "NAD 83" in a GPS receiver. They used a version of "NAD 83" that did not match the "NAD 83" on the chart.

II. Origin of Datums

When we see a map or a globe today we have a fairly good idea of the meaning of the latitude and longitude lines on it. The longitude lines have a zero in Greenwich England and the zero of latitude is the equator. Even today there are subtle difference in the meaning of these statements. In the past there were major differences depending on the country you were from. The coordinates were said to be on different datums.

In the past each major country had its own system of coordinates. The latitude was meant to be the same, basically defined by the spin axis of the earth. But the origin of longitude is completely arbitrary and each country chose a convenient location on its soil. Thus two maps that were of the same area would not have the same coordinates for a city.

In fact both the latitude and longitude would often be different, although the latitude difference would be small. These reference systems were set up by establishing a "good primary reference site" and then doing **relative surveying** outward from it. This was usually an astronomical observatory so the coordinates of this primary point would be well known from celestial observations. The outward surveying had to stop at oceans and other major obstacles. So islands frequently had their own primary reference point that was not as well determined as the ones in major European capital cities.

Because these systems were not quite consistent from country to country, even where there were no survey problems, maps differed based on which country made them. Each map set was on a different Datum.

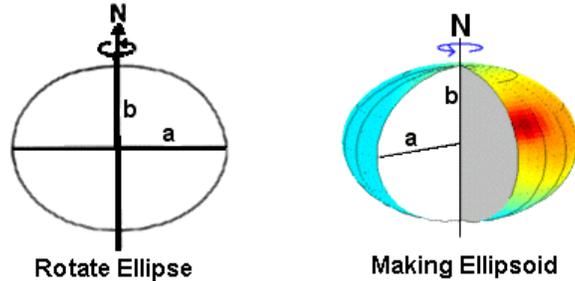
In reality there are many kinds of datums. Here we will only consider two, horizontal datums and vertical datums. Mapmakers distinguish between horizontal and vertical position because different techniques are used to measure heights and horizontal positions.

And things are even more complicated because the world is not quite a sphere.

III. Datums and Ellipsoids

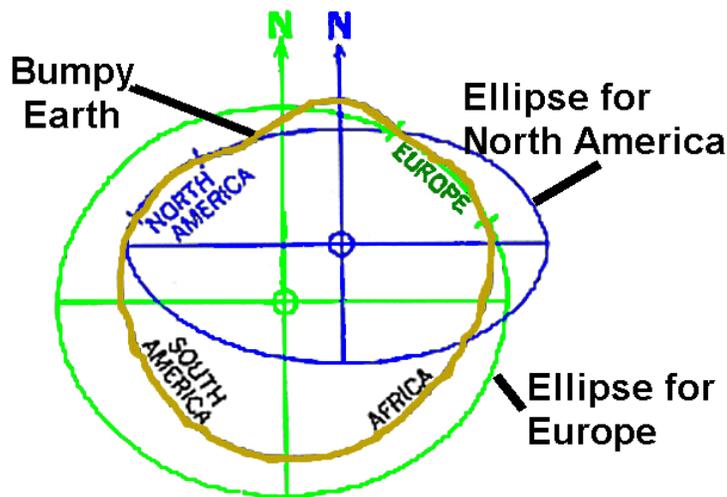
The world is slightly bigger at the equator than at the poles. The distance from the center of the earth to the equator is larger than the distance from the center to the poles by about 23 km. This is a factor of about 1/300 or 0.3 percent. While this is small, it is important for maps of the world or even much smaller regions.

The non-spherical earth is modeled by an ellipse of revolution. This shape is called and **ellipsoid**. (In some books it is called a **spheroid**.) In order to make an ellipsoid model of the earth, take an ellipse and align the shorter axis with the spin axis of the earth. The longer axis will point out the equator. Rotate the ellipse about the shorter, polar, axis to form a solid. This is the ellipsoid.



The shape and size of the ellipsoid representing the earth is very difficult to measure. A good measurement needs celestial observations at many points all over the world. These were unavailable before satellite based surveying began in the 1960's. So countries made due with what they had, usually over a "small region" like Europe.

Each country decided for itself what was the best estimate of the size and shape of the earth. They each had their own ellipsoid. And they also had their own primary reference point. Defining the latitude and longitude of a particular point on the earth defines the origin of the ellipsoid. That is choosing an ellipsoid and primary reference point coordinates gives both the shape and location of the ellipsoid. This model of the earth is needed for the surveying that defines the location of other points on maps.



Different Ellipsoids From Fitting Different Regions of Earth

Ellipsoids made from a "local" area's data, like Europe or North America could be significantly different. This did not bother the mapmakers because they just wanted something to use in making maps that worked well for their area. For example the ellipsoids chosen for Europe and North America are quite different, with the origins being offset about 250 m.

The ellipsoids are usually defined by two quantities. These could be the length of the two axes. However a more common set used in geodesy and surveying is the **semi-major axis** (equatorial axis length) and the **flattening**. The polar axis is also called the **semi-minor axis**. If the equatorial axis is called a, and the polar axis b, then the flattening is defined as

$$f = \frac{a - b}{a} = 1 - \frac{b}{a}$$

A list of the most common ellipsoids in use is given in the following table.

Name	Semi-Major Axis -a (Km)	Semi-Minor Axis - b (km)	1/Flattening
Airy	6377.563	6356.257	299.32
Modified Airy	6377.340	6356.034	299.32
Australian National	6378.160	6356.775	298.25
Bessel 1841	6377.397	6356.079	299.15
Clarke 1866	6378.206	6356.584	294.98
Clarke 1880	6378.249	6356.516	293.46
Everest	6377.276	6356.075	300.80
Fischer 1960	6378.155	6356.773	298.30
Helmert 1906	6378.200	6356.818	298.30
Indonesian 1974	6378.160	6356.774	298.25
International	6378.388	6356.912	297.00
Krassovsky	6378.245	6356.863	298.30
South American 1969	6378.160	6356.774	298.25
WGS 72	6378.135	6356.751	298.26
GRS 80	6378.137	6356.752	298.257
WGS 84	6378.137	6356.752	298.257

A datum in the modern sense is defined by choosing an ellipsoid and then a **primary reference point**. Therefore giving the ellipsoid used is not enough. The **North American Datum of 1927, NAD27**, uses the Clarke 1866 ellipsoid and a point in central Kansas (called Meade's Ranch) as its primary reference point. Note that some maps of some Caribbean islands are also listed as being on NAD27. But this is really a different datum because a different primary reference point was used.

There was a major update of the North American Datum in the early 1980. This resulted in the **North American Datum of 1983 or NAD83**. NAD83 uses Meade's Ranch but a new ellipsoid. The ellipsoid is **GRS80**, which is the same as the WGS84 ellipsoid. (**The World Geodetic System of 1984, WGS84**, was established at about the same time by the US Defense Department.) In essence new data was used to establish better coordinates for the existing major benchmark network in North America.

Thus the brass disks in the ground are the same, only a new database of coordinates for each marker is different. The update was done with many new measurements and computations. But the main result was a new **catalogue of the coordinates** of the survey marks already in the ground.

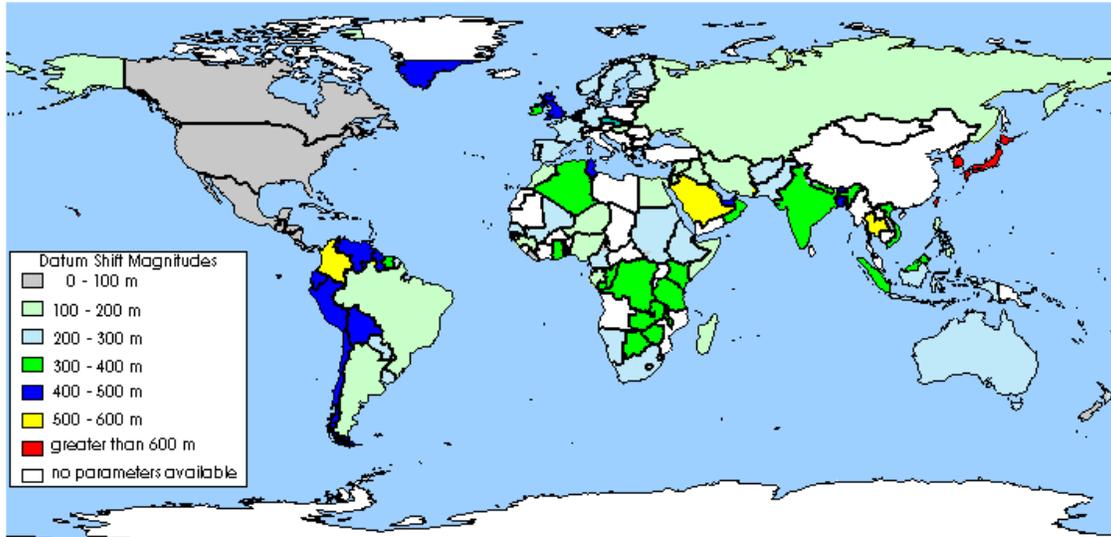
A few of the common datums in use today are:

North American Datum of 1927	NAD27	Clarke 1866
North American Datum of 1983	NAD83	GRS80 = WGS84
North American Vertical Datum of 1929	NAVD29	Clarke 1866
North American Vertical Datum of 1988	NAVD88	GRS80 = WGS84
World Geodetic System of 1984	WGS84	WGS84 = GRS80
South American Datum of 1969	SA1969	South American
European Datum of 1950	EU50	International

IV. Modern Datums and Reference Frames

Today maps almost always use Greenwich England for the longitude origin and try to have the best values for the ellipsoid and primary reference points. However we are still left with older surveys and maps that are on different datums. There are hundreds of them, but only about 50 that have common use.

With the advent of satellite surveying systems, worldwide coordinate systems were needed. This led to the establishment of worldwide systems. The US Defense Department made the first of these in 1966. A later one, called **World Geodetic System 1972, WGS72**, was quite successful. It was used for the Navy Navigation System, which was opened to public use. Later a better system, called WGS84 was generated when the **Global Positioning System (GPS)** required better coordinates.

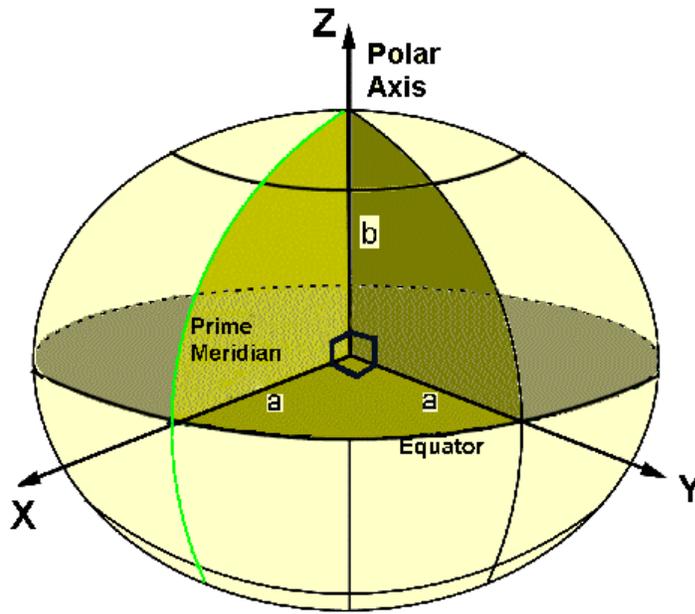


Horizontal Shifts on Common Maps to WGS 84

The above diagram shows the approximate shift in horizontal location between WGS84 and datums used commonly for maps. While the US is now officially on the North American Datum of 1983 (NAD83), which is essentially identical to WGS84, there are still many maps on the older NAD27. This is true even of government printed topographic maps from the US Geological Survey. In other areas of the world the shift are even larger. Kilometer size shifts exist for the Tokyo datum for example.

The science community has been working on a series of world reference systems that are called **International Terrestrial Reference Systems or ITRF's**. The earliest ones were ITRF92 and ITRF94, which was quite good. Modest improvements followed with ITRF97 and ITRF2000. The later two models were so accurate that models of the motion of the crustal plates of the earth had to be included.

The WGS's and ITRF's were basically defined as **Cartesian XYZ systems**. These are a perpendicular axis with the origin at the center of the earth. The Z-axis goes out the north pole. The X-Y plane is the equatorial plane. The positive X-axis defines the origin of longitude. This is called an **Earth Centered, Earth Fixed, ECEF coordinate system**. An ellipsoid was associated with each. This was needed to convert the xyz coordinates to latitude, longitude and height. Of course height had it's own complication which are discussed elsewhere.



Earth Fixed Cartesian Coordinates

X-Y Plane is Equatorial Plane
 X On Prime Meridian
 Z Polar Axis

V. Realizations

A. How Datums Really are Used

There is a second whole side to datums, the practical side. The primary reference point for the US is in central Kansas at a place called Meade's Ranch. It is impractical to begin all surveys in Kansas. The mapping organization of each nation do high quality surveys and establish a network of high accuracy points. These are usually a bronze disk set in concrete or rock.



This disk, along with the coordinates of the small punch mark in the disk center, generates a "realization" of the datum.

B. Pre-Satellite Era Realizations

In the past this work involved measuring angles between points. This was done in a series of triangles or more often four sided figures with the internal diagonals measured. These are called **braced quadradrils**. The primary network of the US prior to satellite surveying begins with a series of these in a large network.



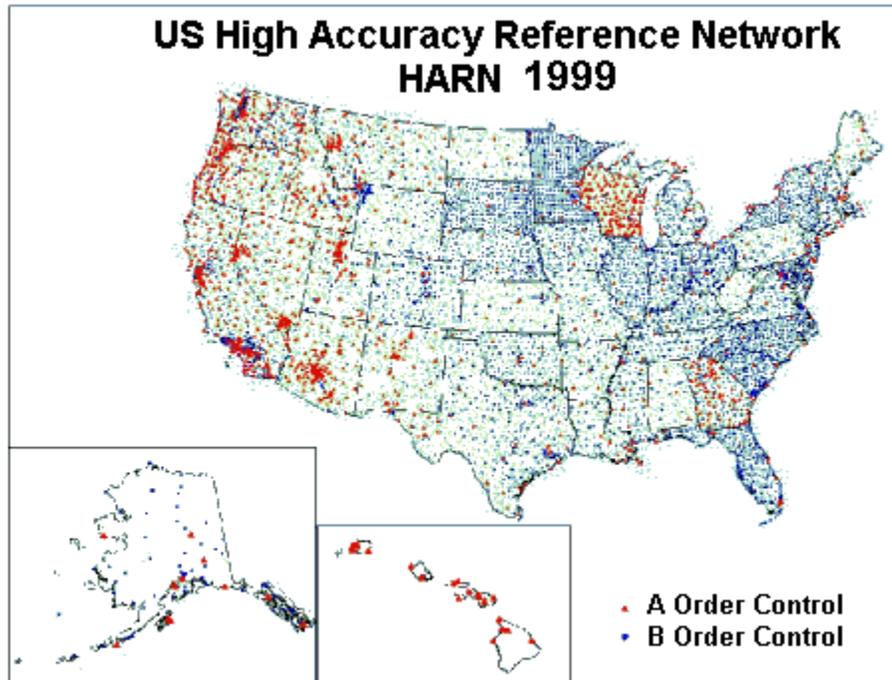
US Control Network Before Satellite Surveying

-  First Order Triangulation
-  Second Order Triangulation
-  First Order Traverse

The US primary reference point is at the center of two of these legs. The points at the intersection of each set of lines are "monumented" and documented. Together these form the primary, or **first order network** for all coordinates in the US. The relative, point to point, nature of classical surveying is clear from this form of network.

C. Satellite Era Realizations

With the coming of **satellite surveying**, reference points could be far apart and essentially disjoint. They were still measured by relative surveying, on much higher accuracies could be obtained over much longer distances. This is clear from the current US primary realization. It looks like a set of points. The density is a function of the amount of work beyond the minimum that each state decided to perform.



D. Datums As Rubber Sheets

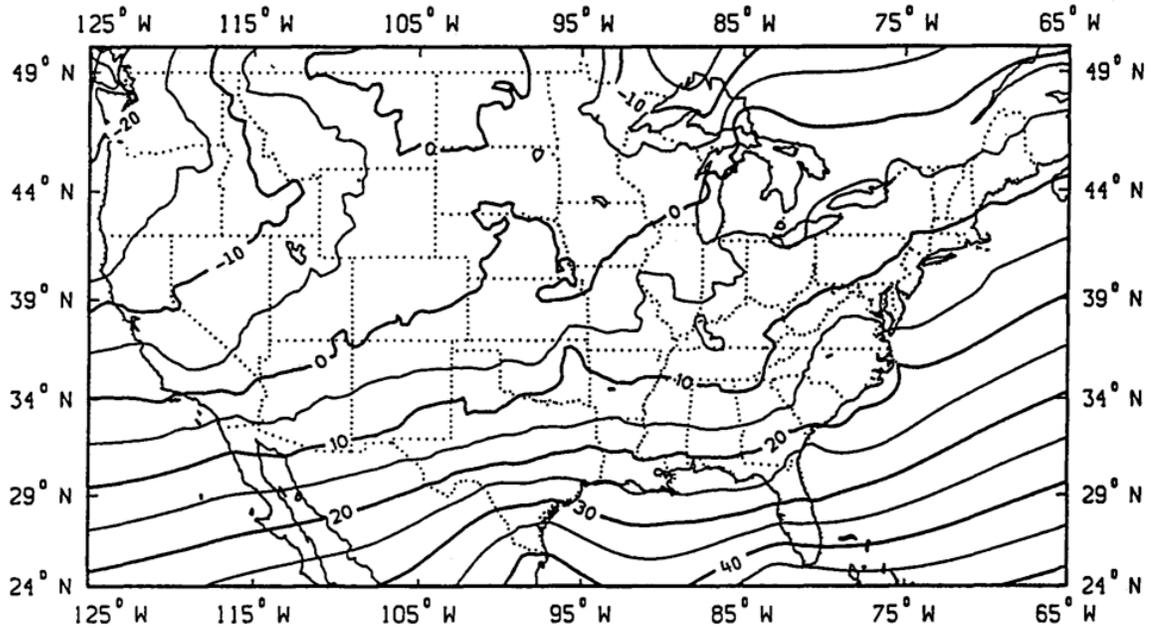
Datums are established by taking a large number of high precision survey measurements at many points over the region of interest and solving for a best set of coordinates. In doing this there must be many more measurements than 2 times the number of points. The factor of 2 comes in because of the two horizontal coordinates at each location.

Even with a ratio of 10 to 1 of the number of measurements to unknown, the solutions can have significant errors. The **random errors** will be minimized, but **systematic errors** will remain. The surveys measurements were almost all relative measurements from point to point. Errors could accumulate. In addition only a small number of distances (called baselines) were commonly done due to the extreme expense.

Any error in a base length would cause a scale error in all measurements dependent on it. This is an example of a systematic error.

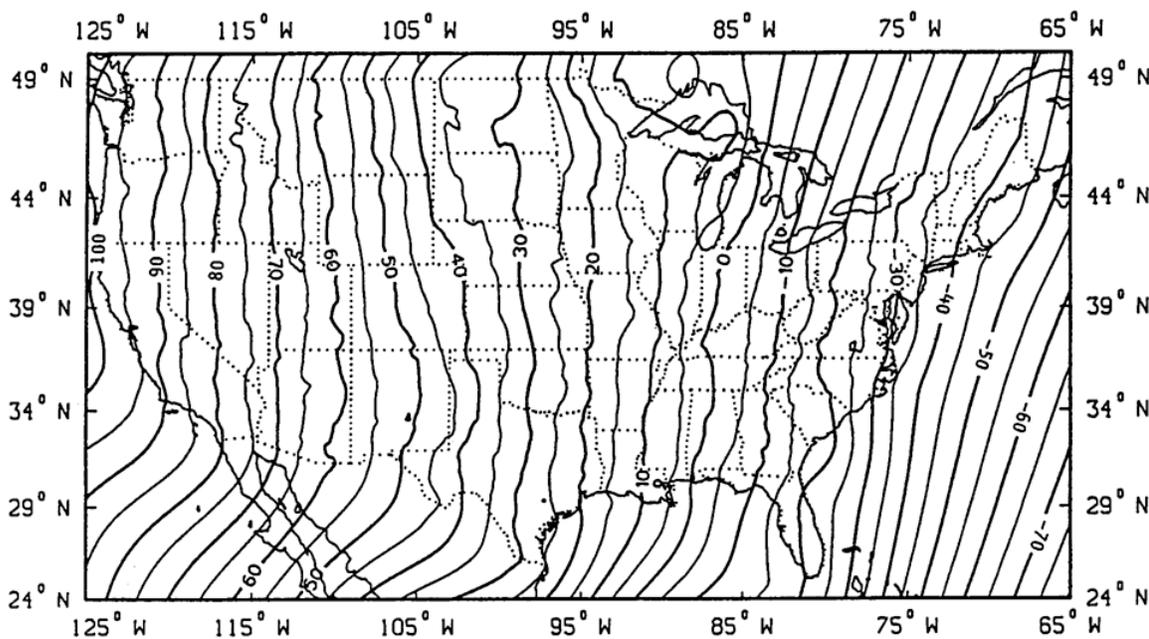
As a result there were often smoothly varying distortions that built up over distances. The resulting distortions often looked like a suspended sheet of rubber with weights at different points. This was very common in large area datums such as the North American Datum of 1927. Of course, the precise location and extent of the hills and valleys in this sheet were unknowable until a more accurate survey was done. For NAD27 this occurred with WGS72, the first extensive satellite based datum.

The North American Datum of 1983, NAD83, was the first large area civilian system based mainly on satellite surveying. This datum was significantly better than the NAD27. The distortions in the old system show up in the **contour plots** of the differences between the latitude and longitude in the two datums. It is clear that the shifts are not constant, but vary systematically from place to place. The latitude shifts have a significant north-south gradient with bends in the west over the Rocky Mountains. Differences vary from -20 to +50 m.



NAD27 to NAD83 Latitude Shift in Meters

The differences in the longitude have a predominate east-west gradient. The values vary from -40 m on the east coast to +100 m on the west coast. It is clear that over a small area map such as a 7.5' quad, the shift will be almost constant. But for the entire US there are significant, systematic variations.



NAD27 to NAD83 Longitude Shift in Meters

VI. What Datum Am I On

Because coordinates in different datums can differ by 100's of meters (or even a kilometer in the far east), it is important to know what datum you are on. There are usually two questions:

What datum is my map or **database** on?

What datum is my navigation system or survey on?

Clearly if the answer is the same, the map/data base and the position sensor can work together. Ships have gone aground when this was not so.

A. Map Datums - Paper and Electronic

The database of a map is usually listed in the **legend of the map**. In fact there are usually both a horizontal and vertical datum listed. Today it is not uncommon to see two horizontal datums listed, one for the original map and one for some overprinting. This is how USGS has updated a lot of topographical maps from NAD27 to NAD83. But you have to read the legend carefully to notice this. In fact there are quite a few military maps issued by NIMA that use the same technique. The maps of the Balkans distributed in the mid 1990's were on the European Datum of 1950 with annotations in the legends on how to shift the positions to WGS 84.

For computer data files the issue is much more difficult. The data from the legends of the maps is usually preserved, but often not displayed. If the map is simply scanned as an image the legend is there, but the data is not usually "**registered**" or set up for computer reading of accurate coordinates. If the map has been entered into a **Geographical Information System (GIS)** it may well be registered, but in this case the legend data is present only in an auxiliary file. This type of legend data is called "**metadata**".

B. Navigation and Survey Equipment Datums

What datum is GPS on? The answer depends on how the GPS receiver generated the solution. **Stand alone GPS** and **Differential GPS (DGPS)** have different answers. And there are both a general answer, and a more precise answer.

1. Standalone GPS Users

For the stand alone user the simple answer is WGS 84. GPS operates by measuring ranges from the satellites to the user. In order to convert these ranges into positions, the locations of the satellites are needed. Having a range and not knowing where it is from is not useful. The time history of the satellite positions is called the **ephemeris** of the satellites. So a stand-alone user is on the datum of the ephemeris he uses.

In the most common case the navigator or surveyor uses the ephemeris that is present on the signal broadcast by the satellites. This is called the **Broadcast Ephemeris (BCE)**. This information is "on WGS84". The quotes denote that this answer is not precisely correct, or is complicated by history.

There have been **Precise Ephemeris (PE's)** available for over 20 years. These are post-fit ephemeris based on a large set of ground stations. These are available from several civilian sources. There is even a consortium that produces a blended set of several PE's. In the past these were used for post fit work because it was a week or more after the fact that the PE's became available. After 2000, they have become available at very short delays. There even is a rapid prediction service that generates projected PE's that are much better than BCE out to a day or so. PE's are usually on the latest International Terrestrial Reference System - ITRF2000 in 2002. If you use these, you are on the ITRF 2000 datum.

The more complex answer for the BCE's puts these also on an ITRF. An ephemeris is computed from GPS observations made at known, fixed locations. The datum of an ephemeris is determined by the coordinates used for the antennas of these observations. These antenna locations for the **Operational Control System (OCS)** stations have been adjusted several times. This has effectively changed the datum of WGS84 as realized by the GSP BCE's.

In order to avoid confusion, the name of the datum was not changed when the adjustments were made. A suffix was added, but not widely used or know outside the geodesy community. The BCE datums were/are:

BCE Datum Name	Implemented	Matches
WGS 84	1980	Original WGS84
WGS 84 (G730)	1-2-1994	ITRF94
WGS 84 (G873)	9-27-1996	ITRF96
WGS 84 (G1150)	1- 2002	ITRF2000

GPS time is counted in weeks after January 1980. The number in the Gnnn is the GPS week number of the change. The first change moved coordinates about a meter. The last three changes have been much smaller. For the general navigator these changes may not be significant. There are some precision applications where these differences are important.

2. Differential GPS Users

With DGPS a reference station at a known location measures the errors in ranges to individual satellites and sends these to remote users over some communication link. In order to compute the errors, the reference station needs to know its position. The datum of the coordinates used for this position becomes the datum of all DGPS solutions that use those corrections. If you use DGPS you are on the datum used by the reference station.

3. Other Electronic Navigation Systems

Other electronic navigation systems are much like DGPS. The coordinates used for the stations that transmit the signals or transpond signals define the datum. Often these are not accurate enough for meter level distinctions to be important. However in Asia, the difference between WGS 84 and other local datums can be large. The user must take care to know which datum his navigation system and maps are on.

4. Surveying

Surveying is much like DGPS. Both **classical surveying** and **satellite surveying** is usually done point to point. In a surveyed network there must be at least one known point. In large surveys there will be more. These know points are "held constant" in the analysis of the data. The datum used for the point(s) held constant defines the datum of the other surveyed points.

This is true of almost all GPS surveying as well as classical survey techniques. Data is taken at the same time with GPS receivers at different locations. The relative locations are then computed.

There is one exception, **Absolute GPS Surveying**. Large government agencies have to establish the primary points in remote areas. In this case a more complex post analysis is done on GPS data taken at a fixed location over several days. The analysis needs Precise Ephemeris to achieve survey quality positions. The answer is on the datum of the PE's used.