Introduction to Datums

James R. Clynch
February 2006

I. What Are Datums – in Geodesy and Mapping?

A datum is the traditional answer to the practical problem of making an accurate map. If you do not have a very accurate location on a benchmark in your area, you make a map by choosing a mark and defining it as the reference point. You then survey outward and make a map. It may not match other's maps from adjacent areas at high accuracy because you are on a different datum, the one you created.

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, define the location of points measured from it. If the reference point is in error by 100 m north, so will all the points using this reference mark. 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.

Why is this important. Will if you use a navigation system (like GPS) that is not set to the map datum, you can be off by 100m (usually) to a kilometer (sometimes). GPS receivers are inherently on the WGS84 datum, but can be set to display locations in several other datums.

A datum can be defined by specifying the ellipsoid, the coordinates of a single point and the direction north. The point ties down the ellipsoid to the physical earth and also implicitly defines the placement of the center of the earth. This location is called the primary reference point. For North America, it is in central Kansas at a place called Meade's Ranch.

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 . You don't have to start all surveys in North America in Kansas.

This means that datums are the reference frames used in the construction of maps.

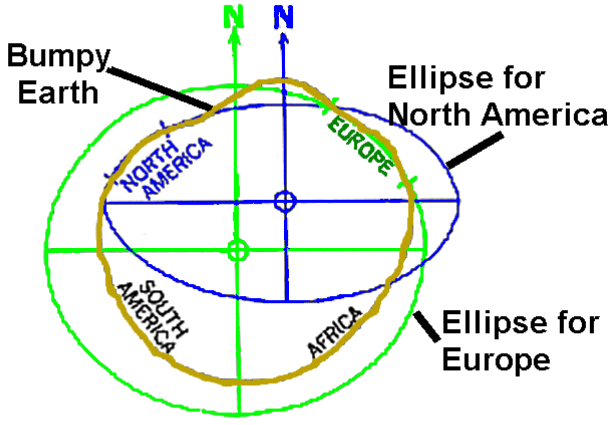
Things are complicated in practice as the same area may have two datums giving each point two different coordinates. In addition datums were often generated by individual countries, and did not match at the boundaries.

While there are about 20 common ellipsoids, most ellipsoids are used in multiple datums. For example the number of datums on some popular ellipsoids are:

Ellipsoid	**Number of Common Datums**
Airy 1830	2
Bessel 1841	7
Clarke 1866	9
Clarke 1880	26
International 1924	47

This understates the true complexity and confusion. For example the North American Datum of 1927 (NAD 27) is listed once for Clarke 1866 in the above table. But NAD 27 is really about 25 different datums. They are all on the Clarke 1866 ellipsoid, but have different primary reference points. The continental NAD 27 is different than that in the Bahamas or the one in Greenland or the one in Cuba ... This issue has not gone away with the newer NAD of 1983. There are different datums called NAD 83 in different areas.

The uses of different ellipsoids arises from the variations in the shape of the true earth and the difficulty of estimating the ellipsoid shape from a "limited" portion of the world, such as Europe of North America. The scientist in each region took the data they had and fit it to an ellipsoid.



Different Ellipsoids From Fitting Different Regions of Earth

This often resulted in different ellipsoid parameters (axis lengths) and center location. The above figure exaggerates the differences between the European datum ellipsoid and

that for North America, but is essentially correct. Often the origin shift is the largest difference in datum ellipsoids. This does not show up in many tables where only the semi-major axis and flattening are given. The origin is really implicitly defined by the coordinates assigned to the primary reference point.

In general if a datum covers areas not directly connected by a survey, such as over a large 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 set the GPS receiver to a version of "NAD83" that did not match the "NAD 83" on the chart.

II. Why and How Datums Happen

When people first began to map accurately using surveys in an area, they sometimes have the choice to extend some existing survey point set or beginning anew. When there is no land connection between the new and old areas, there was no choice until the advent of satellite surveying they had to begin a new datum. If the old surveys were very far away, a new datum was almost always begun.

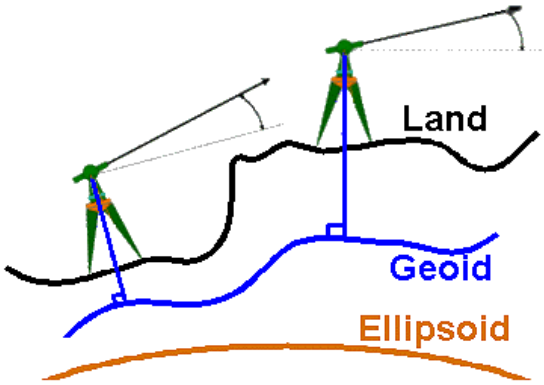
Most countries defined their own map reference system – or their datum. This means that at political boundaries the maps would not quite line up. Today there are regional (North American, European) datums as well as global ones. But maps still exist on different datums, sometimes even for the same area.

Relative Errors Dominate

Datums are realized through a set of physical points and the associated coordinates (latitude and longitude) for these “marks”. The relative position between a mark and adjacent or close marks is established using surveying techniques. All accurate surveying consists of making measurements between points – thus the measurements are relative, not absolute. (Celestial navigation gives absolute locations, but it is more inaccurate.)

Errors build up as a network of points is extended. The errors are largest where the area is mountainous. This occurs both because surveying is difficult there, but also because the mountains cause minor changes in the gravity field of the earth.

Transit Theodolite



Classical Surveying
Follows Geoid

The surveyor uses gravity to define up, and hence the horizontal. With small changes in gravity due mountains etc, this causes the survey to diverge from the model ellipsoid of the earth.

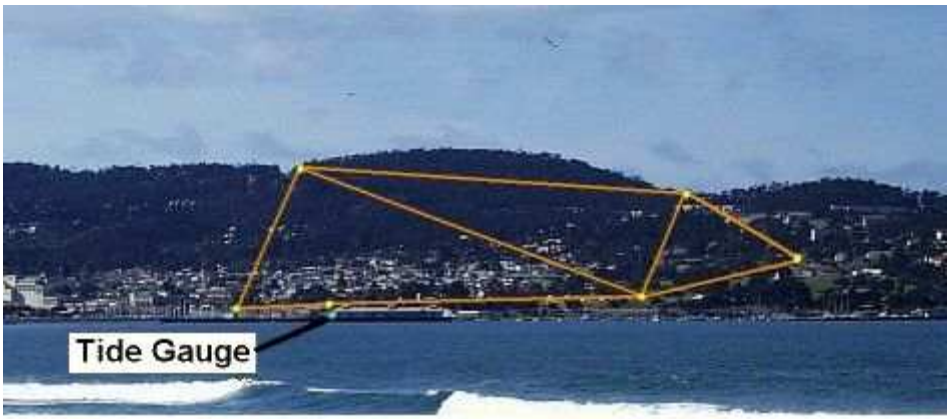
Datums and More Datums

Occasionally when new survey techniques come along the networks are re-surveyed. A new datum is thus produced. It usually consists of the same physical marks, but new coordinates. For example there was a major adjustment of the survey data in North America during the 1920's. These efforts lead to the publishing of the North American Datum of 1927. Satellite surveying and electronic distance measuring equipment pointed out the distortions in this datum. A new datum based on new measurement was published – the North American Datum of 1983. These are often called NAD 27 and NAD 83 respectively.

III. Example of Datum / Datums

A "datum" is defined in various ways, mathematically, for the surveyor, and practically. It can be thought of as a collection of surveyed points whose locations are accurately know with respect to each other. Here a hypothetical datum will be discussed to illustrate some of the common issues.

A set of control points that might be used in the Monterey, CA area are shown below on a picture.



A Local Datum's Control Points

The tide gauge and it's ground reference mark are real. Notice that there is a point very near the tide gauge. This is common to tie the tide measurements to the land measurements. Here is one of the reference markers for this tide gauge.

Benchmark “Tide” in Monterey CA

It is a bronze disk inserted into a bridge foundation. the name of the agency and the name/number of the "mark" are also given. (Here the mark shows National Ocean Survey, Mark 3450M, established in 1983)

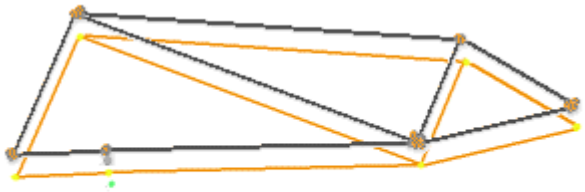
The above diagram inside the picture shows the points. The lines connecting the points are the things actually measured in the survey. The following diagram shows the same network without the background picture of Monterey.

A Local Datum's Control Points

Collections of control points like this are used to define the locations for mapping and are the user view of a datum. Geodesy books, however, usually begin with the modern mathematical definition and call this collection a "realization" of the datum.

Multiple Datums – Confusion and Error

However one collection, or datum, may not fit some other set (datum). It is in error with respect to the second collection of points. This occurs periodically as datums are updated with new information and survey techniques. It happens about twice a century in the US. In the last major datum change in the US, points typically "moved" by about 100 m (300 ft).



Control Points in Two Datums

Mismatched datums are also commonly happen when two sets of control points are separated by large distances or over an ocean. Accurate use of a map depends on knowing the datum it is on. On modern topographic maps the datum is often given in the map legend.

IV. WGS 84 / NAD 83 - A Standard Datum in Several Names

With the advent of satellite navigation systems, it was possible to construct a world wide datum. These are labeled Geodetic Systems. These were generated by the US Department of Defense beginning in 1966. They constructed World Geodetic System 1966 (WGS 66), which was not very good. It was followed by WGS 72 which was much better and then WGS 84. The Soviet Union also constructed a series of these, called SRS's with the last being SRS 90. It should be noted that these WGS's include information about the gravity field as well as an ellipsoid and a realization of the "datum".

WGS 84 is the current standard in the west. The US civilian survey organization, the National Geodetic Survey (NGS) generated a new datum for North America at the same time (NAD 83). It is essentially identical to WGS 84.

The world geodesy community has been generating these geodetic reference systems since the 1990's. These are labeled International Terrestrial Reference Systems (ITRF's). There have been ITRF 92, ITRY 96, ITRF 96, ITRF 98 etc. As these became refined, the changes became smaller. They were converging on the real world. In fact, the current versions include a model of the motion of the crustal plates. This is necessary to preserve the accuracy. (Monterey California, for example, is on the Pacific Plate and moving north west at 6 cm per year with respect to the North American Plate only a dozen kilometers inland).

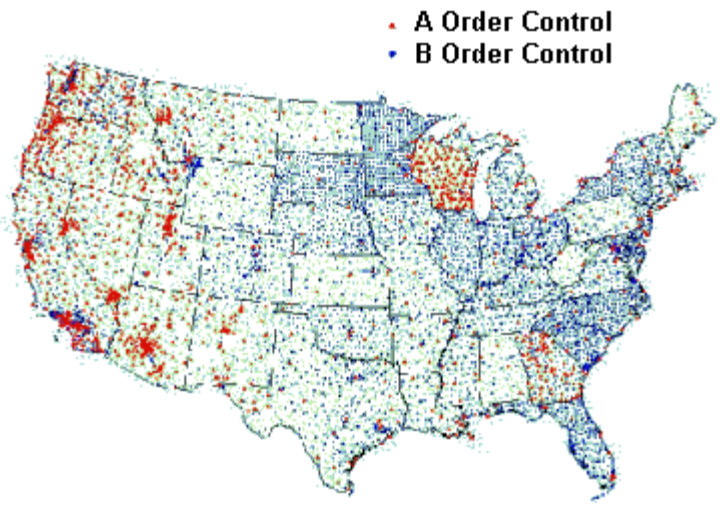
The US civilian agencies have adopted new a new datum in 2000. The US Department of Defense has essentially done the same thing, but has kept the name WGS 84. Rather than change the coordinates of all its know bench marks, it has adjusted the coordinates of the reference sites used to generate the orbits with the Global Positioning System. Thus the positions obtained using the over the air ephemeris and no Differential GPS corrections have changed. GPS counts time in weeks since 1980. The new datums are labeled

WGS 84 (G730) implemented in January 1994 (GPS Week 730),
WGS 84 (G873) implemented in September 1996, and
WGS 84 (G1150) implemented in October 2002.

These brought WGS into alignment with ITRF 94, ITRF 96 and ITRF 2000 at the few cm level.

The civilian agencies in the US have been using differential GPS survey techniques to generate a realization of a new network at better than one part per million (1 ppm or 1 cm over 10 km). They have achieved an order of magnitude better (0.1 ppm) in most areas. They have a goal of a High Accuracy Reference Network (HARN) survey marker within 200 km of anyplace in the US.

This contains most of the markers used to define NAD 83, in effect allowing a redefinition of that datum. A new NAD 2000 is being released. Here is the real primary control network of the US - the HARN network. This is the realization – how the datum is really used.



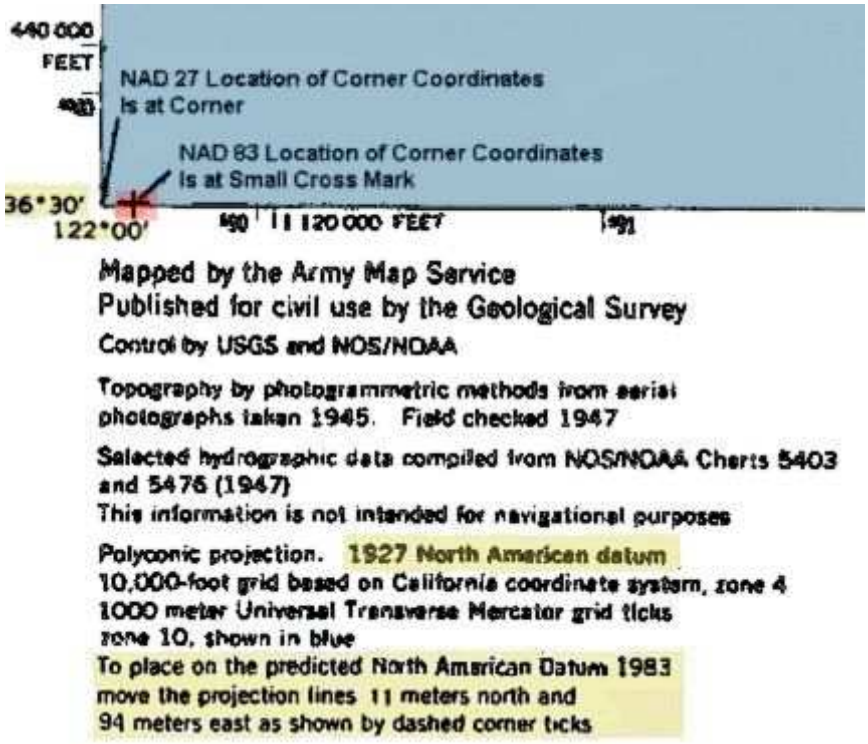
US Primary
Geodetic Control Network 1999

The official, high accuracy, coordinates of these points are on file with the US National Geodetic Survey (NGS). They can be retrieved from the NGS web site.

Note: The vertical networks have typically been separate from the horizontal network. The NAD 83 was followed by the North American Vertical Datum of 1988 (NAVD 88). Vertical networks are very closely tied to the gravitational model used in these systems.

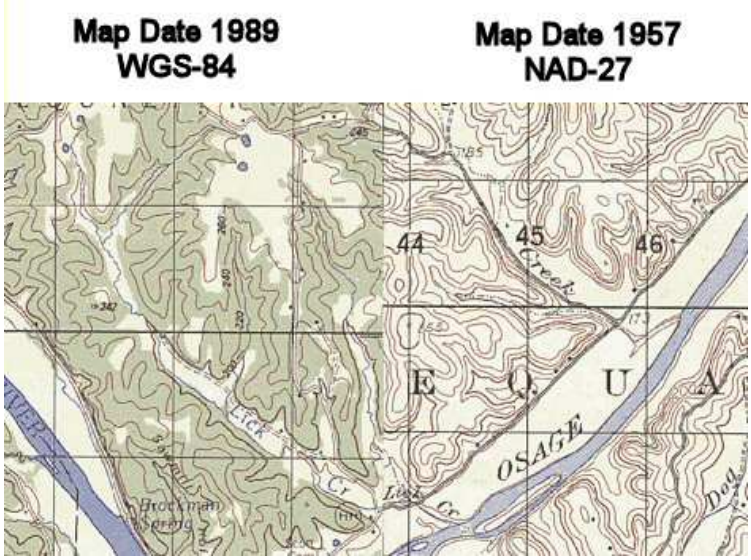
V. Map Legends and Datums

If you examine the legend of a topographic map, it will list the datum use to generate it. On US Geodetic Survey maps this is in the lower left corner. Here is the corner of the map covering Monterey, CA. Notice that it says that the map is on NAD 83, but ... look closely.



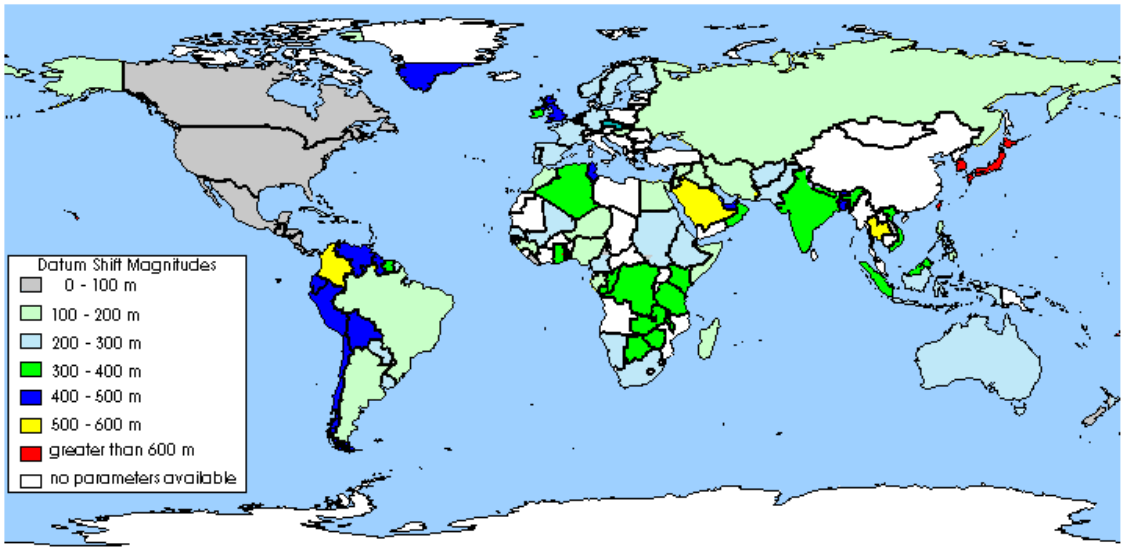
This is a example of an older map being updated using an overprinting. Many US topographic maps are the same as the NAD 27 maps with a notation in the legend on how to adjust the coordinates to NAD 83. There is also often a small cross in the corner giving the location of the corner coordinates in the new datum.

The different coordinates for the same point in different datums can often be significant. For example the difference between NAD 27 and NAD 83 can be over 100 meters (300 ft) as it is here. The following shows two small area (7.5 minute series) topographic maps from central Missouri. One is on the newer NAD83, the other still on NAD27. The maps have been cut and aligned with by the features. Notice that the coordinates are now off. The difference is about 20 m in the North/South and 210 m East/West.



Map Segments in Two Datums Showing Latitude Offset

Datum shifts can be very large. In one case, that of the Tokyo Datum and the commonly used World Geodetic System 1984 (WGS84) the difference is measured in kilometers or miles.



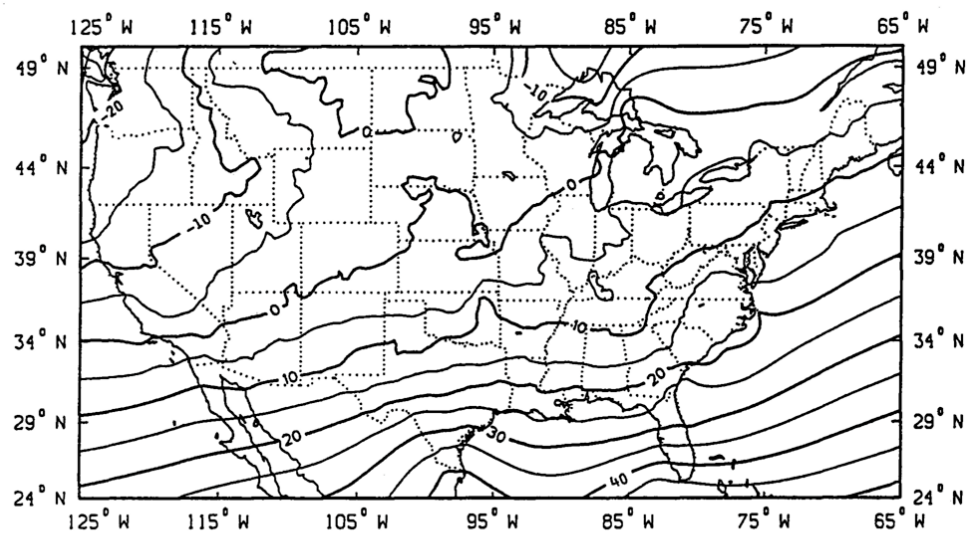
Horizontal Shifts on Common Maps to WGS 84

Because WGS 84 is so important today, a map of the difference between it and common local datums has been generated. The differences are color coded in bins to make it easier to use. WGS84 is the default datum for most GPS receivers. Clearly in some areas of the world, you would be way off using this setting with a locally generated map.

VI. Datums as Rubber Sheets

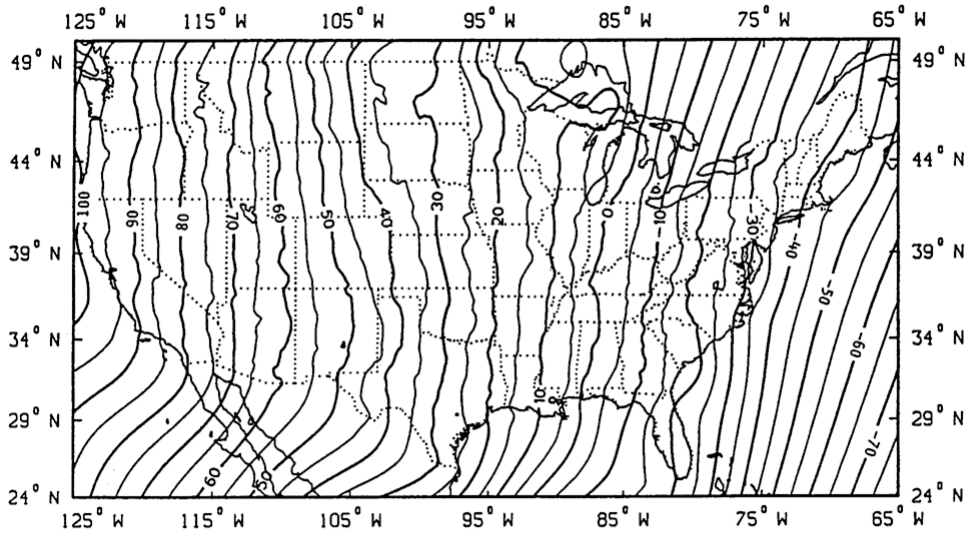
All older datums have significant distortions. These arise from the accumulation of survey errors. Over a small area, the difference between two datums is essentially an offset in north, east, and height. But this offset is not constant over large areas. For example, the NAD 27 datum has significant systematic distortions. The current standard for the US is NAD83. This is predominately based on satellite surveys, that are about two orders of magnitude better than the NAD27 surveys.

The errors in NAD27, when measured by comparing with NAD83 is a slowly varying function of location. The old NAD27 is like a rubber sheet with weights on it. The best way to display this distortion is to plot the differences as contour plots.



NAD27 to NAD83 Lattitude Shift in Meters

The latitude shift shows a progression from the northwest at about -20m to the southeast where it is 40 m. . The difference in the shift is 60 meters. The difference in the longitude is larger.



NAD27 to NAD83 Longitude Shift in Meters

The latitude difference is about –40 m on the east coast and increased as you move westward where it reaches 100 m on the west coast. There is a similar difference in the height, with the range being -15 to –40 m.

Clearly, using one offset for the US, or even one for the eastern US and one for the western US, leaves significant errors. Using one or two offsets is what most GPS receivers do when you ask them to display NAD 27 coordinates. The results are not adequate for navigation.