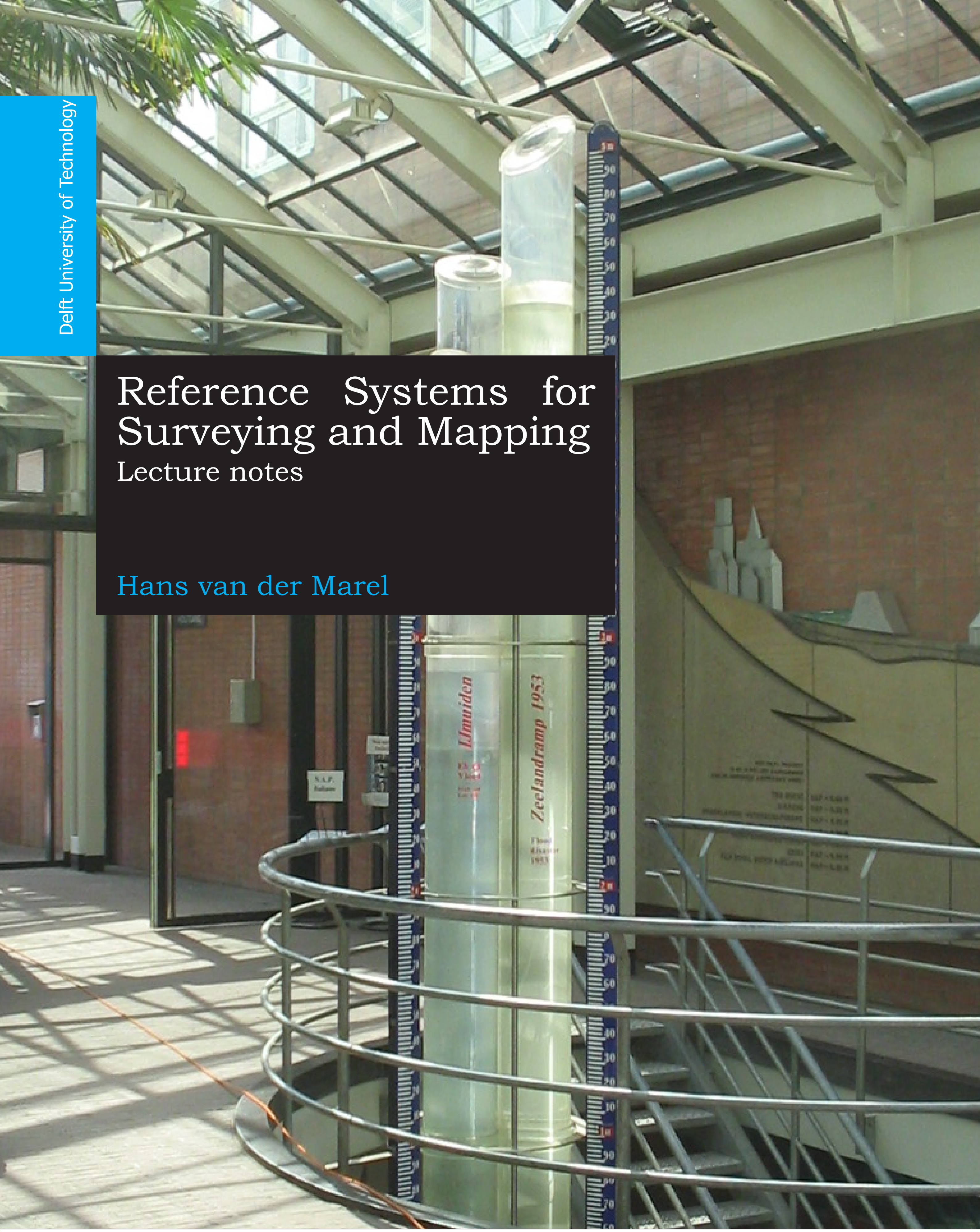
Delft University of Technology



Reference Systems for Surveying and Mapping

Lecture notes

Hans van der Marel

Delft University of Technology

Challenge the future

The front cover shows the NAP (Amsterdam Ordnance Datum) ”datum point” at the Stopera, Amsterdam (picture M.M.Minderhoud, Wikipedia/Michiel1972).

H. van der Marel

Surveying and Mapping
Lecture notes CTB3310 (Reference Systems for Surveying and Mapping)

August 2014

Publisher:
Faculty of Civil Engineering and Geosciences
Delft University of Technology
P.O. Box 5048
Stevinweg 1
2628 CN Delft
The Netherlands

The text has been type set using the MikTex 2.9 implementation of LaTeX. Graphs and diagrams were produced, if not mentioned otherwise, with Matlab and Inkscape.

Contents

1 Introduction

Surveying and mapping deal with the description of the shape of the Earth, spatial relationships between objects near the Earth surface, and data associated these. Mapping is concerned with the (scaled) portrayal of geographic features and visualization of data in a geographic framework. Mapping is more than the creation of paper maps: contemporary maps are mostly digital, allow multiple visualizations and analysis of the data in Geographic Information Systems (GIS). Surveying is concerned with accurately determining the terrestrial or three-dimensional position of points and the distances and angles between them. Points describe the objects and shapes that appear on maps and in Geographic Information Systems. These points are usually on the surface of the Earth. They are used to connect topographic features, boundaries for ownership, locations of buildings, location of subsurface features (pipe-lines), and they are often used to monitor changes (deformation, subsidence). Points may only exist on paper, or in a CAD system, and represent points to be staked out for construction work.

To describe the position of points a mathematical framework is needed. This mathematical framework consist of a *coordinate reference system*. A coordinate system uses one or more numbers, or *coordinates*, to uniquely determine the position of a point in a 2D or 3D Euclidean space. In this reader several of these mathematical frameworks are described and how they are used in surveying and mapping.

In Chapters 3.1 and 3 two and three dimensional Cartesian coordinate systems are introduced. Although straightforward, 3D Cartesian coordinates are not very convenient for describing positions on the surface of the Earth. It is actually more convenient to use curvilinear coordinates, or, to project the curved surface of the Earth on a flat plane. Curvilinear coordinates, known as geographic coordinates, or latitude and longitude, are discussed in Chapter 4. The *map projections*, which result in easy to use 2D Cartesian coordinates, are covered in Chapter 5.

Coordinate conversions and datum transformations are discussed in Chapter 6. This topic can be somewhat bewildering for the inexperienced user because there are so many different coordinate types and datums in use, but fortunately any transformation can be decomposed into a few elementary coordinate conversions and a datum transformation.

The height always plays a special role in coordinate reference systems. Height is also closely associated with the flow of water and gravity. The height coordinate systems are discussed in Chapter 7. Finally in Chapter 8 several important and commonly used reference systems are described. They include the International Terrestrial Reference System (ITRS), the well know World Geodetic System (WGS84) used by GPS, the European Terrestrial Reference System (ETRS89), and the Dutch triangulation system (RD) and Dutch ”Normaal Amsterdams Peil” (NAP) height system.

2

2D Coordinate systems

2.1. 2D Cartesian coordinates

The position of a point P_i on a plane surface can be described by two coordinates, x_i and y_i, in a two dimensional (2D) Cartesian coordinate system, as illustrated in Figure 2.1a. The axes in the 2D Cartesian coordinate system, named after the 17th century mathematician René Descartes, are perpendicular (orthogonal), have the same scale, and meet in what is called the *origin*. The Cartesian coordinate system is *right-handed*, meaning, with the positive x-axis pointing right, the positive y-axis is pointing up[1]. Therefore, fixing or choosing one axis, determines the other axis. The coordinates (x_i, y_i) are defined as the distance from the origin to the perpendicular projection of the point P_i onto the respective axes. The point P_i can also be represented by a position vector $\mathbf{r}_i$ from the origin to the point P_i,

$$\mathbf{r}_i = x_i\mathbf{e}_x + y_i\mathbf{e}_y \ , \tag{2.1}$$

with $\mathbf{e}_x$ and $\mathbf{e}_y$ the unit vectors defining the axis of the Cartesian system ($\mathbf{e}_x \perp \mathbf{e}_y$).

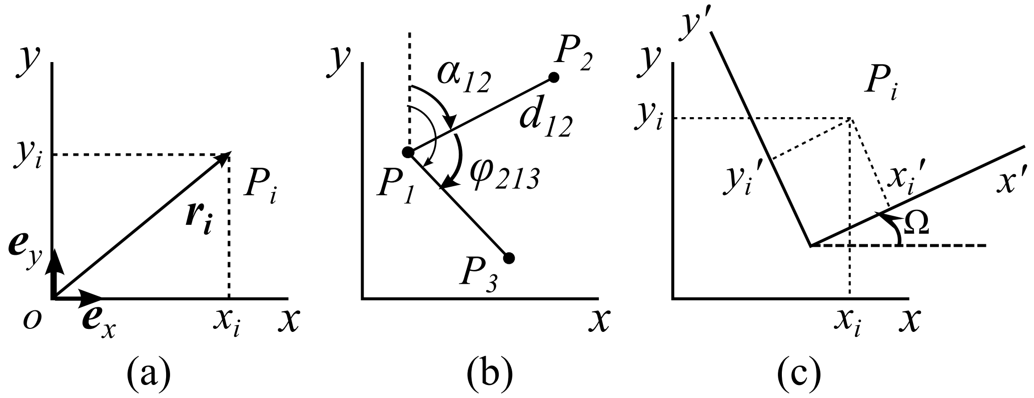


Figure 2.1: 2D Cartesian coordinate system (a), definition of azimuth, angle and distance (b) and a 2D coordinate transformation (c).

For surveying and mapping the distance d_{12} and azimuth α_{12} between two points P_1 and

[1]An easy way to remember this is the *right hand rule*: place your right hand on the plane with the thumb pointing up (in the direction of a "z-axis"), the fingers now point from the x-axis to the y-axis.

P_2 are defined as,

$$d_{12} = \|\mathbf{r}_2 - \mathbf{r}_1\| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
$$\alpha_{12} = \arctan \frac{x_2 - x_1}{y_2 - y_1} \tag{2.2}$$

The azimuth α is given in angular units (degrees, radians, gon) while the distance d is expressed in length units (meters), see also Appendix A. The angle $\angle P_2 P_1 P_3$ between points P_2, P_1 and P_3 is,

$$\varphi_{213} = \angle P_2 P_1 P_3 = \alpha_{13} - \alpha_{12} = \arccos \frac{< \mathbf{r}_2 - \mathbf{r}_1, \mathbf{r}_3 - \mathbf{r}_1 >}{\|\mathbf{r}_2 - \mathbf{r}_1\| \|\mathbf{r}_3 - \mathbf{r}_1\|} \tag{2.3}$$

with $< \mathbf{u}, \mathbf{v} >$ the dot (inner) product of two vectors. See Figure 2.1b. The corresponding distance ratio is defined as d_{12}/d_{13}. Note that in surveying and mapping the azimuth, or bearing, is defined differently than in mathematics.

In land surveying the x-axis is usually (roughly) oriented in the East direction and the y-axis in the North direction. Therefore, the x- and y-coordinates are also sometimes called *Easting* and *Northing*. The azimuth, or bearing, is referred to the North direction. This can either be the geographic North, magnetic North, or as is the case here, to the so-called *grid* North: the direction given by the y-axis. The azimuth angle is defined as the angle of the vector $\mathbf{r}_{12}$ with the North direction and is counted clockwise, i.e. for the azimuth a *left-handed* convention is used, see Figure 2.1b. In mathematics the x- and y-coordinate often called abscissa and ordinate, and angles are counted counter-clockwise from the x-axis, with $\theta_{12} = \arctan \frac{y_2 - y_1}{x_2 - x_1}$ following the mathematical text book definition of tangent. Thus $\alpha_{12} = 90 - \theta_{12}$.

Another possibility for describing the position of a point P_i in a 2D Cartesian coordinate system is by its polar coordinates, which are the azimuth α_{oi} and distance d_{oi} to the point P_i from the origin of the coordinate system. For many types of surveying instruments and measurements it is often convenient to make use of polar coordinates. For instance, with a tachymeter the distance and direction measurements are polar coordinates in a 2D local coordinate system, with origin in the instrument and y-axis in an arbitrary, yet to be determined, direction (representing the zero readng of the instrument).

2.2. 2D coordinate transformations

The 2D Cartesian coordinate system is defined by the origin of the axis, the direction of one of the axis (the second axis is orthogonal to the first) and the scale (the same for both axis). This becomes immediately clear when a second Cartesian coordinate system is considered with axis x' and y', see Figure 2.1c. The coordinates (x'_i, y'_i) for point P_i in the new coordinate system are related to the coordinates (x_i, y_i) in the original system through a rotation with rotation angle Ω, a scale change by a scale factor s and a translation by two origin shift parameters (t_x, t_y),

$$\begin{pmatrix} x'_i \\ y'_i \end{pmatrix} = s \begin{pmatrix} \cos\Omega & \sin\Omega \\ -\sin\Omega & \cos\Omega \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix} \tag{2.4}$$

The transformation is a so-called *conformal* [2] transformation whereby angles $\angle AOB$ between points A, O and B are preserved, i.e. angles are not changed by the transformation. This also

[2] Other types of transformations are *affine* and the more general *polynomial* transformations. An affine transformation involves a rotation, scale change in both x- and y-direction, and a translation. Affine transformations introduce so-called *sheering* between the coordinate axis. Angles are not necessarily preserved in an affine transformation, but lines remain straight, and parallel lines remain parallel after an affine transformation. Also ratios

means that shapes are preserved. A conformal transformation is therefore also known as a *similarity transformation*. Distances are not necessarily preserved in a conformal or similarity transformation, unless the scale factor s is one, but ratios of distances d_{OA}/d_{OB} between three points are preserved.

The transformation involves four parameters: two translations t_x and t_y, a rotation Ω, and, a scale factor s. This means that any 2D Cartesian coordinate system is uniquely defined by four parameters. Note that translation, rotation and scale only describe relations between coordinate systems, so there is always one coordinate system that is used as a starting point. Translation, rotation and scale change are relative concepts. However, a 2D Cartesian coordinate system can also be defined uniquely by assigning coordinates for (at least) two points. Assigning coordinates for two points uniquely defines a 2D Cartesian system: two points represent four parameters (coordinates) from which the position, orientation and scale of the axis can be constructed. The distance between two points defines the scale, from the azimuth between two points follows the orientation of the y-axis, and the coordinates themselves define where the origin is.

This means that a coordinate reference systems can be realized by selecting a number of points in the field and assigning coordinates to them. In its simplest form you could stake out a marker, assign this marker the coordinates (0,0), stake out a second marker and assign this the coordinates (0,1), which defines the y-axis and length scale. But you also could have assigned different coordinates, thereby defining a different reference frame. Instead of assigning the rather arbitrary value 1 for the y-coordinates of the second point, you could also have used a measured distance between the two points involved in the definition. This implies that the scale of our freshly defined coordinates system is determined by the scale of the measuring device (e.g. a tape measure) and also any measurement errors that were made in this measurements are included in the definition of scale. All this works well for defining a local coordinate system, but what about a national system, or that of a neighboring or previously realized project? In order to access any other system you should include at least two (observable) points for which coordinates in the other system are known. These could be points that have been established by other organizations, such as a Cadaster or mapping agency which publishes the coordinates of many reference markers. It could also be points you have established yourself using for instance GPS measurements.

of distances between points lying on a straight line are preserved. A polynomial transformation is a non-linear transformation which involves quadratic and often higher order terms of the coordinates. This means that straight lines are not necessarily preserved. Polynomial transformations are sometimes used to match satelllite and aerial imagery onto a 2D Cartesian coordinate system, or to transform grid coordinates between map projections.

3

3D Coordinate systems

3.1. 3D Cartesian coordinates

The position of an object in 3D space can be described in several ways. One of the most straightforward ways is to give the position by three coordinates X, Y, Z in a Cartesian coordinate system. See also Figure 3.1a. The coordinates are defined with respect to a reference point, or origin (with coordinates 0,0,0), which can be selected arbitrarily. For a global *geocentric* Terrestrial coordinate system it is however convenient to choose the origin at the center of the Earth. Further, the direction of one of the axis is chosen to coincide with the Earth rotation axis, while the other axis is based on a conventional definition of the zero meridian. The third axis completes the pair to make an orthogonal set of axes. The scale along the axes is simply tied to the SI definition of the meter (Appendix A). Capital letters X, Y, Z are used for the coordinates to set it apart from the 2D coordinate system in section , but also because this is the usual notation for coordinates in a 3D global Terrestrial coordinate system with the origin at the center of the Earth.

Instead of a global geocentric Terrestrial system also a local 3D Cartesian coordinate system can be defined, with the Y-axis pointing in the North direction, the X-axis in the up direction, and the X-axis completing the pair and therefore pointing in the East direction, and with the origin somewhere on the surface of the Earth. This type of system is referred to as *topocentric* coordinate system. For the coordinates it is common to use the capital letters E, N, U (East, North, Up) instead of X, Y, Z.

In both examples, the geocentric as topocentric coordinates, the coordinate system is somehow tied to the Earth, but this is not necessary. An another common variant is where the coordinate system is tied to an instrument or sensor. Sometimes the third axis may be aligned to the direction of the gravity vector, as is typical for a theodolite, but the thrid axis may also be tied to the observing platform (boat, car, plane) and have a more or less arbitrary orientation with respect to the Earth gravity field.

The 3D topocentric Cartesian coordinate system can be considered a straightforward extension of a 2D Cartesian coordinate system. Just image in Figure 2.1 a z-axis from the origin pointing outside the paper towards you. Coordinates, position vectors, distances, and angles are defined in a similar fashion. The 3D position vector for a point P_i with coordinates (X_i,Y_i,Z_i) given by

$$\mathbf{r}_i = X_i \mathbf{e}_X + Y_i \mathbf{e}_Y + Z_i \mathbf{e}_Z \ , \tag{3.1}$$

with $\mathbf{e}_X$, $\mathbf{e}_Y$ and $\mathbf{e}_Z$ the unit vectors defining the axis of the Cartesian system. This is illustrated

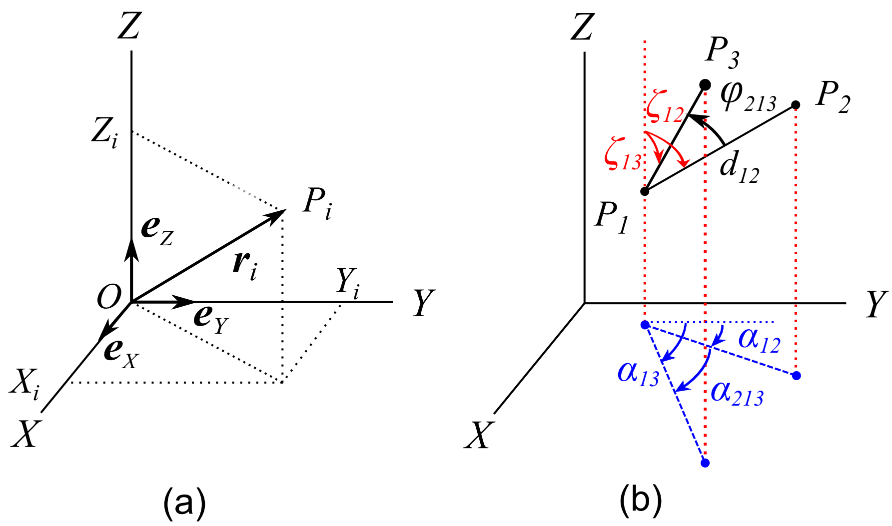


Figure 3.1: 3D Cartesian coordinate system (a) and definition of azimuth α_{12}, horizontal angle α_{213}, vertical angle ζ_{12}, angle φ_{213} and distance d_{12} (b).

in Figure 3.1a. As shown in Figure 3.1b, the distance d_{12} between two points P_1 and P_2 is

$$d_{12} = \|\mathbf{r}_2 - \mathbf{r}_1\| = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2 + (Z_2 - Z_1)^2} \quad , \tag{3.2}$$

and angle $\angle P_2P_1P_3$ between points P_2, P_1 and P_3 is,

$$\varphi_{213} = \angle P_2P_1P_3 = \arccos \frac{< \mathbf{r}_2 - \mathbf{r}_1, \mathbf{r}_3 - \mathbf{r}_1 >}{\|\mathbf{r}_2 - \mathbf{r}_1\| \; \|\mathbf{r}_3 - \mathbf{r}_1\|} \tag{3.3}$$

with $< \mathbf{u}, \mathbf{v} >$ the dot (inner) product of two vectors. For a topocentric system, the azimuth α_{12} and vertical angle ζ_{12} between points P_1 and P_2 can be defined as,

$$\begin{aligned} \alpha_{12} &= \arctan \frac{X_2 - X_1}{Y_2 - Y_1} \\ \zeta_{12} &= \arctan \frac{Z_2 - Z_1}{\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}} \end{aligned} \tag{3.4}$$

See also Figure 3.1b. Note that this definition only makes sense for topocentric systems where the Z-axis is oriented in the up direction and Y-axis to the North. Eq. 3.4 cannot be used for global geocentric Terrestrial coordinate systems. The angle φ_{213} of Eq. 3.3 is not the same as the horizontal angle α_{213} in Figure 3.1b. The horizontal angle is defined as $\alpha_{213} = \alpha_{13} - \alpha_{12}$.

The 3D Cartesian coordinate system is defined by the origin of the axes, the direction of two axes (the third axis is orthogonal to the other two) and the scale, which is the same for all axes. This becomes immediately clear when a second Cartesian coordinate system is considered with axes X', Y' and Z'. The coordinates (X'_i, Y'_i, Z'_i) for point P_i in the new coordinate system are related to the coordinates (X_i, Y_i, Z_i) in the original system through a 7-parameter transformation, consisting of three rotations, three translations and a scale factor, see Section 3.2. This transformation is again a conformal or similarity transformation, whereby angles $\angle AOB$ and distance ratio's between points A, O and B are preserved, i.e. shapes are not changed by the transformation. In the transformation 7 parameters are involved. This means that any 3D Cartesian coordinate system is uniquely defined by 7 parameters. Note that again translation, rotation and scale only describe relations between coordinate systems, which means that there is always one coordinate system that is used as a starting point. However, a 3D Cartesian coordinate system can also be defined uniquely by assigning coordinates for (at least) three points. Assigning coordinates for two points uniquely defines six degrees of freedom, which leaves one degree of freedom (a rotation) which needs to be resolved by one coordinate of a third point (although one coordinate is sufficient for the definition, approximate values for the other two coordinates are needed for numerical reasons). This means that a 3D-coordinate reference system can be realized by selecting at least 3 points and assigning at least 7 coordinates to them.

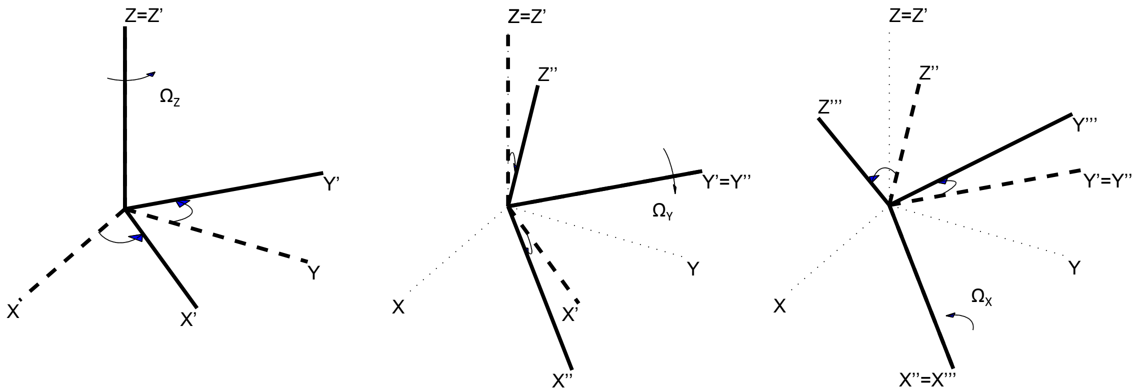


Figure 3.2: Definition of rotation angles for the 7-parameter similarity transformation. The figure on the left shows a rotation Ω_z about the Z-axis, the middle figure a rotation Ω_y about the newly obtained Y'-axis, and the figure on the right a rotation Ω_x about the final X''-axis.

3.2. 7-parameter similarity transformation

To transform 3D Cartesian coordinates from a source to target coordinate system *a 7-parameter similarity transformation* is used. The transformation consists of three translations (t_x, t_y, t_z), three rotations $(\Omega_x, \Omega_y, \Omega_z)$ and a differential scale factor μ,

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = (1+\mu)\cdot \mathbf{R}(\Omega_x, \Omega_y, \Omega_z)\cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} \tag{3.5}$$

with (X, Y, Z) the coordinates in the source coordinate system and (X', Y', Z') the coordinates in the target coordinate system. The differential scale factor μ is often a small number and is sometimes expressed in parts-per-million (ppm), with $1\text{ppm} = 10^{-6}$. The scale factor $(1+\mu)$ is then close to one. The translation vector (t_x, t_y, t_z) has to be added to the source coordinates after rotation. The translation vector gives the coordinates of the origin of the source coordinate system with respect to the target coordinate system. The rotation matrix $\mathbf{R}(\Omega_x, \Omega_y, \Omega_z)$ is defined as

$$\mathbf{R}(\Omega_x, \Omega_y, \Omega_z) = \mathbf{R}_1(\Omega_x)\cdot\mathbf{R}_2(\Omega_y)\cdot\mathbf{R}_3(\Omega_z) = \begin{pmatrix} \cos\Omega_z\cos\Omega_y & \sin\Omega_z\cos\Omega_y & -\sin\Omega_y \\ \cos\Omega_z\sin\Omega_y\sin\Omega_x - \sin\Omega_z\cos\Omega_x & \sin\Omega_z\sin\Omega_y\sin\Omega_x + \cos\Omega_z\cos\Omega_x & \cos\Omega_y\sin\Omega_x \\ \cos\Omega_z\sin\Omega_y\cos\Omega_x + \sin\Omega_z\sin\Omega_x & \sin\Omega_z\sin\Omega_y\cos\Omega_x - \cos\Omega_z\sin\Omega_x & \cos\Omega_y\cos\Omega_x \end{pmatrix} \tag{3.6}$$

with Ω_x, Ω_y and Ω_z rotations around the x-, y- and z-axis respectively, and $\mathbf{R}_i(\Omega_i)$ the corresponding Euler rotation matrices describing rotations around the coordinate axis, with

$$\mathbf{R}_3(\Omega_z) = \begin{pmatrix} \cos\Omega_z & \sin\Omega_z & 0 \\ -\sin\Omega_z & \cos\Omega_z & 0 \\ 0 & 0 & 1 \end{pmatrix}, \; \mathbf{R}_2(\Omega_y) = \begin{pmatrix} \cos\Omega_y & 0 & -\sin\Omega_y \\ 0 & 1 & 0 \\ \sin\Omega_y & 0 & \cos\Omega_y \end{pmatrix}, \\ \mathbf{R}_1(\Omega_x) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\Omega_x & \sin\Omega_x \\ 0 & -\sin\Omega_x & \cos\Omega_x \end{pmatrix} \tag{3.7}$$

whereby the right-handed rotation is positive, which is, when viewed along the axis towards the origin, a counter-clockwise rotation. See Figure 3.2. This sense of rotation is the same as was used in the 2-dimensional case, see Eq. 2.4. Imagine in Figure 2.1c a z-axis pointing out of the paper, then the rotation matrix of Eq. 2.4 is essentially $\mathbf{R}_3(\Omega_z)$ (which does not change the z-axis or z-coordinates) and the rotation angle Ω of Eq. 2.4 is actually Ω_z in the 3D-transformation. The rotation Ω_z around the z-axis is actually a rotation of the x-axis (and also y-axis) by Ω_z. The complete rotation $\mathbf{R}(\Omega_x, \Omega_y, \Omega_z)$ is thus the product of first a rotation around the z-axis, followed by a rotation around the new y-axis, and finally a rotation around the then current x-axis. Changing the order of the Euler rotations in Eq. 3.6 will result in different rotation angles. This is typical for Euler rotations.

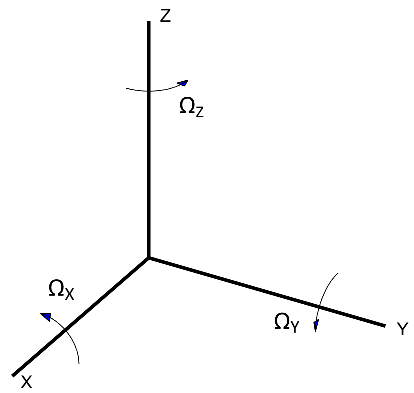


Figure 3.3: Definition of infinitesimal small rotation angles Ω_x, Ω_y and Ω_z for the *Helmert* transformation.

In datum transformations the rotation angles are often quite small with $\cos\Omega \simeq 1$ and $\sin\Omega \simeq \Omega$ (with Ω in radians). The rotation matrix $\mathbf{R}(\Omega_x, \Omega_y, \Omega_z)$ is then

$$\mathbf{R}(\Omega_x, \Omega_y, \Omega_z) \simeq \begin{pmatrix} 1 & -\Omega_z & \Omega_y \\ \Omega_z & 1 & -\Omega_x \\ -\Omega_y & \Omega_x & 1 \end{pmatrix} \tag{3.8}$$

with the rotation angles as defined in Figure 3.3. The 7-parameter similarity transformation of Eq. 3.5 in it's simplified form is

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} + \begin{pmatrix} \mu & -\Omega_z & \Omega_y \\ \Omega_z & \mu & -\Omega_x \\ -\Omega_y & \Omega_x & \mu \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{3.9}$$

This transformation is also known as the 7-parameter *Helmert* transformation (the 3-parameter Helmert transformation only includes the translation). This transformation is reversible: changing the sign of the seven transformation parameters results in the inverse transformation.

The reader should be aware that often different conventions are used for the sign of the rotation parameters. The convention that is used in this reader is that a positive rotation is a counter-clockwise rotation when viewed in the direction of the origin, and this convention is applied to a rotation of the axis of the coordinate system. Other conventions define transformations not based on rotation of the axis, but are based on rotations of the position vector, resulting in an opposite sign for the rotation angles or other signs for the small angle terms in the rotation matrices. It is always a good idea to check that the transformation formulas provided together with published transformation parameters use the same sign-convention as the software you are using.

The 7-parameter similarity transformation uses rotations about the origin of the source system. This may result in numerical problems for networks of points that are confined to a small regions on the Earth surface, such as coordinates of a national reference system. In this case there will be a high correlation between the translations and rotations in the derivation of the parameter values for the standard 7-parameter transformation. Therefore, instead of rotations being derived around the origin of the system which is near the geocenter, rotations are derived around a point somewhere within the domain of the network. For this type of transformation three additional parameters, the coordinates of the rotation point, are required to describe the transformation. These additional parameters can be chosen freely, or by convention, and do not have the same role in the derivation of parameter values for the other 7-parameters. The transformation essentially remains a 7-parameter transformation, with 7 degrees of freedom, although an extra 3 parameters are needed in the specification. The transformation formula is

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = (1+\mu) \cdot \mathbf{R}(\Omega'_x, \Omega'_y, \Omega'_z) \cdot \begin{pmatrix} X - X_0 \\ Y - Y_0 \\ Z - Z_0 \end{pmatrix} + \begin{pmatrix} t'_x \\ t'_y \\ t'_z \end{pmatrix} + \begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \end{pmatrix} \tag{3.10}$$

with (X_0, Y_0, Z_0) the coordinates of the selected rotation point. This transformation is not reversible in the sense that the same parameter values, with different signs, can be used for the reverse transformation. This is because the coordinates for the rotation point are changed by the transformation. However, in practice sometimes the same coordinates are used, but this results in cumulative errors after repeated transformations. Eq. 3.10 uses the same sign convention as Eq. 3.5, but whereas many publications use for Eq. 3.5 the same sign convention as this reader, most publications use for Eq. 3.10 the opposite convention. You are warned.

4

Spherical and ellipsoidal coordinate systems

Although straightforward, Cartesian coordinates are not very convenient for representing positions on the surface of the Earth. Take a global terrestrial coordinate system, with origin at the center of mass of the Earth, the Z-axis is coinciding with the Earth rotation axis, the X-axis is based on a conventional definition of the zero meridian and the Y-axis completing the pair to make an orthogonal set of axis, and scale tied to the SI definition of the meter. Considering that the Earth radius is about $6,400$ km, then to represent positions on the surface of the Earth 7 digits (before the decimal point) would be needed for each of the three coordinates. For representing positions on the surface of the Earth it is actually more convenient to use curvilinear coordinates defined on a sphere or ellipsoid approximating the Earth's surface.

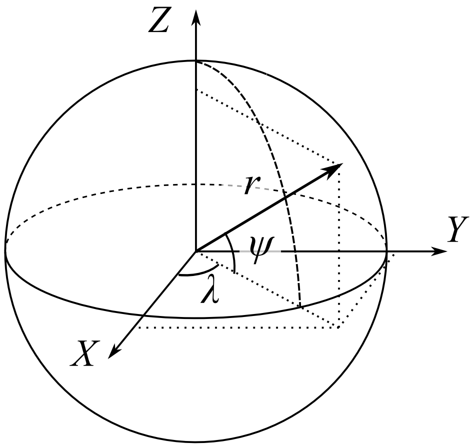


Figure 4.1: Spherical and cartesian coordinates.

4.1. Spherical coordinates

For example, assuming a sphere with radius R approximating the Earth surface, spherical coordinates ψ, λ and r (with $r = R + h'$, and $R = 6,371$ km the mean radius of the Earth) can be defined, see Figure 4.1. The relationship between Cartesian and spherical coordinates is given by,

$$\begin{aligned} X &= r\cos\psi\cos\lambda \\ Y &= r\cos\psi\sin\lambda \\ Z &= r\sin\psi \end{aligned} \tag{4.1}$$

The inverse relationship is given by,

$$\begin{aligned} \psi &= \arctan(\frac{Z}{\sqrt{X^2+Y^2)}}) \\ \lambda &= \arctan(\frac{Y}{X}) \\ r &= \sqrt{X^2+Y^2+Z^2)} \end{aligned} \tag{4.2}$$

The spherical coordinates ψ and λ can be used to represent positions on the sphere. In this case the sphere is a coordinate surface (surface on which one of the coordinates is constant), with ψ the *geocentric latitude* and λ the *longitude* of the point. In Eqs. 4.1 and 4.2 we abstained from using the expression $R + h'$, with R the radius of the sphere and h' the height above the reference sphere. In particular, we abstained from using h' as a third coordinate. Instead we used the geocentric radius or distance r of the point. This is because the sphere is not a very good approximation of the surface of the Earth and heights defined with respect to the sphere are meaningless (e.g. Mount Everest would have a height of 20 km, and the ocean surface in the Arctic a height of -10 km).

4.2. Geographic coordinates (ellipsoidal coordinates)

As shown by Newton (Principia, 1687) a rotating self-gravitating fluid body in equilibrium takes the form of an oblate ellipsoid. The oblate ellipsoid, or simply ellipsoid, is a much better approximation for the shape of the Earth than a sphere. An ellipsoid is the three dimensional surface generated by the rotation of an ellipse about its shorter axis. Two parameters are required to describe the shape of an ellipsoid. One is invariably the equatorial radius, which is the *semi-major axis*, a. The other parameter is either the polar radius or *semi-minor axis*, b, or the *flattening*, f, or the *eccentricity*, e. They are related by

$$f = \frac{a-b}{a} \;,\; e^2 = 2f - f^2 = \frac{a^2-b^2}{a^2} \;,\; b = a(1-f) = a\sqrt{1-e^2} \tag{4.3}$$

For the Earth the semi-major axis a is about $6,378$ km and semi-minor axis b about $6,357$ km, a 21 km difference. The flattening is of the order $1/300$, which is indistinguishable in illustrations if drawn to scale (illustrations, such as in this text, always exaggerate the flattening). Also, since f is a very small number, instead of f often the *inverse flattening* $1/f$ is given.

The position of a point with respect to an ellipsoid is given in terms of *geographic* or *geodetic* latitude φ, longitude λ and height h above the ellipsoid, see Figure 4.2. The relationship between Cartesian and geographic coordinates is given by,

$$\begin{aligned} X &= (\bar{N}+h)\cos\varphi\cos\lambda \\ Y &= (\bar{N}+h)\cos\varphi\sin\lambda \\ Z &= (\bar{N}(1-e^2)+h)\sin\varphi \end{aligned} \tag{4.4}$$

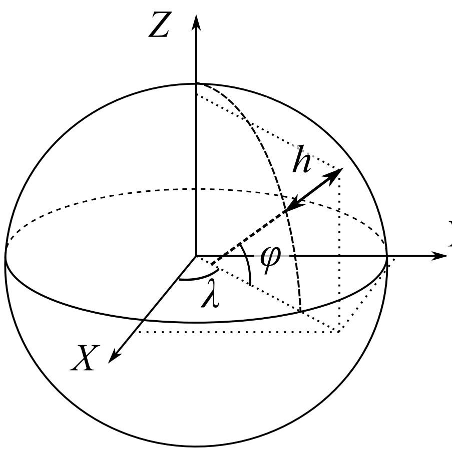


Figure 4.2: Ellipsoidal and Cartesian coordinates. The ellipsoidal latitude φ is also known as *geodetic* or *geographic* latitude. The ellipsoidal coordinates φ and λ are also called geographic coordinates.

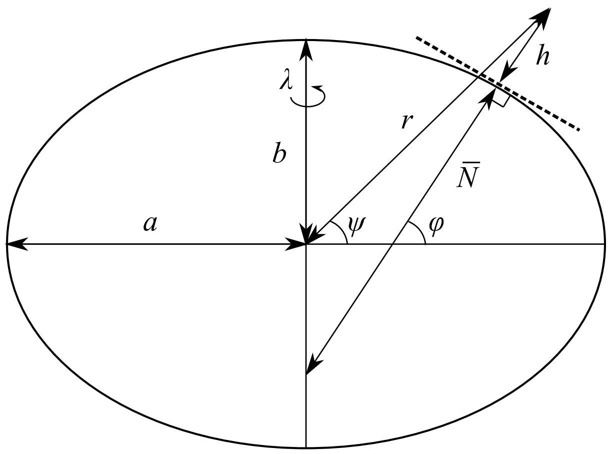


Figure 4.3: Ellipsoidal, geodetic or geographic latitude φ, geocentric (or spherical) latitude ψ, radius of curvature $\bar{N} = \bar{N}(\varphi)$, radius r, ellipsoidal height h, semi-major axis a and semi-minor axis b of the ellipsoid.

The inverse relationship is given by,

$$
\begin{aligned}
\varphi &= \arctan\left(\frac{Z + e^2\bar{N}\sin\varphi}{\sqrt{X^2+Y^2}}\right) \\
\lambda &= \arctan\left(\frac{Y}{X}\right) \\
h &= \frac{\sqrt{X^2+Y^2}}{\cos\varphi} - \bar{N}
\end{aligned}
\tag{4.5}
$$

$\bar{N}$ in Eqs. 4.4 and 4.5 is the radius of curvature in the prime vertical, as shown in Figure 4.3. The radius of curvature for an ellipsoid depends on the location on the ellipsoid. It is a function of the geographic latitude and is different in north-south and east-west direction.

The radius of curvature in the prime vertical, $\bar{N} = \bar{N}(\varphi)$, the radius of curvature in the (north-south) meridian, $\bar{M} = \bar{M}(\varphi)$, are

$$\begin{aligned}\bar{N}(\varphi) &= \frac{a}{\sqrt{1 - e^2 \sin^2 \varphi}} \\ \bar{M}(\varphi) &= \frac{a(1-e^2)}{(1 - e^2 \sin^2 \varphi)^{3/2}}\end{aligned} \tag{4.6}$$

with radius of curvature $\bar{N}$ (east-west) normal to $\bar{M}$ (north-south). On the equator the radius of curvature in east-west is equal to the semi-major axis a, with $\bar{N}(0^0) = a$, while the radius of curvature in north-south is smaller than the semi-minor axis, with $\bar{M}(0^0) = a(1-e^2) = b(1-f) = b^2/a$. On the poles the radius of curvature $\bar{N}(\pm 90^0) = \bar{M}(\pm 90^0) = a/\sqrt{(1-e^2)} = a^2/b$ is larger than the semi-major axis a, see Figure 4.4.

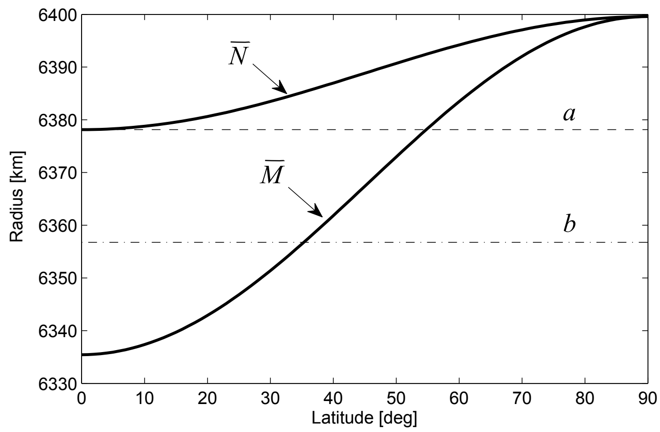


Figure 4.4: Radius of curvature $\bar{N}(\varphi)$ and $\bar{M}(\varphi)$ as function of latitude φ. The dashed lines represent the semi-major axis a and semi-minor axis b.

The geographic latitude in Eq. 4.5 must be computed by an iteration process as the geographic latitude φ appears both in the left and right hand side of the equation. Also note that the radius of curvature $\bar{N}$ in Eq. 4.6, which is a function of the geographic latitude (that still needs to be computed by Eq. 4.5), can be computed by the same iteration process. The iterative procedure, whereby in the first iteration $N' = \bar{N} \sin \varphi$ is approximated by Z, reads

$$\begin{aligned}& N'_0 = Z \\ & \text{for } i = 1, 2, \ldots \\ & \quad \varphi_i = \arctan\left(\frac{Z + e^2 N'_{i-1}}{\sqrt{X^2 + Y^2}}\right) \\ & \quad \bar{N}_i = \frac{a}{\sqrt{1 - e^2 \sin^2 \varphi_i}} \\ & \quad N'_i = \bar{N}_i \sin \varphi_i\end{aligned} \tag{4.7}$$

Usually four iterations are sufficient. For points near the surface of the Earth φ can also be computed using a direct method of B.R. Bowring (Survey Review, 181, July 1976, p. 323-327),

$$\varphi = \arctan\left(\frac{Z + e'^2 b \sin^3 \mu}{\sqrt{X^2 + Y^2} - e^2 a \cos^3 \mu}\right) \tag{4.8}$$

with e'^2, also called the *second eccentricity*, and μ given by

$$
\begin{aligned}
e'^2 &= \frac{e^2}{1-e^2} = \frac{a^2-b^2}{b^2} \\
\mu &= \arctan\left(\frac{aZ}{b\sqrt{X^2+Y^2}}\right)
\end{aligned}
\tag{4.9}
$$

The error introduced by this method is negligible for points between -5 and 10 km from the Earth surface, and, certainly much smaller than the error after four iterations with the iterative method.

The relation between the geocentric latitude ψ and the geodetic (or geographic) latitude φ for a point on the surface of the Earth, see Figure 4.3, is

$$
\psi(\varphi) = \arctan\left(\frac{\bar{N}(1-e^2)+h}{\bar{N}+h}\tan\varphi\right) \simeq \arctan((1-e^2)\tan\varphi) \mid h \ll \bar{N} \tag{4.10}
$$

The geodetic and geocentric latitudes are equal at the equator and poles. The maximum difference of $\varphi - \psi$ is approximately 11.5 minutes of arc at a geodetic latitude of $45^0 5'$. The geocentric and geodetic longitude are always the same. However, it is important not to confuse geocentric and geodetic latitude, which otherwise could result in an error in position of up to 20 km.

The normal, or vertical, to the ellipsoidal surface is the coordinate line that corresponds to h and $\bar{N}(\varphi)$. The ellipsoidal normal at the observation point is given by the unit direction vector $\bar{\mathbf{n}}$

$$
\bar{\mathbf{n}} = \begin{pmatrix} \cos\varphi\cos\lambda \\ \cos\varphi\sin\lambda \\ \sin\varphi \end{pmatrix} \tag{4.11}
$$

The ellipsoidal normal does not pass through the centre of the ellipsoid, except at the equator and at the poles. In general, the ellipsoidal normal does not coincide with the true vertical, $\mathbf{n}$, or plumb-line given by the direction of the local gravity field, $\mathbf{g}$, at that point. The gravity is the resultant of the gravitational acceleration and the centrifugal acceleration at that point, see also Appendix B. The direction of the true vertical $\mathbf{n}$ is given by the *astronomical* latitude ϕ and longitude Λ,

$$
\mathbf{n} = -\frac{\mathbf{g}}{g} = \begin{pmatrix} \cos\phi\cos\Lambda \\ \cos\phi\sin\Lambda \\ \sin\phi \end{pmatrix} \tag{4.12}
$$

with $g = \|\mathbf{g}\|$. The astronomical latitude ϕ is the angle between the equatorial plane and the true vertical at a point on the surface, the ellipsoidal, geodetic or geographic latitude φ is the angle between the equatorial plane and the ellipsoidal normal. A similar distinction exists for the astronomical latitude Λ and ellipsoidal longitude λ. The angle between the directions of the ellipsoidal normal and true vertical at a point are called the *deflection of the vertical*. The deflection of the vertical is divided in two components, defined as,

$$
\begin{aligned}
\xi &= \phi - \varphi \\
\eta &= (\Lambda - \lambda)\cos\varphi
\end{aligned}
\tag{4.13}
$$

Astronomic latitude ϕ and longitude Λ are obtained from astronomical observations to stars whose position (declination δ and right ascension α) in a Celestial reference system are accurately known, or from gravity observations using gravimeters[1]. The deflection of the vertical is usually only a few seconds of arc, whereby the largest values occur in mountainous areas and in areas with large gravity anomalies.

4.3. Topocentric coordinates, azimuth and zenith angle

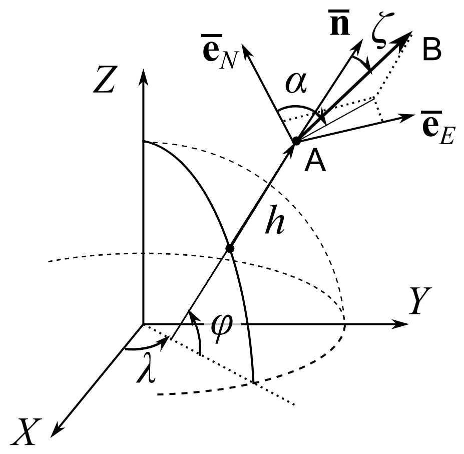


Figure 4.5: Local left-handed topocentric system in point A, with ellipsoidal coordinates (φ, λ, h), with the local azimuth α and zenith angle ζ for the direction $\vec{AB}$. The vertical (ellipsoidal normal vector) $\bar{\mathbf{n}}$ is the z-axis of the local left-handed system , $\bar{\mathbf{e}}_E$ is the y-axis and is orthogonal to the plane of the meridian and positive to the East, and $\bar{\mathbf{e}}_N = \bar{\mathbf{n}} \times \bar{\mathbf{e}}_N$,in the plane of the meridian, is the x-axis completing the topocentric system.

It is often necessary to have also relations for differential Cartesian coordinates (dX, dY, dZ) and differential ellipsoidal coordinates $(d\varphi, d\lambda, dh)$. This is useful when observations are made at eccentric stations, or to transform velocities in the Cartesian system to velocities in the ellipsoidal system, or to propagate error estimates (standard deviations, variances, co-variances) from the Cartesian system into the ellipsoidal system, or vice-versa. Likewise, it is useful to express the differential ellipsoidal coordinates not in radians, but instead in meters, using the relations,

$$\begin{aligned} dN &= (\bar{M}(\varphi) + h)d\varphi \\ dE &= (\bar{N}(\varphi) + h)\cos\varphi d\lambda \end{aligned} \tag{4.14}$$

with dN the differential in North-South (latitude) direction, with positive direction to the North, and dE the differential in East-West (longitude) direction. $\bar{M}(\varphi)$ and $\bar{N}(\varphi)$ are the meridian radius of curvature and radius of curvature in the prime vertical as given by Eq. 4.6 and Figure 4.4. $d\varphi$ and $d\lambda$ must be given in units of radians, but dN and dE are in units of meters. dN and dE are often referred to as *Northing* and *Easting*. The relation between the differential coordinates (dX, dY, dZ) and (dN, dE, dh) is,

$$\begin{aligned} \begin{pmatrix} dX \\ dY \\ dZ \end{pmatrix} &= \begin{pmatrix} -\sin\varphi\cos\lambda & -\sin\lambda & \cos\varphi\cos\lambda \\ -\sin\varphi\sin\lambda & \cos\lambda & \cos\varphi\sin\lambda \\ \cos\varphi & & \sin\varphi \end{pmatrix} \begin{pmatrix} dN \\ dE \\ dh \end{pmatrix} \\ &= \begin{pmatrix} \bar{\mathbf{e}}_N & \bar{\mathbf{e}}_E & \bar{\mathbf{n}} \end{pmatrix} \begin{pmatrix} dN \\ dE \\ dh \end{pmatrix} \end{aligned} \tag{4.15}$$

[1]Astronomical latitude is not to be confused with declination, the coordinate astronomers used in a similar way to describe the locations of stars north/south of the celestial equator, nor with ecliptic latitude, the coordinate that astronomers use to describe the locations of stars north/south of the ecliptic.

with $\bar{\mathbf{e}}_N$, $\bar{\mathbf{e}}_E$ and $\bar{\mathbf{n}}$ the three axis of a left-handed local topocentric system centered at the observation point (φ, λ, h). The inverse relation of Eq. 4.15 is,

$$\begin{pmatrix} dN \\ dE \\ dh \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{e}}_N & \bar{\mathbf{e}}_E & \bar{\mathbf{n}} \end{pmatrix}^{-1} \begin{pmatrix} dX \\ dY \\ dZ \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{e}}_N & \bar{\mathbf{e}}_E & \bar{\mathbf{n}} \end{pmatrix}^{T} \begin{pmatrix} dX \\ dY \\ dZ \end{pmatrix} \tag{4.16}$$

The vectors $\bar{\mathbf{e}}_N$, $\bar{\mathbf{e}}_E$ and $\bar{\mathbf{n}}$ form the three axis of a left-handed local topocentric system centered at the observation point (φ, λ, h), with z-axis along the normal of the ellipsoid $\bar{\mathbf{n}}$, y-axis $\bar{\mathbf{e}}_E$ orthogonal to the plane of the meridian and positive to the East, and x-axis $\bar{\mathbf{e}}_N = \bar{\mathbf{n}} \times \bar{\mathbf{e}}_N$ in the plane of the meridian completing the topocentric system. The system is left-handed because azimuth is counted by convention from the North and towards the East. For example, in a local topocentric system a point to the North has azimuth α of 0^0, and a point to the East $+90^0$, see Figure 4.5. The azimuth α and zenith angle ζ are defined by,

$$\begin{pmatrix} dN \\ dE \\ dh \end{pmatrix} = s \begin{pmatrix} \cos\alpha \sin\zeta \\ \sin\alpha \sin\zeta \\ \cos\zeta \end{pmatrix} \tag{4.17}$$

with s the slant range $\sqrt{dN^2 + dE^2 + dh^2}$. The inverse relation is,

$$\begin{aligned} \alpha &= \arctan(\frac{dE}{dN}) = \arctan(\frac{< \mathbf{e}_E, \mathbf{s} >}{< \mathbf{e}_N, \mathbf{s} >}) \\ \zeta &= \arctan(\frac{dh}{\sqrt{dN^2 + dE^2}}) = \arccos(\frac{dh}{s}) = \arccos(\frac{< \bar{\mathbf{n}}, \mathbf{s} >}{s}) \\ s &= \sqrt{dN^2 + dE^2 + dh^2} = \sqrt{dX^2 + dY^2 + dZ^2} = \sqrt{< \mathbf{s}, \mathbf{s} >} \end{aligned} \tag{4.18}$$

with $\mathbf{s} = (dX, dY, dZ)^T$. Compare this to Eq. 3.4.

For practical purposes the differentials in Eqs. 4.14, 4.15, 4.16, 4.17 and 4.18 can be replaced by finite differences $(\Delta X, \Delta Y, \Delta Z)$, $(\Delta N, \Delta E)$, $\Delta\varphi$, $\Delta\lambda$ and Δh. The algorithms that involve $\Delta\varphi$, $\Delta\lambda$ and Δh are only valid for small values. However, the algorithms that involve only $(\Delta X, \Delta Y, \Delta Z)$, $(\Delta N, \Delta E, \Delta h)$, or, the azimuth α and zenith angle ζ can be used also with large values provided $(\Delta N, \Delta E, \Delta h)$ are interpreted as coordinates in a local topocentric left-handed system. The latter comes in very useful for computation of azimuth and zenith angles of Earth satellites for a point on the surface of the Earth.

4.4. Practical aspects of using latitude and longitude

The ellipsoidal height differs by not more than 100 m from a equipotential surface, or a true height coordinate surface, and the ellipsoidal normals agree with the true vertical to within a few seconds of arc. There are no other simple rotational shapes that would match the true Earth better than an ellipsoid.

The ellipsoidal, geodetic, or geographical, latitude and longitude are therefore the most common representation to describe the position of points on the Earth. And invariably, when we use latitude and longitude without any further reference, this is almost always the ellipsoidal, geodetic or geographic latitude and longitude! However, the geodetic latitude φ should never be confused with the geocentric or spherical latitude ψ, or astronomical altitude ϕ, which are two different types of coordinates. Also you should not confuse geodetic longitude λ with astronomical longitude Λ. For the ellipsoidal height h it is a different story. Because the ellipsoid is off from an equipotential surface by up to 100 meters, the ellipsoidal height h is not a suitable coordinate for a height reference system. But, contrary to Cartesian coordinates, with ellipsoidal coordinates we can easily separate between 'horizontal' and 'vertical' coordinates. And, as you will see in Chapter 7 on height systems, we can replace the ellipsoidal height h by the orthometric height H, with $H = h - N(\varphi, \lambda)$ and $N(\varphi, \lambda)$ the so-called geoid height. This leaves us with a couple of options to represent positions

- Use Cartesian coordinates X, Y and Z to represent positions in three dimensions
- Use geographic coordinates and/or height
  - Ellipsoidal latitude φ and longitude λ to represent positions on the surface of the earth, with, or, without
  - Ellipsoidal height h or orthometric height H to represent the vertical dimension

  The latitude and longitude can be used with, and, without height information, or vice versa. In case no height information is provided it may be assumed that the positions are on the ellipsoid or another reference surface. The reference surface can be a theoretical surface, such as the ellipsoid, or modeled by a digital (terrain) model (with heights given on a regular grid or as a series of (base) functions).

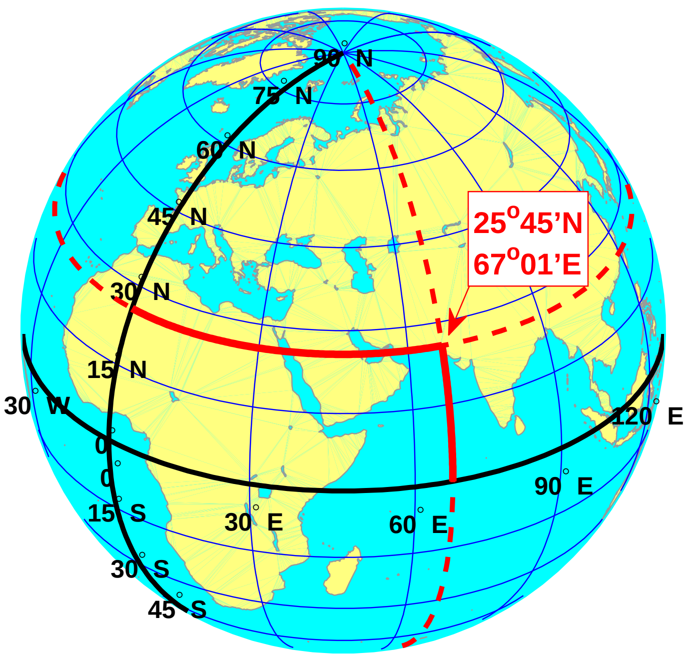


Figure 4.6: Latitude and longitude grid as seen from outer space (orthographic azimuthal projection). The prime meridian (through Greenwich) and equator are in black with latitude and longitude labels. The meridian and parallel through Karachi, Pakistan, $25^o45'$N $67^o01'$E, are the dotted lines in red. Meridians are *great circles* with constant longitude that run from North to South. Parallels are *small circles* with constant latitude (the equator is also a great circle).

The mesh formed by the lines of constant latitude and constant longitude forms a graticule that is linked to the rotation axis of the Earth, as is shown in Figure 4.6. The poles are where the axis of rotation of the Earth intersects the reference surface. Meridians are lines of constant longitude that run over the reference surface from the North Pole to the South pole. By convention, one of these, the Prime Meridian, which passes through the Royal Observatory, Greenwich, England, is assigned zero degrees longitude. The longitude of other places is given as the angle East or West from the Prime Meridian, ranging from 0^0 at the Prime Meridian to 180^0 Eastward (written 180^0 E or $+180^0$) and 180^0 Westward (written 180^0 W or -180^0) of the Prime Meridian. The plane through the center of the Earth and orthogonal to the rotation axis intersects the reference surface in a great circle called the equator. Planes parallel to the equatorial plane intersect the surface in circles of constant latitude; these are the parallels. The equator has a latitude of 0^0, the North pole has a latitude of 90^0 North (written 90^0 N or $+90^0$), and the South pole has a latitude of 90^0 South (written 90^0 S or -90^0). The latitude of an arbitrary point is the angle between the equatorial plane and the radius to that point.

Degrees of longitude and latitude can be sub-divided into 60 minutes of arc, each of which is divided into 60 seconds of arc. A longitude or latitude is thus specified in sexagesimal notation as $23^0 27' 30''$ [EWNS]. The seconds can include a decimal fraction. An alternative representation uses decimal degrees, whereby degrees are expressed as a decimal fraction: 23.45833^0 [EWNS]. Another option is to express minutes with a fraction: $23^0 27.5'$ [EWNS]. The [EWNS] suffix can be replaced by a sign: the convention is to use a negative sign for West and South, and a positive sign for East and North. Further, for calculations decimal degrees may be converted to radians. Note that the longitude is singular at the Poles and calculations that are sufficiently accurate for other positions, may be inaccurate at or near the Poles. Also the discontinuity at the $\pm 180^0$ meridian must be handled with care in calculations, for example when subtracting or adding two longitudes.

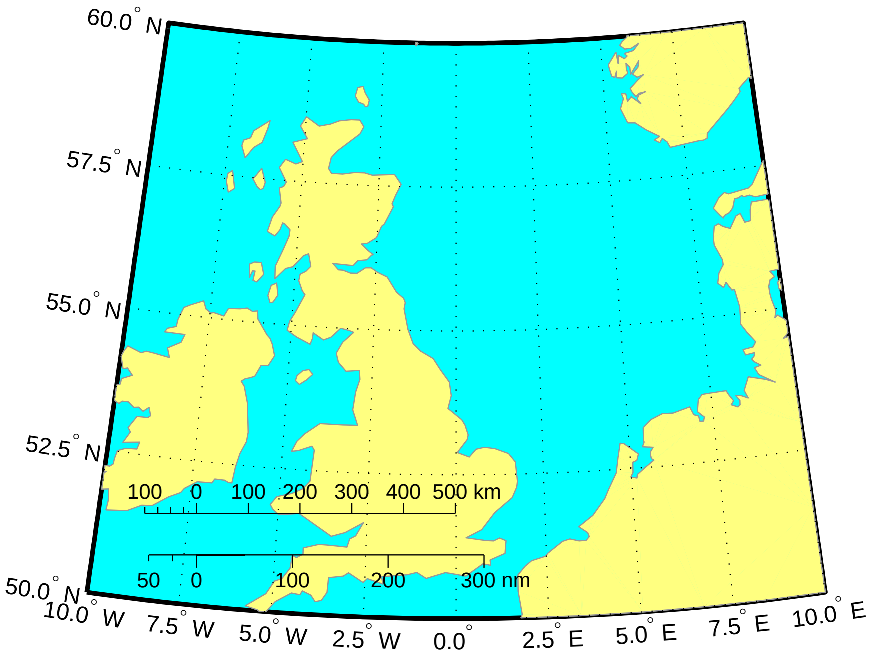


Figure 4.7: The latitude and longitude grid over the North-Sea in an equidistant cylindrical projection of uniform scale. One degree of latitude is about 60 nm (Nautical miles), or more precisely 111.2 km at 50^o and 111.4 km at 60^o latitude. However, one degree of longitude is much shorter, it varies between 71.7 km on the 50^o parallel and just 55.8 km on the 60^o parallel. This is because the meridians converge to the North.

One minute of arc of latitude measured along the meridian corresponds to one nautical mile ($1,852$ m). The nautical mile, which is a non-SI unit, is very popular with navigators in shipping and aviation because of its convenience when working with nautical charts (which often have a varying scale): a distance measured with a chart divider can be converted to nautical miles using the chart's latitude scale. This only works with the latitude scale, but not the longitude scale, which follows directly from Eq. 4.14 (on account of the term $\cos\varphi$, which result in the meridians converging at the poles), as is shown in Figure 4.7 for the North-Sea area. From Eq. 4.14 it also follows that one degree of arc of latitude measured along the meridian is between 110.57 km at the equator and 111.69 km at the poles[2]. Thus, at 52^0 latitude, one arc-second ($1''$) along the meridian corresponds to roughly 30.9 m and one arc-second along the 52^0 parallel to roughly 19.0 m.

Latitude and longitude are angular measures that work well to pin-point a position, but, calculations using the latitude and longitude can be quite involved. For example, the computation of distance, angles and surface area, is far from straightforward and very different from computations using two-dimensional Cartesian coordinates. In general users are left with two options: (1) use spherical or ellipsoidal computations, or, (2) first map the latitude and longitude to two-dimensional Cartesian coordinates x and y, and then do all the computations in the two-dimensional (map) plane. The second option involves a so-called *map projection*. Computations on the sphere or ellipsoid are discussed in Section 4.5, map projections are discussed in Section 5.

4.5. Spherical and ellipsoidal computations

Distances have different meanings. For instance, the distance between an observer in Delft and a satellite orbiting the Earth is the straight line distance computed from the 3D Cartesian coordinates of both points. If the coordinates of the observer are given in geographical coordinates, these are first converted into Cartesian coordinates; something for which also the height above the ellipsoid is needed (unless the station is assumed to lie on the ellipsoid). On the other hand, for the distance between two places on the ellipsoid, say Delft (NL) and San Diego (CA, USA), the shortest distance over the sphere or ellipsoid is required, and not the straight line distance.

The equivalent of a straight line in Euclidean geometry for spherical and ellipsoidal geometry is the shortest path between points on a sphere or ellipsoid, which is called *geodesic*. On a sphere geodesics are great circles. This is illustrated in Figure 4.8. Similar geometric concepts are defined in spherical and ellipsoidal geometry as in Euclidean geometry, replacing straight lines by great circles and geodesics. For instance, in spherical geometry angles are defined between great circles, resulting in spherical trigonometry.

The solution of many problems in geodesy and navigation, as well as in some branches of mathematics, involve finding solutions of two main problems:

Direct (first) geodetic problem Given the latitude φ_1 and longitude λ_1 of point P1, and the azimuth α_1 and distance s_{12} from point P1 to P2, determine the latitude φ_2 and longitude λ_2 of point P2, and azimuth α_2 in point P2 to P1.

Inverse (second) geodetic problem Given the latitude φ_1 and longitude λ_1 of point P1, and latitude φ_2 and longitude λ_2 of point P2, determine the distance s_{12} between point P1 and P2, azimuth α_1 from P1 to P2, and azimuth α_2 from P2 to P1.

On a sphere the solutions to both problems are (simple) exercises in spherical trigonometry. On an ellipsoid the computation is much more involved. Work on ellipsoidal solutions was carried out by for example Legrendre, Bessel, Gauss, Laplace, Helmert and many others after them.

[2] In surveying and geodesy a circle is not divided in 360^0 but in 400 gon or grad. This has the added advantage that one gon (grad) measured along the meridian conresponds to 100 km, and one milli-gon (milli-grad) to 10 cm, and a decimilligrad (10^{-7} grad) corresponds to 1 cm. However, this "decimal" system for angular measurement never gained a big following outside surveying. But, be aware, quite often surveying equipment uses *gon* or *grad* to measure arcs instead of degrees.

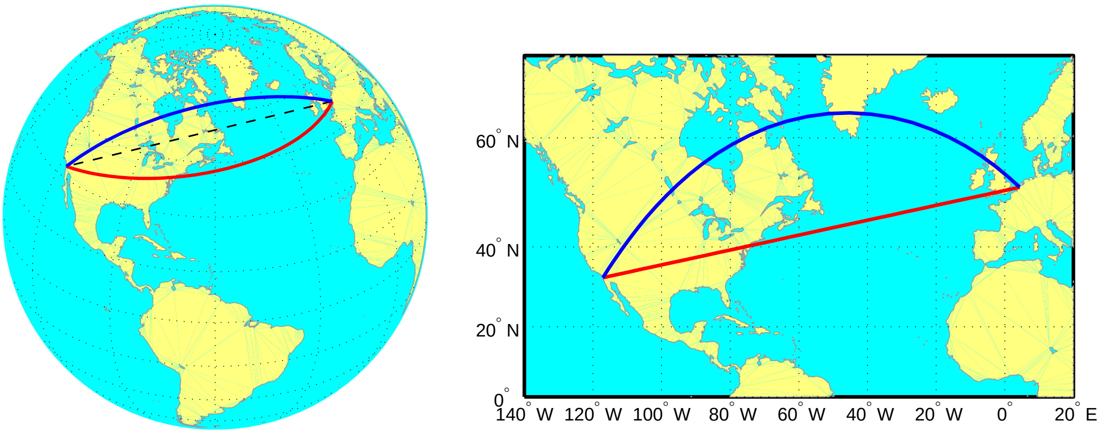


Figure 4.8: Great circle (blue), rhumb line (red) and straight line (black dashed) distance between Delft, NL, 52^oN 4.37^oE and San Diego, CA, USA, 32.8^oN 117.1^oW. The plot on the left uses an orthographic azimuthal projection, with the Earth as seen from outer space, while the plot on the right uses the Mercator projection. The great circle, rhumb line and straight line distances are 9005, 10077 and 8294 km respectively. The straight line (black dashed) passes through the Earth lower mantle, with a deepest point of 1529 km below the Hudson Strait, Northern Canada, 61.0^oN 71.7^oW. This is also the half-way point for a traveler following the great circle route (blue), which is the shortest route over the Earth surface from Delft to San Diego. The course a traveler is steering on this route varies between NW (313.5^o) when leaving Delft and SSW (212.1^o) when arriving in San Diego. A rhumb line on the other hand crosses meridians always at the same angle. A traveler following the rhumb line (red) from Delft to San Diego would have to steer a constant WSW course (257.8^o). Rhumb lines become straight lines in a Mercator projection.

The starting point is writing the geodesic as a differential equation relating an elementary segment with azimuth α and length ds to differential ellipsoidal coordinates $(d\varphi, d\lambda)$,

$$\begin{aligned} \frac{d\varphi}{ds} &= \frac{\cos\alpha}{\bar{M}(\varphi)} \\ \frac{d\lambda}{ds} &= \frac{\sin\alpha}{\bar{N}(\varphi)\;\cos\varphi} \end{aligned} \tag{4.19}$$

with $\bar{M}(\varphi)$ the meridian radius of curvature and $\bar{N}(\varphi)$ the radius of curvature in the prime vertical as given by Eq. 4.6 and Figure 4.4, and with $\bar{N}(\varphi)\cos\varphi$ the radius of the circle of latitude φ. See also Eq. 4.14 which gives similar relations for Northing dN and Easting dE. These equations hold for any curve. For specific curves the variation of the azimuth $d\alpha$ must be specified in relation to ds. For example, for the rumbline, the curve that makes equal angles with the local meridian, $d\alpha/ds = 0$. For the geodesic this relation is

$$\frac{d\alpha}{ds} = \sin\varphi\frac{d\lambda}{ds} = \frac{\tan\varphi}{\bar{N}(\varphi)}\sin\alpha \tag{4.20}$$

Eqs. 4.20 and 4.19 form a complete set of differential equations for the geodesic. These differential equations can be used to solve the direct and inverse geodetic problems numerically. Other solutions involve evaluating integral equations that can be derived from these differential equations. In geodetic applications where f is small, the integrals are typically evaluated as a series or using iterations. The treatment of this complicated topic goes beyond the level of this reader.

On a sphere the solution of the direct and inverse geodetic problem can be found using spherical trigonometry resulting in closed formula. These formula are important for navigation.

Finding the *course* and *distance* through spherical trigonometry is a special application of the inverse geodetic problem. The inital and final course α_1 and α_2, and distance s_{12} along the great circle, are

$$\begin{aligned}
\tan\alpha_1 &= \frac{\sin\lambda_{12}}{\cos\phi_1\tan\phi_2 - \sin\phi_1\cos\lambda_{12}} \\
\tan\alpha_2 &= \frac{\sin\lambda_{12}}{-\cos\phi_2\tan\phi_1 + \sin\phi_2\cos\lambda_{12}} \\
\cos\sigma_{12} &= \sin\phi_1\sin\phi_2 + \cos\phi_1\cos\phi_2\cos\lambda_{12}
\end{aligned} \tag{4.21}$$

with $\lambda_{12} = \lambda_2 - \lambda_1$. The distance is given by $s_{12} = R\,\sigma_{12}$, where σ_{12} is the central angle (in radians) between the two points and R the Earth radius.For practical computations the quadrants of the arctangens are determined by the signs of the numerator and denominator in the tangent formulas (e.g., using the `atan2` function). Using the mean Earth radius yields distances to within 1% of the geodesic distance on the WGS-84 ellipsoid.

Finding *way-points*, the positions of selected points on the great circle between P1 and P2, through spherical trigonometry is a special application the direct geodetic problem. Given the initial course α_1 and distance s_{12} along the great circle, the latitude and longitude of P2 are found by,

$$\begin{aligned}
\tan\phi_2 &= \frac{\sin\phi_1\cos\sigma_{12} + \cos\phi_1\sin\sigma_{12}\cos\alpha_1}{\sqrt{(\cos\phi_1\cos\sigma_{12} - \sin\phi_1\sin\sigma_{12}\cos\alpha_1)^2 + (\sin\sigma_{12}\sin\alpha_1)^2}} \\
\tan\lambda_{12} &= \frac{\sin\sigma_{12}\sin\alpha_1}{\cos\phi_1\cos\sigma_{12} - \sin\phi_1\sin\sigma_{12}\cos\alpha_1} \\
\tan\alpha_2 &= \frac{\sin\alpha_1}{\cos\sigma_{12}\cos\alpha_1 - \tan\phi_1\sin\sigma_{12}}
\end{aligned} \tag{4.22}$$

with $\sigma_{12} = s_{12}/R$ the central angle in radians and R the Earth radius, and $\lambda_2 = \lambda_1 + \lambda_{12}$.

Computations on the sphere, let alone the ellipsoid, are quite complicated. Other tasks than the direct and inverse geodetic problem, such as the computation of the area on a sphere or ellipsoid, which is simple in a 2D Cartesian geometry, require even more complicated computations. Instead a different approach can be taken, which consists of a mapping of the latitude and longitude $(\varphi,\lambda)_i$ to *grid* coordinates $(x,y)_i$ in a 2D Cartesian geometry, known as *map projection*.

5 Map projections

Map projections are used in both cartography and geodesy. The output of a map projection in cartography is usually a small scale map, on paper, or in a digital format. The required accuracy of the mapping is low and a sphere may be safely used as the surface to be mapped. In cartography it is more about appearance and visual information than accuracy of the coordinates. In geodesy a map projection is more a mathematical device that transfers the set of geographical coordinates (φ, λ) into a set of planar coordinates (y, x) without loss of information. The relation can therefore also be inverted (i.e. undone). It implies that an ellipsoid should be used as the surface to be mapped. This also applies for medium and large scale maps, and coordinates that are held digitally in a Geographic Information System (GIS) or other information system. In this reader a *map projection* is defined as the mathematical transformation

$$\begin{aligned} y &= f(\varphi, \lambda, h) \\ x &= g(\varphi, \lambda, h) \end{aligned} \tag{5.1}$$

whereby h is implicitly given as zero ($h = 0$), meaning points are first projected on the ellipsoid. The coordinates (y, x) are called *map* or *grid* coordinates. The grid coordinates are often referred to as Northing (y) and Easting (x).

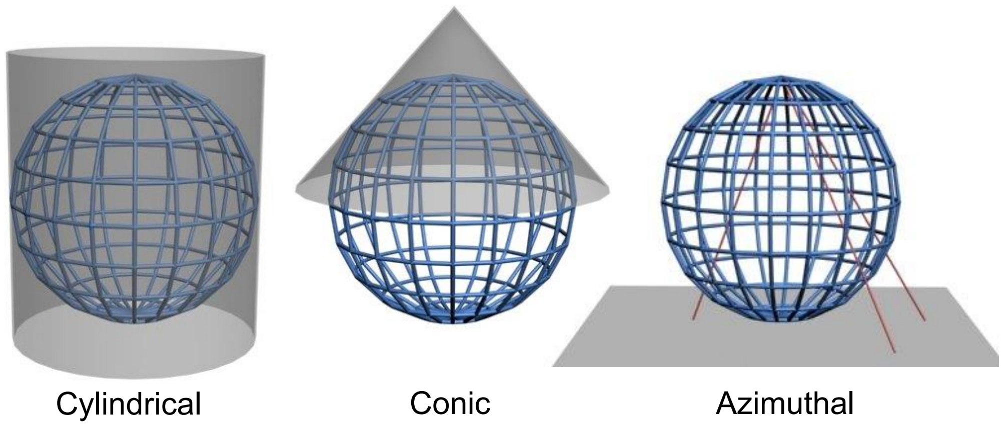


Figure 5.1: Cylindrical, Conic and Azimuthal map projection types (Source: Wikimedia Commons).

Many different map projections are in use all over the world for different applications and for good reasons. However, having many different types of map projections and grid coordinates, may sometimes also result in confusion about what coordinates are actually used or given. Some software packages may support many of these map projections, but it is virtually impossible to support them all. Other softwares are specifically written for one specialized map projection, and give wrong results when using coordinates from a different type of projection.

Map projections can be grouped into four groups depending on the nature of the projection surface, see Figure 5.1,

Cylindrical map projections The plane of projection is a cylinder wrapped around the Earth. Cylindrical projections are easily recognized for its shape: maps are rectangular and meridians and parallels are straight lines crossing at right angles. A well known projection is the Mercator projection.

Conic map projections The plane of projection is a cone wrapped around the Earth. Parallels become arcs of concentric circles. Meridians, on the other hand, converge to the North or South. Often used for regions of large east-west extent.

Azimuthal map projections The plane of projection is a plane tangent to the Earth. A well known projection is the stereographic projection (which is used for instance by the Dutch RD system).

Miscellaneous projections Mostly used for cartographic purposes.

Any projection can be applied in the normal, oblique and transverse position of the cylinder, cone or plane, as shown in Figure 5.2 for a cylinder. In the normal case the axis of projection, the axis of the cylinder and cone, or normal to the plane, coincides with the minor axis of the ellipsoid.An example is the Mercator projection, with the equator the line of contact of the cylinder. In the transverse case the axis of projection is in the equatorial plane (orthogonal to the minor axis), for example, in the Universal Transverse Mercator (UTM) projection small strips are mapped on a cylinder wrapped around the poles and with a specific meridian as line of contact. In the oblique case the axis of projection does not coincide with the semi-minor axis or equatorial plane.

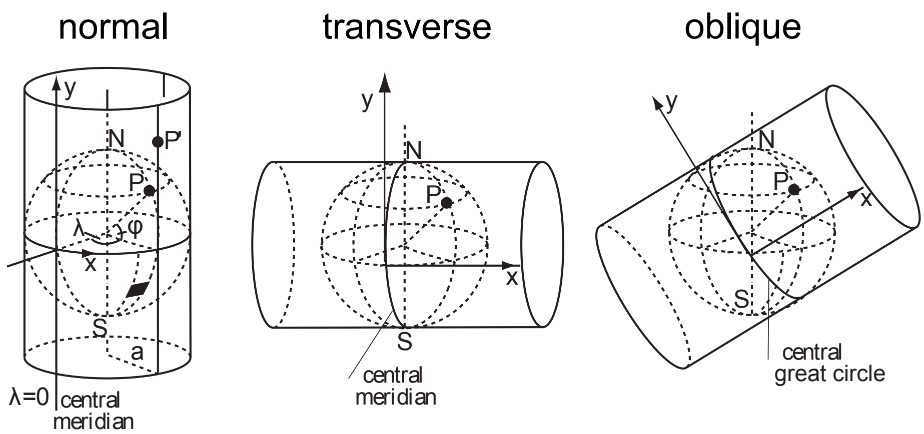


Figure 5.2: Normal, transverse and oblique projection for a cylinder (Source: Wikimedia Commons).

Some distortion in the geometrical elements, distance, angles and area, is inevitable in map projections. In this respect map projections are divided into

Conformal projections Preserves the angle of intersection of any two curves.

Equal Area (equivalent) projections Preserves the area or scale.

Equi Distance (conventional) projections Preserves distances.

Map projections may have one or two of these properties, but never all three together. In geodesy conformal mappings are preferred. A conformal mapping may be considered a similarity transformation in an infinitesimal small region. A conformal mapping differs only from a similarity transformation in the plane in that its scale is not constant but varying over the area to be mapped. For cartographic purposes, e.g. employing geostatistics, equal area mappings may be better suited.

In some projections an intermediate sphere is introduced. These are called double projections; the first step is a conformal mapping onto a sphere, the second step is the subsequent projection from the sphere onto a plane. This is also the basis for the Dutch map projection: the first step is a conformal Gauss projection on the sphere, the second step a stereographic projection onto a plane tangential to the ellipsoid with the center at Amersfoort.

In order to specify a map projection the following information is required

- Name of the map projection or EPSG dataset coordinate operation method code
- Latitude of natural origin or standard parallel (φ_0) for cylindrical and azimuthal projections, or, the latitude of first standard parallel (φ_1) and second standard parallel (φ_2) for conic projections
- Longitude of natural origin (the central meridian) (λ_0)
- Optional scale factor at natural origin (on the central meridian)
- False Easting and Northing

The false Easting and Northing are used to offset the planar coordinates (x, y) in order to prevent negative values. The International Association of Oil & Gas Producers (OGP) maintains a geodetic parameter dataset of common coordinate conversions, datum transformations and map projections. This is known as the EPSG dataset (EPSG stands for European Petroleum Survey Group), whereby each coordinate operation or transformation is identified by an unique number. In the EPSG Dataset codes are assigned to coordinate reference systems, coordinate transformations, and their component entities (datums, projections, etc.). Within each entity type, every record has a unique code (`http://www.epsg.org`). The EPSG website also provides the equations for the various mappings that have been stored in the EPGS database[1]. Software for map projections is provide for instance by the PROJ.4 package (`https://trac.osgeo.org/proj/`).

Map projections are usually equations that provide a relationship between latitude and longitude on the one hand, and planar grid coordinates on the other hand. However, sometimes the transformation to planar coordinates, and vice versa, may be supplemented by tabulated values in the form of a *correction grid* to account for local distortions in the planar grid coordinates. This is often the case when the planar grid has been based on first order networks established in the 19^{th} and early 20^{th} century using triangulations, pre-dating the more accurate satellite based techniques in use today. These older measurements, although quite an achievement in their time, typically resulted in long wavelength ($> 30km$) distortions in the first order networks, which were the basis for all other (secondary and lower order) measurements, and are therefore present in all planar grid coordinates. In order for satellite data, which are not related to the first order networks, to be transformed into planar grid coordinates and to used together with already existing data, many national mapping agencies decided to adopt a conventional correction grid to their planar coordinates. So, if the planar

[1]OGP Publication 373-7-2, Geomatics Guidance Note number 7, part 2, June 2013, `http://www.epsg.org`

coordinates are converted into latitude and longitude (to be used together with other satellite data), the correction grid corrects for distortions in the planar grid coordinates. If, on the other hand, latitude and longitude is converted to grid coordinates, (conventional) distortions are re-introduced so that the satellite data, expressed in grid coordinates, matches existing datasets.

Working with planar grid coordinates to compute distances, angles and area's is much more convenient than using geographical coordinates. However, one should be aware that in the map projection small distortions are introduced. For example, an azimuth computed from grid coordinates may not be referring to true North because of *meridian convergence* in azimuthal and conic projections. Meridian convergence is defined as the angle meridians make with respect to the grid y-axis. Also, sometimes corrections need to made for distances and surface area's. These corrections are usually quite small and well known. If they become too large it may be necessary to reduce the area of the projection, e.g. by defining different zones, each with a different natural origin or central meridian (or parallel). This approach is for instance used by the popular Universal Transverse Mercator (UTM) projection, which uses between 80^oS and 84^oN latitude 60 zones, each of 6^o width in longitude, centered around a central meridian. However, the Netherlands falls in two zones, 31N and 32N, which is not very convenient and may explain why the UTM projection is not used very often in the Netherlands except off-shore on the North Sea. UTM has also been the projection of choice for the European Datum 1950 (ED50).

6

Datum transformations and coordinate conversions

In the previous chapters several types of coordinate systems and representations have been introduced, such as Cartesian coordinates, geographic coordinates and grid (map) coordinates, including operations that can be performed on them. Coordinates systems also have a geodetic *datum*. The geodetic datum, or *datum*, specifies how a coordinate system is linked to the Earth: it consists of parameters that describe how to define the origin of the coordinate axis, how to orient the axis, and how scale is defined. Two types of coordinate operations have to be distinguished

coordinate conversions These are conversions from Cartesian into geographical coordinates, geographical coordinates into grid (map) coordinates, geocentric Cartesian into topocentric, geographic into topocentric, etc., and vice versa. These are operations that operate on coordinates from the same *datum*. In general, one type of coordinates can be converted into another, without introducing errors or loss of information, as long as no change of datum is involved.

datum transformation This changes the *datum* of the reference system, i.e. how the coordinate axis are defined and how the coordinate system is linked to the Earth. Datum transformations typically involve a 7-parameter similarity transformation between Cartesian 3D coordinates.

See Figure 6.1.

Coordinate conversions depend only on the chosen parameters for the reference ellipsoid, such as the semi-major axis a and flattening f, and the chosen map projection and projection parameters. Once these are selected and remain unchanged coordinate conversions are unambiguous and without loss of precision. Table 6.1 gives the parameters for three commonly used ellipsoids in the Netherlands. It is common that different datums use different ellipsoids and map projections. Therefore, coordinate transformations between two different datums not only involve a 7-parameter similarity transformation, but often also a change in the reference ellipsoid, type of map projection and projection parameters.

Datum transformations are transformations between coordinates of two different reference systems. Usually this is a 7-parameter similarity transformation between Cartesian coordinates of both systems, as shown in Figure 6.1, but if a dynamic Earth is considered with moving tectonic plates and stations, the similarity transformation can be time dependent with 14 instead of 7 parameters. Affine or polynomial transformations between geographic or grid coordinates

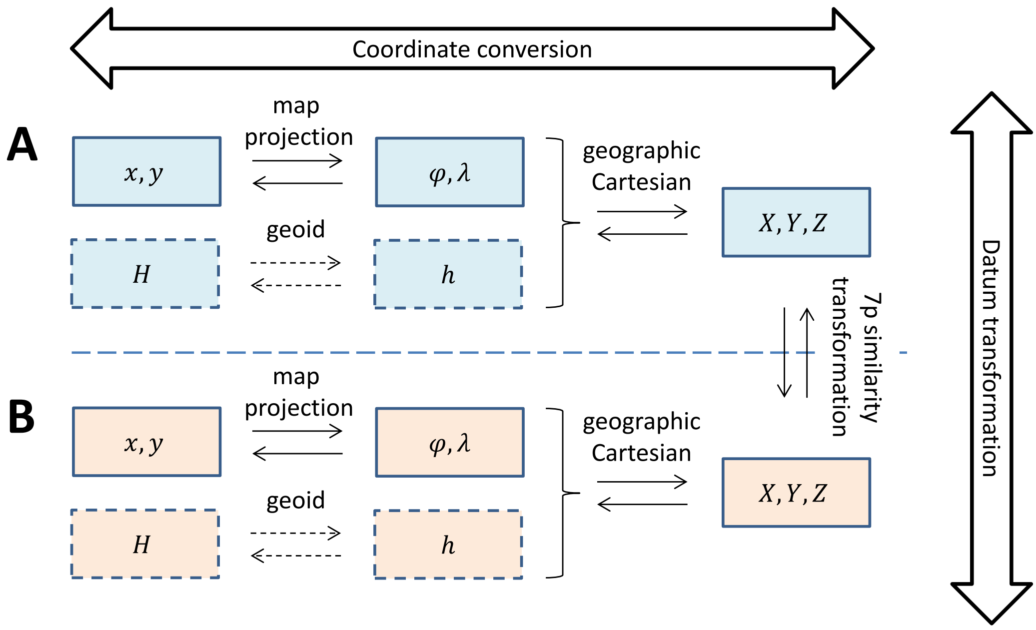


Figure 6.1: Coordinate conversions and datum transformations. Horizontal operations represent coordinate conversions. The vertical operations are datum transformations from system A to B. Not shown in this diagram are polynomial transformations (approximations) directly between map coordinates or geographic coordinates of the two systems.

Ellipsoid	a [m]	$1/f$ [-]	GM $[m^3/s^2]$
Bessel (1841)	6 377 397.155	299.152 812 8	
GRS80	6 378 137	298.257 222 100...	3 986 005 10^8
WGS84	6 378 137	298.257 223 563	3 986 004.418 10^8

Table 6.1: Common ellipsoids, with semi-major axis a, inverse flattening $1/f$, and if available, associated value for GM. The full list of ellipsoids is much longer. The very small difference in the flattening between WGS84 and GRS80 results in very tiny differences of at most 0.105 mm and can be neglected for all practical purposes.

of both systems are also possible, but not shown in Figure 6.1. These affine or polynomial transformations are mostly course approximations. A big difference with coordinate conversions is that the parameters for the datum transformation are often empirically determined and thus subject to measurement errors. More specific, three possibilities need to be distinguished. The first possibility is that the datum transformation parameters are conventional. This means they are chosen and therefore not stochastic. The datum transformation is then just some sort of coordinate conversion (which are also not stochastic). The second possibility is that the datum transformation parameters are given, but have been derived by a third party through measurements. What often happens is that this third party does new measurements and updates the transformation parameters occasionally or at regular intervals. This is also related to the concepts of *reference system* and *reference frames*. Reference frames are considered (different) realizations of the same reference system, with different coordinates assigned to the points in the reference frame, and often with different realizations of the transformation parameters. The station coordinates and transformation parameters are stochastic, so new measurements, mean new estimates that are different from the previous estimates. The third possibility is that there is no third party that has determined the transformation parameters, and you as user, have to estimate them using at least three common points in both systems. In this case you will need coordinates from the other reference system. Keep in mind that the coordinates from the external reference system all come from the same realization, or, reference frame.

The conversion from 3D coordinates to 2D grid (map) or geographic coordinates in Figure 6.1 is very straightforward: this is accomplished by simply dropping the height coordinate. The reverse, from 2D to 3D, is indeterminate. This is an issue when 2D coordinates (geographic or grid coordinates) have to be transformed into another datum or reference system, as this involved 3D Cartesian coordinates. However, in practice this issue is resolved easily by creating an artificial ellipsoidal height h, for instance by setting the ellipsoidal height $h = 0$. The resulting height in the new system will of course be meaningless, and has to be dropped, but as long as the chosen ellipsoidal height is within a few km of the actual height the error induced in the horizontal positioning will be small.

Many different datums and reference ellipsoids have been used in the history of geodesy. At the end of the 19th and beginning of the 20th century many countries developed their own national coordinate system, choosing an ellipsoid of revolution that best fitted the area of interest. In this pre-satellite era this meant doing astronomical observations to determine the origin and orientation of the ellipsoid. This resulted in many different ellipsoids and datums. In the 1950's USA initiated work on ED50 (European Datum 1950) which had as goal to link the various European datums and create a European reference system primarily for NATO applications. ED50 became also popular for off-shore work and to define the European borders. The satellite era saw the development of a number of world-wide reference systems, such as WGS60 and WGS72 which were based on Transit/Doppler measurements, with the most recent version WGS84 based on GPS in 1987. Later the International Terrestrial Reference Frame (ITRF) and the European Terrestrial Reference System ETRS89 were established which are more accurate than WGS84. These global reference frames also made it possible for the first time to determine accurate datum transformation parameters for the national reference frames that were established in the 19th and early 20th century.

With the advent of GPS and other space geodetic techniques the newer reference ellipsoids and datums are all very well aligned to the center of mass and rotation axis of the Earth. These geocentric reference ellipsoids are usually within 100m of the geoid world-wide. In pre-satellite days the reference ellipsoids were devised to give a good fit to the geoid only over the limited area of a survey, and it is therefore no surprise that there are significant differences in shape and orientation between the older and newer ellipsoids, resulting in large datum transformation parameters for the old systems. This also means that there are significant differences between latitude and longitude defined on one of the older legacy ellipsoids with respect to the satellite based datums. Confusing the datums of the latitude and longitude may result in significant positioning errors and could result in very hazardous situations.

It is therefore very important with coordinates (does not matter if they are Cartesian, geographic or grid coordinates) to always specify the reference system and reference frame they belong to. Also, for measurements on a dynamic Earth, it is important to document the measurement epoch. The reference system and reference frame of the coordinates, and the measurement epoch, are very important meta data for coordinates which should never be omitted. Failure to record or provide this important meta data will almost always result in confusion and may result in unnecessary costs.

7

Vertical reference systems

Until now the focus has been on the position of points on the Earth surface (location), using for instance geographic latitude and longitude on a reference ellipsoid, or x- and y-coordinates in a map projection. Now it is time to turn our attention to specifying the height, or elevation, of points.

The elevation of a point can only be expressed with respect to a another point or reference surface. In theory, it is possible to use the radius to the Center of Mass (CoM) of the Earth - also the origin of most 3D coordinates systems - as a measure for elevation. This is however only practical for Earth satellites, but not very practical for points on the surface of the Earth. Instead it will be much more convenient to use the height above a reference ellipsoid, as we have seen in Section 4.2, or to use a different - more physical - definition of height.

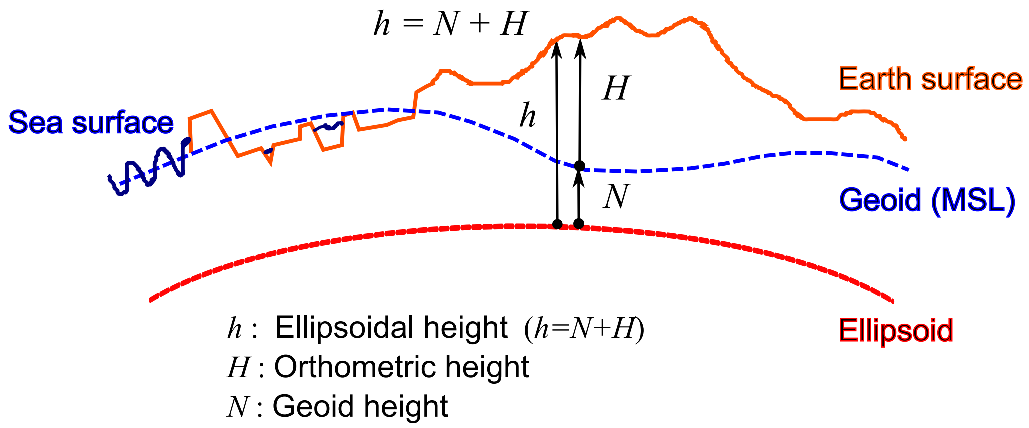


Figure 7.1: Relation between ellipsoidal height h, orthometric height H and geoid height N.

The main drawback of ellipsoidal height is that surfaces of constant ellipsoidal height are not necessarily equipotential surfaces. Hence, in an ellipsoidal height system, it is possible that water flows from a point with low 'height' to a point with a higher 'height'. This defies one on the main purposes of height measurements: defining water levels and water flow. Also, ellipsoidal heights are a relatively new concept, which can only be measured using space geodetic techniques such as GPS. Since heights play an important role in water management a necessary requirement is that water always flows from point with higher height to points with lower height.

In fact, instead of working with heights and height differences which are expressed in meters, it is actually more appropriate to use the gravitational potential W or potential differences ΔW. For an discussion of gravity and potential numbers the reader is referred to Appendix B. Here it suffices to recapitulate from Appendix B that an equipotential surface, with potential W_0 such that it more or less coincides with mean sea-level over Sea, fulfills all the requirements for a reference surface for the height. The unit of potential W is $[Nm/kg]$ which equals $[m^2/s^s]$. From Eq. B.2 and B.4 follows that the potential difference ΔW and height difference ΔH are related,

$$\Delta W = -g\ \Delta H \tag{7.1}$$

with g the gravitational acceleration in $[m/s^2]$. Note that the gravitation acceleration is not constant but depends on the location on Earth and height. *Orthometric heights* are defined by the inverse of Eq. 7.1,

$$H_{\text{orthometric}} = -\frac{1}{g}(W - W_0)\ \ . \tag{7.2}$$

with W_0 the potential of the chosen reference equipotential surface. *Normal heights* are based on the normal gravity γ instead of the gravitational acceleration g,

$$H_{\text{normal}} = -\frac{1}{\gamma}(W - W_0) \tag{7.3}$$

with γ the /em normal gravity from a normal (model) gravity field that matches gravitational acceleration for a selected reference ellipsoid with uniform mass equal to the mass of the Earth. In order to distinguish orthometric and normal heights from ellipsoidal heights we use a capital H for orthometric and normal heights, and a lower case h for ellipsoidal heights.

The relation between orthometric (normal) height H and ellipsoidal height h is given by the following approximation

$$h = N + H \tag{7.4}$$

with N the height of the geoid above the ellipsoid. This is illustrated in Figure 7.1. This approximation is valid near the surface of the Earth. In fact, some of the smaller effects, or the difference between normal and orthometric height, are often lumped with the geoid height into N, which then strictly speaking is a correction surface for transforming orthometric (or normal) height to ellipsoidal heights.

The zero point, or datum point, for the heights depends on the choice of W_0. This datum point is often defined based on tide-gauge data such that the geoid is close to mean sea-level. In Figure 7.2 the reference tide-gauges used for different European countries are shown, together with the difference in the height datums. The differences have been computed from the European re-adjustment of precise levellings. The effects of using different tide-gauges, and the differences between mean sea-level for the North Sea, Baltic Sea, Mediteranian and Black Sea are clearly visible. Also some countries, for instance Belgium, did not use mean sea-level to define their height datum but used low water spring as reference.

It is not necessary to use mean sea-level as reference surface for all applications. In particular for hydrography, it is more common to use the Lowest Astronomical Tide (LAT) as reference surface. This is *not* an equipotential surface as this reference surface also depends on the tidal variations.

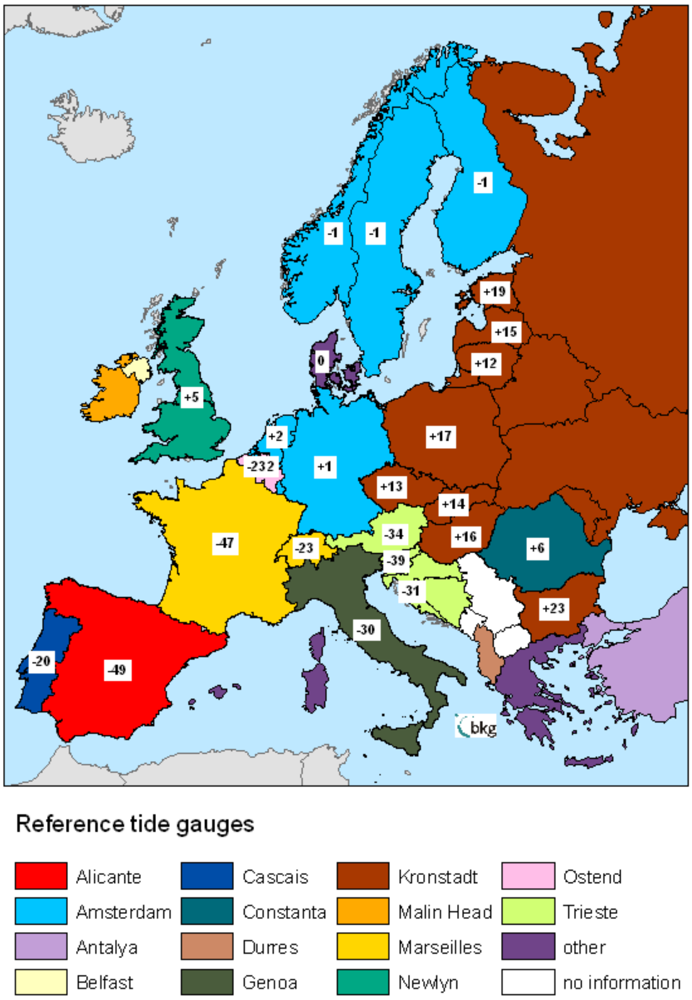


Figure 7.2: Differences between national height datums for Europe and reference tide-gauges (Source: BKG `http://www.bkg.bund.de`).

8
Common reference systems and frames

In this chapter a number of common reference systems and frames are discussed. These include the International Terrestrial Reference System (ITRS), which is realized through the International Terrestrial Reference Frames (ITRF) and the well know world wide WGS84 system used by GPS. Then the focus is shifted to regional reference systems and frames, with the European Terrestrial Reference System ETRS89 as our prime example. In the last part of this chapter the Dutch triangulation system RD and height system NAP, and their relation to the ETRS89, is discussed. This chapter does not pretend to give an overview of all available systems. Instead the focus is on reference systems and frames that are in use in the Netherlands, with the remark that many other countries have undergone similar developments and chosen similar solutions.

8.1. International Terrestrial Reference System and Frames

The International Terrestrial Reference System (ITRS) is a global reference system co-rotating with the Earth. It is realized through International Terrestrial Reference frames, which provides coordinates of a set of points located on the Earth's surface. It can be used to describe plate tectonics, regional subsidence or displacements in a global context, or to represent the Earth when measuring its rotation in space. The ITRF is maintained through an international network of space geodetic observatories and an international network of GNSS (GPS) tracking stations. The ITRF is the most accurate terrestrial reference frame to date. Therefore, it is frequently used as the basis for other reference frames, or, as an intermediate to describe relations between coordinate systems. For instance, the well known WGS84, used by GPS, is directly linked to the ITRF.

The definition of the International Terrestrial Reference System (ITRS) is based on IUGG resolution No 2 adopted in Vienna, 1991. As a consequence, the ITRS is

- a geocentric non-rotating system, with the center of mass being defined for the whole Earth, including oceans and atmosphere,
- the unit of length is the meter and its scale is consistent with the Geocentric Coordinate Time (TCG) by appropriate relativistic modeling,
- the time evolution of the orientation is ensured by using a no-net-rotation condition with regards to horizontal tectonic motions over the whole Earth,

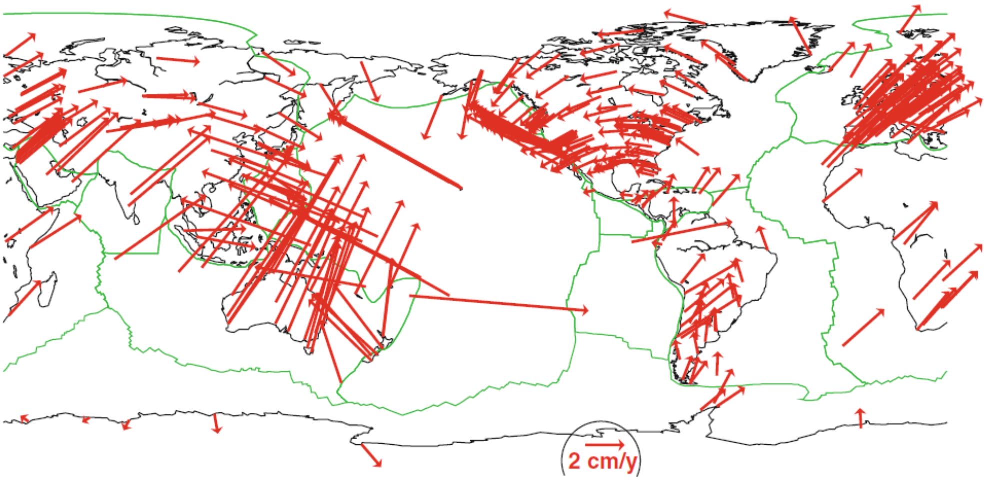


Figure 8.1: ITRF2008 velocity field with major plate boundaries shown in green (Figure from `http://itrf.ensg.ign.fr/`).

- the initial orientation is given by the Bureau International de l'Heure (BIH) orientation at 1984.0.

The ITRS is realized through a number of International Terrestrial Reference Frames (ITRF). Each individual ITRF contains station positions and velocities, often together with full variance matrices, computed using observations from space geodesy techniques such as Very Long Baseline Interferometry (VLBI), Lunar Laser Ranging(LLR), Satellite Laser Ranging (SLR), Global Positioning System (GPS) and Doppler Orbitography and Radiopositioning Integrated by Satellite system (DORIS). The stations are located on sites covering every continent and tectonic plate on Earth. To date there are twelve realizations of the ITRS: ITRF88, ITRF89, ITRF90, ITRF91, ITRF92, ITRF93, ITRF94, ITRF96, ITRF97, ITRF2000, ITRF2005 and ITRF2008. ITRF2013 is currently under preparation. The numbers in the ITRF designation specify the last year of data that were used. For example for ITRF97, which was published in 1999, space geodetic observations available up to and including 1997 were used, while for ITRF2000 an additional three years of observations, up to and including 2000, were used.

The realization of the ITRS is an on going activity resulting in periodic updates of the ITRF reference frames. These updates reflect

- improved precision of the station positions and velocities due to the availability of a longer time span of observations, which is in particular important for the velocities,
- improved datum definition due to the availability of more observations and better models,
- discontinuities in the time series due to earthquakes and other geophysical events,
- newly added and discontinued stations,
- and occasionally a new reference epoch.

All ITRF model the secular Earth's crust changes. The position at a specific epoch is given by

$$\mathbf{r}(t) = \mathbf{r}(t_0) + \dot{\mathbf{r}} \cdot (t - t_0) \tag{8.1}$$

The ITRF2008 velocities are given in Figure 8.1. The velocities are of the order of a few centimeters per year up to a decimeter per year for some regions. This means that for most applications velocities cannot be ignored. It also implies that when coordinates are distributed it is equally important to provide the epoch of observation to which the coordinates refer. Higher frequencies of the station displacements, e.g. due to solid Earth tides and tidal loading effects, can be computed using models specified in the IERS conventions, chapter 7 (`http://www.iers.org/IERS/EN/DataProducts/Conventions/conventions.html`).

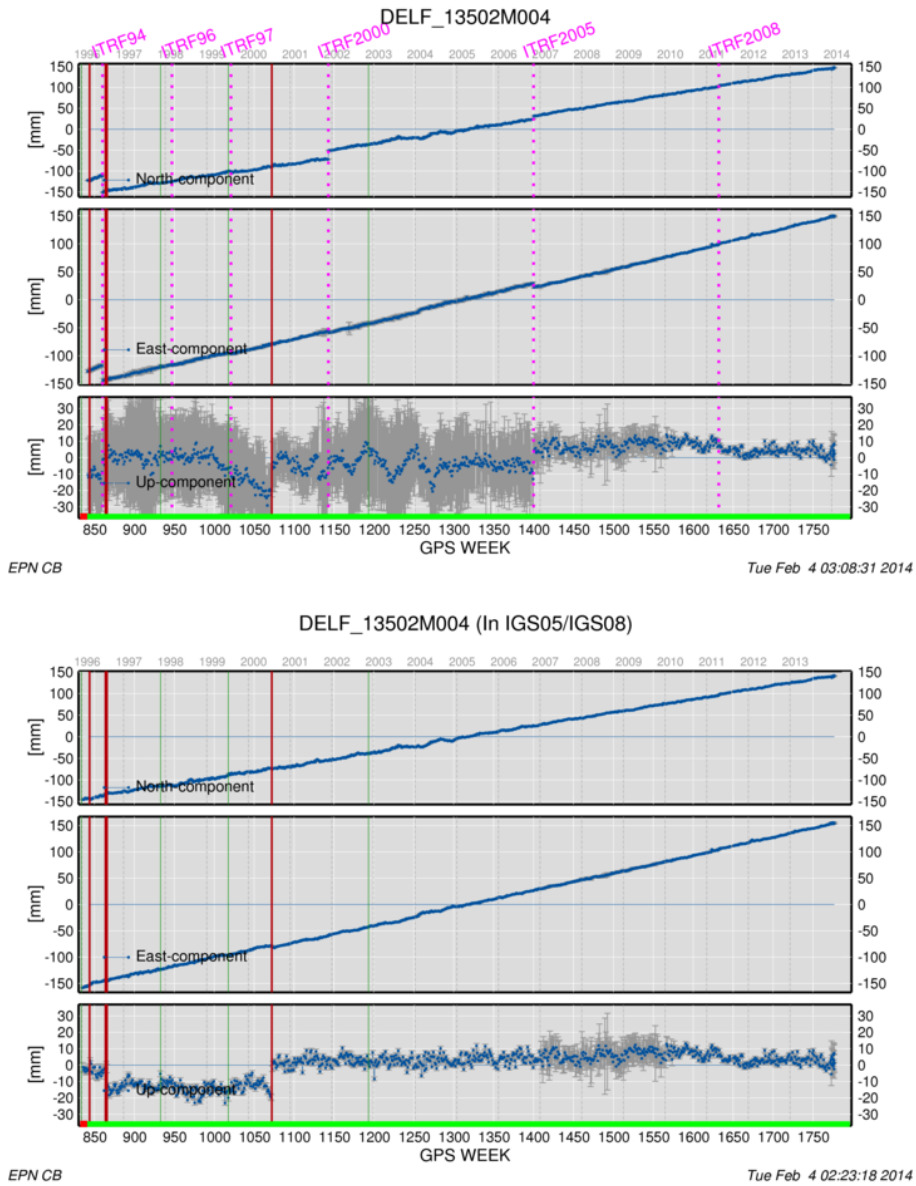


Figure 8.2: Time series of station positions of a GPS receiver in Delft from 1996-2014. The top figure shows the time series in the ITRFyy reference frame that was current at the time the data was collected. The bottom figure shows the data after re-processing in the IGS05/IGS08 reference frame that is based on the most recent ITRF2008 frame. The vertical red lines indicate equipment changes (Figures from `http://www.epncb.oma.be/`).

The datum of each ITRF is defined in such a way as to maintain the highest degree of continuity with past realizations and observational techniques. The ITRF origin and rates are essentially based on the Satellite Laser Ranging (SLR) time series of Earth orbiting satellites. The ITRF orientation is defined in such a way that there are null rotation parameters and null rotation rates with respect to ITRF2000, whereas for ITRF2000 is no net rotation with respect to the NNR-NUVEL1A plate tectonic model is used. These conditions are applied over a core network of selected stations. The ITRF scale and scale rate are based on the VLBI and SLR scales/rates. The role of GPS in the ITRF is mainly to tie the sparse networks of VLBI and SLR stations together, and provide stations with a global coverage of the Earth.

Although the goal is to ensure continuity between the ITRF realizations as much as possible there are transformations involved between different ITRF that reflect the differences in datum realization. Each transformation consists of 14 parameters, a 7-parameter similarity transformation for the positions involving a scale factor, three rotation and three translations, and a 7-parameter transformation for the velocities involving a scale rate, three rotation rates and three translation rates. The translation parameters and formula are given in `http://itrf.ensg.ign.fr/doc_ITRF/Transfo-ITRF2008_ITRFs.txt`. The transformation formula is essentially Eq. 3.9.

In Figure 8.2 a time series of station positions for a GPS receiver in Delft is shown. The top figure shows a couple of features: (1) the secular motion of the point, (2) jumps in the coordinate time series and velocity whenever a new ITRF is introduced, and (3) discontinuities due to equipment changes (mainly antenna changes). The bottom figure shows the time series after reprocessing in the most recent reference frame. The jumps due to changes in the reference frame have disappeared and the day-to-day repeatability has been improved considerably due to improvements in the reference frame and processing strategies. One feature is not shown in Figure 8.2, and that is the effect of earthquakes. Stations which are located near plate-boundaries would experience jumps and post-seismic relaxation effects in the time series due to earthquakes. Although geophysically very interesting, this makes stations near plate boundaries less suitable for reference frame maintenance. Figure 8.3 shows the same time series, but in the European ETRF2000 reference frame, which is discussed in Section 8.3.

The International Terrestrial Reference Frame (ITRF) is maintained by the International Earth Rotation and Reference Systems Service (IERS). For more information see also `http://itrf.ensg.ign.fr/`. The IERS is also responsible for the International Celestial Reference Frame (ICRF) and the Earth Orientation Parameters (EOPs) that connect the ITRF with the ICRF. The observational techniques are organized in services, such as the International GNSS Service (IGS), International Laser Ranging Service (ILRS) and International VLBI Service (IVS). For instance, the IGS is a voluntary organization of scientific institutes that operates together a tracking network of over 300 stations, several analysis and data centers, and a central bureau. The main product of IGS are precise orbits for GNSS satellites (including GPS), satellite clock errors and station positions all in the ITRF. Web-based Precise Point Positioning (PPP) services, which utilize IGS orbits and clocks, allow GPS users to directly compute positions in ITRF. At the same time many regional and national institutions have densified the IGS network to provide dense regional and national networks of station coordinates in the ITRF.

When working with the ITRF it is typical to provide coordinates as Cartesian coordinates. However, the user is free to convert these into geographic coordinates. The recommended ellipsoid for ITRS is the GRS80 ellipsoid. This is the same ellipsoid as used for instance by WGS84.

8.2. World Geodetic System 1984 (WGS84)

The USA Department of Defense (DoD) World Geodetic System 1984 (WGS84) is probably by far the best known terrestrial reference system. Which is understandable considering the popularity of Global Positioning System (GPS) receivers, but it is also somewhat surprising considering that WGS84 is a US military system.

In general WGS84 is identical to the ITRS and its realizations ITRFyy at one meter level.

Therefore, in practice, when precision does not really matter and the user is satisfied with coordinates at the one meter level, coordinates in ITRS, or derivatives of ITRS (like the European ETRS89 which is discussed next in Section 8.3), are sometimes simply referred to as "WGS84".

For civilian users WGS84 coordinates are only obtainable through the use of GPS. The only WGS84 realization available to civilian users are the GPS broadcast orbits as civilian users have no direct access to DoD tracking sites or tracking data. This means that for civilian users the accuracy of WGS84 is restricted to the accuracy of the GPS broadcast orbits, which is of the order of a few meters. Users may try to improve the accuracy to a few decimeters by taking averages of GPS station positions over several days, but then if accuracy is really an issue it would be much better to switch to ITRF or ETRS89.

In fact there are different WGS84 realizations. Until GPS week G730 WGS84 was based on the US Navy Doppler Transit Satellite System. Newer realizations of WGS84 are coincident with ITRF at about 10-centimeter level. For these realizations there are no official transformation parameters. The newer realizations are adjusted occasionally in order to update the tracking station coordinates for plate velocity. These updates are identified by the GPS week, i.e. WGS84(G730, G873 and G1150).

The use of WGS84 should be avoided for applications other than for hiking, regular navigation, and other non-precision applications. The WGS84 is not suited for applications requiring decimeter, centimeter or millimeter accuracy.

8.3. European Terrestrial Reference System 1989 (ETRS89)

The European Terrestrial Reference System 1989 (ETRS89) is the standard coordinate system for Europe. It is the reference system of choice for all international geographic and geodynamic projects in Europe. The system also forms the backbone for many national reference systems. Although the ITRS plays an important role in studies of the Earth's geodynamics it is less suitable for use as an European georeferencing system. This is because in ITRS all points in Europe exhibit a more-or-less similar velocity of a few centimeters per year, as was shown in Figures 8.1 and 8.2.

The ETRS89 terrestrial reference system is coincident with ITRS at the epoch 1989.0 and fixed to the stable part of the Eurasian Plate. The year in the name ETRS89 refers explicitly to the time the system was coincident with ITRS[1]. ETRS89 is accessed through the EUREF Permanent GNSS Network (EPN), a science-driven network of continuously operating GPS reference stations with precisely known station positions and velocities in the ETRS89, or through one of many national or commercial GPS networks which realize ETRS89 on a national scale.

Station velocities in ETRS89 are generally very small because ETRS89 is fixed to the stable part of the Eurasian plate. Compared to ITRS, with station velocities in the order of a few centimeter/year, station velocities in ETRS89 are typically smaller than a few mm/year. This is clearly illustrated in Figure 8.3 for the GPS station in Delft that was also used for Figure 8.2. Of course, there are exceptions in geophysically active areas, but for most practical applications, one may ignore the velocities. This makes ETRS89 well suited for land surveying, high precision mapping and geographic reference systems (GIS) applications. Also, ETRS89 is well suited for the exchange of geographic data sets between European national and international institutions and companies.

The ETRS89 system is realized in several ways, and like with ITRS, realizations of a system are called reference frames. By virtue of the ETRS89 definition, which ties ETRS89 to ITRS at

[1]Sometimes people believe the coordinates should be at epoch 1989.0, but this is not necessary, as coordinates can be given at any epoch.

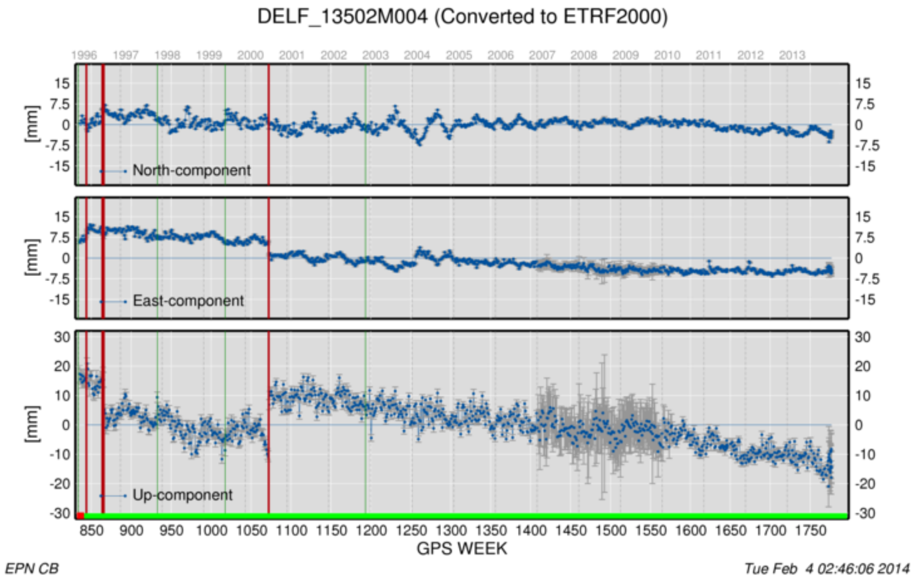


Figure 8.3: Time series of station positions of a GPS receiver in Delft from 1996-2014 in the European reference frame ETRF2000. The vertical red lines indicate equipment changes (Figure from `http://www.epncb.oma.be/`).

epoch 1989.0 and the Eurasian plate, for each realization of the ITRS (called ITRFyy), also a corresponding frame in ETRS89 can be computed. These frames are labelled ETRFyy. Since each realization also reflects improvements in the datum definition of ITRF, which results in small jumps in the coordinate time series, the EUREF Technical Working Group recommends not to use the ETRF2005 and ETRF2008 for practical applications. Instead the ETRF2000 is adopted as a conventional frame of the ETRS89 system. However, in order to benefit from improvements in the quality of the ITRF2005 and ITRF2008 station positions and velocities the ETRF2000 have been updated twice, resulting in a new set of positions and velocities called ETRF2000(R05) and ETRF2000(R08). ETRF2000(R05) and ETRF2000(R08) reflect improvements in the station positions and velocities, but the datum definition is left as much as possible unchanged with respect to ETRF2000, and thereby minimizing the effects of changes in station coordinates in ETRS89 at large.

Another way to realize ETRS89 is by using GNSS campaign measurements or a network of permanent stations. From 1989 onwards many national mapping agencies have organized GPS campaigns to compute ETRS89 coordinates for stations in their countries, and then link their national networks to ETRS89. Later on these campaigns were replaced by networks of permanent GPS receivers. These provide users with downloadable GPS data and coordinates in ETRS89 that they can use together with their own measurements. The permanent networks also provide 24/7 monitoring of the reference frames. An example is the Active GPS Reference System for the Netherlands (AGRS.NL), which was established in 1997. Nowadays the Dutch Kadaster, and two other commercial providers, operate a real-time network RTK services (NETPOS, 06-GPS, LNRNET) that provide GPS data and corrections in real-time which allows instantaneous GPS positioning in ETRS89 at the centimeter level. All Dutch network RTK services are certified by the Dutch Kadaster and thus provide a national realization of ETRS89 linked to the ETRF2000 reference frame. Similar services are operated in many other European countries.

When working with ETRS89 it is typical to provide coordinates as either Cartesian coordinates, or, geographic coordinates and ellipsoidal height (using the GRS80 ellipsoid). There is no European standard for the type of map projection to be used, so the user can still select a favorite map projection depending on the application at hand. In several countries, including the Netherlands, a conventional transformation from ETRS89 to grid (map) coordinates that resemble the old national systems are provided.

The ETRS89 was established in 1989 and is maintained by the sub-commission EUREF (European Reference Frame) of the International Association of Geodesy (IAG). ETRS89 is supported by EuroGeographics and endorsed by the European Union (EU). All contributions to the EPN are voluntary, with more than 100 European agencies and universities involved. The reliability of the EPN network is based on extensive guidelines guaranteeing the quality of the raw GPS data to the resulting station positions and on the redundancy of its components. The GPS data is also used for a wide range of scientific applications such as the monitoring of ground deformations, sea level, space weather and numerical weather prediction. On other continents solutions similar to ETRS89 have been adopted.

A service to convert coordinates from ITRS to ETRS89, and vice versa, is provided at `http://www.epncb.oma.be/_productsservices/coord_trans/`. Specifications for the transformation procedure and reference frame fixing can be found here `http://etrs89.ensg.ign.fr/memo-V8.pdf`.

8.4. Dutch Triangulation System (RD)

The Dutch Triangulation System (RD), in Dutch *Rijksdriehoeksstelsel*, has a history dating from the 19th Century. Following a century of traditional triangulations GPS started to replace triangulation measurements in 1987. The increasing use of GPS resulted in a redefinition of RD in 2000, whereby from 2000 onwards RD was linked directly to ETRS89 through a transformation procedure called RDNAPTRANS.

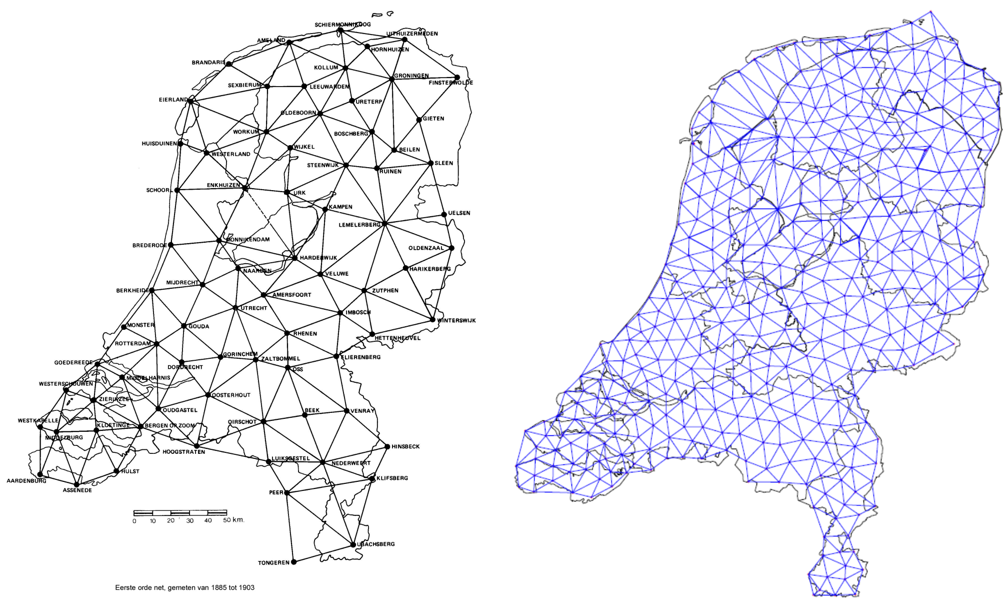


Figure 8.4: First order triangulation network for the Netherlands of 1903 (left) and GPS base network of 1997 (right). Figure from de Bruijne et al., 2005.

8.4.1. RD1918

The first-order triangulation grid was measured in the years between 1885 and 1904 (Figure 8.4). The church tower of Amerfoort was selected as origin of the network and as reference ellipsoid the ellipsoid of Bessel (1841) was chosen. The scale was derived from a distance measurement on a base near Bonn, Germany. Between 1896 and 1899 geodetic-astronomic measurements were carried out at thirteen points throughout the Netherlands in order to derive the geographical longitude and latitude of the origin in Amersfoort and the orientation of the grid. As map projection an oblique stereographic double projection was selected (Heuvelink, 1918). The projection consists of a Gauss-Schreiber conformal projection of Bessel's ellipsoid (1841) onto a sphere, followed by a oblique stereographic projection of the sphere to a tangential plane, as shown in Figure 8.5.

The stereographic projection is a perspective projection from the point antipodal to the central point in Amersfoort on a plane parallel to the tangent at Amersfoort. This projection is conformal, which means the projection is free from angular distortion, and that lines intersecting at any specified angle on the ellipsoid project into lines intersecting at the same angle on the projection. Therefore, meridians and parallels will intersect at 90^o angles in the projection, but, except for the central meridian through Amersfoort, meridians will converge slightly to the North and do not have constant x-coordinate in RD. This is known as *meridian convergence*. This projection is not equal-area. Scale is true only at the intersection of the projection plane with the sphere and is constant along any circle around the center point in Amerfoort. However, by letting the tangential projection plane intersect the sphere, the scale distortions at the edges of the projection domain will be within reasonable limits.

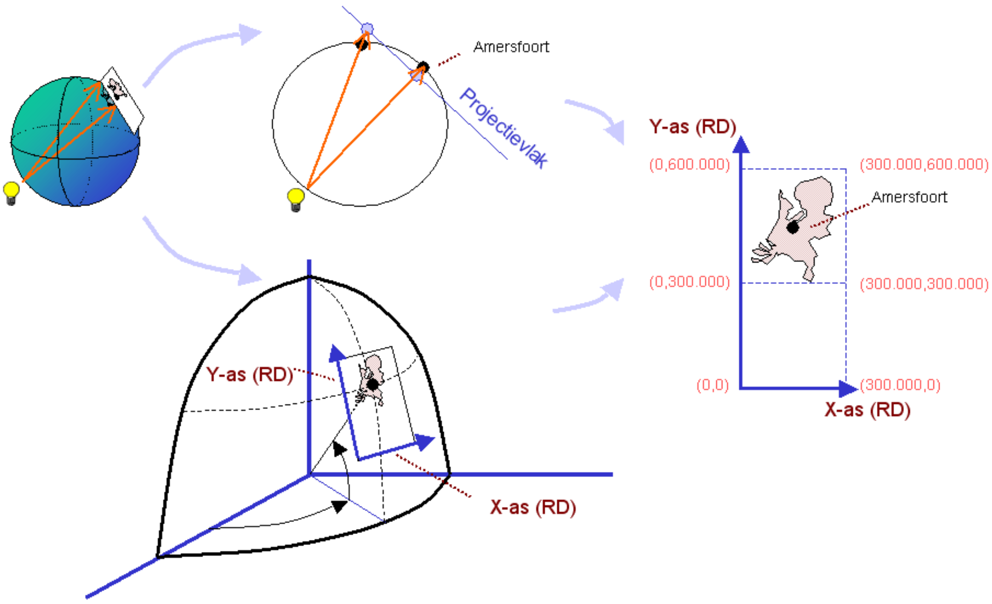


Figure 8.5: RD double projection (Bessel ellipsoid → Sphere → Plane) and definition of RD coordinates (Figure T.Nijeholt at nl.wikibooks).

During the years between 1898-1928 a densification programme was carried out which resulted in the publication of 3732 triangulation points. At the time of publication already 365 points had disappeared or were disrupted. To prevent further reduction in points and maintain the network the "Bijhoudingsdienst der Rijksdriehoeksmeting" was established at the Dutch Cadastre. From 1960 to 1978 a complete revision was carried out and the RD system was also connected to neighboring countries, which resulted in a reference frame with roughly 6000 points at mutual distances of 2.5-4 km. To prevent confusion between old and new coordinates the new coordinates were shifting the 155 km to the West and 463 km to the South (False Easting and Northing). This resulted in only positive coordinates and y-coordinates that are always larger than the x-coordinates.

Starting in 1993 a so-called GPS base network of 418 points was established to offer GPS users a convenient way to connect to the RD system. See Figure 8.4. Most of traditional triangulation points, many of which are church spires or towers, are not accessible to GPS measurements. The points in the GPS base network have an unobstructed view of the sky and are easily accessible by car. The GPS base network points are located at distances of 10 to 15 km from each other, which is well suited for GPS baseline measurements. The points in the GPS base network have been connected to neighboring RD points to determine RD coordinates and by second-order levelling to neighboring NAP benchmarks to determine heights. In addition, the point in the GPS base network were connected by GPS measurements to points in the European ETRS89 system. As a result the GPS base network points have measured coordinates both in RD/NAP and ETRS89. This made it possible - for the first time - to study systematic errors in the RD system. It was found that the RD system of 1918 has systematic errors of up to 25 cm with significant regional correlations, as shown in Figure 8.6 for the province of Friesland. In the age of GPS this may seem as a large number, but in 1918 this was an excellent accomplishment.

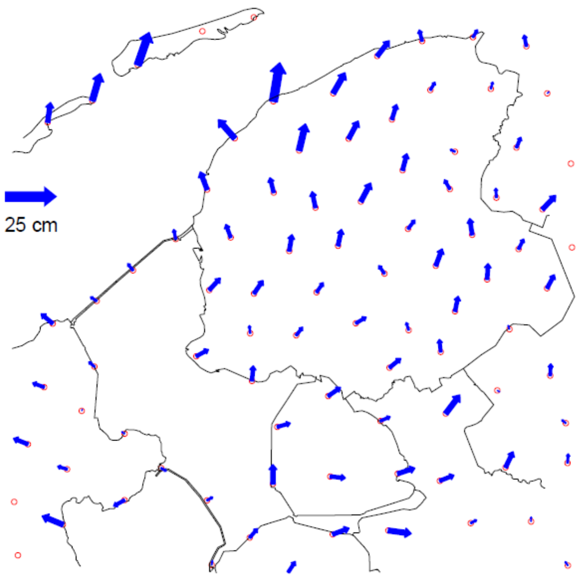


Figure 8.6: Differences between RD and ETRS89 based coordinates for the GPS Kernnet in the province of Friesland, showing significant regional correlation between the vectors (Figure from de Bruijne et al., 2005).

The systematic errors in the RD system were never an issue until the introduction of GPS. Before GPS, all measurements were connected to nearby triangulation points which could almost always be found within a radius of 3-4 km, and users never noticed large discrepancies unless the triangulation points were damaged. However, with GPS it became routine to measure over distances of 15 km up to 100 km to the nearest GPS basenet point or permanent GPS receiver, and then systematic errors in RD will become noticeable. This led to a major revision in the definition of RD in 2000.

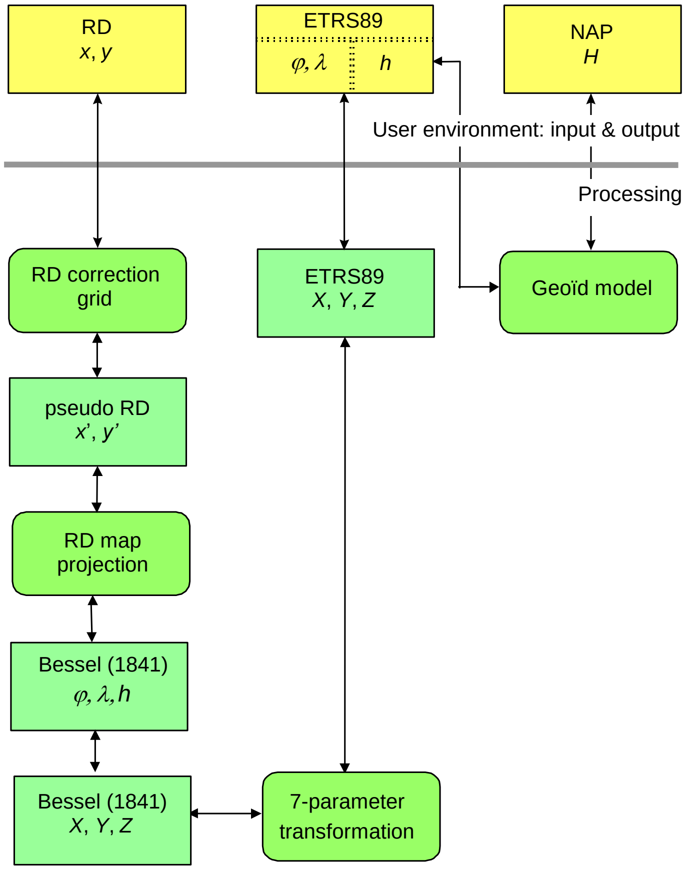


Figure 8.7: RDNAPTRANS transformation procedure. The figure outlines the relationships and transformations between ETRS89, RD2000 and NAP (Figure after de Bruijne et al., 2005). The coordinates below the line, with the exception of Cartesian coordinates in ETRS89, are used only for computational purposes and should never be published or distributed to other users.

8.4.2. RD2000

In 2000 a new definition of the RD grid was adopted and assigned the name RD2000. This definition replaces Heuvelink's (1918) definition, which is since then referred to as RD1918.

In the new definition RD2000 is based on ETRS89. Within this new definition two types of coordinates are allowed: (1) Cartesian or geographic coordinates in ETRS89, and (2) RD x- and y-coordinates. The big difference is that the RD coordinates are now obtained by a conventional transformation from the ETRS89 coordinates. The transformation has been assigned the name RDNAPTRANS. This definition was chosen to minimize the impact for users. GPS users can happily work with ETRS89, and if they wish, transform their coordinates to RD at the very last stage. Owners of large databases with geographic information in RD have their investments protected and do not need to make changes.

The new definition has not changed the published RD coordinates significantly. In addition, the European ETRS89 frame was introduced as the three-dimensional reference frame for the Netherlands. This was effected by the publication of the ETRS89 coordinates along with RD coordinates.

The RDNAPTRANS transformation procedure of Figure 8.7 is an essential part of the RD2000 definition. It has four main elements:

1. 7-parameter transformation from ETRS89 to an intermediate system defined on the Bessel, 1841, ellipsoid, including conversions from Cartesian to geographic coordinates), resulting in latitude, longitude and height on the Bessel, 1841, ellipsoid.

2. a map projection using the same constants and definitions as RD1918, inclusive a false Easting and Northing of 155 km and 463 km. The projected coordinates are referred to as "pseudo RD".

3. a conventional correction grid for the x- and y-coordinates in RD, which 'corrects' the pseudo RD coordinates of the previous step for the systematic distortions in the old RD-grid. The corrections are obtained by interpolation in the correction grid.

4. NLGEO2004 geoid for the conversion between NAP heights and height on the GRS80 ellipsoid, which will be discussed next in Section 8.5.

The transformation procedure works in both directions, and, both for 2D and 3D coordinates. In case no heights are available the points are assumed to lie on the ellipsoid with ellipsoidal height of zero. In these cases geographical latitude and longitude can be used, but heights, as well as 3D Cartesian coordinates are meaningless. Outside the transformation procedure the use of geographic coordinates on the Bessel-1841 ellipsoid and pseudo-RD coordinates is not recommended. For geographic coordinates solely ETRS89 coordinates should be used within the Netherlands. For RD-coordinates only coordinates that *include* the systematic distortions should be used. Failing to do so may result in pollution and errors of existing databases based on RD.

Since 2000 two minor revisions of RD2000 occurred. These were related to changes in the European reference frame, and only affected the 7-parameter transformation. The modified transformation procedures are referred to as RDNAPTRANS2004 and RDNAPTRANS2008. Similar minor revisions may be possible in the future , as there is a need to maintain a close link with the most up to date realizations of ETRS89 as well as to retain as constant as possible RD coordinates.

8.4.3. Future of RD coordinates

Although the RDNAPTRANS transformation procedure is well documented and example source code in C and Matlab is available free of charge, this transformation procedure is only supported by a few Geographic Information System (GIS) packages. The correction grid is often not directly supported by software and the map projection chosen for RD is considered to be exotic. This, combined with the fact that the 7-parameter transformation entails a significant shift and rotation, has sparked a discussion whether RD coordinates should be replaced by a different map projection. This discussion is on-going but could result in a recommendation for one (or more?) alternative map projections (besides the current RDNAPTRANS procedure which will be retained for legacy purposes). At the heart of the discussion is that many users find it difficult to work directly with geographic coordinates (latitude and longitude) and prefer working with rectangular 2D grid coordinates, but lack the expertise and software to do the conversion to RD. However, it will be unlikely that RD coordinates will disappear completely as there are a large number of legacy products based on RD.

8.5. Amsterdam Ordnance Datum - Normaal Amsterdams Peil (NAP)

The Amsterdam Ordnance Datum, in Dutch *Normaal Amsterdams Peil (NAP)*, is the official reference system for heights in the Netherlands. It is also the datum for the European Vertical Reference System (EVRS).

The history of the Dutch height datum goes back to a bolt installed in Amsterdam's shipbuilding district as early as 1556. A century later, in 1682, eight stone datum points were incorporated in the then new locks along the IJ waterway, defining a height datum that was called *Amsterdamse (Stadts)peyl*. This datum was extended during the 18th and beginning of the 19th Century to include the then Zuiderzee and the large rivers, and in 1818, King William I decreed the use of the *Amsterdams Peil (AP)* as the general reference point for water levels. At that time many different height datums were in use in the Netherlands which needed to be connected through levellings.

Figure 8.8: Levelling lines of the 5th precise levelling in the Netherlands, 1996-1999 (Figure from de Bruijne et al., 2005).

The 1st national precise levelling dates from the period 1875-1885, including 410 already existing points and 2100 km of continuous levelling lines. The datum was based on five remaining stone datum points in the Amsterdam locks. To distinguish the newly derived heights from previous results the name *Normaal Amsterdams Peil (NAP)*, the Amsterdam Ordnance Datum, was introduced. The 2nd precise levelling was carried out between 1926-1940. At this time only two of the original stone datum points in the Amsterdam locks remained. Therefore, this levelling included the construction of 18 new 1st-order and 29 new 2nd-order underground reference points in - presumably - stable geological strata throughout the Netherlands. The total length was 4592 km and included new routes, such as the one over the Afsluitdijk. Work on the 3rd precise levelling started in 1950. A large number of points were lost during the war and needed to be replaced. The 3rd levelling was also needed to get insight into the reliability of the underground reference points. During this levelling a trial was made using hydrostatic levelling to connect islands and to cross some major rivers. During the 3rd levelling only one stone datum point in Amsterdam remained. The level of this last stone datum point, which soon would be lost due to construction work, was transferred to a new underground reference point on the Dam Square in Amsterdam and assigned the height NAP +1.4278 m. This datum point is in a certain sense symbolic, as the height datum is actually defined based on the underground reference points in geologically more stable locations.

The 4th precise levelling was carried out between 1965-1978. Tests with hydrostatic levellings during the third levelling yielded extremely accurate results, and for this reason hydrostatic levelling was extended to 7 large loops connecting some 30 underground reference points, which was further extended by optical measurements. It also saw the installation of additional underground posts (nulpalen) in the vicinity of "tide" gauges (water level gauges) to provide a stable level for water level measurements.

In the 1990's it became clear that motions in the Netherlands' subterranean strata have a major influence on the NAP grid. Geophysical models indicate that Post-Glacial uplift of Scandinavia results in a slight tilting of the subterranean strata in the Netherlands, with the West of the Netherlands sinking by approximately 3 cm per century. This was confirmed by analysis of precise levelling measurements, but the uncertainty in the data was very high, and until then the height of the underground datum points had never been adjusted. Because of policy oriented issues, related to the protection from floods, more insight was needed into the height changes of the underground datum points. For this reason the 5th precise levelling was carried out between 1996-1999, see Figure 8.8. This was the first time that a combination of optical and hydrostatic levelling, satellite positioning (GPS) and gravitational measurements, were used. It also include ice levelling measurements on the IJsselmeer and the Markermeer. The levelling measurements still constitute the basis for the primary NAP grid. The gravity measurements constituted the 2nd measurement epoch of the Dutch gravitational grid. They served to get an independent insight into subterranean movements. The GPS measurements served to enhance the levelling net over greater distances, and to connect the levelled NAP heights to ETRS89. The network of the 5th precise levelling was also connected to the German and Belgian networks. These connections played an important role in the establishment of a European Vertical Reference System (EVRS) which uses the same "Amsterdam" datum as the NAP grid.

The primary NAP grid is comprised of about 300 underground points and 70 posts (nulpalen). The underground points are not accessible to the public., but provide an as stable as possible basis for measurements of the secondary NAP grid. The secondary NAP grid consists mainly of bronze bolts, with a head of between 20 - 25 mm in diameter, that are fitted to a building or other structure with an appropriate stability. The heights of these bronze bolts have been determined by levelling loops with an average length of 2 km with a precision better than 1 mm/km. A bronze marker is installed after every kilometer. There are about 35,000 of these markers (peilmerken) installed in the Netherlands. The heights of the markers are published by RWS (`http://www.rws.nl/zakelijk/databestanden/normaal_amsterdams_peil/`). These markers serve as the basis for height determination by consulting engineers, water boards, municipalities, provinces, state, and other authorities,

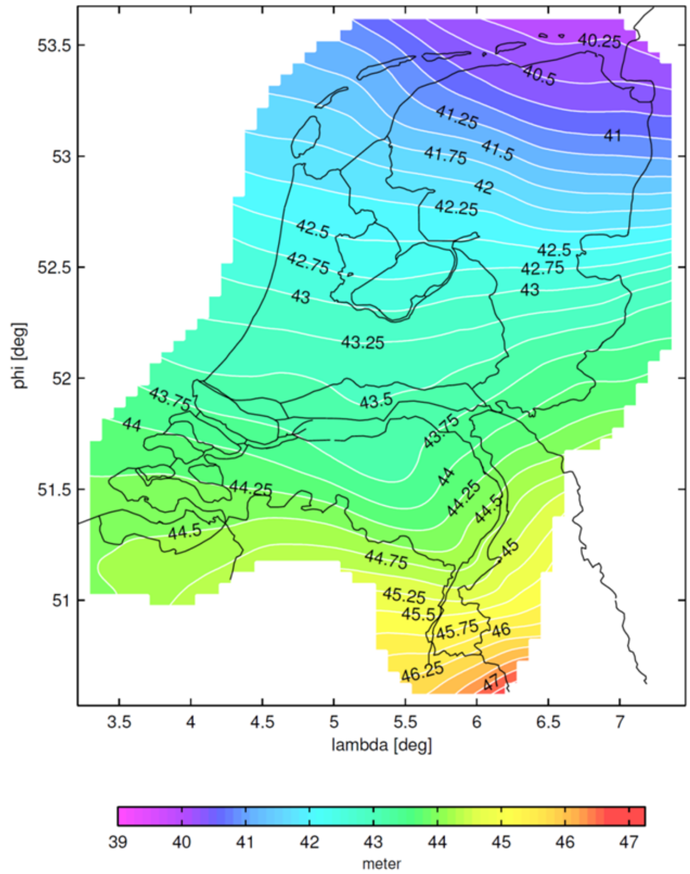


Figure 8.9: NLGEO2004 geoid for the Netherlands, with geoid height in [m] with respect to the GRS80 ellipsoid in ETRS89 (Figure from de Bruijne et al., 2005).

whereby one of these markers can almost always be found within a distance of 1 km.

GPS has not replaced levelling as much as it did with triangulation. There are two reasons for this; (1) GPS height are not as accurate as the horizontal positions, (2) levelled (orthometric) height and GPS (ellipsoidal) height are different things. See Chapter 7 for an explanation. Therefore, the dense NAP grid will not be outdated by GPS in the foreseeable future, like it did for the RD grid, and certainly not for applications requiring millimeter accuracy.

However, GPS can be used to obtain heights with an accuracy of about $1-2$ cm using the RDNAPTRANS procedure, as outlined in Figure 8.7. The transformation from ellipsoidal height to NAP heights, and vice versa, requires a correction for the geoid height.

Calculation of a geoid requires gravitational measurements over - in principle - the entire Earth. The larger scales depend mainly on satellite data, but for the highest precision at regional and national scales gravity measurements in and around the area of interest are needed. The first Dutch geoid, with a relative precision of 1 decimeter, became available in 1985. In order to improve this geoid in the period of 1990-1994 some 13,000 relative gravitational measurements were carried out in a grid of almost 8,000 points (1 point per 5 km2) in the Netherlands. The resulting geoid, which became available in 1996, had a precision of one to a few centimeters. This was the first accurate geoid model of the Netherlands. This model was improved in 2004, resulting in the NLGEO2004 model that is used by RDNAPTRANS, see Figure 8.9. The improvements resulted from using additional gravitational measurements on Belgian and German territory and a set of 84 GPS / levelling points from the 5th precise levelling to connect NAP and ETRS89. The NLGEO2004 model has a precision better than 1 cm in geoid height. The relative precision for two points close together is approximately 3.5 mm, increasing to 5 mm for two points separated by a distance of 50 to approximately 7 mm for two points separated by a distance of 120 km (de Bruijne et al., 2005). Therefore, the accuracy of GPS determined NAP heights using RDNAPTRANS will largely depend on the precision of the GPS measurement.

9
Further reading

Altamimi Z., X. Collilieux, L. Métivier, ITRF2008: an improved solution of the International Terrestrial Reference Frame, Journal of Geodesy, vol. 85, number 8, page 457-473, doi:10.1007/-s00190-011-0444-4, 2011.

A. de Bruijne, J. van Buren, A. Kösters and H. van der Marel, De geodetische referentiestelsels van Nederland; Definitie en vastlegging van ETRS89, RD en NAP en hun onderlinge relatie, Nederlandse Commissie voor Geodesie 43, Delft, 2005, 132 pagina's. ISBN-13: 978 90 6132 291 7. ISBN-10: 90 6132 291 X.

A. de Bruijne, J. van Buren, A. Kösters and H. van der Marel, Geodetic reference frames in the Netherlands; Definition and specification of ETRS89, RD and NAP, and their mutual relationships, Netherlands Geodetic Commission 43, Delft, 2005, 132 pages. ISBN 90 6132 291 X.

IERS Conventions (2010). Gérard Petit and Brian Luzum (eds.). (IERS Technical Note ; 36) Frankfurt am Main: Verlag des Bundesamts für Kartographie und Geodäsie, 2010. 179 pp., ISBN 3-89888-989-6. All software and material associated with IERS Conventions (2010) can be found at `http://tai.bipm.org/iers/conv2010/conv2010.html` and at `http://maia.usno.navy.mil/conv2010/`.

C.D. de Jong, G. Lachapelle, S. Skone, I.A. Elema, Hydrography, Series on Mathematical Geodesy and Positioning, 2nd edition, VSSD, Delft, 2010, 351 pp. ISBN 978-90-407-2359-9. (Chapter 4: Coordinate systems).

OGP, EPSG Guidance Note 7 Part 1; Using the EPSG Dataset, OGP Publication 373-7-1, August 2012, `http://www.epsg.org`.

OGP, EPSG Guidance Note 7 Part 2; Coordinate Conversions and Transformations including Formulas, , OGP Publication 373-7-2, June 2013, `http://www.epsg.org`.

A

Quantity, dimension, unit

Christian Tiberius

The measurement of any quantity involves *comparison* with some (precisely) defined unit value of the quantity. The statement that a certain distance is 25 metres, means that it is 25 times the length of the unit metre. A quantity shall be used (defined), with appropriate unit, that provides a reproducible standard.

quantity	dimension	unit
distance	length	metre [m]
time duration	time	second [s]
mass	mass	kilogram [kg]

Table A.1: Three fundamental quantities with the symbol for the unit indicated between square brackets.

As listed in Table A.1, the three fundamental quantities are distance, time duration and mass. The units are given in the Système Internationale (SI), the International System of Units. This system was first established in 1889 by the Bureau International des Poids et Mesures (BIPM). The official *definitions* read:

- *the metre* is the length of the path travelled by light in vacuum during a time interval of $1/299792458$ of a second
- *the second* is the duration of 9192631770 periods of radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom
- *the kilogram* is the unit of mass; it is equal to the mass of the international prototype of the kilogram

The metre, in its above definition, is linked to the second, via the speed of light in vacuum c, which equals 299792458 m/s. The speed of light in vacuum is a physical constant, and its determination remains a permanent challenge to physicists; the uncertainty at present is at the 1 m/s level. Originally the meter was established by the end of the 18th century in France; it was thought to be one ten-millionth part of the meridional quadrant of the Earth (a meridian passing through both poles).

The angle does not appear as a quantity in Table A.1. A supplementary unit (actually the quantity is dimension- and unitless) is the radian [rad] for plane angles. A full circle

designation	symbol	power
Tera	T	10^{12}
Giga	G	10^{9}
Mega	M	10^{6}
kilo	k	10^{3}
milli	m	10^{-3}
micro	μ	10^{-6}
nano	n	10^{-9}
pico	p	10^{-12}

Table A.2: Most common powers of ten and their designations.

corresponds to 2π rad. Other units for angles are grades (gon), a circle is 400 grad (the centesimal system), and degrees, a circle is 360 deg, or 360°.

To handle a wide range of magnitudes, standard prefixes are available for the above units according to the decimal system (as for instance 'kilo' and 'milli' for meter), see Table A.2.

Mass is an intrinsic property of an object that measures its resistance to acceleration, i.e. it is a measure of the object's inertia. Force is a derived quantity. It has dimension 'mass multiplied by length divided by time-squared'; the unit is Newton [N] which equals [kg m/s^2].

B

Gravity and gravity potential

Christian Tiberius, Hans van der Marel

Gravity, the main force experienced on Earth, causes (free) objects to change their positions, as to decrease their potential. According to the first two laws by Newton, the force prescribes the acceleration of the object, and this acceleration is the second time derivative of the position. And when there is no resulting force acting on an object, it is not subject to any acceleration and the object will either remain in rest, or be in uniform motion (constant velocity) along a straight line.

The two basic elements of Newtonian mechanics are mass and force. As introduced with Table A.1, mass is an intrinsic property of an object that measures its resistance to acceleration.

In an inertial coordinate system, Newton's second law states that the (net external) force (vector) $\mathbf{F}$ equals mass m times acceleration (vector) $\ddot{\mathbf{x}}$

$$\mathbf{F} = m\ddot{\mathbf{x}}$$

The acceleration vector $\ddot{\mathbf{x}}$ is the second derivative with respect to time of the position coordinates vector $\mathbf{x}$. The unit of the derived quantity force is is Newton [N] and equals [kg m/s^2].

Earth's gravity field

The Earth's gravitational field, represented by acceleration vector $\mathbf{g}$ (with direction and magnitude), varies with location on Earth, as well as above the Earth's surface. This acceleration vector has special symbol $\mathbf{g}$, instead of the general $\mathbf{a}$ or $\ddot{\mathbf{x}}$. Acceleration has unit [m/s^2].

Weight is defined as the force of gravity on an object. Its magnitude is (also) not a constant value over the Earth or above it. If the force of gravity is the only one force acting on an object, the object is said to be in free fall with acceleration $g = ||\mathbf{g}||$ and its apparent weight is zero. A satellite orbiting the Earth is (in the ideal situation) in free fall; the acceleration $\mathbf{g}$, directed towards the center of Earth, makes the satellite maintaining the circular orbit. An object located extremely far from the Earth (and any other body) would be truely weightless (but still have the same mass).

The magnitude of weight is given by a spring scale. The spring is designed to balance the force of gravity. The spring scale converts force (in [N]) into mass (in [kg]) on the display by assuming a magnitude of g equal to 9.81 m/s^2, (approximately everywhere) on and near the Earth's surface.

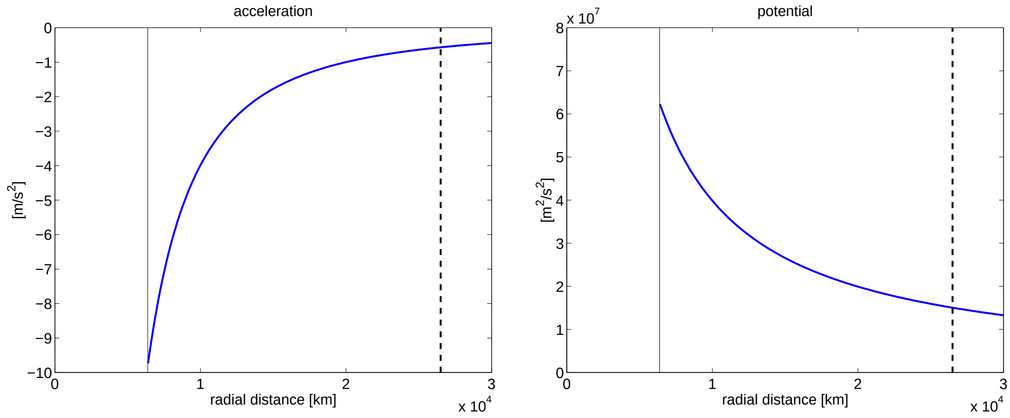


Figure B.1: The (size of) acceleration g (left) and the potential W (right) both approach zero as radial distance r goes to infinity. The curves start at the Earth's surface (the thin vertical line at $r = 6378$ km. The dashed line indicates the radial distance of the orbit of a GPS satellite; the acceleration (across track) is less than 1 m/s², at a speed of almost 4 km/s.

For an ideal spherical(ly layered) Earth (or when all if its mass were concentrated at its center), the gravitational force exerted on an object with mass m a distance r away, is given by

$$\mathbf{F} = -\frac{GMm}{r^2}\frac{\mathbf{r}}{r} \tag{B.1}$$

where G is the universal gravitational constant ($G = 6.6726 \cdot 10^{-11}$ Nm^2/kg^2), M the mass of the Earth ($M \approx 5.98 \cdot 10^{24}$ kg), $\mathbf{r}/r$ the unit direction vector from the Earth's center toward the object, and r the (radial) distance from the Earth's center to the object outside the Earth. The force (vector) $\mathbf{F}$, exerted by the Earth on the object with mass m is directed from the object toward the Earth's center.

The gravitational attraction consequently reads

$$\mathbf{g} = -\frac{GM}{r^2}\frac{\mathbf{r}}{r} \tag{B.2}$$

The force (B.1) has been divided by mass m to obtain the acceleration (B.2), which consequently could be interpreted as the force per unit mass. The magnitude of the gravitational acceleration $g = ||\mathbf{g}||$ (B.2) decreases with increasing height above the Earth's surface, and reduces to zero at infinite distance, see Figure B.1 (at left). The acceleration (vector) $\mathbf{g}$ also points toward the Earth's center.

Gravity Potential

The work done by a force equals the in- or dot-product of the force vector $\mathbf{F}$ and the displacement vector $d\mathbf{s}$, according to

$$W = \int_A^B \mathbf{F} \cdot d\mathbf{s} \tag{B.3}$$

The (tangential component of the) force is integrated along the path travelled by the object from A to B. Work, a scalar, is expressed in joule [J], the unit of energy, and equals J = Nm.

Taking the force per unit mass in (B.3), which is actually the acceleration (B.2), and interpreting the work (B.3) as a difference in energy $W_B - W_A$ after and before the force carrying the object from A to B, causes W in (B.3) to be the *potential energy* per unit mass of the gravitational force, or the potential of gravitation for short. The word 'potential' expresses that energy *can* be, but not necessarily *is*, delivered by the force.

With $m = 1$ kg, (B.1) and (B.3), the potential of gravitation becomes

$$W = \frac{GM}{r} \tag{B.4}$$

for a pointmass or spherical Earth. In this (geodetic) sense, the potential is expressed in [Nm/kg], which equals [$\mathrm{m}^2/\mathrm{s}^2$]. The potential is a function of the radial coordinate r, and the integration constant has been chosen such that the potential is zero at infinite distance, see also Figure B.1 at right.

From a physics perspective W in (B.4) presents the work done per unit mass. In physics it is common practice to define the potential energy function (with symbol U), such that the work done by a conservative force equals the decrease in the potential energy function (that is, use an opposite sign, for instance in (B.4)).

According to (B.4), the potential W is constant $W = W_o$, when radial distance r is constant, that is, on spherical surfaces around the Earth, all centered at the middle of the Earth. Surfaces where the gravity potential W is constant are equipotential surfaces, and the gravity vector g is everywhere orthogonal to them (dictating the local level, according to which geodetic instruments are set up). The surface of *reference* in a vertical sense for physical phenomena on Earth like water flow is the *geoid*, the equipotential surface at mean sea level.

As an introduction the (shape of the) Earth and its gravity field have been treated here as being (perfectly) spherical. Reality (and an adequate model thereof) is much more complicated. As a second approximation the Earth is taken to be a rotational ellipsoid (oblateness of the Earth) and subsequently the inhomogeneous distribution of mass within (and on) the Earth and the presence of heavenly bodies are considered.

The acceleration experienced on Earth in practice (and hence observable) consists of, first, gravitational acceleration due to the mass of Earth, as discussed above, but also of Sun and Moon (tidal acceleration) and secondly, centrifugal acceleration due to the Earth's rotation. Two additional contributions are the inertial acceleration of rotation and the Coriolis acceleration, which is absent if the object (or measurement equipment) is in rest, or in free fall. Gravity is then commonly defined as the sum of gravitational acceleration, but discounting the part due to the attraction of Sun and Moon, and centrifugal acceleration.

Gravimetry

With levelling, increments (distances) are measured along the (local) direction of gravity; the (vertical) center line of the instrument is a tangent line to the geoid. Gravity determines the direction of the height system; the plane perpendicular to the vector of gravity (locally) represents points at equal height.

The purpose of gravimetry is to eventually describe the geoid with respect to a chosen (geometric) reference body, for instance a rotating equipotential ellipsoid. It comes to determination of the geoid height.

The Earth gravity potential W itself (in an absolute sense) can not be observed. Inferences about the potential have to be made through measurements mainly of the first order (positional) derivative, that is through the gravity vector $\mathbf{g}$ of which direction and magnitude can be observed. The direction of gravity can be observed by astronomical measurements (latitude and longitude). The magnitude of gravity can be observed by absolute measurements (a pendulum or a free falling object, or by relative measurements (with a spring gravimeter).

A satellite falling around the Earth can also be looked upon as an accelerometer, as its orbit is primarily governed by the Earth's gravity field.

Gradiometers measure second order (positional) derivatives of the gravity potential, for instance in a satellite by two (or more) accelerometers at short distance. They sense the difference in acceleration (differential accelerometry). A satellite tandem mission, where two satellites closely go together, has a similar purpose.

Finally it should be noted that strictly speaking the separation made between geometric and (physical) gravimetric observables is not a distinct one. They are unified in the theory of general relativity: the path of a light ray for instance (as used for electronic measurements of distance) will bent as it travels through a (strong) gravity field.

Colophon

This reader has been produced for the 3rd year BSc course CTB3310 *Surveying and Mapping* in the Civil Engineering curiculum at the Delft University of Technology.

Title: Reference Systems for Surveying and Mapping

Author: Hans van der Marel

Contributors: Hans van der Marel, Christian Tiberius

Layout: Hans van der Marel, Anneleen Oyen

E-mail: `h.vandermareltudelft.nl`

Published by:

Delft University of Technology
Dept. Geoscience and Remote Sensing, Faculty of Civil Engineering and Geosciences, Stevinweg 1, 2628 CN Delft, The Netherlands.
`http://www.grs.citg.tudelft.nl`

Release history:

1.0	8 Feb 2014	Initial edition
1.1	21 Feb 2014	First release for 2014 CTB3310 course - minor corrections, several new figures, TU style
1.2a	28 Aug 2014	Interim release - minor corrections - new figure 3.1, figure 7.1 redrawn
1.3	[undefined]	Next major release

File name: `TRS_1-2a.tex`