UNIT-III

(Part – 1)
SPATIAL REFERENCING AND POSITIONING

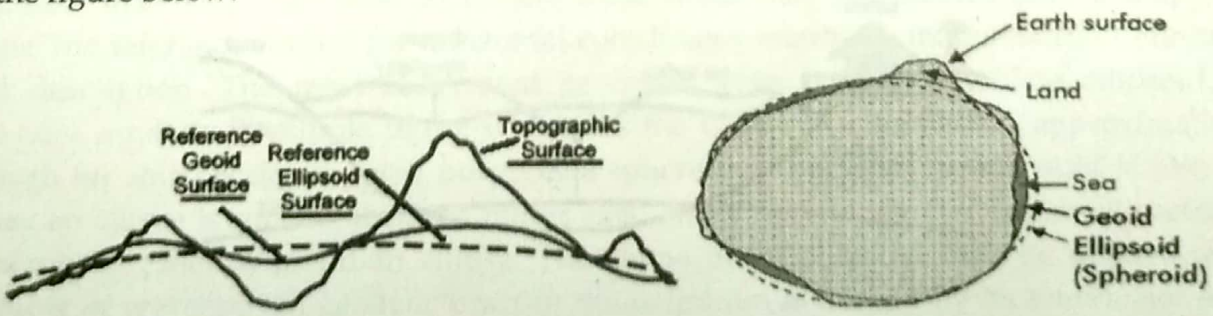
Early GIS mainly handles spatially referenced data from a single country, usually derived from paper maps published by the country's mapping organization. Nowadays, GIS users are combining spatial data from a given country with global spatial data sets, reconciling spatial data from published maps with coordinates established with satellite positioning techniques and integrating their spatial data with that from neighboring countries. To perform these kinds of tasks successfully, GIS users need to understand basic spatial referencing concepts.

4.1 SPATIAL REFERENCING

What makes GIS distinguished from other information system is their ability to combine spatially referenced data and to combine spatial data from different sources that use different spatial reference systems.

- **Reference Surfaces For Mapping**

The surface of the Earth is anything but uniform. The oceans can be treated as reasonably uniform, but the surface or topography of the land masses exhibits large vertical variations between mountains and valleys. These variations make it impossible to approximate the shape of the Earth with any reasonably simple mathematical model. Two main reference surfaces have been established to approximate the shape of the Earth. One reference surface is called the Geoid, the other reference surface is the ellipsoid as shown in the figure below.



The Geoid and the vertical datum

Imagine that the entire Earth's surface is covered by water. If ignored tidal and current effects on this 'global ocean', the resultant water surface is affected only by gravity. This has an effect on the shape of this surface because the direction of gravity– more commonly known as plumb line–is dependent on the mass distribution inside the Earth. Due to irregularities or mass anomalies in this distribution the 'global ocean' results in an undulated surface. This surface is called the Geoid. The plumb line through any surface point is always perpendicular to it.

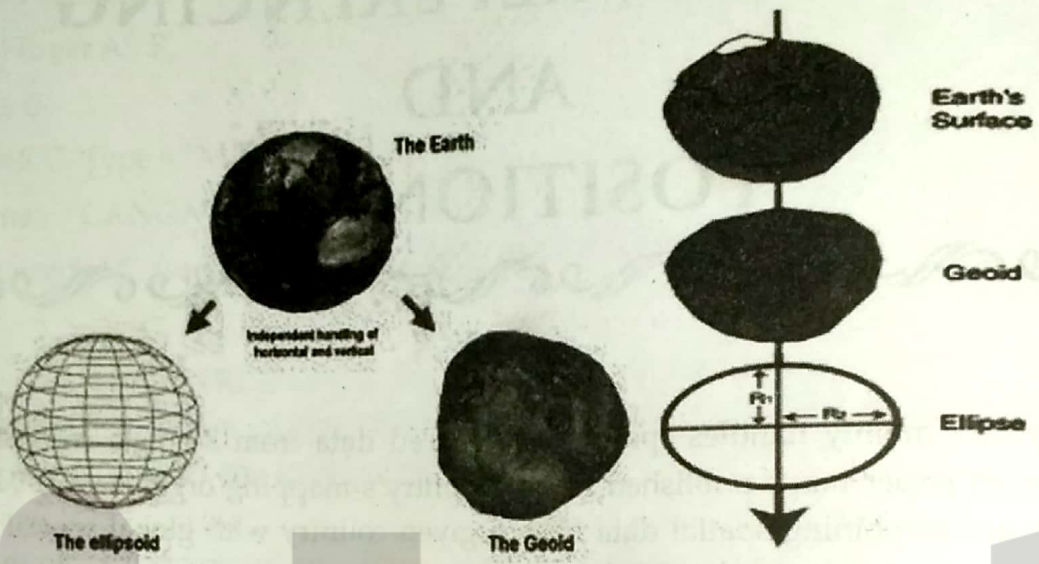
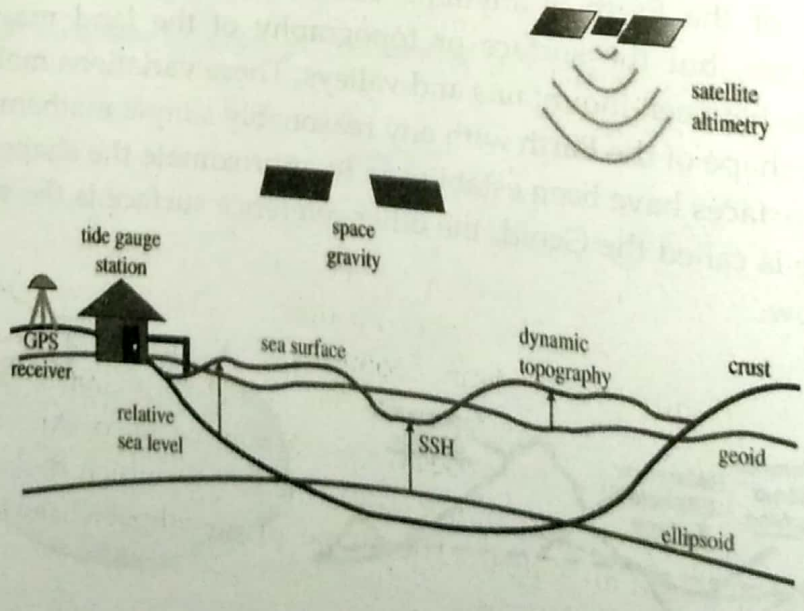


Figure : The Ellipsoid and Geoid reference model of the Earth surface.

The Geoid is used to describe heights. In order to establish the Geoid as reference for heights, the ocean's water level is registered at coastal places over several years using tide gauges (mareographs). Averaging the registrations largely eliminates variations of the sea level with time. The resulting water level represents an approximation to the Geoid and is called the mean sea level. The height of a point in Mumbai with respect to the tide gauge is measured using a technique known as geodetic levelling. The result of this process will be the height above local mean sea level for the Mumbai tidal gauge station point. The height determined with respect to a tide-gauge station is known as the orthometric height i.e. height H above the Geoid.



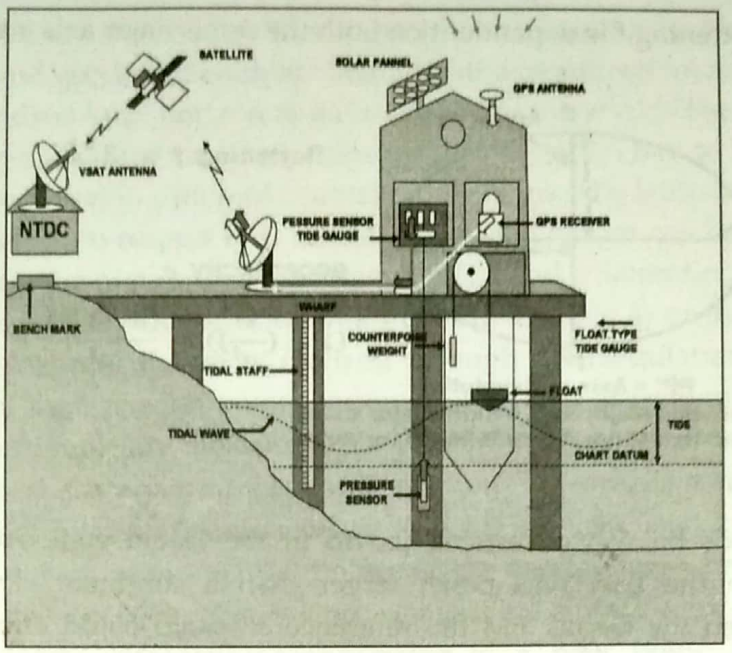


Figure : Measuring MSL using geodetic leveling

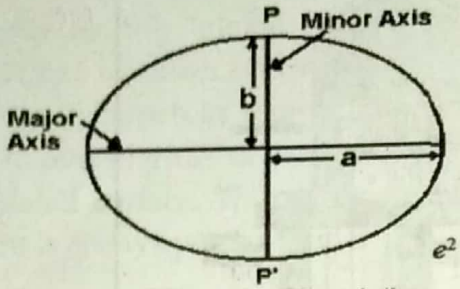
There are several realizations of local mean sea levels (also called local vertical datums) in the world. They are parallel to the Geoid but offset by up to a couple of meters. This offset is due to local phenomena such as ocean currents, tides, coastal winds, water temperature and salinity at the location of the tide gauge. Care must be taken when using heights from another local vertical datum. For example, this might be the case in the border area of adjacent nations. Even within a country, heights may differ depending on to which tide gauge, mean sea level point, they are related. As an example, the mean sea level from the Atlantic to the Pacific coast of the USA increases by 0.6 to 0.7 m. The tide gauge (zero height) of the Netherlands differs -2.34 meters from the tide gauge (zero height) of the neighboring country Belgium.

The local vertical datum is implemented through a leveling network. A leveling network consists of benchmarks, whose height above mean sea level has been determined through geodetic leveling. The implementation of the datum enables easy user access. The surveyors do not need to start from scratch (i.e. from the Amsterdam tide-gauge) every time they need to determine the height of a new point. They can use the benchmark of the leveling network that is closest to the point of interest.

The ellipsoid

The physical surface, called Geoid, is used as a reference surface for heights. Also a reference surface for the description of the horizontal coordinatesof points of interest is required. This will later used to project these horizontal coordinates onto a mapping plane, the reference surface for horizontal coordinates requires a mathematical definition and description. The most convenient geometric reference is the oblate ellipsoid. It provides a relatively simple figure which fits the Geoid to a first order approximation, though for small scale mapping purposes a sphere may be used. An ellipsoid is formed when an ellipse is rotated about its minor axis. This ellipse which defines an ellipsoid or spheroid is called a meridian ellipse. The shape of an ellipsoid may be defined in a number of ways, but in geodetic practice the definition is usually by its semi-major axis

and flattening, Flattening f is dependent on both the semi-major axis a and the semi-minor axis b.



PP' = Axis of Revolution
a = One-Half Major Axis = Semi-Major Axis
b = One-Half Minor Axis = Semi-Minor Axis

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

eccentricity e

$$e^2 = (1 - (\frac{b^2}{a^2})) = \frac{(a^2 - b^2)}{a^2} = 2f - f^2$$

Local ellipsoids have been established to fit the Geoid well over an area of local interest, which in the past was never larger than a continent. This meant that the differences between the Geoid and the reference ellipsoid could effectively be ignored, allowing accurate maps to be drawn in the vicinity of the datum. Global reference ellipsoids is contrast to local ellipsoids, which apply only to a specific country or localized area of the Earth's surface, global ellipsoids approximate the Geoid as a mean earth ellipsoid.

The International Union for Geodesy and Geophysics (IUGG) plays a central role in establishing these reference figures. In 1924, the IUGG in Madrid introduced the ellipsoid determined by Hayford in 1909 as the international ellipsoid. In 1967 of the IUGG in Luzern, the 1924 reference system was replaced by the Geodetic Reference System 1967 (GRS 1967). It represents a good approximation to the mean Earth figure.

For some time, the Geodetic Reference System 1967 was used in the planning of new geodetic surveys. For example, the Australian Datum (1966) and the South American datum (1969) are based upon this ellipsoid. However, in 1979 in Canberra the IUGG recognized that the GRS 1967 no longer represented the size and shape of the Earth to an adequate accuracy. Consequently, it was replaced by the Geodetic Reference System 1980 (GRS80) ellipsoid and WGS 84 afterward.

Ellipsoid	Semi-major axis	1/flattening
Airy 1830,	6377563.396	299.3249646
Bessel 1841	6377397.155	299.1528128
Clarke 1880,	6378249.145	293.465
International 1924	6378338	297
GRS 80	6378137	298.257222101
WGS	6378137	298.257223563

The local horizontal datum

Ellipsoids have varying position and orientations. An ellipsoid is positioned and oriented with respect to the local mean sea level by adopting a latitude (ϕ) and longitude (λ) and ellipsoidal height (h) of a so-called fundamental point and an azimuth to an

additional point. We say that this defines a local horizontal datum. Notice that the term horizontal datum and geodetic datum are being treated as equivalent and interchangeable words. Several hundred local horizontal datums exist in the world. The reason is obvious: Different local ellipsoids with varying position and orientation had to be adopted to best fit the local mean sea level in different countries or regions. The latitude and longitude (ϕ, λ) of any other point with respect to this local horizontal datum can be determined using geodetic positioning techniques, such as triangulation and trilateration. The result of this process will be the geographic (or horizontal) coordinates (ϕ, λ) of the new point in the Datum. A local horizontal datum is realized through a triangulation network. Such a network consists of monumental points forming a network of triangular mesh elements. The angles in each triangle are measured in addition to at least one side of a triangle; the fundamental point is also a point in the triangulation network. The angle measurements and the adopted coordinates of the fundamental point are then used to derive geographic coordinates (ϕ, λ) for all monumented points of the triangulation network. Within this framework, users do not need to start from scratch (i.e. from the fundamental point) in order to determine the geographic coordinates of a new point. They can use the monument of the triangulation network that is closest to the new point. The extension and re-measurement of the network is nowadays done through satellite measurements.

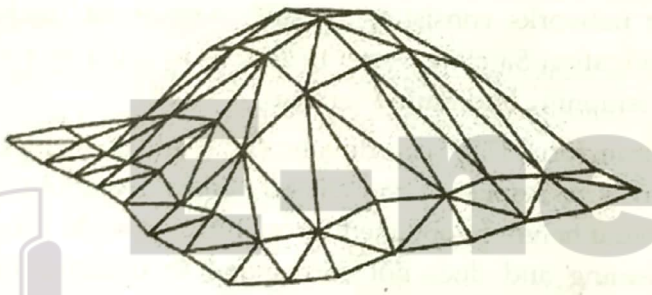


Figure : Triangulation Network

The global horizontal datum

Local horizontal datums have been established to fit the Geoid well over the area of local interest, which in the past was never larger than a continent. With increasing demands for global surveying activities are underway to establish global reference surfaces. The objective is to make geodetic results mutually comparable and to provide coherent results also to other disciplines like astronomy and geophysics.

The most important global (geocentric) spatial reference system for the GIS community is the International Terrestrial Reference System (ITRS). It is a three dimensional coordinate system with a well-defined origin (the centre of mass of the Earth) and three orthogonal coordinate axes (X, Y, Z). The Z-axis points towards a mean Earth north pole. The X-axis is oriented towards a mean Greenwich meridian and is orthogonal to the Z-axis. The Y -axis completes the right-handed reference coordinate system. The ITRS is realized through the International Terrestrial Reference Frame (ITRF), a distributed set of ground control stations that measure their position continuously using GPS . Constant re-measuring is needed because of the involvement of new control stations and ongoing geophysical processes that deform the Earth's crust at measurable global, regional and local scales. These deformations cause positional differences in time, and have resulted in more than one realization of the ITRS. Examples are the ITRF96 or the ITRF2000. The ITRF96 was established at the 1st of January, 1997. This means that the

measurements use data up to 1996 to fix the geocentric coordinates (X, Y and Z in meters) and velocities (positional change in X, Y and Z in meters per year) at the different stations. The velocities are used to propagate the measurements to other epochs (times). The trend is to use the ITRF everywhere in the world for reasons of global compatibility.

GPS uses the World Geodetic System 1984 (WGS84) as its reference system. It has been refined on several occasions and is now aligned with the ITRF to within a few centimeters worldwide. Global horizontal datums, such as the ITRF2000 or WGS84, are also called geocentric datums because they are geocentrically positioned with respect to the centre of mass of the Earth. They became available only recently (roughly after the 1960's), with advances in extra-terrestrial positioning techniques.

Since the size and shape of satellite orbits is directly related to the centre of mass of the Earth, observations of natural or artificial satellites can be used to pinpoint the centre of mass of the Earth, and hence the origin of the ITRS. This technique can also be used for the realization of the global ellipsoids and datums at the accuracy level required for large-scale mapping. To implement the ITRF in a region, a densification of control stations is needed to ensure that there are enough coordinated reference points available in the region. These control stations are equipped with permanently operating satellite positioning equipment (i.e. GPS receivers and auxiliary equipment) and communication links. Examples for networks consisting of such permanent tracking stations are the Indian Regional Navigation Satellite System (IRNSS) in India is a system that provides accurate real-time positioning and timing services.

We can easily transform ITRF coordinates (X, Y and Z in meters) into geographic coordinates (ϕ, λ, h) with respect to the GRS80 ellipsoid without the loss of accuracy. However, the ellipsoidal height h, obtained through this straight forward transformation, has no physical meaning and does not correspond to intuitive human perception of height, therefore use the height H, above the Geoid. If all published maps are also globally referenced by that time, the underlying spatial referencing concepts will become transparent and hence redundant for GIS users.

Hundreds of existing local horizontal and vertical datums are still used because they form the basis of map products all over the world. Tools to transform coordinates from local horizontal datums to a global horizontal datum and vice versa are required. The organizations that usually develop transformation tools and make them available to the user community are provincial or National Mapping Organizations (NMOs) and cadastral authorities.

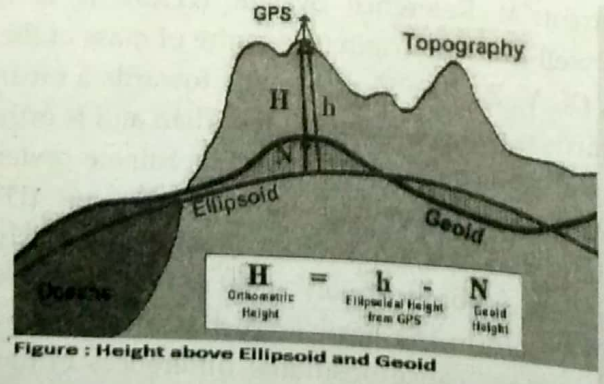
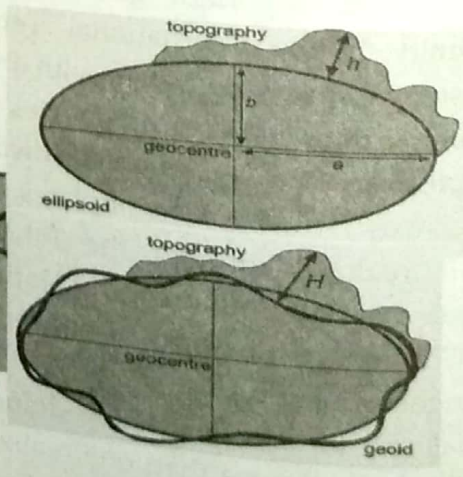


Figure : Height above Ellipsoid and Geoid



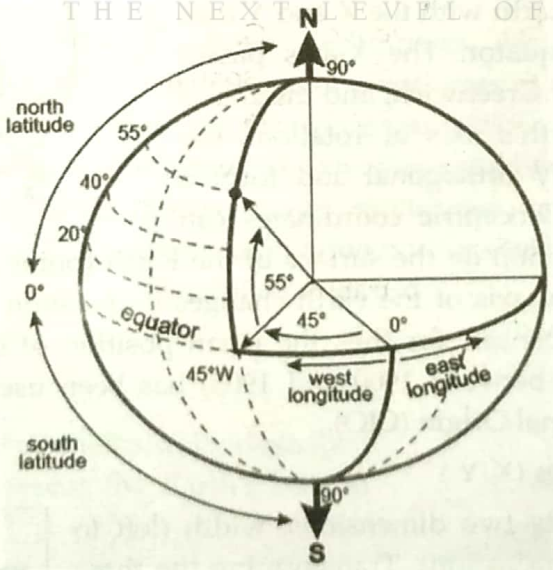
- **Coordinate systems**

Different kinds of coordinate systems are used to position data in space. Spatial (or global) coordinate systems are used to locate data either on the Earth's surface in a 3D space, or on the Earth's reference surface in a 2D space. The geographic coordinate system in 2D and 3D space and the geocentric coordinate system, also known as the 3D Cartesian coordinate system. Planar coordinate systems on the other hand are used to locate data on the flat surface of the map in a 2D space.

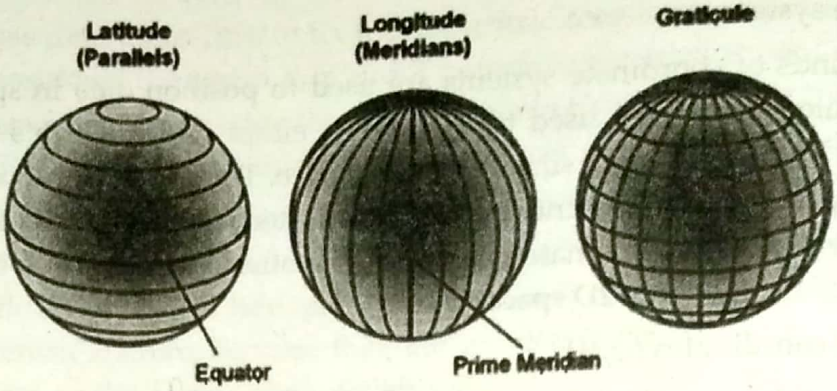
2D Geographic coordinates (ϕ, λ)

The most widely used global coordinate system consists of lines of geographic latitude (phi or ϕ or φ) and longitude (lambda or λ). Lines of equal latitude are called parallels. They form circles on the surface of the ellipsoid4. Lines of equal longitude are called meridians and they form ellipses (meridian ellipses) on the ellipsoid. The latitude (ϕ) of a point P is the angle between the ellipsoidal normal through P ′ and the equatorial plane. Latitude is zero on the equator (ϕ = 0◦), and increases towards the two poles to maximum values of ϕ = +90◦ (N 90◦) at the North Pole and ϕ = -90◦ (S 90◦) at the South Pole. The longitude (λ) is the angle between the meridian ellipse which passes through Greenwich and the meridian ellipse containing the point in question. It is measured in the equatorial plane from the meridian of Greenwich (λ = 0◦) either eastwards through λ = + 180◦ (E 180◦) or westwards through λ = -180◦ (W 180◦).

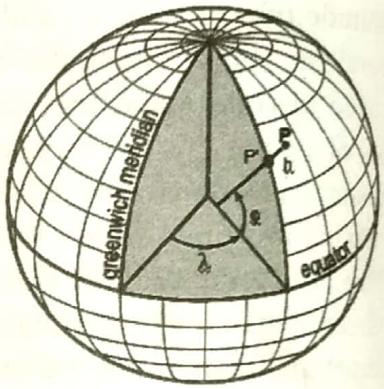
Latitude and longitude represent the geographic coordinates (ϕ, λ) of a point P ′ (Figure 4.10) with respect to the selected reference surface. They are always given in angular units. For example, the coordinates for University of Mumbai in Mumbai are, value of ϕ is 19.073212° and λ is 72.854195°.



The graticule on a map represents the projected position of the geographic coordinates (ϕ, λ) at constant intervals, or in other words the projected position Graticule of selected meridians and parallels. The shape of the graticule depends largely on the characteristics of the map projection and the scale of the map.

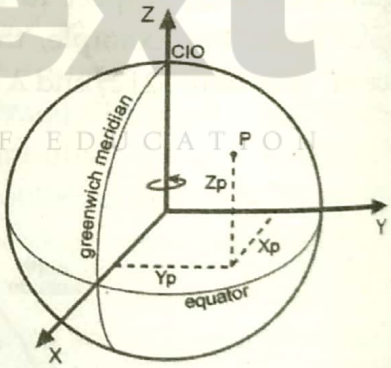


3D Geographic coordinates (ϕ, λ, h)



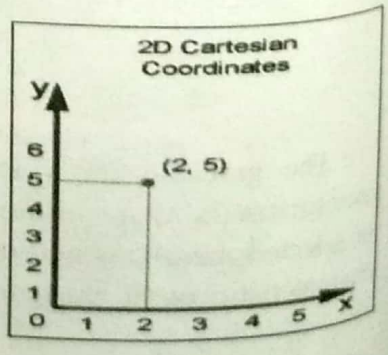
3D geographic coordinates (ϕ, λ, h) are obtained by introducing the ellipsoidal height h to the system. The ellipsoidal height (h) of a point is the vertical distance of the point in question above the ellipsoid. It is measured in distance units along the ellipsoidal normal from the point to the ellipsoid surface. 3D geographic coordinates can be used to define a position on the surface of the Earth (point P).

3D Geocentric coordinates (X, Y, Z)



An alternative method of defining a 3D position on the surface of the Earth is by means of geocentric coordinates (X, Y, Z), also known as 3D Cartesian coordinates. The system has its origin at the mass-centre of the Earth with the X and Y axes in the plane of the equator. The X-axis passes through the meridian of Greenwich, and the Z-axis coincides with the Earth's axis of rotation. The three axes are mutually orthogonal and form a right-handed system. Geocentric coordinates can be used to define a position on the surface of the Earth (point P in Figure). It should be noted that the rotational axis of the earth changes its position over time (referred to as polar motion). To compensate for this, the mean position of the pole in the year 1903 (based on observations between 1900 and 1905) has been used to define the so-called Conventional International Origin (CIO).

2D Cartesian coordinates (X, Y)



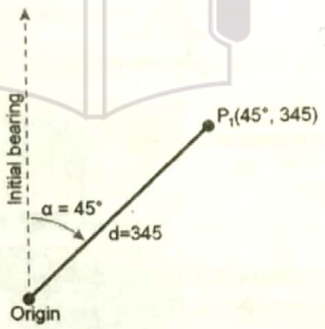
A flat map has only two dimensions: width (left to right) and length (bottom to top). Transforming the three dimensional Earth into a two-dimensional map is subject of map projections and coordinate transformation. Like in several other cartographic applications, two-dimensional Cartesian coordinates (x, y), also known as planar rectangular coordinates, are used to describe the location of any point unambiguously. The two coordinates x and y for point P, specify any location P on the map.

The 2D Cartesian coordinate system is a system of intersecting perpendicular lines, which contains two principal axes, called the X-axis and Y -axis. The horizontal axis is usually referred to as the X-axis and the vertical the Y –axis, the X-axis is also sometimes called Easting and the Y -axis the Northing. The intersection of the X and Y -axis forms the origin. The plane is marked at intervals by equally spaced coordinate lines, called the map grid. The coordinates x=0 and y=0 are given to the origin. However, sometimes large positive values are added to the origin coordinates. This is to avoid negative values for the x and y coordinates in case the origin of the coordinate system is located inside the area of interest. The point which then has the coordinates x=0 and y=0 is called the false origin.

The grid on a map represents lines having constant 2D Cartesian coordinates. It is almost always a rectangular system and is used on large and medium scale maps to enable detailed calculations and positioning. The map grid is usually not used on small scale maps. Scale distortions that result from transforming the Earth's curved surface to the map plane are so great on small-scale maps that detailed calculations and positioning are difficult.

2D Polar coordinates (α, d)

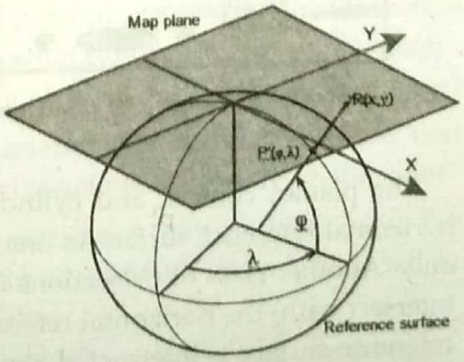
Polar coordinate is the distance "d" from the origin to the point concerned and the angle α between a fixed (or zero) direction and the direction to the point. The angle α is called azimuth or bearing and is measured in a clockwise direction. It is given in angular units while the distance d is expressed in length units. Bearings are always related to a fixed direction (initial bearing) or a datum line. In principle, this reference line can be chosen freely. However, in practice three different directions are widely used: True North, Grid North and Magnetic North. The corresponding bearings are called: true (or geodetic) bearing, grid bearing and magnetic (or compass) bearing.



Polar coordinates are often used in land surveying. For some types of surveying instruments it is advantageous to make use of this coordinate system. The development of precise remote distance measurement techniques has led to the virtually universal preference for the polar coordinate method in detailed surveys.

- **Map projections**

A map projection is a mathematically described technique of how to represent the Earth's curved surface on a flat map.



To represent parts of the surface of the Earth on a flat paper map or on a computer screen, the curved horizontal reference surface must be mapped onto the 2D mapping plane. The reference surface for large-scale mapping is usually an oblate ellipsoid, and for small-scale mapping, a sphere. Mapping onto a 2D mapping plane means transforming each point on the reference surface

with geographic coordinates (ϕ, λ) to a set of Cartesian coordinates (x, y) representing positions on the map plane

The actual mapping cannot usually be visualized as a true geometric projection, directly onto the mapping plane. This is achieved through mapping equations. A forward mapping equation transforms the geographic coordinates (ϕ, λ) of a point on the curved reference surface to a set of planar Cartesian coordinates (x, y), representing the position of the same point on the map plane :

$$(x, y) = f(\phi, \lambda)$$

The corresponding inverse mapping equation transforms mathematically the planar Cartesian coordinates (x, y) of a point on the map plane to a set of geographic coordinates (ϕ, λ) on the curved reference surface:

$$(\phi, \lambda) = f(x, y)$$

Classification of map projections

A large number of map projections have been developed, each with its own specific qualities. These qualities in turn make resulting maps useful for certain purposes. By definition, any map projection is associated with scale distortions.

There is simply no way to flatten out a piece of ellipsoidal or spherical surface without stretching some parts of the surface more than others. The amount and which kind of distortions a map will have depends on the type of the map projection that has been selected. Some map projections can be visualized as true geometric projections directly onto the mapping plane, in which case we call it an azimuthal projection, or onto an intermediate surface, which is then rolled out into the mapping plane. Typical choices for such intermediate surfaces are cones and cylinders. Such map projections are then called conical, and cylindrical, respectively.

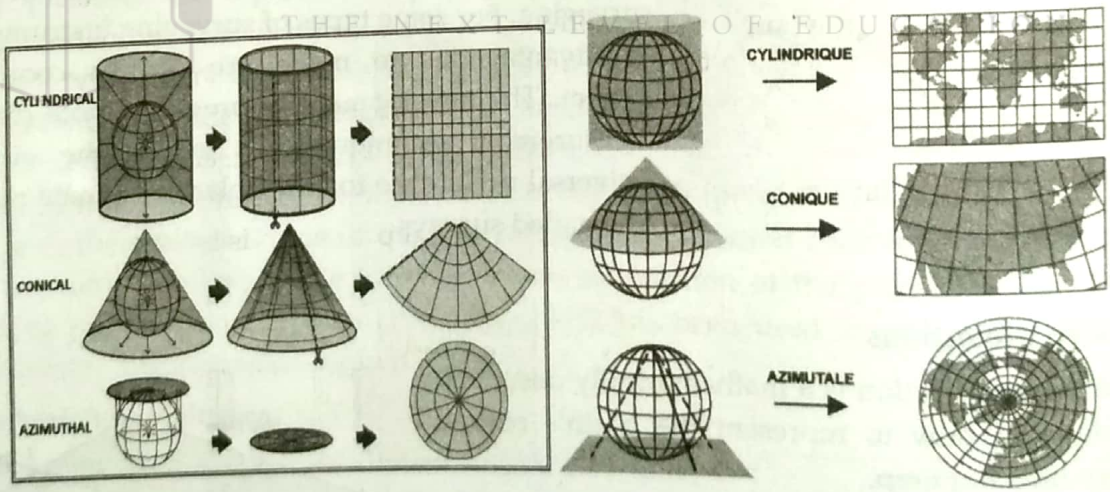


Figure : Types of Map Projection

The planar, conical, and cylindrical surfaces are all tangent surfaces; they touch the horizontal reference surface in one point (plane) or along a closed line (cone and cylinder) only. Another class of projections is obtained if the surfaces are chosen to be secant to (to intersect with) the horizontal reference surface; illustrations are in Figure below. Then, the reference surface is intersected along one closed line (plane) or two closed lines (cone and

cylinder). Secant map surfaces are used to reduce or average out scale errors because the line(s) of intersection are not distorted on the map.

In the geometric depiction of map projections, the symmetry axes of the plane, cone and cylinder coincide with the rotation axis of the ellipsoid or sphere, i.e. a line through N and S pole. In this case, the projection is said to be a normal projection. The other cases are transverse projections (symmetry axis in the equator) and oblique projections (symmetry axis is somewhere between the rotation axis and equator of the ellipsoid or sphere). These cases are illustrated in Figure below.

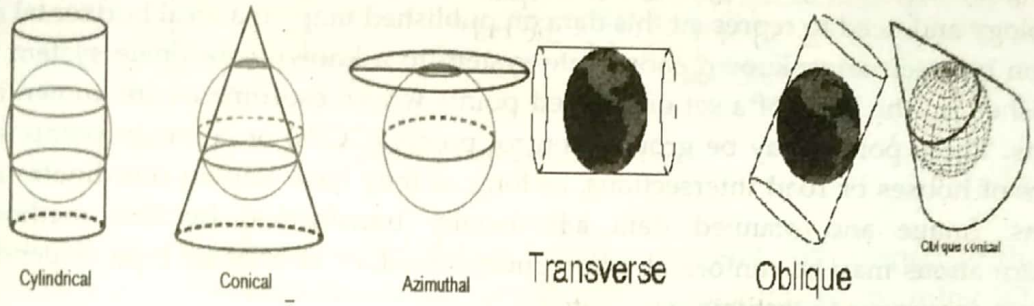


Figure : Cases of Map Projection

The Universal Transverse Mercator (UTM) uses a transverse cylinder, secant to the horizontal reference surface. UTM is an important projection used worldwide. The projection is a derivation from the Transverse Mercator projection (also known as Gauss-Kruger or Gauss conformal projection). The UTM divides the world into 60 narrow longitudinal zones of 6 degrees, numbered from 1 to 60. The narrow zones of 6 degrees (and the secant map surface) make the distortions small enough for large scale topographic mapping.

Cylindrical projections are typically used to map the world in its entirety. Conical projections are often used to map the different continents, while the normal azimuthal projection may be used to map the polar areas. Transverse and Oblique aspects of many projections can be used for most parts of the world. The general shape of the mapping area should match with the distortion pattern of a specific projection. If an area is approximately circular it is possible to create a map that minimizes distortion for that area on the basis of an azimuthal projection. The cylindrical projection is best for a rectangular area and a conic projection for a triangular area.

So far, we have not specified how the curved horizontal reference surface is projected onto the plane, cone or cylinder. How this is done determines which kind of distortions the map will have compared to the original curved reference surface. The distortion properties of a map are typically classified according to what is not distorted on the map:

- In a conformal map projection the angles between lines in the map are identical to the angles between the original lines on the curved reference surface. This means that angles (with short sides) and shapes (of small areas) are shown correctly on the map.
- In an equal–area (equivalent) map projection the areas in the map are identical to the areas on the curved reference surface (taking into account the map scale), which means that areas are represented correctly on the map.
- In an equidistant map projection the length of particular lines in the map are the same as the length of the original lines on the curved reference surface .

- **Coordinate transformations**

Map and GIS users are mostly confronted in their work with transformations from one two-dimensional coordinate system to another. This includes the transformation of polar coordinates delivered by the surveyor into Cartesian map coordinates or the transformation from one 2D Cartesian (x, y) system of a specific map projection into another 2D Cartesian (x, y) system of a defined map projection. Datum transformations are transformations from a 3D coordinate system (i.e. horizontal datum) into another 3D coordinate system. These kinds of transformations are also important for map and GIS users. They are usually collecting spatial data in the field using satellite navigation technology and need to represent this data on published map on a local horizontal datum. Relation between an unknown coordinate system to a known coordinate system can be established on the basis of a set of selected points whose coordinates are known in both systems. These points may be ground control points (GCPs) or common points such as corners of houses or road intersections, as long as they have known coordinates in both systems. Image and scanned data are usually transformed by this method. The transformations may be conformal, affine, polynomial, or of another type, depending on the geometric errors in the data set.

2D Polar to 2D Cartesian transformations

The transformation of polar coordinates (α, d), into Cartesian map coordinates (x, y) is done when field measurements, angular and distance measurements are transformed into map coordinates. The equation for this transformation is:

$x = d(\sin(a))$

$y = d(\cos(a))$

The inverse equation is: $a = \tan^{-1}\frac{X}{Y}$

$d^2 = x^2 + y^2$

A more realistic case makes use of a translation and a rotation to transform one system to the other.

Changing map projection

Forward and inverse mapping equations are normally used to transform data from one map projection to another. The inverse equation of the source projection is used first to transform source projection coordinates (x, y) to geographic coordinates (ϕ, λ). Next, the forward equation of the target projection is used to transform the geographic coordinates (ϕ, λ) into target projection coordinates (x , y). The first equation takes us from a projection A into geographic coordinates. The second takes us from geographic coordinates (ϕ, λ) to another map projection B. These principles are illustrated in Figure below.

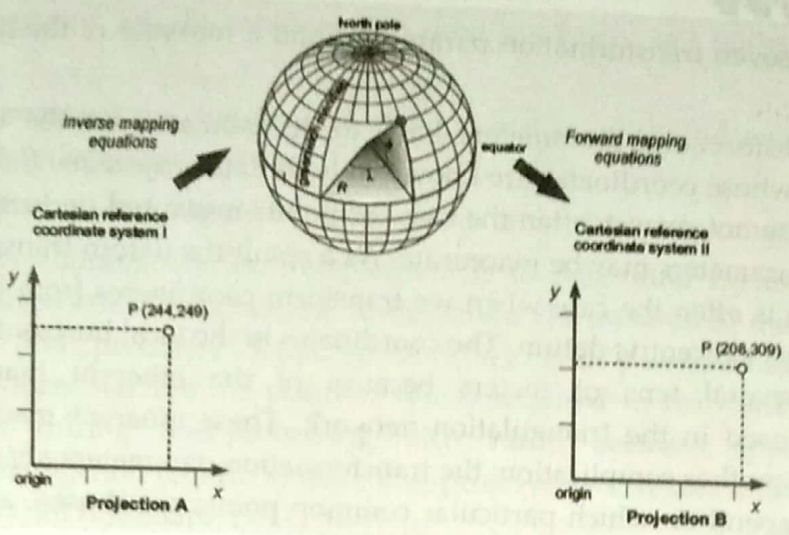


Figure : Map Transformation

Historically, a GIS has handled data referenced spatially with respect to the (x, y) coordinates of a specific map projection. For GIS application domains requiring 3D spatial referencing, a height coordinate may be added to the (x, y) coordinate of the point. The additional height coordinate can be a height H above mean sea level, which is a height with a physical meaning. These (x, y, H) coordinates can be used to represent the location of objects in a 3D GIS.

Datum transformations

A change of map projection may also include a change of the horizontal datum. This is the case when the source projection is based upon a different horizontal datum than the target projection. If the difference in horizontal datums is ignored, there will not be a perfect match between adjacent maps of neighboring countries or between overlaid maps originating from different projections. It may result in up to several hundred meters difference in the resulting coordinates. Therefore, spatial data with different underlying horizontal datums may need a so-called datum transformation.

The inverse equation of projection A is used first to take us from the map coordinates (x, y) of projection A to the geographic coordinates (ϕ, λ, h) in datum A. A height coordinate (h or H) may be added to the (x, y) map coordinates. Next, the datum transformation takes us from these coordinates to the geographic co-ordinates (ϕ, λ, h) in datum B. Finally, the forward equation of projection B is used to take us from the geographic coordinates (ϕ, λ, h) in datum B to the map coordinates (x , y) of projection B.

Mathematically a datum transformation is feasible via the geocentric coordinates (x, y, z), or directly by relating the geographic coordinates of both datum systems. The latter relates the ellipsoidal latitude (ϕ) and longitude (λ), and possibly also the ellipsoidal height (h), of both datum systems. We can easily transform geographic coordinates (ϕ, λ, h) into geocentric coordinates (x, y, z), and the other way around. The datum transformation via the geocentric coordinates implies a 3D similarity transformation. Essentially, this is a transformation between two orthogonal 3D Cartesian spatial reference frames together with some elementary tools from adjustment theory. The transformation is usually expressed with seven parameters: three rotation angles (α, β, γ), three origin shifts (X_0, Y_0, Z_0) and one scale factor (s). The input in the process are coordinates of points in datum A and coordinates of the same points in datum B. The output is an

estimate of the seven transformation parameters and a measure of the likely error of the estimate.

Datum transformation parameters have to be estimated on the basis of a set of selected points whose coordinates are known in both datum systems. If the coordinates of these 5 points are not correct–often the case for points measured on local datum system– the estimated parameters may be inaccurate. As a result the datum transformation will be inaccurate. This is often the case when we transform coordinates from a local horizontal datum to a global geocentric datum. The coordinates in the local horizontal datum may be distorted by several tens of meters because of the inherent inaccuracies of the measurements used in the triangulation network. These inherent inaccuracies are also responsible for another complication: the transformation parameters are not unique. Their estimate will depend on which particular common points are chosen, and they also will depend on whether all seven transformation parameters, or only a sub-set of them, are estimated.

	Parameter	National set	Provincial set	NIMA set
Scale	s	$1 - 8.3 \cdot 10^{-6}$	$1 - 9.2 \cdot 10^{-6}$	1
Angles	α	+1.04	+0.32	
	β	+0.35	+3.18	
	γ	−3.08	−0.91	
Shifts	X_0	−581.99 m	−518.19 m	−635 m
	Y_0	−105.01 m	−43.58 m	−27 m
	Z_0	−414.00 m	−466.14 m	−450 m

4.2 SATELLITE-BASED POSITIONING

Satellites are used in geocentric reference systems, and increase the level of spatial accuracy substantially. They are critical tools in geodetic engineering for the maintenance of the ITRF. They also play a key role in mapping, surveying, and in a growing number of applications requiring positioning techniques. Nowadays, for fieldwork that includes spatial data acquisition, the use of satellite-based positioning is considered indispensable.

Satellite-based positioning was developed and implemented to address military needs, somewhat analogously to the early development of the internet. The technology is now widely available for civilians use. The requirements for the development of the positioning system were:

- Suitability for all kinds of military use: ground troops and vehicles, aircraft and missiles, ships;
- Requiring only low-cost equipment with low energy consumption at the receiver end;
- Provision of results in real time for an unlimited number of users concurrently;
- Support for different levels of accuracy (military versus civilian);
- Around-the-clock and weather-proof availability;
- Use of a single geodetic datum;
- Protection against intentional and unintentional disturbance, for instance, through a design allowing for redundancy.

A satellite-based positioning system set-up involves implementation of three hardware segments :

1. The space segment, i.e. the satellites that orbit the Earth, and the radio signals that they emit,
2. The control segment, i.e. the ground stations that monitor and maintain the space segment components, and
3. The user segment, i.e. the users with their hard- and software to conduct positioning.

In satellite positioning, the central problem is to determine values (X, Y, Z) of a receiver that receives satellite signals, i.e. to determine the position of the receiver with a stated accuracy and precision. Required accuracy and precision depends on the application; timeliness, i.e. are the position values required in real time or can they be determined later during post-processing, also varies between applications. Some applications like navigation require kinematic approaches which take into account the fact that the receiver is not stationary, but is moving.

- **Absolute positioning**

The working principles of absolute, satellite-based positioning are fairly simple:

1. A satellite, equipped with a clock, at a specific moment sends a radio message that includes :
   a) The satellite identifier,
   b) Its position in orbit, and
   c) Its clock reading.
2. A receiver on or above the planet, also equipped with a clock, receives the message slightly later, and reads its own clock.

From the time delay observed between the two clock readings, and knowing the speed of radio transmission through the medium between (satellite) sender and receiver, the receiver can compute the distance to the sender, also known as the satellite's pseudorange. The pseudorange of a satellite with respect to a receiver is its apparent distance to the receiver, computed from the time delay with which its radio signal is received. Such a computation determines the position of the receiver to be on a sphere of radius equal to the computed pseudorange. If the receiver instantaneously would do the same with a message of another satellite that is positioned elsewhere, the position of the receiver is restricted to an- other sphere. The intersection of the two spheres, which have different centres, determines a circle as the set of possible positions of the receiver. If a third satellite message is taken into consideration, the intersection of three spheres determines at most two positions, one of which is the actual position of the receiver. The overall procedure is known as trilateration: the determination of a position based on three distances.

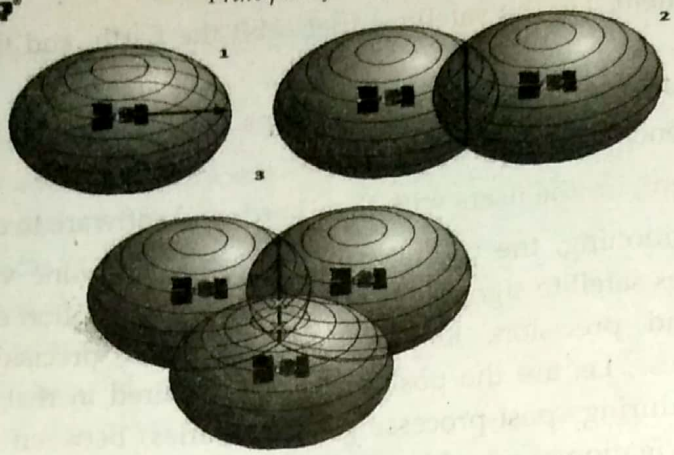


Figure : Pseudoranging using 1, 2 and 3 satellites.

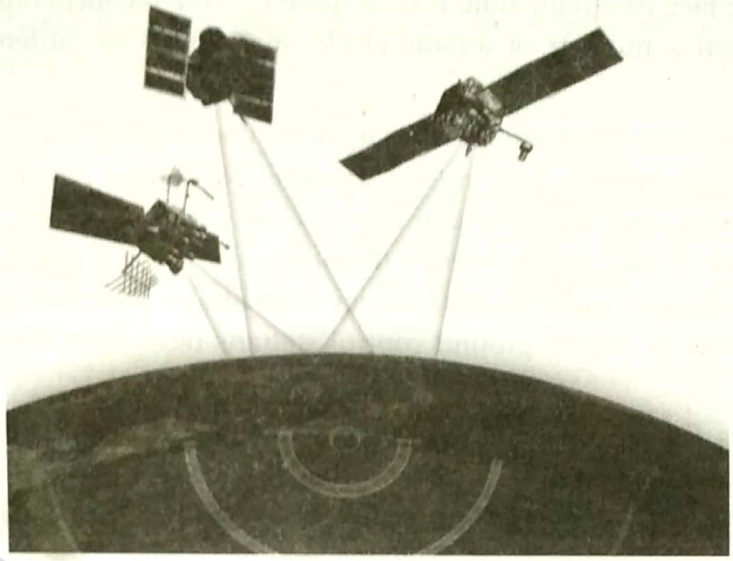
It would appear therefore that the signals of three satellites would suffice to determine a positional fix for our receiver. In theory this is true, but in practice it is not. The reason is that we have made the assumption that all satellite clocks as well as our receiver clock are fully synchronized, where in fact they are not. The satellite clocks are costly, high-precision, atomic clocks that we can consider synchronized for the time being, but the receiver typically has cheaper, quartz clock that is not synchronized with the satellite clocks. An additional unknown parameter, called as the synchronization bias of the receiver clock, i.e. the difference in time reading between it and the satellite clocks.

A set of unknown variables has now become (X, Y, Z, Δt) representing a 3D position and a clock bias. By including the information obtained from a fourth satellite message, This will result in the determination of the receiver's actual position (X, Y, Z), as well as its receiver clock bias Δt, and if we correct the receiver clock for this bias we effectively turn it into a high-precision, atomic clock as well. Obtaining a high-precision clock is a fortunate side-effect of using the receiver, as it allows the design of experiments distributed in geographic space that demand high levels of synchrony. One such application is the use of wireless sensor networks for various natural phenomena like earthquakes, meteorological patterns or in water management.

Another application is in the positioning of mobile phone users making an emergency call. Often the caller does not know their location accurately. The telephone company can trace back the call to the receiving transmitter mast, but this may be servicing an area with a radius of 300 m to 6 km. That is too inaccurate a position for an emergency ambulance to go to. However, if all masts in the telephony network are equipped with a satellite positioning receiver (and thus, with a very good, synchronized clock) the time of reception of the call at each mast can be recorded. The time difference of arrival of the call between two nearby masts determines a hyperbola on the ground of possible positions of the caller; if the call is received on three masts, we would have two hyperbolas, al lowing intersection, and thus 'hyperbolic positioning'. With current technology the (horizontal) accuracy would be better than 30 m.

When only three and not four satellites are 'in view', the receiver is capable of falling back from the above 3D positioning mode to the inferior 2D positioning mode. With the relative abundance of satellites in orbit around the earth, this is a relatively rare situation, but it serves to illustrate the importance of 3D positioning. If a 3D fix has already been obtained, the receiver simply assumes that the height above the ellipsoid has not changed since the last 3D fix. If no fix had yet been obtained, the receiver assumes that it is

positioned at the geocentric ellipsoid adopted by the positioning system, i.e. at height h=0.8 In the receiver computations, the ellipsoid fills the slot of the missing fourth satellite sphere, and the unknown variables can therefore still be determined. Clearly in both of these cases, the assumption for this computation is flawed and the positioning results in 2D mode will be unreliable—much more so if no previous fix had been obtained and one's receiver is not at all near the surface of the geocentric ellipsoid.



Time, clocks and world time

While latitude was determined with a sextant from the position of the Sun in the sky, they carried clocks with them to determine the longitude of their position. Early ship clocks were unreliable, having a drift of multiple seconds a day, which could result in positional error of a few kilometers.

Before any notion of standard time existed, villages and cities simply kept track of their local time determined from position of the Sun in the sky. When trains became an important means of transportation, these local time systems became problematic as the schedules required a single time system. Such a time system needed the definition of time zones: typically as 24 geographic strips between certain longitudes that are multiples of 15°. This all gave rise to Greenwich Mean Time (GMT). GMT was the world time standard of choice. It was a system based on the mean solar time at the meridian of Greenwich, United Kingdom, which is the conventional 0-meridian in geography.

GMT was later replaced by Universal Time (UT), a system still based on meridian crossings of stars, but now of far away quasars as this provides more accuracy than that of the Sun. It is still the case that the rotational velocity of our planet is not constant and the length of a solar day is increasing. So UT is not a perfect system either. It continues to be used for civil clock time, but it is officially now replaced by International Atomic Time (TAI). UT actually has various versions, amongst which are UT0, UT1 and UTC. UT0 is the Earth rotational time observed in some location. Because the Earth experiences polar motion as well, UT0 differs between locations. If we correct for polar motion, we obtain UT1, which is identical everywhere. It is still a somewhat erratic clock because of the earlier mentioned varying rotational velocity of the planet. The uncertainty is about 3 msec per day.

Coordinated Universal Time (UTC) is used in satellite positioning, and is maintained with atomic clocks. By convention, it is always within a margin of 0.9 sec of UT1, and twice annually it may be given a shift to stay within that margin. This occasional shift of a leap second is applied at the end of June 30 or preferably at the end of December 31. The last minute of such a day is then either 59 or UTC 61 seconds long. So far, adjustments have always been to add a second. UTC time can only be determined to the highest precision after the fact, as atomic time is determined by the reconciliation of the observed differences between a number of atomic clocks maintained by different national time bureaus.

Position with clocks using satellite signals is calculated using conversion factor in the speed of light, approximately $3 \cdot 10^8$ m/s in vacuum. No longer can multiple seconds of clock bias be allowed, and this is where atomic clocks come in. They are very accurate time keepers, based on the exactly known frequency with which specific atoms (Cesium, Rubidium and Hydrogen) make discrete energy state jumps. Positioning satellites usually have multiple clocks on board; ground control stations have even better quality atomic clocks.

Atomic clocks, however, are not flawless: their timing tends to drift away from true time somewhat, and they too need to be corrected. The drift, and the change in drift over time, are monitored, and are part of the satellite's navigation message, so that they can be corrected for.

- **Errors in Absolute Positioning**

Errors related to the space segment

As a first source of error, the operators of the control segment may intentionally deteriorate radio signals of the satellites to the general public, to avoid optimal use of the system by the enemy, for instance in times of global political tension and war. This selective availability—meaning that the military forces allied with the control segment will still have access to undisturbed signals—may cause error that is an order of magnitude larger than all other error sources combined.

Secondly, the satellite message may contain incorrect information. Assuming that it will always know its own identifier, the satellite may make two kinds of error:

1. **Incorrect clock reading**

Even atomic clocks can be off by a small margin, and since Einstein, we know that travelling clocks are slower than resident clocks, due to a so-called relativistic effect. If one understands that a clock that is off by 0.000001 sec causes a computation error in the satellite's pseudorange of approximately 300 m, it is clear that these satellite clocks require very strict monitoring.

2. **Incorrect orbit position**

The orbit of a satellite around our planet is easy to describe mathematically if both bodies are considered point masses, but in real life they are not. For the same reasons that the Geoid is not a simply shaped surface, the Earth's gravitation field that a satellite experiences in orbit is not simple either. Moreover, it is disturbed by solar and lunar gravitation, making its flight path slightly erratic and difficult to forecast exactly.

Both types of error are strictly monitored by the ground control segment, which is responsible for correcting any errors of this nature, but it does so by applying an agreed

upon tolerance. A control station can obviously compare results of positioning computations like discussed above with its accurately known position, flagging any unacceptable errors, and potentially labeling a satellite as temporarily 'unhealthy' until errors have been corrected, and brought to within the tolerance. This may be done by uploading a correction on the clock or orbit settings to the satellite.

Errors related to the medium

The medium between sender and receiver may be of influence to the radio signals. The middle atmospheric layers of stratosphere and mesosphere are relatively harmless and of little hindrance to radio waves, but this is not true of the lower and upper layer. They are, respectively:

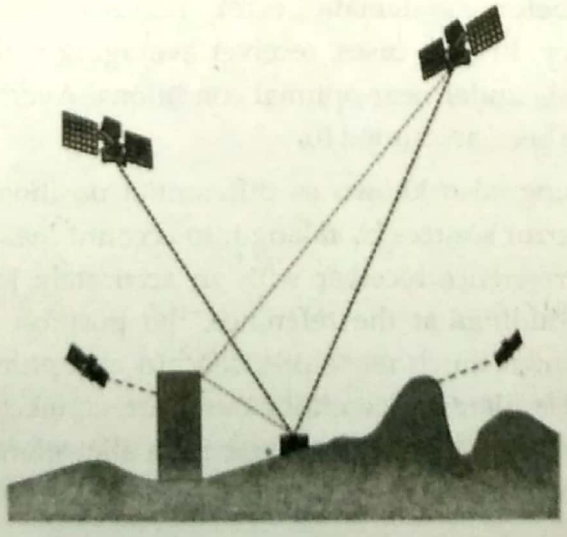
- **The troposphere :** the approximate 14 km high airspace just above the Earth's surface, which holds much of the atmosphere's oxygen and which envelopes all phenomena that we call the weather. It is an obstacle that delays radio waves in a rather variable way.

- **The ionosphere :** the most outward part of the atmosphere that starts at an altitude of 90 km, holding many electrically charged atoms, thereby forming a protection against various forms of radiation from space, including to some extent radio waves. The degree of ionization shows a distinct night and day rhythm, and also depends on solar activity.

The ionosphere is a more severe source of delay to satellite signals, which obviously means that pseudoranges are estimated larger than they actually are. When satellites emit radio signals at two or more frequencies, an estimate can be computed from differences in delay incurred for signals of different frequency, and this will allow for the correction of atmospheric delay, leading to a 10–50% improvement of accuracy. If this is not the case, or if the receiver is capable of receiving just a single frequency, a model should be applied to forecast the delay, typically taking into account the time of day and current latitude of the receiver.

Errors related to the receiver's environment

The error occurring when a radio signal is received via two or more paths between sender and receiver, some of which typically via a bounce off of some nearby surface, like a building or rock face. The term applied to this phenomenon is multi-path; when it occurs the multiple receptions of the same signal may interfere with each other. Multi-path is a difficult to avoid error source.



All of the above error sources have an influence on the computation of a satellite's pseudorange. In accumulation, they are called the user equivalent range error (UERE). Some error sources may be at work for all satellites being used by the receiver, for instance, selective availability and the atmospheric delay, while others may be specific to one satellite, for instance, incorrect satellite information and multi-path.

Errors related to the relative geometry of satellites and receiver

There is one more source of error that is unrelated to individual radio signal characteristics, but that rather depends on the combination of the satellite signals used for positioning. Of importance is their constellation in the sky from the receiver perspective. The sphere intersection technique of positioning will provide more precise results when the four satellites are nicely spread over the sky. The error source is known as geometric dilution of precision (GDOP). GDOP is lower when satellites are just above the horizon in mutually opposed compass directions. However, such satellite positions have bad atmospheric delay characteristics, so in practice it is better if they are at least 15° above the horizon. When more than four satellites are in view, modern receivers use 'least-squares' adjustment to calculate the best positional fix possible from all of the signals.

These errors are not all of similar magnitude. GDOP functions not so much as an independent error source but rather as a multiplying factor, decreasing the precision of position and time values obtained. The procedure that we discussed above is known as absolute, single-point positioning based on code measurement. It is the fastest and simplest, yet least accurate way of determining a position using satellites. It suffices for recreational purposes and other applications that require horizontal accuracy not under 5–10 m. Typically, when encrypted military signals can also be used, on a dual-frequency receiver the achievable horizontal accuracy is 2–5 m. Below, we discuss other satellite-based positioning techniques with better accuracies.

- **Relative Positioning**

One technique to remove errors from positioning computations is to perform many position computations, and to determine the average over the solutions. Many receivers allow the user to do so. It should however be clear from the above that averaging may address random errors like signal noise, selective availability (SA) and multi-path to some extent, but not systematic sources of error, like incorrect satellite data, atmospheric delays, and GDOP effects. These sources should be removed before averaging is applied. It has been shown that averaging over 60 minutes in absolute, single-point positioning based on code measurements, before systematic error removal, leads only to a 10–20% improvement of accuracy. In such cases, receiver averaging is therefore of limited value, and requires long periods under near-optimal conditions. Averaging is a good technique if systematic errors have been accounted for.

In relative positioning, also known as differential positioning, one tries to remove some of the systematic error sources by taking into account measurements of these errors in a nearby stationary reference receiver with an accurately known position. By using these systematic error findings at the reference, the position of the target receiver of interest will become known much more precisely. In an optimal setting, reference and target receiver experience identical conditions and are connected by a direct data link, allowing the target to receive correctional data from the reference. In practice, relative positioning allows reference and target receiver to 70–200 km apart, and they will

essentially experience similar atmospheric signal error. For each satellite in view, the reference receiver will determine its pseudorange error. After all, its position is known with high accuracy, so it can solve any pseudorange equations to determine the error. Subsequently, the target receiver, having received the error characteristics will apply the correction for each of the four satellite signals that it uses for positioning, it can narrow down its accuracy to the 0.5–5 m range.

- **Network positioning**

Network positioning is an integrated, systematic network of reference receivers covering a large area like a continent or even the whole globe.

The organization of such a network can take different shapes, augmenting an already existing satellite-based system. A general architecture consists of a network of reference stations, strategically positioned in the area to be covered, each of which is constantly monitoring signals and their errors for all positioning satellites in view. One or more control centres receive the reference station data, verify this for correctness, and relay (uplink) this information to a geostationary satellite. The satellite will retransmit the correctional data to the area that it covers, so that target receivers, using their own approximate position, can determine how to correct for satellite signal error, and consequently obtain much more accurate position fixes. With network positioning, accuracy in the submetre range can be obtained. Typically, advanced receivers are required, but the technology lends itself also for solutions with a single advanced receiver that functions in the direct neighborhood as a reference receiver to simple ones.

- **Code versus phase measurements**

So far, we have assumed that the receiver determines the range of a satellite by measuring time delay on the received ranging code. There exists a more advanced range determination technique known as carrier phase measurement. This typically requires more advanced receiver technology, and longer observation sessions. Carrier phase measurement can currently only be used with relative positioning, as absolute positioning using this method is not yet well developed.

The technique aims to determine the number of cycles of the (sine-shaped) radio signal between sender and receiver. Each cycle corresponds to one wavelength of the signal, which in the applied L-band frequencies is 19–24 cm. Since this number of cycles cannot be directly measured, it is determined, in a long observation session, from the change in carrier phase with time. This happens because the satellite is orbiting itself. From its orbit parameters and the change in phase over time, the number of cycles can be derived.

With relative positioning techniques, a horizontal accuracy of 2 mm–2 cm can be achieved. This degree of accuracy makes it possible to measure tectonic plate movements, which can be as big as 10 cm per year in some locations on the planet.

- **Positioning technology**

GPS

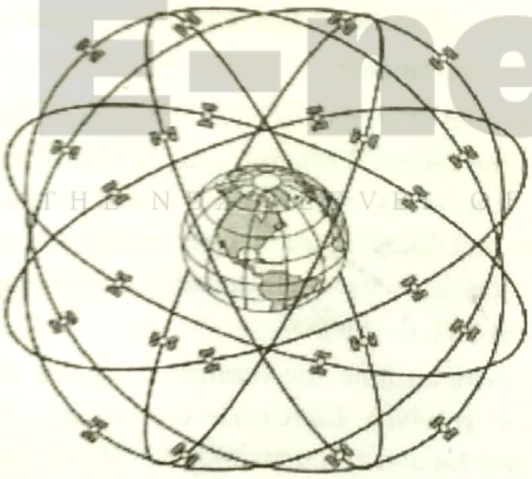
The NAVSTAR Global Positioning System (GPS) was declared operational in 1994, providing Precise Positioning Services (PPS) to US and allied military forces as well as US government agencies, and Standard Positioning Services (SPS) to civilians throughout the

world. Its space segment nominally consists of 24 satellites, each of which orbits the Earth in 11h58m at an altitude of 20,200 km. There can be any number of satellites active, typically between 21 and 27. The satellites are organized in six orbital planes, somewhat irregularly spaced, with an angle of inclination of 55–63° with the equatorial plane, nominally having four satellites each (see Figure 4.28). This means that a receiver on Earth will have between five and eight (sometimes up to twelve) satellites in view at any point in time. Software packages exist to help in planning GPS surveys, identifying expected satellite set-up for any location and time.

GPS's control segment has its master control in Colorado, US, and monitor stations in a belt around the equator, namely in Hawaii, Kwajalein Atoll in the Marshall Islands, Diego Garcia (British Indian Ocean Territory) and Ascension Island (UK, southern Atlantic Ocean).

The NAVSTAR satellites transmit two radio signals, namely the L1 frequency at 1575.42 MHz and the L2 frequency at 1227.60 MHz. There are also a third and fourth signal, but they are not important for our discussion here. The first two signals consist of:

- The carrier waves at the given frequencies,
- A coarse ranging code, known as C/A, modulated on L1,
- An encrypted precision ranging code, known as P(Y), modulated on L1and L2, and
- A navigation message modulated on both L1 and L2.



The NAVSTAR GPS consists of 24 GPS satellites operated by the U.S. Department of Defense

The role of L2 is to provide a second radio signal, thereby allowing dual-frequency receivers a way of determining fairly precisely the actual ionospheric delay on satellite signals received. The role of the ranging codes is two-fold:

1. To identify the satellite that sent the signal, as each satellite sends unique codes, and the receiver has a look-up table for these codes, and
2. To determine the signal transit time, and thus the satellite's pseudorange.

The navigation message contains the satellite orbit and satellite clock error information, as well as some general system information. GPS also carries a fifth, encrypted military signal carrying the M-code. GPS uses WGS84 as its reference system. It

has been refined on several occasions and is now aligned with the ITRF at the level of a few centimetres worldwide. GPS has adopted UTC as its time system. In the civil market, GPS receivers of varying quality are available, their quality depending on the embedded positioning features: supporting single- frequency or dual-frequency, supporting only absolute or also relative positioning, performing code measurements or also carrier phase measurements. Leica and Trimble are two of the well-known brands in the high-precision, professional surveying domain; Magellan and Garmin, operate in the lower price, higher volume consumer market range, amongst others for recreational use in outdoor activities. Many of these are single frequency receivers, doing only code measurements, though some are capable of relative positioning. This includes the new generation of GPS-enabled mobile phones.

GLONASS

What GPS is to the US military, is GLONASS to the Russian military, specifically the Russian Space Forces. Both systems were primarily designed on the basis of military requirements. The big difference between the two is that GPS generated a major interest in civil applications, thus having an important economic impact. The GLONASS space segment consists of nominally 24 satellites, organized in three orbital planes, with an inclination of 64.8° with the equator. Orbiting altitude is 19,130 km, with a period of revolution of 11 hours 16 min. GLONASS uses the PZ–90 as its reference system, and like GPS uses UTC as time reference, though with an offset for Russian daylight.

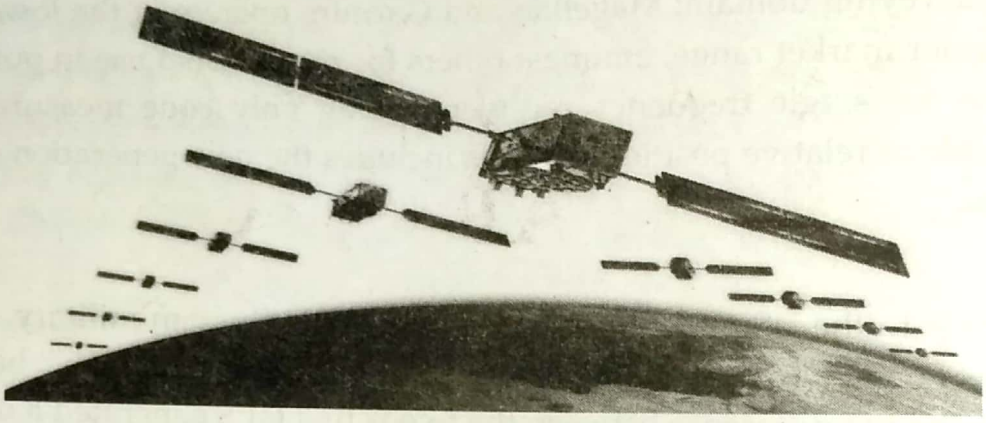
GLONASS radio signals are somewhat similar to that of GPS. Satellites use different identifier schemes and their navigation message use other parameters. They also use different frequencies: GLONASS L1 is at approximately 1605 MHz, and L2 is at approximately 1248 MHz. Otherwise, the GLONASS system performance is rather comparable to that of GPS.

Galileo

Galileo is the name of this EU system. The vision is that satellite-based positioning will become even bigger due to the emergence of mobile phones equipped with receivers, perhaps with some 400 million users by the year 2015. The completed system will have 27 satellites, with three in reserve, orbiting in one of three, equally spaced, circular orbits at an elevation of 23,222 km, inclined 56° with the equator. This higher inclination, when compared to that of GPS, has been chosen to provide better positioning coverage at high latitudes, such as northern Scandinavia where GPS performs rather poorly.

The EU and the US agreed to make Galileo and GPS compatible by adoption of interchangeable satellite signal set-ups. The effect of this agreement is that the Galileo/GPS

tandem satellite system will have so many satellites in the sky (close to 60) that a receiver can almost always find an optimal constellation in view. This will be especially useful in situations where in the past bad signal reception happened : in built-up areas and forests, for instance. It will also bring the implementation of a Global Navigation Satellite System (GNSS) closer as positional accuracy and reliability will improve. Such a system helps to implement fully automated air and road traffic.



IRNSS & GAGAN

The Indian Regional Navigation Satellite System (IRNSS) is an independent satellite based regional system developed indigenously by India on par with US-based GPS, Russia's GLONASS and Galileo developed by Europe. It was renamed "Navic" (Navigation with Indian Constellation). Terrestrial, Aerial & Marine Navigation, Disaster Management, Vehicle tracking & fleet management, Integration with mobile phones, Precise Timing, Mapping & Geodetic data capture, Terrestrial navigation aid and Visual & voice navigation for drivers are some of its essential applications.

GPS Aided GEO Augmented Navigation (GAGAN) is a step by the Indian Government towards initial Satellite-based Navigation Services in India. The Airports Authority of India (AAI) and Indian Space Research Organization (ISRO) have collaborated to develop the GPS Aided Geo Augmented Navigation (GAGAN) as a regional Satellite Based Augmentation System (SBAS). The GAGAN's goal is to provide a navigation system to assist aircraft in accurate landing over the Indian airspace and in the adjoining area and applicable to safety-to-life civil operations. GAGAN is inter-operable with other international SBAS systems. GAGAN Payload is now operational. The satellites GSAT-8 and GSAT-10 satellites have the GAGAN payloads. The third payload of the system will be launched with GSAT-15 satellite which is scheduled for launch in late of 2015. One essential component of the GAGAN project is the study of the ionospheric behavior over the Indian region. GAGAN ionospheric algorithm was developed by ISRO.