The core of GIScience
a process-based approach

ITC Educational textbook series

UNIVERSITY OF TWENTE

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

ITC

Chapter 3

Spatial referencing and satellite-based positioning

Richard Knippers
Klaus Tempfli

Introduction

In the early days of geoinformation science, spatially referenced data usually originated within national boundaries, i.e. these data were derived from printed maps published by national mapping organizations. Nowadays, users of geoinformation are combining spatial data from a given country with global spatial data sets, reconciling spatial data from published maps with coordinates established by satellite positioning techniques, and integrating their spatial data with that from neighbouring countries.

To perform these kinds of tasks successfully, we need to understand basic spatial referencing concepts. Section 3.1 discusses the relevance and actual use of reference surfaces, coordinate systems and coordinate transformations. We will explain the principles of spatial referencing as applied to mapping, the traditional application of geoinformation science. These principles are generally applicable to all types of geospatial data.

Section 3.2 discusses satellite-based systems of spatial positioning. The development of these global positioning systems has made it possible to unambiguously determine any position in space. This and related developments have laid the foundations for the integration of all spatial data within a single, global 3D spatial-reference system, which we may expect to emerge in the near future.

3.1 Spatial referencing

One of the defining features of geoinformation science is its ability to combine spatially referenced data. A frequently occurring issue is the need to combine spatial data from different sources that use different spatial reference systems. This section provides

a broad background of relevant concepts relating to the nature of spatial reference systems and the translation of data from one spatial referencing system into another.

3.1.1 Reference surfaces

The surface of the Earth is far from uniform. Its oceans can be treated as reasonably uniform, but the surface or topography of its 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. Consequently, two main reference surfaces have been established to approximate the shape of the Earth : one is called the *Geoid*, the other the *ellipsoid*; see Figure 3.1.

geoid and ellipsoid

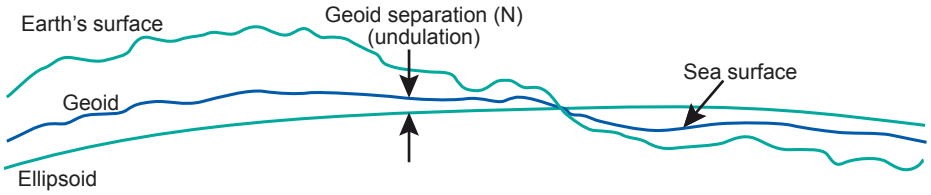

Figure 3.1
The Earth's surface and two reference surfaces used to approximate it: the Geoid; and a reference ellipsoid. The Geoid separation (N) is the deviation between the Geoid and the reference ellipsoid.

The Geoid and the vertical datum

We can simplify matters by imagining that the entire Earth's surface is covered by water. If we ignore effects of tides and currents 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 the plumb line—is dependent on the distribution of mass inside the Earth. Owing to irregularities or mass anomalies in this distribution, the surface of the *global ocean* would be undulating. The resulting surface is called the Geoid (Figure 3.2). A plumb line through any surface point would always be perpendicular to the surface.

plumb line

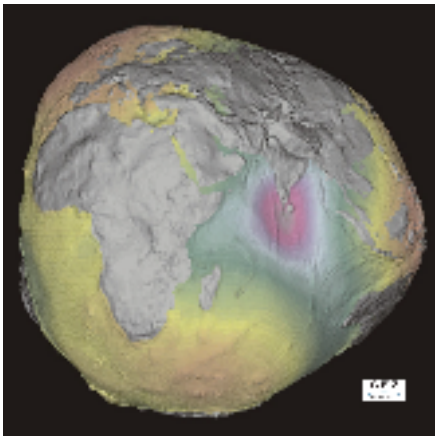

Figure 3.2
The Geoid, exaggerated to illustrate the complexity of its surface. Image: GFZ German Research Centre for Geosciences.

The Geoid is used to describe *heights*. In order to establish the Geoid as a reference for heights, the ocean's water level is registered at coastal locations over several years using tide gauges (mareographs). Averaging the registrations largely eliminates variations in sea level over time. The resultant water level represents an approximation to the Geoid and is termed mean sea level.

mean sea level

For the Netherlands and Germany, local mean sea level is related to the Amsterdam Tide Gauge (zero height). We can determine the height of a point in Enschede with

respect to the Amsterdam Tide Gauge using a technique known as geodetic levelling (Figure 3.3). The result of this process will be the height of the point in Enschede above local mean sea level. Height determined with respect to a tide gauge station is known as *orthometric height* (height H above the Geoid) .

Several definitions of local mean sea levels (also called local vertical datums) appear throughout the world. They are parallel to the Geoid but offset by up to a couple of metres to allow for 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 . This might be the case, for example, in areas along the border of adjacent nations.

local vertical datums

Even within a country, heights may differ depending on the location of the tide gauge to which mean sea level is related. As an example, the mean sea level from the Atlantic to the Pacific coast of the U.S.A. differs by 0.6–0.7 m. The tide gauge (zero height) of the Netherlands differs -2.34 m from the tide gauge (zero height) of neighbouring Belgium.

The local vertical datum is implemented through a levelling network (Figure 3.3a), which consists of benchmarks whose height above mean sea level has been determined through geodetic levelling . The implementation of the datum enables easy user access. Surveyors, for example, do not need to start from scratch (i.e. from the Amsterdam tide gauge) each time they need to determine the height of a new point. They use the benchmark of the levelling network that is closest to the new point (Figure 3.3b).

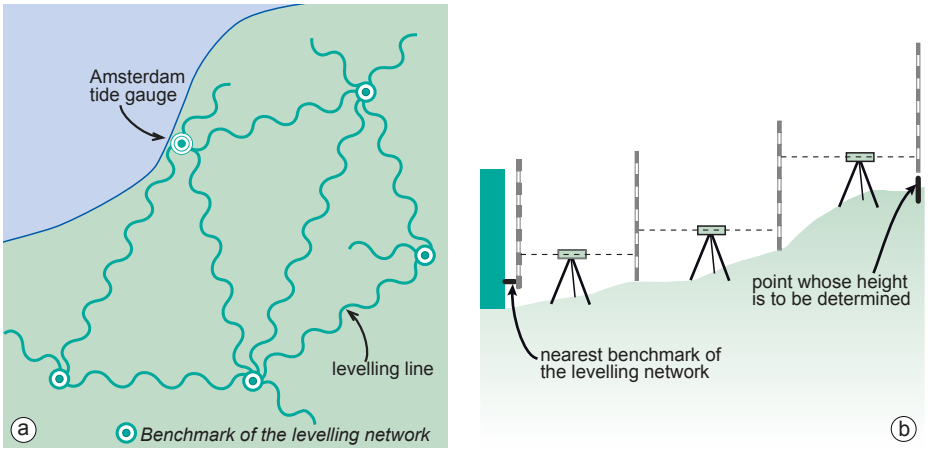

Figure 3.3
A levelling network implements a local vertical datum: (a) network of levelling lines starting from the Amsterdam Tide Gauge, showing some of the benchmarks; (b) how the orthometric height (H) is determined for some point by working from the nearest benchmark.

As a result of satellite gravity missions, it is currently possible to determine height (H) above the Geoid to centimetre levels of accuracy. It is foreseeable that a global vertical datum may become ubiquitous in the next 10–15 years. If all geodata, for example maps, were to use such a global vertical datum, heights would become globally comparable, effectively making local vertical datums redundant for users of geoinformation.

The ellipsoid

We have defined a physical surface, the Geoid, as a reference surface for heights. We also need, however, a reference surface for the description of the *horizontal coordinates* of points of interest. Since we will later want 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

horizontal coordinates

is the *oblate ellipsoid* (Figure 3.4). It provides a relatively simple figure that fits the Geoid to a first-order approximation (for small-scale mapping purposes we may use the *sphere*). An ellipsoid is formed when an ellipse is rotated around its minor axis. This ellipse, which defines an ellipsoid or *spheroid*, is called a meridian ellipse (notice that ellipsoid and spheroid are used here to refer to the same).

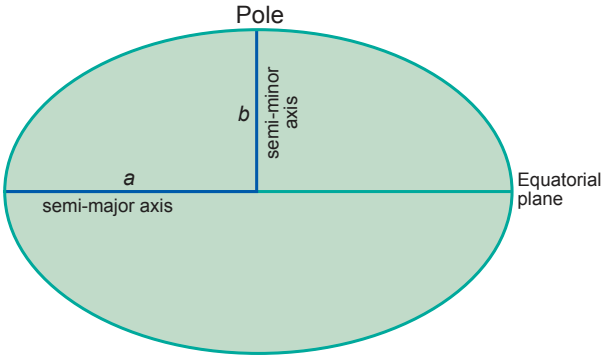

Figure 3.4
An oblate ellipsoid, defined by its semi-major axis a and semi-minor axis b.

The shape of an ellipsoid may be defined in a number of ways, but in geodetic practice it is is usually defined by its semi-major axis and flattening (Figure 3.4). Flattening f is dependent on both the semi-major axis a and the semi-minor axis b:

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

The ellipsoid may also be defined by its semi-major axis a and its eccentricity e, which can be expressed as:

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

Given one axis and any one of the other three parameters, the other two can be derived. Typical values of the parameters for an ellipsoid are:

$$a = 6378135.00\ m,\ b = 6356750.52\ m,\ f = \frac{1}{298.26},\ e = 0.08181881066$$

local ellipsoids

Many different sorts of ellipsoids have been defined. Local ellipsoids have been established to fit the Geoid (mean sea level) 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 (Figure 3.5).

global ellipsoids

With increasing demands for global surveying, work is underway to develop global reference ellipsoids. In 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 general assembly of the IUGG in Madrid introduced the ellipsoid determined by Hayford in 1909 as the international ellipsoid. According to subsequently acquired knowledge, however, the values for this ellipsoid give an insufficiently accurate approximation. At the 1967 general assembly of the IUGG in Luzern, the 1924 reference system was replaced by the Geodetic Reference System 1967 (GRS 1967) el-

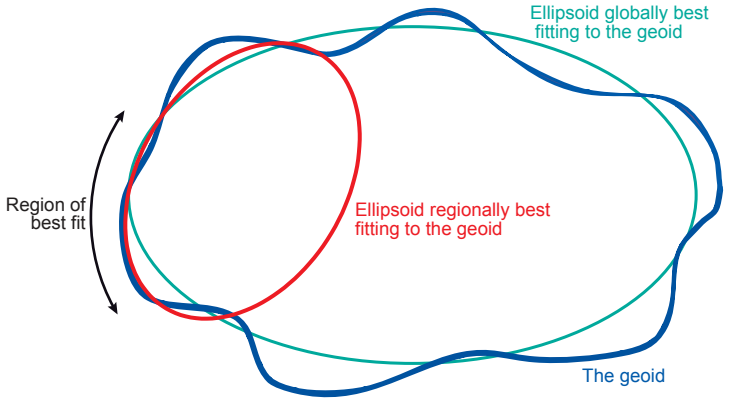

Figure 3.5
The Geoid, its global best-fit ellipsoid, and a regional best-fit ellipsoid for a chosen region. Adapted from: Ordnance Survey of Great Britain. *A Guide to Coordinate Systems in Great Britain.*

lipsoid. Later, in 1980, this was in turn replaced by the Geodetic Reference System 1980 (GRS80) ellipsoid.

Name	*a* (m)	*b* (m)	*f*
International (1924)	6378388.	6356912.	1:297.000
GRS 1967	6378160.	6356775.	1:298.247
GRS 1980 & WGS84	6378137.	6356752.	1:298.257

Table 3.1
Three global ellipsoids defined by a semi-major axis a, semi-minor axis b, and flattening f. For all practical purposes, the GRS80 and WGS84 can be considered to be identical.

The local horizontal datum

Ellipsoids have varying positions 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 what is called a fundamental point and an azimuth to an additional point. We say that this defines a *local horizontal datum*. Note that the term horizontal datum and geodetic datum are treated as equivalent and interchangeable terms.

Several hundred local horizontal datums exist in the world. The reason for this is obvious: different local ellipsoids of varying position and orientation had to be adopted to provide a best fit of the local mean sea level in different countries or regions. The Potsdam Datum, the local horizontal datum used in Germany is an example of a local horizontal datum. The fundamental point is located in Rauenberg and the underlying ellipsoid is the Bessel ellipsoid (a = 6,377,397.156 m, b = 6,356,079.175 m). We can determine the latitude and longitude (ϕ, λ) of any other point in Germany with respect to this local horizontal datum using geodetic positioning techniques, such as triangulation and trilateration. The result of this process will be the geographic (or horizontal) coordinates (ϕ, λ) of a new point in the Potsdam Datum.

triangulation networks

A local horizontal datum is determined through a triangulation network. Such a network consists of monumented points that form a network of triangular mesh elements (Figure 3.6). The angles in each triangle are measured, in addition to at least one side of the 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.

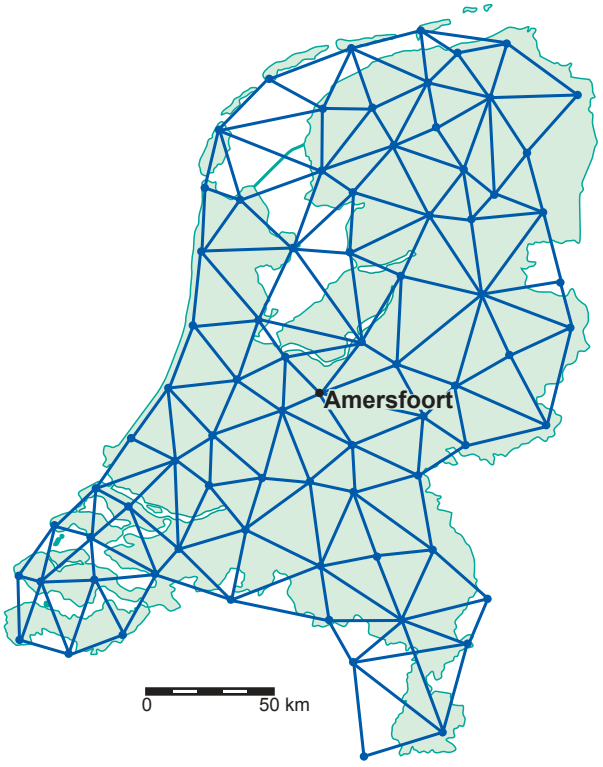

Figure 3.6
The old primary triangulation network in the Netherlands was made up of 77 points (mostly church towers). The extension and re-measurement of the network is done nowadays through satellite measurements. Adapted from original figure by "Dutch Cadastre and Land Registers", now called *het Kadaster*.

The global horizontal datum

With increasing demands for global surveying, activities are underway to establish global reference surfaces. The motivation in this is to make geodetic results mutually compatible and to be able to provide coherent results to other disciplines, e.g. astronomy and geophysics.

The most important global (geocentric) spatial reference system for the geoinformation community is the International Terrestrial Reference System (ITRS) . This 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 North Pole. The X-axis is oriented towards the mean Greenwich meridian and is orthogonal to the Z-axis. The Y-axis completes the right-handed reference coordinate system (Figure 3.7a).

ITRS

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 (Figure 3.7b). Constant re-measuring is needed because of the addition of new control stations and ongoing geophysical processes (mainly tectonic plate motion) that deform the Earth's crust at measurable global, regional and local scales. These deformations cause positional differences over time and have resulted in more than one realization of the ITRS. Examples are the ITRF96 and the ITRF2000. The ITRF96 was established on 1 January 1997, which means that the measurements use data acquired up to 1996 to fix the geocentric coordinates (X, Y and Z in metres) and velocities (posi-

ITRF

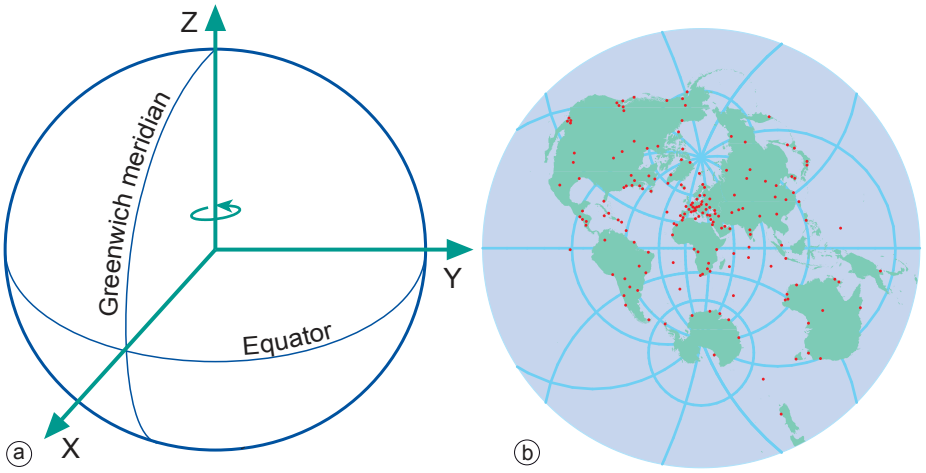

Figure 3.7
(a) The International Terrestrial Reference System (ITRS) and; (b) the International Terrestrial Reference Frame (ITRF) visualized as a distributed set of ground control stations (represented by red dots).

tional change in X, Y and Z in metres per year) of the different stations. The velocities are used to propagate measurements to other epochs (times). The trend is to use the ITRF everywhere in the world for reasons of global compatibility.

geocentric datums

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 centimetres worldwide. Global horizontal datums, such as 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, since the 1960s), as a result of advances in extra-terrestrial positioning techniques. [1]

Since the range and shape of satellite orbits are 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[2]. This technique can also be used for the realization of global ellipsoids and datums at levels of accuracy required for large-scale mapping.

To implement the ITRF in a particular region, a densification of control stations is needed to ensure that there are enough coordinated reference points available in that region. These control stations are equipped with permanently operating satellite positioning equipment (i.e. GPS receivers and auxiliary equipment) and communication links. Examples of (networks consisting of) such permanent tracking stations are the Actief GNSS Referentie Systeem Nederland (AGRS) in the Netherlands and the Satellitenpositionierungsdienst der deutschen Landesvermessung (SAPOS) in Germany.

We can transform ITRF coordinates (X, Y and Z in metres) 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 straightforward transformation has no physical meaning and is contrary to our intuitive human perception of height. Therefore, we use instead the height, H, above the Geoid (see Figure 3.8). It is foreseeable that global 3D spatial referencing in terms of (ϕ, λ, H) could become ubiquitous in the next 10–15 years. If, by then, all published maps are also globally referenced the underlying spatial referencing concepts will become transparent and, hence, irrelevant to users of geoinformation.

3D spatial referencing

[1] Extra-terrestrial positioning techniques include, for example, Satellite Laser Ranging (SLR), Lunar Laser Ranging (LLR), Global Positioning System (GPS), and Very Long Baseline Interferometry (VLBI).

[2] In the case of an idealized spherical Earth, it is one of the focal points of the elliptical orbits.

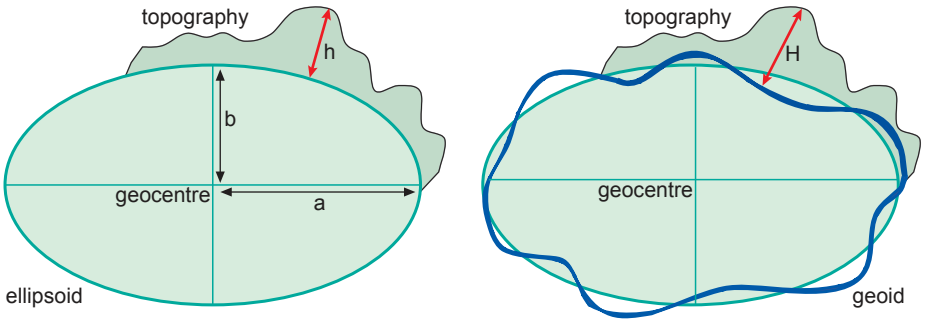

Figure 3.8
Height h above the geocentric ellipsoid, and height H above the Geoid. h is measured orthogonal to the ellipsoid, H orthogonal to the Geoid.

Hundreds of existing local horizontal and vertical datums are still relevant because they form the basis of map products all over the world. For the next few years we still have to deal with both local and global datums, until the former are eventually phased out. During the transition period, we will need tools to transform coordinates from local horizontal datums to a global horizontal datum and vice versa (see Subsection 3.1.4). 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.

3.1.2 Coordinate systems

spatial coordinate systems

planar coordinate systems

Spatial data are special, because they are spatially referenced. Different kinds of coordinate systems are used to position data in space. Here we distinguish between *spatial* and *planar* coordinate systems. *Spatial* (or global) coordinate systems locate data either on the Earth's surface in a 3D space or on the Earth's reference surface (ellipsoid or sphere) in a 2D space. *Planar* coordinate systems, on the other hand, locate data on the flat surface of a map in a 2D space. Initially the 2D Cartesian coordinate system and the 2D polar coordinate system will be examined. This will be followed by a discussion of the geographic coordinate system in a 2D and 3D space and the geocentric coordinate system, also known as the 3D Cartesian coordinate system.

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 ellipsoid.[3]. Lines of equal longitude are called meridians and form ellipses (meridian ellipses) on the ellipsoid (Figure 3.9)

The latitude (ϕ) of a point P (Figure 3.10) is the angle between the ellipsoidal normal through P' and the equatorial plane. Latitude is zero on the Equator ($\phi = 0°$), and increases towards the two poles to maximum values of $\phi = +90°$ (N 90°) at the North Pole and $\phi = -90°$ (S 90°) at the South Pole.

The longitude (λ) of the point is the angle between the meridian ellipse that passes through Greenwich and the meridian ellipse containing the point in question. It is measured on the equatorial plane from the meridian of Greenwich ($\lambda = 0°$), either eastwards through $\lambda = +180°$ (E 180°) or westwards through $\lambda = -180°$ (W 180°).

Latitude and longitude represent the geographic coordinates (ϕ, λ) of a point P' (Figure 3.10) with respect to the selected reference surface. They are always given in angular units. For example, the coordinates for the City Hall in Enschede are:[4]

[3] The concept of geographic coordinates can also be applied to a sphere.
[4] This latitude and longitude refers to the Amersfoort datum. The use of a different reference surface will

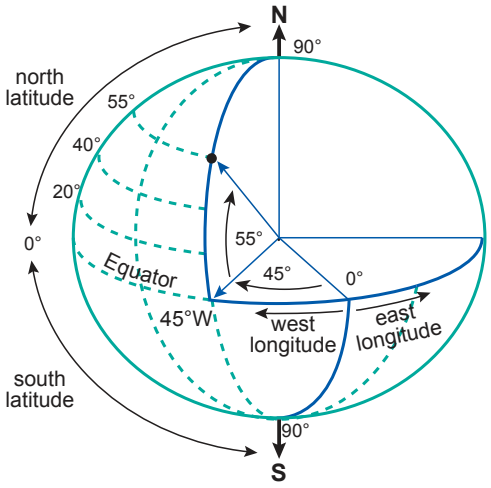

Figure 3.9
Latitude (ϕ) and longitude (λ) angles express the position of points in the 2D geographic coordinate system.

$$\phi = 52^\circ 13' 26.2'' N,\ \lambda = 6^\circ 53' 32.1'' E$$

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 of selected meridians and parallels (Figure 3.13). 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 ellipsoidal height (h) into 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 in Figure 3.10).

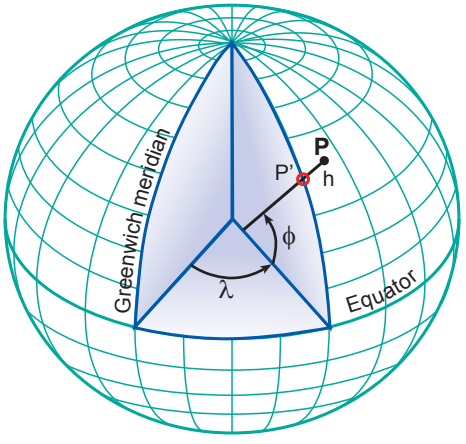

Figure 3.10
The angles of latitude (ϕ) and longitude (λ) and the ellipsoidal height (h) represent the 3D geographic coordinate system.

result in different angles of latitude and longitude.

3D geocentric coordinates (X, Y, Z)

An alternative method for defining a 3D position on the surface of the Earth is to use geocentric coordinates (X, Y, Z), also known as *3D Cartesian coordinates*. The system's origin lies at the Earth's centre of mass, with the X and Y axes on 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 3.11).

polar motion

The rotational axis of the Earth, however, changes 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) is used to define what is referred to as the "Conventional International Origin" (CIO).

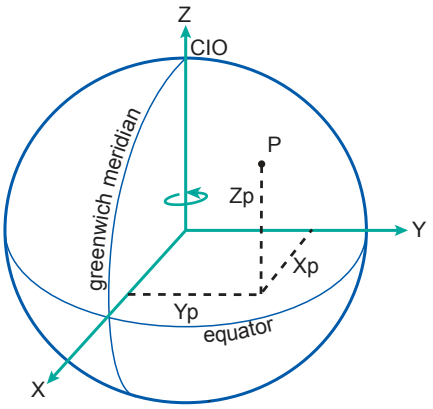

Figure 3.11
An illustration of the 3D geocentric coordinate system (see text for further explanation).

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 onto a two-dimensional map is the subject matter of map projections and coordinate transformations (Subsection 3.1.3 and Subsection 3.1.4). Here, as for several other cartographic applications, *two-dimensional Cartesian coordinates* (x, y), also known as *planar rectangular coordinates*, describe the location of any point unambiguously.

The 2D Cartesian coordinate system is one of intersecting perpendicular lines with the X-axis and the Y-axis as principal axes. The X-axis (the *Easting*) is the horizontal axis and the Y-axis (the *Northing*) is the vertical axis with an intersection at the *origin*. The plane is marked at intervals by equally-spaced coordinate lines that together form the *map grid* . Given two numerical coordinates x and y for point P, one can unambiguously specify any location P on the map (Figure 3.12).

false origin

Usually, the origin is assigned the coordinates $x = 0$ and $y = 0$. Sometimes, however, large positive values are added to the origin coordinates. This is to avoid negative values for the x and y coordinates in cases where the origin of the coordinate system is located inside the specific area of interest. The point that then has the coordinates $x = 0$ and $y = 0$ is called the *false origin*. The Rijksdriehoekstelsel (RD) in the Netherlands is an example of a system with a false origin. The system is based on the azimuthal double stereographic projection (see Section 3.1.3), with the Bessel ellipsoid used as reference surface. The origin was shifted from the projection centre (Amersfoort) towards the southwest(false origin)to avoid negative coordinates inside the country (see Figure 3.13).

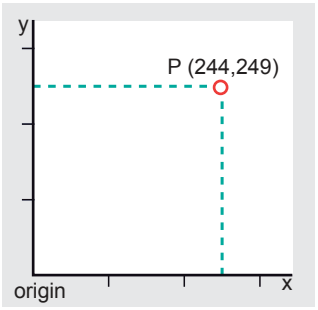

Figure 3.12
An illustration of the 2D Cartesian coordinate system (see text for further explanation).

The grid on a map represents lines having constant 2D Cartesian coordinates (Figure 3.13). 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 (about 1:1,000,000 or smaller). Scale distortions that result from transforming the Earth's curved surface to the mapping plane are so great on small-scale maps that detailed calculations and positioning become difficult.

map grid

2D Polar coordinates (α, d)

Another way of defining a point in a plane is by using polar coordinates. This 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*

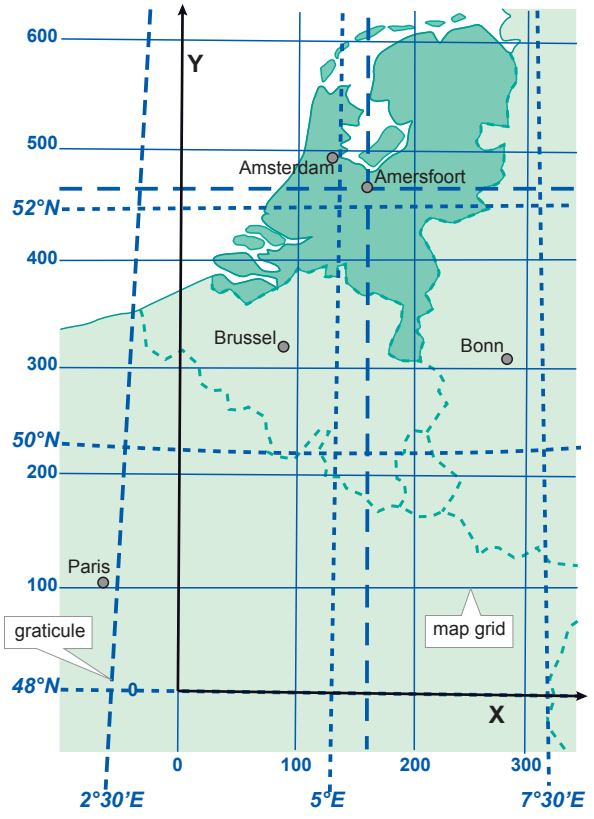

Figure 3.13
The coordinate system of the Netherlands represented by the map grid and the graticule. The origin of the coordinate system has been shifted (the false origin) from the projection centre (Amersfoort) towards the southwest.

and is measured in a clockwise direction. It is given in angular units while the distance d is expressed in length units.

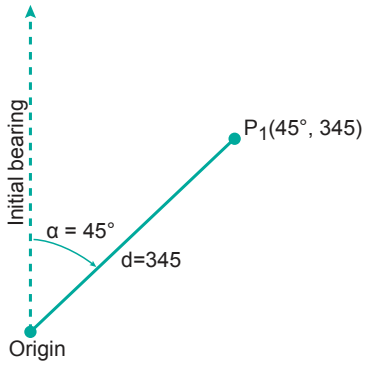

Figure 3.14
An illustration of the 2D Polar coordinate system (see text for further explanation).

Bearings are always related to a fixed direction (initial bearing) or a datum line. In principle, this reference line can be chosen freely. Three different, widely used fixed directions are: *True North*, *Grid North* and *Magnetic North*. The corresponding bearings are true (or geodetic) bearings, grid bearings and magnetic (or compass) bearings, respectively.

polar coordinates

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 a virtually universal preference for the polar coordinate method for detailed surveys.

3.1.3 Map projections

Maps are one of the world's oldest types of document. In the days that our planet was thought to be *flat*, a map was simply a miniature representation of a part of the world. To represent the specifically curved Earth's surface, a map needs to be a flattened representation of a part of the planet. Map projection concerns itself with ways of translating the curved surface of the Earth into a flat, 2D map.

> *Map projection* is a mathematically described technique for representing the Earth's curved surface on a flat map.

To represent parts of the surface of the Earth on a flat, printed map or a computer screen, the curved horizontal reference surface must be mapped onto a 2D mapping plane. The reference surface for large-scale mapping is usually an oblate ellipsoid; for small-scale mapping it is a sphere.[5] 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) that represent positions on the map plane (Figure 3.15).

mapping equations

The actual mapping cannot usually be visualized as a true geometric projection, directly onto the mapping plane. Rather, it 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:

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

[5] In practice, maps at scales of 1:1,000,000 or smaller can use the mathematically simpler sphere without the risk of large distortions. At larger scales, the more complicated mathematics of ellipsoids is needed to prevent large distortions occurring on the map.

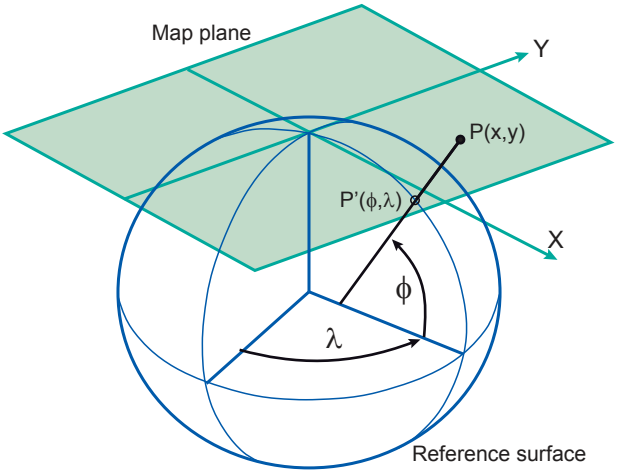

Figure 3.15
Example of a map projection in which the reference surface with geographic coordinates (ϕ, λ) is projected onto the 2D mapping plane with 2D Cartesian coordinates (x, y).

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:

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

The Mercator projection (spherical assumption) [106], a commonly used mapping projection, can be used to illustrate the use of mapping equations. The *forward mapping equation* for the Mercator projection is:[6]

$$x = R(\lambda - \lambda_0)$$
$$y = R \ln \tan \left(\frac{\pi}{4} + \frac{\phi}{2} \right)$$

The *inverse mapping equation* for the Mercator projection is:

$$\phi = \frac{\pi}{2} - 2 \arctan \left(e^{-\frac{y}{R}} \right)$$
$$\lambda = \frac{x}{R} + \lambda_0$$

Classification of map projections

Many map projections have been developed, each with its own specific qualities. It is these qualities that make the resulting maps useful for certain purposes. By definition, any map projection is associated with scale distortions. There is simply no way to flatten an ellipsoidal or spherical surface without stretching some parts of the surface more than others. The amount and kind of distortions a map has depends on the type of map projection.

scale distortions

Some map projections can be visualized as true geometric projections directly onto the mapping plane—known as an azimuthal projections—or onto an intermediate surface,

[6] When an ellipsoid is used as a reference surface, the equations are considerably more complicated than those introduced here. R is the radius of the spherical reference surface at the scale of the map; ϕ and λ are given in radians; λ_0 is the central meridian of the projection; $e = 2.7182818$, the base of the natural logarithms, not the eccentricity.

which is then rolled out onto the mapping plane. Typical choices for such intermediate surfaces are cones and cylinders. These map projections are called conical or cylindrical projections, respectively. Figure 3.16 shows the surfaces involved in these three classes of projection.

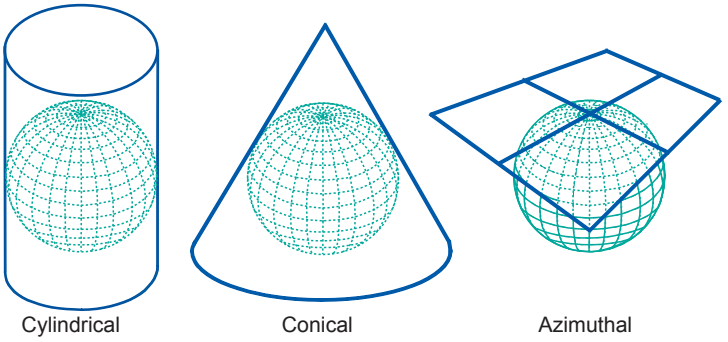

Figure 3.16
Classes of map projections

The azimuthal, conical, and cylindrical surfaces in Figure 3.16 are all *tangent* surfaces, i.e. they touch the horizontal reference surface at one point (azimuthal), 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; see Figure 3.17. Then, the reference surface is intersected along one closed line (azimuthal) or two closed lines (cone and cylinder). Secant map surfaces are used to reduce or average out scale errors since the line(s) of intersection are not distorted on the map.

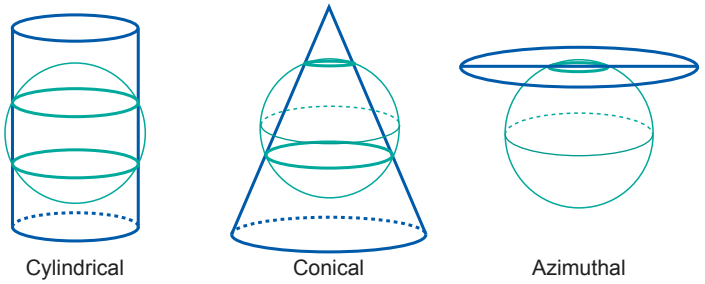

Figure 3.17
Three classes of secant projection

normal, transverse, and oblique projections

In the geometric depiction of map projections in Figures 3.16 and 3.17, the symmetry axes of the plane, cone and cylinder coincide with the rotation axis of the ellipsoid or sphere, i.e. a line through the North and South poles. In this case, the projection is said to be a *normal projection*. The other cases are *transverse projections* (symmetry axis in the equatorial plane) and *oblique projections* (symmetry axis is somewhere between the rotation axis and the equator of the ellipsoid or sphere); see Figure 3.18.

The Universal Transverse Mercator (UTM) is a system of map projection that is used worldwide. It is derived from the Transverse Mercator projection (also known as Gauss-Kruger or Gauss conformal projection). UTM uses a transverse cylinder secant to the horizontal reference surface. It 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.

Normal cylindrical projections are typically used to map the world in its entirety. Conical projections are often used to map individual continents, whereas the normal az-

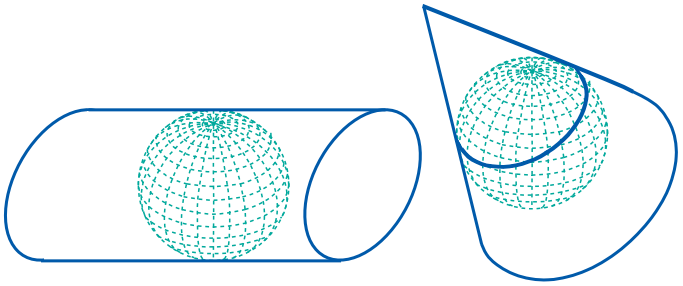

Figure 3.18
A transverse cylindrical and an oblique conical projection

imuthal projection may be used to map polar areas. Transverse and oblique aspects of many projections can be used for most parts of the world.

It is also important to consider the shape of the area to be mapped. Ideally, the general shape of the mapping area should be well-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. Cylindrical projection is best for a rectangular area and conic projection for a triangular area.

So far, we have not specified *how* the curved horizontal reference surface is projected onto a plane, cone or cylinder. *How* this is done determines what 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:

distortion properties

- With 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.
- With 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.
- With 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 (taking into account the map scale).

A particular map projection can exhibit only one of these three properties. No map projection can be both conformal and equal-area, for example.

The most appropriate type of distortion for a map depends largely on the purposes for which the map will be used. Conformal map projections represent angles correctly, but as the region becomes larger they show considerable area distortions (Figure 3.19). Maps used for the measurement of angles (e.g. aeronautical charts, topographic maps) often make use of a conformal map projection such as the UTM projection.

Equal-area projections, on the other hand, represent areas correctly, but as the region becomes larger, considerable distortions of angles and, consequently, shapes occur (Figure 3.20). Maps that are to be used for measuring area (e.g. distribution maps) are often made using an equal-area map projection.

The equidistant property is achievable only to a limited degree. That is, true distances can be shown only from one or two points to any other point on the map, or in certain directions. If a map is true to scale along the meridians (i.e. no distortion in the

Figure 3.19
The Mercator projection, a cylindrical map projection with conformal properties. The area distortions are significant towards the polar regions.

Figure 3.20
The cylindrical equal-area projection, i.e. a cylindrical map projection with equal-area properties. Distortions of shapes are significant towards the poles.

North–South direction), we say that the map is *equidistant along the meridians* (e.g. an equidistant cylindrical projection) (Figure 3.21). If a map is true to scale along all parallels we say the map is *equidistant along the parallels* (i.e. no distortion in the East–West direction). Maps for which the area and angle distortions need to be reasonably acceptable (several thematic maps) often make use of an equidistant map projection.

As these discussions indicate, a particular map projection can be classified. An example would be the classification "conformal conic projection with two standard parallels", which means that the projection is a conformal map projection, that the intermediate surface is a cone, and that the cone intersects the ellipsoid (or sphere) along two parallels. In other words, the cone is secant and the cone's symmetry axis is parallel to the rotation axis. (This would amount to the middle projection displayed in Figure 3.17). This projection is also referred to as "Lambert's conical projection" [47].

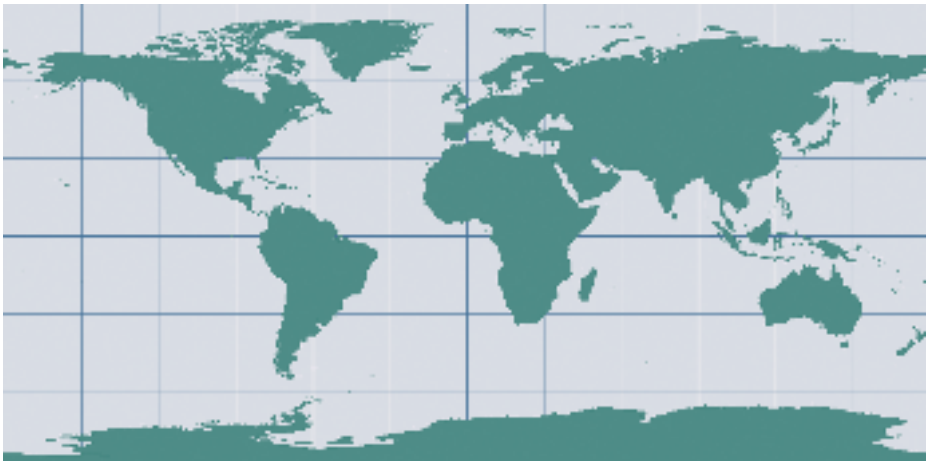

Figure 3.21
The equidistant cylindrical projection (also called Plate Carrée projection), a cylindrical map projection with equidistant properties. The map is equidistant (true to scale) along the meridians. Both shape and area are reasonably well preserved.

3.1.4 Coordinate transformations

Users of geoinformation often need transformations from a particular 2D coordinate system to another system. This includes the transformation of polar coordinates 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. This transformation is based on relating the two systems on the basis of a set of selected points whose coordinates are known in both systems, such as ground control points or common points such as corners of houses or road intersections. 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, i.e. angular and distance measurements, are transformed into map coordinates. The equation for this transformation is:

$$x = d \sin \alpha$$
$$y = d \cos \alpha$$

The inverse equation is:

$$\alpha = \arctan\left(\frac{x}{y}\right)$$
$$d^2 = x^2 + y^2$$

Changing map projection

Forward and inverse mapping equations are normally used to transform data from one map projection into 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 3.22.

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

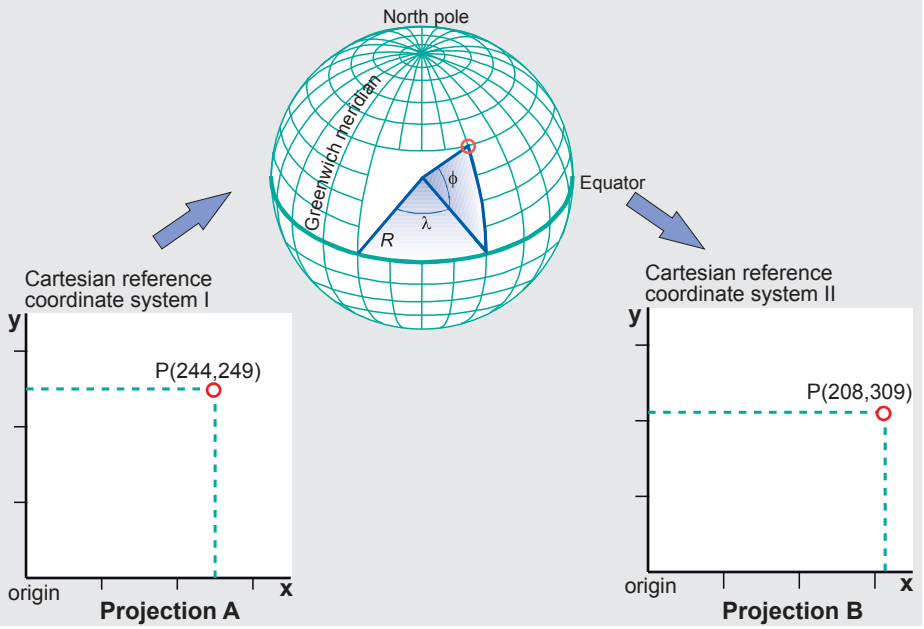

Figure 3.22
The principle of changing from one map projection to another.

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 neighbouring countries or between overlaid maps originating from different projections. It may lead to differences of several hundreds of metres in the resulting coordinates. Therefore, spatial data with different underlying horizontal datums may require *datum transformation*.

Suppose we wish to transform spatial data from the UTM projection to the Dutch RD system, and suppose that the data in the UTM system are related to the European Datum 1950 (ED50), while the Dutch RD system is based on the Amersfoort datum. To achieve a perfect match, in this example the change of map projection should be combined with a datum transformation step; see Figure 3.23.

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) for 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 coordinates (ϕ, λ, h) for datum B. Finally, the forward equation of projection B takes us from the geographic coordinates (ϕ, λ, h) for datum B to the map coordinates (x', y') of projection B.

Mathematically, a datum transformation is feasible via the geocentric coordinates (x, y, z)

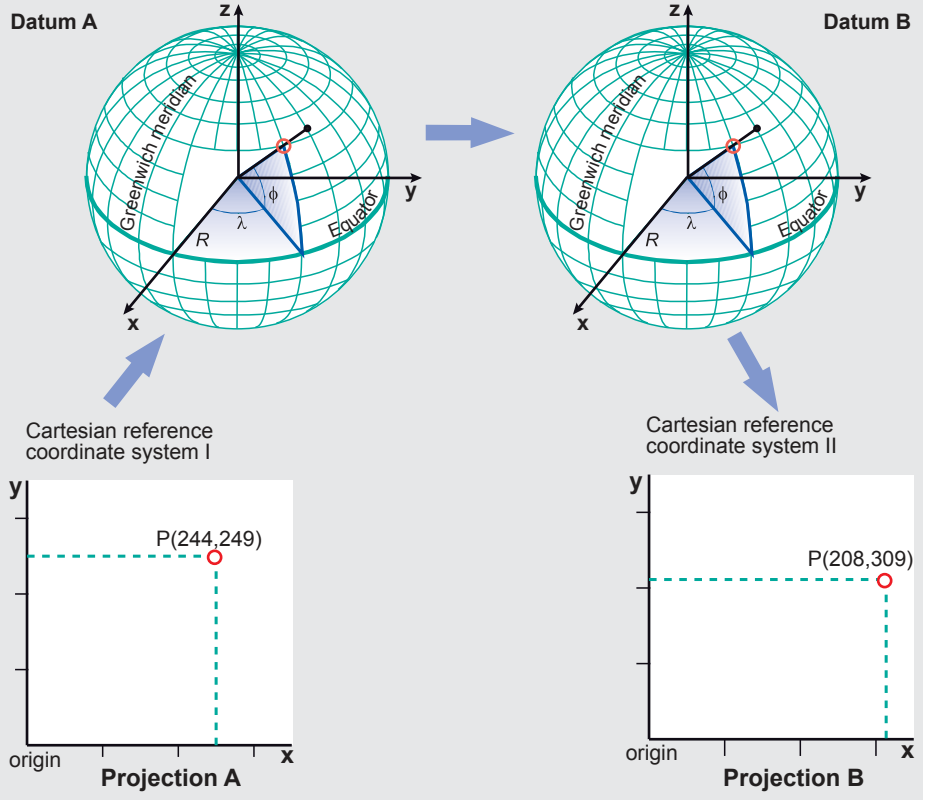

Figure 3.23
The principle of changing from one projection into another, combined with a datum transformation from datum A to datum B.

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 [59].

Geographic coordinates $(\phi,\ \lambda,\ h)$ can be transformed into geocentric coordinates $(x,\ y,\ z)$, and vice versa. The datum transformation via the geocentric coordinates implies a 3D similarity transformation. This is essentially 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 $(\alpha,\ \beta,\ \gamma)$, three origin shifts $(X_0,\ Y_0,\ Z_0)$ and a scale factor (s). The inputs are the coordinates of points in datum A and coordinates of the same points in datum B. The output are estimates of the seven transformation parameters and a measure of the likely error of the estimate.

datum transformation parameters

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 points are not correct—often the case for points measured on a local datum system—the estimated parameters may be inaccurate and hence the datum transformation will be inaccurate.

Inaccuracies often occur 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 metres 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 estimation depends on the particular choice of common points and whether all

seven transformation parameters, or only some of them, are estimated.

The example in Table 3.2 illustrates the transformation of the Cartesian coordinates of a point in the state of Baden-Württemberg, Germany, from ITRF to Cartesian coordinates in the Potsdam Datum. Sets of numerical values for the transformation parameters are available from three organizations:

Table 3.2
Three different sets of datum transformation parameters from three different organizations for transforming a point from ITRF to the Potsdam datum.

	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 (m)	X_0	-581.99	-518.19	-635
	Y_0	-105.01	-43.58	-27
	Z_0	-414.00	-466.14	-450

1. The federal mapping organization of Germany (labelled "National set" in Table 3.2) provided a set calculated using common points distributed throughout Germany. This set contains all seven parameters and is valid for whole Germany.
2. The mapping organization of Baden-Württemberg (labelled "Provincial set" in Table 3.2) provided a set calculated using common points distributed throughout the state of Baden-Württemberg. This set contains all seven parameters and is valid only within the state borders.
3. The National Imagery and Mapping Agency (NIMA) of the U.S.A. (labelled "NIMA set" in Table 3.2) provided a set calculated using common points distributed throughout Germany and based on the ITRF. This set contains a coordinate shift only (no rotations, and scale equals unity). This set is valid for whole Germany.

The three sets of transformation parameters vary by several tens of metres, for reasons already mentioned. The sets of transformation parameters were used to transform the ITRF cartesian coordinates of a point in the state of Baden-Württemberg. Its ITRF (X, Y, Z) coordinates are:

$$(4,156,939.96\ \text{m},\ 671,428.74\ \text{m},\ 4,774,958.21\ \text{m}).$$

The three sets of transformed coordinates for the Potsdam datum are given in Table 3.3.

Table 3.3
Three sets of transformed coordinates for a point in the state of Baden-Württemberg, Germany.

Potsdam coordinates	National set (m)	Provincial set (m)	NIMA set (m)
X	$4,156,305.32$	$4,156,306.94$	$4,156,304.96$
Y	$671,404.31$	$671,404.64$	$671,401.74$
Z	$4,774,508.25$	$4,774,511.10$	$4,774,508.21$

The three sets of transformed coordinates differ by only a few metres from each other. In a different country, the level of agreement could be a within centimetres, but it can be up to tens of metres of each other, depending upon the quality of implementation of the local horizontal datum.

3.2 Satellite-based positioning

The importance of satellites in spatial referencing has already been mentioned before. Satellites have allowed us to create geocentric reference systems and to increase the level of spatial accuracy substantially. Satellite-based systems are critical tools in geodetic engineering for the maintenance of the ITRF. They also play a key role in mapping and surveying in the field, as well as in a growing number of applications requiring positioning techniques. The setting up a satellite-based positioning system requires the 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 components of the space segment;
3. the *user segment*, i.e. the users, along with the hardware and software they use for positioning.

In satellite positioning, the central problem is to determine the values (X, Y, Z) of a receiver of satellite signals, i.e. to determine the position of the receiver with the accuracy and precision required. The degree of accuracy and precision needed depends on the application, as does timeliness, i.e. are the position values required in real time or can they be determined later during post-processing. Finally, some applications, such as navigation, require kinematic approaches, which take into account the fact that the receiver is not stationary, but moving.

Some fundamental aspects of satellite-based positioning and a brief review of currently available technologies follows.

3.2.1 Absolute positioning

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

1. A satellite, equipped with a clock, sends a radio message at a specific moment 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.
3. 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*. This *pseudorange* is the apparent distance from satellite to receiver, computed from the time delay with which its radio signal is received.

Such a computation places the position of the receiver on a sphere of radius equal to the computed pseudorange (see Figure 3.24a). If, instantaneously, the receiver were to do the same with a message from another satellite positioned elsewhere, the position of the receiver would be placed on another sphere. The intersection of the two spheres,

trilateration

which have different centres, describes a circle as being the set of possible positions of the receiver (see Figure 3.24b). If a third satellite message is taken into consideration, the three spheres intersect at, at most, two positions, one of which is the actual position of the receiver. In most, if not all practical situations where two positions result, one of them is a highly unlikely position for a signal receiver, thus narrowing down the true position of the receiver. The overall procedure is known as *trilateration*: the determination of a position based on three distances.

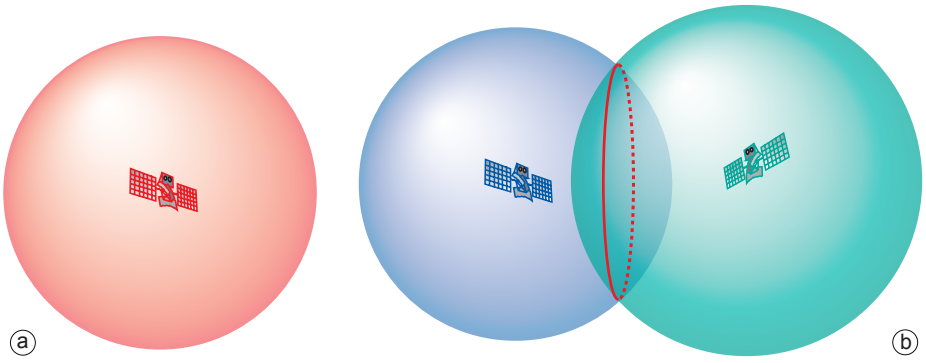

Figure 3.24
Pseudorange positioning: (a) With just one satellite, the receiver position is somewhere on a sphere, (b) With two satellites, the position is located where the two spheres intersect, i.e. in a circle. Not shown: with three satellites, its position is where the three spheres intersect.

clock bias

It would appear, therefore, that the signals of three satellites would be sufficient to determine a *positional fix* for our receiver. In theory this is true, but in practice it is not. The reason being that satellite clocks and the receiver clock are never exactly synchronized. Satellite clocks are costly, high-precision, atomic clocks that we can consider synchronized for the time being, but the receiver typically uses a far cheaper, quartz clock that is not synchronized with satellite clocks. This brings an additional unknown variable into play, namely the synchronization bias of the receiver clock, i.e. the difference in time readings between it and the satellite clocks.

3D positioning

Our set of unknown variables has now become $(X, Y, Z, \Delta t)$ representing a 3D position and a clock bias. The problem can be solved by including the information obtained from a fourth satellite message, (see Figure 3.25). 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 synchronicity. One such application is the use of wireless sensor networks for researching natural phenomena such as earthquakes or meteorological patterns, and for water management.

The positioning of mobile phone users making an emergency call is yet another application. Often callers do 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 ranging from 300 m to 6 km. That is far too inaccurate for emergency services. If all masts in the telephony network are equipped with a satellite positioning receiver (and thus, with a very high-precision synchronized clock), however, 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 describes a hyperbola on the ground of possible positions of the caller. If the call is received on three masts, two hyperbolas are described, allowing intersection and thus "hyperbolic positioning". With current technology the (horizontal) accuracy would be better than 30 m.

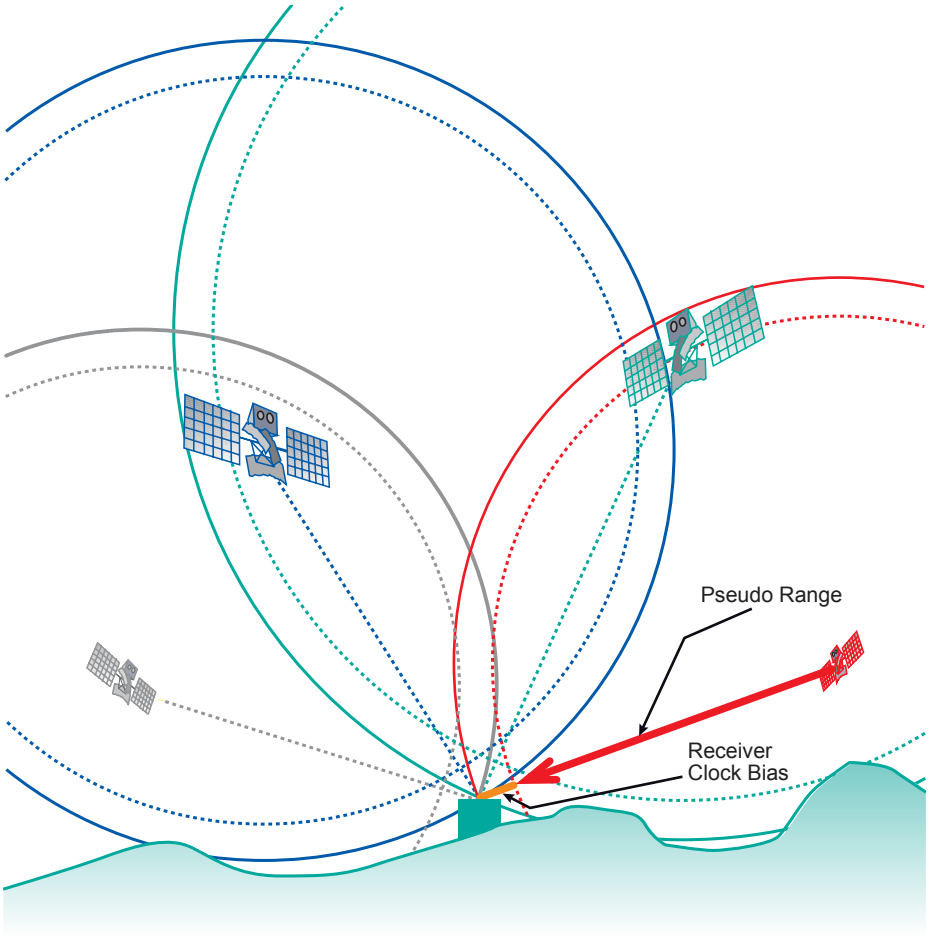

Figure 3.25
Four satellites are needed to obtain a 3D position fix. Pseudoranges are indicated for each satellite as dotted circles, representing a sphere; the actual range is represented as a solid circle, which is the pseudorange plus the range error caused by receiver clock bias.

Returning to satellite-based positioning, 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.

2D positioning mode

If a 3D fix had 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 been obtained, the receiver assumes that it is positioned at the geocentric ellipsoid adopted by the positioning system, i.e. at height $h = 0$.[7] 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 upon which this computation is based is flawed and the resulting positioning 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.

[7] Any receiver is capable of transforming a coordinate $(X,\ Y,\ Z)$, using a straightforward mathematical transformation, into an equivalent coordinate $(\phi,\ \lambda,\ h)$, where h is the height above the geocentric ellipsoid.

Time, clocks and world time

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

Greenwich Mean Time

GMT was later replaced by Universal Time (UT), a system still based on meridian crossings of stars, albeit distant quasars, as this approach provides more accuracy than that based on 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 civilian clock time, but it has now officially been replaced by International Atomic Time (TAI). UT actually has various versions, among them UT0, UT1 and UTC. UT0 is the Earth's rotational time observed at 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. Nevertheless, UT1 is still a somewhat erratic clock system because of the varying rotational velocity of the planet, as mentioned above. The degree of uncertainty is about 3 ms 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 s of UT1, and twice annually it may be shifted to stay within that margin. This occasional shift of a *leap second* is applied at the end of 30 June or, preferably, at the end of 31 December. The last minute of such a day is then either 59 or 61 seconds long. So far, adjustments have always entailed adding 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.

In recent years, we have learned to measure distance, and therefore also position, with clocks, by using satellite signals, the conversion factor being the speed of light, approximately 3×10^8 m s^{-1} in a vacuum. As a consequence, multiple seconds of clock bias could no longer be accepted, and this is where atomic clocks are at an advantage. They are very accurate time keepers, based on the exact frequencies at 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

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

3.2.2 Errors in absolute positioning

Before we continue discussing other modes of satellite-based positioning, let us take a close look at the potential for error in absolute positioning. Users of receivers are required to be sufficiently familiar with the technology in order to avoid real operating blunders such as poor receiver placement or incorrect receiver software settings, which can render positioning results virtually useless. We will skip over many of the physical and mathematical details underlying these errors; they are only mentioned

here to raise awareness and understanding among users of this technology. For background information on the calculation of positional error (specifically, the calculation of RMSE or *root mean square error*), see Subsection 5.3.2.

Errors related to the space segment

As a first source of error, operators of the control segment may, for example in times of global political tension or war, intentionally deteriorate radio signals from satellites to the general public to avoid optimal use of the system by a perceived enemy. This *selective availability*—meaning that military forces allied with the control segment *will* still have access to undisturbed signals—may cause error that has an order of magnitude larger than all other error sources combined.[8]

A second source occurs if the satellite signal contains 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 thanks to Einstein we know that moving clocks are slower than stationary clocks, due to a relativistic effect. If one understands that a clock that is off by 0.000001 s causes an computation error in the satellite's pseudorange of approximately 300 m, it becomes 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 gravitation pull of the Earth that a satellite experiences in orbit is not simple either. Moreover, satellite orbits are also disturbed by solar and lunar gravitation, making flight paths 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 such as those discussed above with its accurately *known* position, flagging any unacceptable errors and potentially labelling a satellite as temporarily "unhealthy" until those errors have been corrected and brought back within the agreed tolerance limits. This may be done by uploading a correction to the clock or the satellite's orbit settings.

Errors related to the medium

A third source may be due to the influence of the *medium* between sender and receiver on the satellite's radio signals. The middle atmospheric layers of the stratosphere and mesosphere are relatively harmless and of little hindrance to radio waves, but this is not true of the lower and upper layers of the atmosphere:

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

- *The ionosphere*: the part of the atmosphere that is farthest from the Earth's surface. It starts at an altitude of 90 km and holds many electrically charged atoms,

[8] Selective availability was stopped at the beginning of May 2000; in late 2007 the White House decided to remove selective availability capabilities all together. However, when deemed necessary, the US government still has a range of capabilities and technology available to implement regional denial of service of civilian GPS signals in an area of conflict, effectively producing the same result.

thereby forming a protective "shield" 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 varies with solar activity.

The ionosphere is a more severe source of delay for satellite signals, which obviously means that pseudoranges are estimated as being 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, which enables correction for atmospheric delay, leading to a 10–50% improvement of accuracy. If this is not the case, or if the receiver is capable of receiving only a single frequency, a model should be applied to forecast the (especially ionospheric) delay; typically the model takes into account the time of day and current latitude of the receiver.

Errors related to the receiver's environment

Fourth in the list of sources of error is that which occurs when a radio signal is received via two or more paths between sender and receiver, typically caused by the signal bouncing off some nearby surface such as 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 (see Figure 3.26). Multi-path is a source of error that is difficult to avoid.

multi-path error

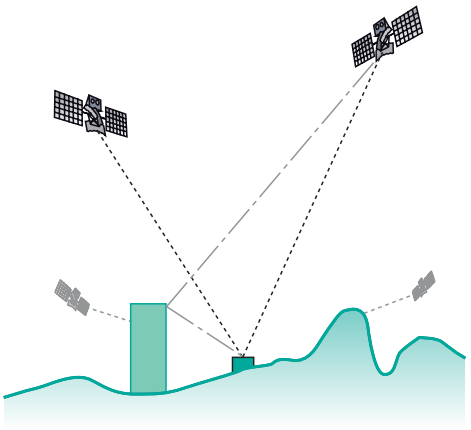

Figure 3.26
At any point in time, a number of satellites will be above the receiver's horizon. But not all of them will be "in view" (e.g. the satellites on the far left and right); and for others, multi-path signal reception may occur.

All of the above sources of error influence computation of a satellite's pseudorange. Cumulatively, they are called the *user equivalent range error* (UERE). Some error sources may affect all satellites being used by a particular receiver, e.g. selective availability and atmospheric delay, while others may be specific to one satellite, for instance, incorrect satellite information and multi-path.

range error

Errors related to the relative geometry of satellites and receiver

There is one more source of error, which is unrelated to individual radio signal characteristics: rather, this error is the result of the combination of signals from satellites used for positioning. The constellation of satellites in the sky from the receiver's perspective is the controlling factor in these cases. Referring to Figure 3.27, the sphere-intersection technique of positioning provides more precise results when the four satellites are evenly spread over the sky; the satellite constellation of Figure 3.27b is preferred over that of 3.27a. This source of error is know 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 char-

geometric dilution of precision

acteristics, 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 possible positional fix from all the signals. This gives a better solution that obtained just using the "best four", as was done previously.

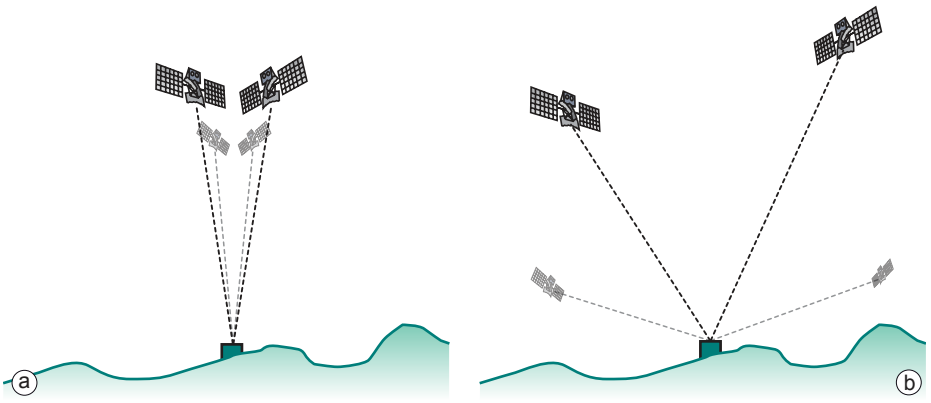

Figure 3.27
Geometric dilution of precision. The four satellites can be in a poor constellation for positioning(a) or in a better constellation (b).

Table 3.4
Indication of typical magnitudes of error in absolute satellite-based positioning

satellite clock (m)	2
satellite position (m)	2.5
ionospheric delay (m)	5
tropospheric delay (m)	0.5
receiver noise (m)	0.3
multi-path (m)	0.5
Total RMSE Range error (m): $\sqrt{2^2 + 2.5^2 + 5^2 + 0.5^2 + 0.3^2 + 0.5^2} =$	5.97

These errors are not all of similar magnitude. An overview of some typical values (without selective availability) is provided in Table 3.4. 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, means of determining a position using satellites. It suffices for recreational purposes and other applications that require horizontal accuracies to within 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.

3.2.3 Relative positioning

One technique for trying to remove errors from positioning computations is to perform many position computations, and to determine the average over all solutions. Many receivers allow the user to do this. It should, however, be clear from the above that *averaging* may address *random* errors such as signal noise, selective availability (SA) and multi-path to some extent, but not *systematic* sources of error, such as 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 min in absolute, single-point positioning based on code measurements, before systematic error removal, leads to only a 10–20% improvement of accuracy. In such cases, receiver

random and systematic error

averaging is therefore of limited value and requires near-optimal conditions for long periods. 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 sources of systematic error by taking into account measurements of these errors in a nearby stationary *reference receiver* that has an accurately known position. By using these systematic error findings for the reference receiver, the position of the *target receiver* of interest can be determined much more precisely.

In an optimal setting, the 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 be 70–200 km apart; they will experience essentially similar atmospheric signal error. Selective availability can also be addressed in this away.

For each satellite in view, the reference receiver will determine its pseudorange error. After all, its position is known to a high degree of 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 satellite signals that it uses for positioning. In doing so, it can improve its accuracy to within 0.5–1 m.

The discussion above assumes we needed positioning information in real time, which called for a data link between reference and target receiver. But various uses of satellite-based positioning do not need real time data, making post-processing of the recorded positioning data suitable. If the target receiver records time and position accurately, correctional data can be used later to improve the accuracy of the originally recorded data.

Finally, mention should be made of the notion of *inverted relative positioning*. The principles are still the same as above, but with this technique the target receiver does not correct for satellite pseudorange error, but rather uses a data link to upload its positioning/timing information to a central repository, where the corrections are applied. This can be useful in cases where many target receivers are needed and budget does not allow them to be expensive.

3.2.4 Network positioning

Now that the advantages of relative positioning have been discussed, we can move on to the notion of *network positioning*: an integrated, systematic network of reference receivers covering a large area, perhaps an entire continent or even the whole globe.

The organization of such a network can take different shapes, augmenting an already existing satellite-based system. Here we discuss a general architecture, consisting of a network of *reference stations*, strategically positioned in the area to be covered, each of them 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 any 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 sub-metre 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 neighbourhood as a reference receiver to simple ones.

3.2.5 Code versus phase measurements

Up until this point, we have assumed that the receiver determines the range of a satellite by measuring time delay of 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. Currently, carrier phase measurement can 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 L-band frequencies used is 19–24 cm. Since the number of cycles of the signal cannot be measured directly, it is determined (in a long observation session) from the change in carrier phase over time. Such a change occurs because the satellite is orbiting. 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 large as 10 cm per year for some locations on the planet.

3.2.6 Positioning technology

This section provides information on currently available satellite-based positioning technology. At present, two satellite-based positioning systems are operational—GPS and GLONASS—and a third is in the implementation phase—Galileo. These systems are US, Russian and European, respectively. Any of them, but especially GPS and Galileo, will be improved over time and will be augmented with new techniques.

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; civilians throughout the world have access to Standard Positioning Services (SPS). The GPS space segment nominally consists of 24 satellites, each of which orbits our planet in 11 h 58 min 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, at an angle of inclination of 55–63° to the equatorial plane; nominally four satellites orbit in each plane (see Figure 3.28). This means that a receiver on Earth will have between five and eight (rarely, even up to 12) satellites in view at any moment in time. Software packages exist to help in planning GPS surveys, identifying the expected satellite set-up for any location and time.

The NAVSTAR satellites transmit two radio signals, an L1 frequency of 1575.42 MHz and an L2 frequency of 1227.60 MHz. There is also a third and fourth signal, but these are not important for the discussion here. The role of the L2 signal is to provide a second radio signal, thereby providing a way, with (more expensive) dual-frequency receivers, of determining fairly precisely the actual ionospheric delay of the satellite signals received.

GPS uses WGS84 as its reference system,which has been refined on several occasions and is now aligned with the ITRF at the level of a few centimetres worldwide. (See also Section 3.1.1.) GPS has adopted UTC as its time system.

WGS84

For civilian applications, GPS receivers of varying quality are available, their quality depending on the embedded positioning features: supporting single- or dual-frequencies; supporting only absolute or also relative positioning; performing code

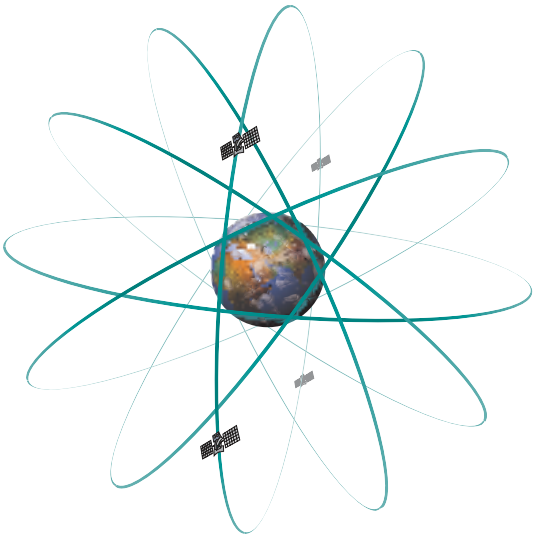

Figure 3.28
Constellation of satellites in the GPS system; here four satellites are shown in only one orbital plane.

measurements or also carrier phase measurements.

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, but GLONASS did not significantly develop civil applications as GPS did and thus it is commercially less important.

GLONASS's space segment consists nominally of 24 satellites, organized in three orbital planes, at an inclination of 64.8° to the Equator. Its orbiting altitude is 19,130 km, with a period of revolution of 11 h 16 min.

GLONASS uses the PZ-90 as its reference system and, like GPS, uses UTC as its time reference, albeit with an offset for Russian daylight.

GLONASS's radio signals are somewhat similar to those of GPS, differing only in the details: the frequency of GLONASS's L1 signal is approximately 1605 MHz (changes are underway), and its L2 signal approximately 1248 MHz; otherwise, GLONASS's system performance is rather comparable with that of GPS.

Galileo

In the 1990s, the European Union (EU) judged that it needed its own satellite-based positioning system, to become independent of the GPS monopoly and to support its own economic growth by providing services of high reliability under civilian control.The EU system is named Galileo.

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 development of the system has experienced substantial delays; currently European ministers insist that Galileo should be up and running by the end of 2013.

When completed, Galileo will have 27 satellites, with three in reserve, orbiting in one of three, equally spaced, circular orbits at an elevation of 23,222 km and inclined at 56° to the Equator. This higher inclination (when compared to that of GPS) has been chosen to provide better positioning coverage at high latitudes, such as in northern

Scandinavia, where GPS performs rather poorly.

In June 2004, the EU and the US agreed to make Galileo and GPS compatible by adopting interchangeable set-ups for their satellite signals. The effect of this agreement is that a 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 signal reception was poor, in built-up areas and forests, for instance. It will also bring the implementation of a Global Navigation Satellite System (GNSS) closer, since positional accuracy and reliability will improve. Such a system would bring the ultimate development of fully automated air and road traffic control systems much closer. Automatic aircraft landing, for instance, requires a horizontal accuracy in the order of 4 m, and a vertical accuracy of less than 1 m. Currently, these requirements cannot be reliably met.

The Galileo Terrestrial Reference Frame (GTRF) will be a realization of the ITRS and will be set up independently from that of GPS so that one system can back up the other. Positional differences between WGS84 and GTRF will be at worst only a few centimetres.

The Galileo System Time (GST) will closely follow International Atomic Time (TAI), with a time offset of less than 50 ns for 95% of the time over any period of the year. Information on the actual offset between GST and TAI, and between GST and UTC (as used in GPS), will be broadcast in the Galileo satellite signal.

Satellite-based augmentation systems

Satellite-based augmentation systems (SBAS) aim to improve the accuracy and reliability of satellite-based positioning (see Subsection 3.2.4) in support of safety-critical navigation applications, such as aircraft operations near airfields. Typically, these systems make use of an extra, now geostationary, satellite that has a large service area, for example a continent, and which sends differential data about standard positioning satellites that are currently in view in its service area. If multiple ground reference stations are used, the quality of the differential data can be quite good and reliable. Usually this satellite will use radio signals of the same frequency as those in use by the positioning satellites, so that receivers can receive the differential code without problem.

Not all advantages of satellite augmentation will be useful for all receivers. For consumer market receivers, the biggest advantage, as compared to standard relative positioning, is that SBAS provides an ionospheric correction grid for its service area, from which a correction specific for the location of the receiver can be retrieved. This is not true in relative positioning, where the reference station determines the error it experiences and simply broadcasts this information for nearby target receivers to use. With SBAS, the receiver obtains information that is best viewed as a geostatistical interpolation of errors from multiple reference stations.

More advanced receivers will be able to deploy also other differential data such as corrections on satellite position and satellite clock drift.

Currently, three systems are operational: for North America WAAS (Wide-Area Augmentation System) is in place; EGNOS (European Geostationary Navigation Overlay Service) for Europe; and MSAS (Multi-functional Satellite Augmentation System) for eastern Asia. The ground segment of WAAS consists of 24 control stations, spread over North America; that of EGNOS has 34 control stations. These three systems are compatible, guaranteeing international coverage.

Usually signals from the geostationary SBAS satellites (under various names, such as AOR, Artemis, IOR, Inmarsat, MTSAT) can be received even outside their respective

service areas. But the use of these signals there must be discouraged, as they will not help improve positional accuracy. Satellite identifiers, as shown by the receiver, have numbers above 30, setting them apart from standard positioning satellites.

Though originally intended to improve the safety of aircraft landings, SBAS, with its horizontal accuracy to within 3 m, has many other uses. At this level of accuracy, vehicle position can be determined to a specific road lane, and "automatic pilots" become a possibility.