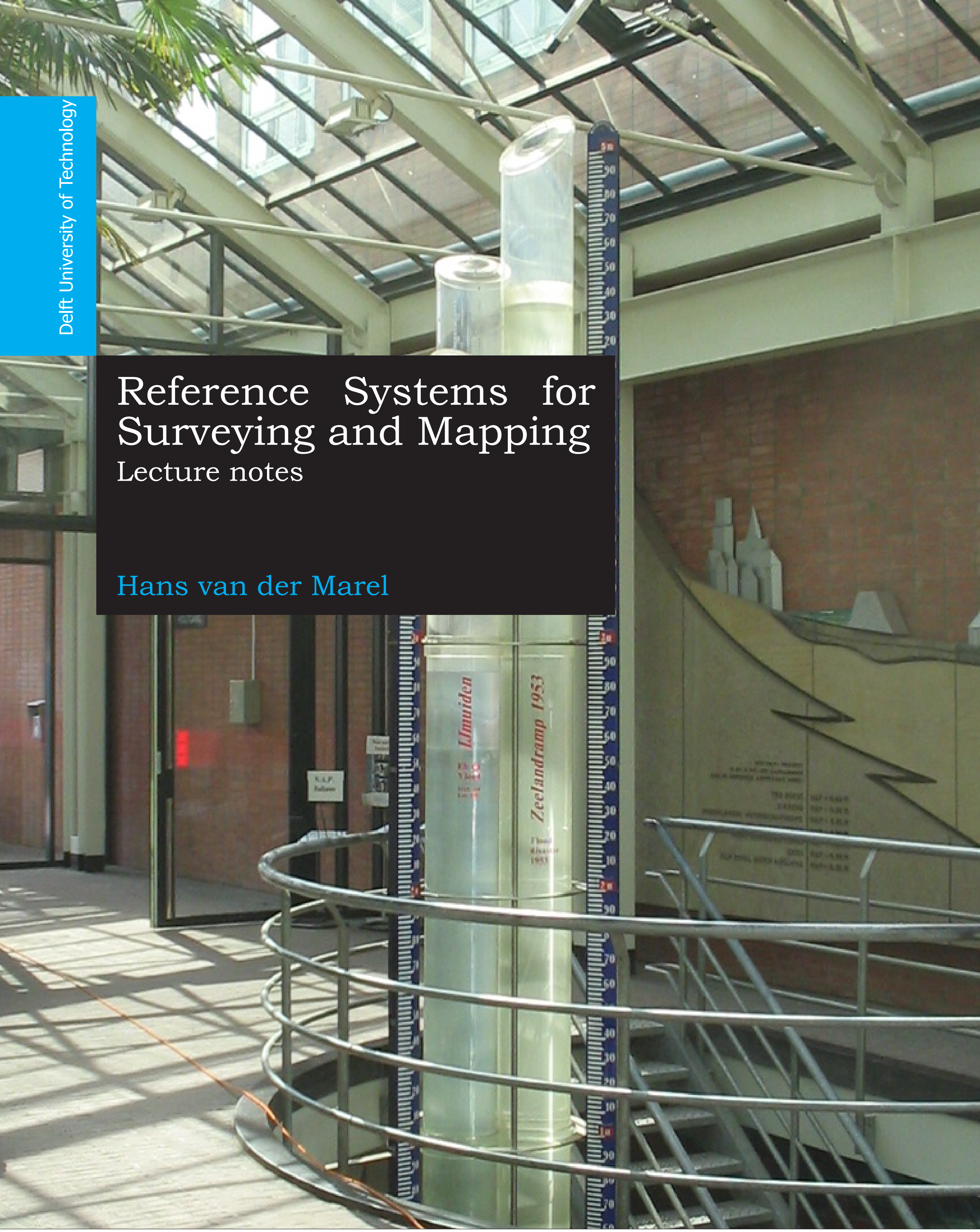
Delft University of Technology

Reference Systems for Surveying and Mapping

Lecture notes

Hans van der Marel



TU Delft 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

Lecture notes on Reference Systems for Surveying and Mapping:
CTB3310 Surveying and Mapping
CTB3425 Monitoring and Stability of Dikes and Embankments
CIE4606 Geodesy and Remote Sensing
CIE4614 Land Surveying and Civil Infrastructure

February 2020

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.

Preface

This reader on reference systems for surveying and mapping has been initially compiled for the course *Surveying and Mapping* (CTB3310) in the 3rd year of the BSc-program for Civil Engineering. The reader is aimed at students at the end of their BSc program or at the start of their MSc program, and is used in several courses at Delft University of Technology.

With the advent of the Global Positioning System (GPS) technology in mobile (smart) phones and other navigational devices almost anyone, anywhere on Earth, and at any time, can determine a three–dimensional position accurate to a few meters. With some modest investments, basically using the same GPS equipment, the Internet, and correction signals from a network of reference GPS receivers, determined individuals and professional users alike can achieve with relative ease three–dimensional positions with centimeter accuracy. A feat that until recently was achievable only to a small community of land surveyors and geodesists.

Our increased ability to collect accurate positional data, but also advances in Geographic Information Systems (GIS) and adoption of open–data policies for sharing many geographic datasets, has resulted in huge amounts of georeferenced data available to users.

However, sharing positional information is not always easy: "*How come my position measurement does not match yours?*", "*You say you have centimeter accuracy, I know I have, and yet we have a hunderd meter difference (you fool)?*". These are just a few frustated outcries you can hear from users (including Civil Engineering students). The reason is simple: users may have opted for different coordinate reference systems (CRS). Positions are relative, given with respect to a specific reference system. There are significant differences between various reference systems that are used, sometimes for historical reasons, sometimes because users selected different options (e.g. map projection) for good reasons. The solution is straightforward, but not simple: knowing the name and identifier for the reference system is key. If you have positional data in the same reference system you are lucky, if not, you have to use coordinate transformations to convert them into the same reference system.

This reader will provide you with the background information and the terminology that is commonly used. For the actual transformations you can use (freely) available software.

This reader keeps a quite informal and introductive style. The subjects are not treated with full mathematical rigor. This reader has been written with an *educational* goal in mind. The idea is to give students in Civil Engineering and Geosciences some *insight* in the reference systems used for surveying and mapping in and outside the Netherlands. Nevertheless, we believe this reader will also be usefull for other students and professionals in any of the geosciences. Sections marked by a [*] contain advanced material that students Civil Engineering may skip for the exam.

Colleague Christian Tiberius is acknowledged for his contributions to Chapter 7 and Appendix A, and for providing many of the exercises. Also I would like to thank Christian Tiberius and Ramon Hanssen for the initial proofreading. Jochem Lesparre and Christian Tiberius are credited for many of the improvements to the current release, and Cornelis Slobbe for the material on the new Dutch quasi–geoid and LAT chart datum. The author does welcome notifications of corrections and suggestions for improvement, as well as feedback in general.

Hans van der Marel
Delft, February 2020

Contents

1

Introduction

Surveying and mapping deal with the description of the shape of the Earth, spatial relationships between objects near the Earth's surface, and data associated to 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* (CRS). 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 2 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 geodetic 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 geodetic datums in use, but fortunately any transformation can be decomposed into a few elementary coordinate conversions and a geodetic 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 8, while in Chapter 7 basic background information on the Earth's gravity field is given.

Finally in Chapter 9 and 10 several important and commonly used reference systems are described. They include the well know World Geodetic System (WGS84) used by GPS, the International Terrestrial Reference System (ITRS), and the European Terrestrial Reference System (ETRS89) in Chapter 9, and the Dutch triangulation system ("Rijksdriehoeksstelsel" RD) and the Dutch height system, the Amsterdam Ordnance Datum ("Normaal Amsterdams Peil" NAP), and Lowest Astronomical Tide (LAT) chart datum, in Chapter 10.

This reader was written close, very close, to $52.0000°$ North latitude. The 52-degrees North latitude happens to run across the campus of Delft University of Technology, just a few meters North of the Civil Engineering and Geosciences faculty building. In 2018, the 52-degrees North line was visualized at the campus with a blue–white line, see Figure 1.1. Check out https://www.delta.tudelft.nl/article/why-blue-line-running-across-campus for a full story how the line was realized and how it has moved over the campus.

Figure 1.1: The blue–white line on the campus of Delft University of Technology visualizes latitude $\varphi = 52.0000000°$ North. The 16-millimetre-wide black line, in the middle of the white band, indicates the 'exact' position of the $52°$ latitude, in the International Terrestrial Reference System (ITRS), on 1 January 2018. Time matters because the $52°$ parallel shifts due to plate tectonics. The width of the black line represents the shift per year. To illustrate the time effect further, six grey lines (not visible in picture) have been painted parallel to the blue line. These grey lines indicate significant events related to Delft events. The first line is 2.79 metres to the North, and marks the foundation of TU Delft in 1842. (Picture courtesy of Conny van Uffelen)

2

2D Cartesian coordinate systems

To describe the position of points on a plane surface, be it a plot of land, a piece of paper, or a computer screen, a two–dimensional (2D) coordinate system need to be defined. One of the best known 2D coordinate reference systems is the 2D Cartesian coordinate system which uses rectangular coordinates. 2D Cartesian coordinates can also be the result of a map projection. Map projections are discussed in Chapter 5.

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 ($\mathbf{e}_x \perp \mathbf{e}_y$).

For surveying and mapping the distance d_{12} and azimuth α_{12} between two points P_1 and P_2 are defined as,

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

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. For practical computations the $\arctan$ in Eq. (2.2) should be replaced with the $\mathrm{atan2}\,(x_2-x_1, y_2-y_1)$ function in order to obtain the right quadrant for the azimuth α_{12} [2]. The angle $\angle P_2P_1P_3$ between points P_2, P_1 and P_3 is,

$$\varphi_{213} = \angle P_2P_1P_3 = \alpha_{13} - \alpha_{12} = \arccos\frac{<\mathbf{r}_2-\mathbf{r}_1, \mathbf{r}_3-\mathbf{r}_1>}{\|\mathbf{r}_2-\mathbf{r}_1\|\|\mathbf{r}_3-\mathbf{r}_1\|} \tag{2.3}$$

[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.

[2]The atan2 function is the four quadrant version of the arctangent function with 2 input values, with $-\pi \leq \mathrm{atan2}\,(dx, dy) \leq \pi$, compared to $-\pi/2 \leq \arctan\frac{dx}{dy} \leq \pi/2$

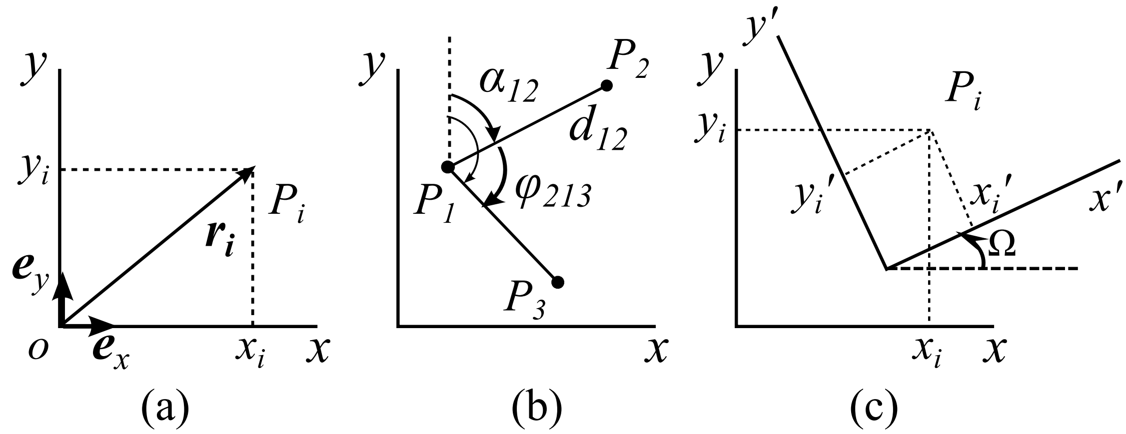


Figure 2.1: 2D Cartesian coordinate system (a), definition of azimuth, angle and distance (b) and a 2D coordinate transformation (c).

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 $\alpha_{12} = \pi/2 - \theta_{12}$ in radians, or $\alpha_{12} = 90^\circ - \theta_{12}$ when expressed in degrees.

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 in the horizontal plane are polar coordinates[3] 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 reading of the instrument).

2.2. 2D coordinate transformations

2.2.1. Shape preserving 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

[3]In 3D, these become spherical coordinates, which is what a tachymeter measures.

(t_x, t_y) via a so–called (2D) similarity transformation,

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

The transformation is a so-called *conformal* transformation whereby angles $\angle AOB$ between points A, O and B are preserved, i.e. angles are not changed by the transformation. This also means that shapes are preserved. A conformal transformation is therefore also known as a *similarity transformation* or 2D Helmert 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.

In the special case that the scale factor s is unity ($s = 1$) both angles and distances are preserved in the transformation. This is called a *congruence* transformation. The transformation involves 3 instead of 4 parameters: two translations t_x and t_y, and a rotation Ω. In this case, a 2D Cartesian coordinate system is defined either by (i) the three transformation parameters with respect to another 2D coordinates system, or (ii) by assigning three coordinates for (at least) two points.

2.2.2. Affine and polynomial transformations

Two other types of transformations, that do not preserve shape, are affine, and the more general polynomial transformations.

An *affine* transformation involves a rotation, scale change separately in both x- and y-direction, and a translation. It can be written as,

$$\begin{pmatrix} x_i' \\ y_i' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix} \quad . \tag{2.5}$$

with the 2-by-2 transformation matrix containing 4 different elements. 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 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. Polynomial transformations are also given by Eq. 2.5, but the transformation parameters are actually functions of the coordinates themselves. This means that straight lines, and shapes, are not necessarily preserved. Polynomial transformations are sometimes used as approximate transformations to match satelllite and aerial imagery onto a 2D Cartesian coordinate system, or as approximate transformation for grid coordinates between two map projections, or to handle non-linear distortions in scanned historical maps.

2.3. Realization of 2D coordinate systems

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

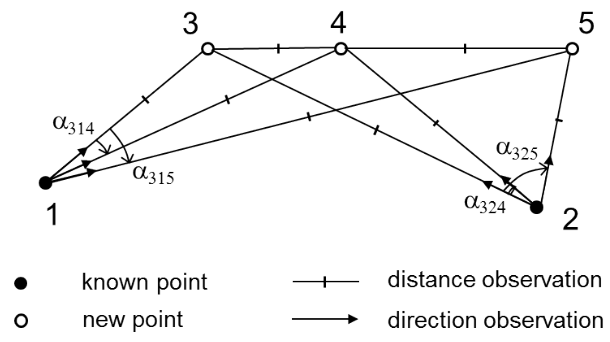


Figure 2.2: Two dimensional survey network with 4 angle measurements and 8 distance measurements.

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 system, when no pre-surveyed points are available, can be established and 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 coordinate system is determined by the scale of the measuring device (e.g. a tape measure) and also any measurement errors that were made in these 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.

2.4. Worked out examples

By means of two worked out examples we will show how a 2D coordinate system is realized in a practical way and how this is related to linear algebra.

2.4.1. 2D coordinate system definition

In this simple example, we show how to assign coordinates to points in the terrain, in order to establish a coordinate system. We will — in a practical way — define the position and orientation of a local 2D survey network. The scale is already implied by the distance measurements (in this example).

Figure 2.2 shows a simple survey network, with 5 points. Between these points, angle and distance measurements have been taken. When the coordinates of points 1 and 2 are *given* (e.g. as a result of an earlier survey), then these angle and distance measurements can be used (and are sufficient) to determine the coordinates of points 3, 4 and 5, for instance through least-squares parameter estimation.

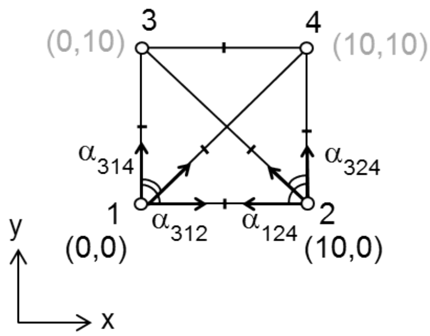


Figure 2.3: Two dimensional simple survey network with 4 angle measurements and 6 distance measurements.

What now, if no coordinates are available a-priori? Then you have to choose some coordinates yourself, in order to establish a so-called *local* network. But, you have to make a considerate choice - you cannot just assign coordinate values to some random points. For instance, if we would assign coordinates to points 1, 2 and 3 (in an arbitrary way), we may cause deformations and distortions of the network — i.e. the coordinates of those points may then not match (at all) the actually observed angles and distances!

The considerate choice requires you to analyse the geometry of the network. As stated in section 2.3, a 2D Cartesian coordinate system is uniquely defined by four parameters: the scale s, orientation Ω, and translations t_x and t_y. The scale is set, in this case, by the distance measurements. What remains to be fixed are the origin and the orientation.

A geometric network, or construction, with angles and distances, like the one in Figure 2.2 provides shape and scale, but not (absolute) position, nor orientation. You can *shift* the network (in two directions), while the angles and distances between the points stay exactly the same, and also, you can *rotate* the network, without altering angles and distances. There are still three degrees of freedom. Distances and angles are *invariant* against translation and rotation. Or, to turn this around, angle and distance measurements lack information about translation and rotation! Hence, you have to supply this!

In this example one could fix the coordinates of point 2, for instance, simply setting it to be the origin $(x_2, y_2) = (0, 0)$ (this fixes two degrees of freedom). And, one could set point 5 exactly along the positive x-axis, hence setting its y-coordinate to zero (this fixes the last degree of freedom). The coordinates of point 5 then become $(x_5, y_5) = (l_{25}, 0)$, where the measured distance l_{25} is used for the x-coordinate. The distance and angle measurements pose a geometric defect with three degrees of freedom in a 2D-network. Hence, three coordinates shall be fixed (no more, no less).

2.4.2. Algebraic analysis

The previous example provides a practical 'recipe' how to establish a 2D coordinate system. In the following we present, by means of a similar simple example, an algebraic analysis of this geometric defect.

The network is shown in Figure 2.3. This artificial network, being just a square, allows for a convenient and simple algebraic analysis. There are four angle measurements α_{314}, α_{312}, α_{123}, and α_{124}. There are six distance measurements l_{12}, l_{13}, l_{14},l_{23},l_{24}, and l_{34}. Forgetting about measurement errors, one can link these measurements to the (unknown) coordinates, by the following system of equations

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{2.6}$$

where the observations are in vector $\mathbf{y}$ on the left, the unknown parameters (the coordinates) in vector $\mathbf{x}$ on the right, and matrix $\mathbf{A}$ relating the two. Often the relation is non-linear, which is consequently approximated with a linearized relation, where we work with increments of observations and parameters (indicated by the Δ-symbol). Further details can be found in the Primer on Mathematical Geodesy (in particular chapters 4 and 5). For the artificial geometry in Figure 2.3, this system of equations becomes:

$$\begin{pmatrix} \Delta\alpha_{314} \\ \Delta\alpha_{312} \\ \Delta\alpha_{123} \\ \Delta\alpha_{124} \\ \Delta l_{12} \\ \Delta l_{13} \\ \Delta l_{14} \\ \Delta l_{23} \\ \Delta l_{24} \\ \Delta l_{34} \end{pmatrix} = \begin{pmatrix} \frac{10}{200} & \frac{10}{200} & 0 & 0 & -\frac{10}{100} & 0 & \frac{10}{200} & -\frac{10}{200} \\ \frac{10}{100} & \frac{10}{100} & 0 & -\frac{10}{100} & -\frac{10}{100} & 0 & 0 & 0 \\ 0 & -\frac{10}{100} & -\frac{10}{200} & \frac{10}{200} & \frac{10}{200} & \frac{10}{200} & 0 & 0 \\ 0 & -\frac{10}{100} & -\frac{10}{100} & \frac{10}{100} & 0 & 0 & \frac{10}{100} & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 & 0 & 0 & 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \Delta x_1 \\ \Delta y_1 \\ \Delta x_2 \\ \Delta y_2 \\ \Delta x_3 \\ \Delta y_3 \\ \Delta x_4 \\ \Delta y_4 \end{pmatrix} \tag{2.7}$$

There are $m = 4+6 = 10$ observations, hence $\mathbf{y}$ is a 10x1-vector, and there are $n = 8$ unknown parameters in vector $\mathbf{x}$. Consequently, matrix $\mathbf{A}$ has dimensions 10x8. The rank of matrix $\mathbf{A}$ is however only 5, not 8. This is the algebraic indication that the measurements leave three degrees of freedom. The *null-space* of matrix $\mathbf{A}$ is not empty. Instead, in this example, the null-space of matrix $\mathbf{A}$ can be spanned by the following three (linearly independent) vectors:

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \; ; \; \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \; ; \; \mathbf{v}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 1 \\ -1 \end{pmatrix} \tag{2.8}$$

These three vectors can be stored together, in 8x3 matrix $\mathbf{V}$, with $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$, for which holds $\mathbf{AV} = \mathbf{0}$. The columns of matrix $\mathbf{V}$ provide a basis for the null-space of matrix $\mathbf{A}$.

Now suppose that $\mathbf{x}$ is a solution to (2.6), then $\mathbf{x}' = \mathbf{x} + \mathbf{V}\boldsymbol{\beta}$, with 3x1-vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$, is also a solution, namely

$$\mathbf{y} = \mathbf{Ax}' = \mathbf{Ax} + \mathbf{AV}\boldsymbol{\beta} = \mathbf{Ax} \tag{2.9}$$

Or, changing the coordinates of the points, in some particular way as imposed by matrix $\mathbf{V}$, does *not* change the observations. Or, the other way around, based on a set of observations, you cannot tell the difference between $\mathbf{x}$ and $\mathbf{x}'$. The null-space of matrix $\mathbf{A}$ being not empty, causes that there is left a certain degree of freedom in the solution.

The vectors $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ can be easily interpreted in this example, see Figure 2.4. Vector $\mathbf{v}_1$ implies an offset to the x-coordinates of all points in the network, meaning that translation parameter t_x is undefined. Vector $\mathbf{v}_2$ implies an offset to the y-coordinates of all points, meaning that t_y is undefined. And eventually, vector $\mathbf{v}_3$ implies a rotation of the network about point 1.

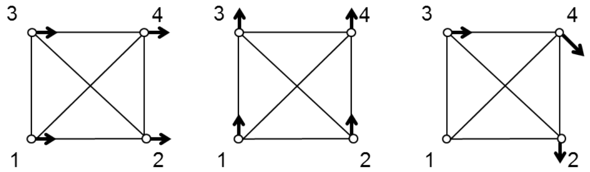


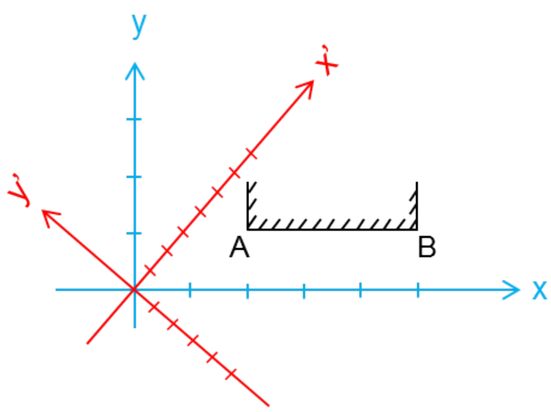
Figure 2.4: Interpretation of the null-space of matrix $\mathbf{A}$, vector $\mathbf{v}_1$ at left, vector $\mathbf{v}_2$ in the middle, and vector $\mathbf{v}_3$ at right.

Applying the earlier practical 'recipe' would cause us to fix the coordinates of point 1 to the origin $(x_1, y_1) = (0, 0)$, and to set the y-coordinate of point 2 to zero, i.e. $(x_2, y_2) = (l_{12}, 0)$. Three coordinates have been fixed, and consequently they can be removed from vector $\mathbf{x}$. Correspondingly, the first, second, and fourth column of matrix $\mathbf{A}$ has to be removed as well.

The interested reader is encouraged to verify that the resulting/reduced matrix $\mathbf{A}$, with dimensions 10x5, has full rank, equal to 5, and an empty null-space. This means that, based on the available measurements, the remaining 5 parameters (coordinates) can be determined, smoothly, for instance through least-squares estimation. The origin, the scale and the orientation of the 2D Cartesian coordinate system have been fixed.

2.5. Problems and exercises

Question 1 Two surveyors measure the facade of a building: points A and B. They both use Euclidean geometry in the local horizontal plane, but they adopt a different coordinate system, see the figure



The coordinates of point A and B in the blue system read $(x_A, y_A) = (2, 1)$, and $(x_B, y_B) = (5, 1)$, and in the red system $(x'_A, y'_A) = (4\sqrt{2}, -\sqrt{2})$, and $(x'_B, y'_B) = (7\sqrt{2}, -4\sqrt{2})$. The coordinates in the two systems are related through a 2–D similarity transformation. Determine, based on the coordinates given for the two points, the transformation parameters, i.e. scale factor s, rotation angle Ω, and translations t_x, t_y.

Answer 1 A clever approach to solving this problem is using Eq. (2.4) on coordinate *differences*

$$\begin{pmatrix} x'_B - x'_A \\ y'_B - y'_B \end{pmatrix} = s \begin{pmatrix} \cos\Omega & \sin\Omega \\ -\sin\Omega & \cos\Omega \end{pmatrix} \begin{pmatrix} x_B - x_A \\ y_B - y_A \end{pmatrix}$$

as the translation parameters cancel. Setting $s\cos\Omega = p$ and $s\sin\Omega = q$, we obtain

$$\begin{pmatrix} 3\sqrt{2} \\ -3\sqrt{2} \end{pmatrix} = \begin{pmatrix} p & q \\ -q & p \end{pmatrix} \begin{pmatrix} 3 \\ 0 \end{pmatrix}$$

from which we can easily solve p and q. Doing so we find $p = \sqrt{2}$ and $q = \sqrt{2}$. From this we can reconstruct that $s = \sqrt{p^2 + q^2}$ and $\Omega = \arctan \frac{q}{p}$, which gives $s = 2$ and $\Omega = \frac{\pi}{4}$. Then using again Eq. (2.4), but now for just one of the points, e.g. A, we have

$$\begin{pmatrix} 4\sqrt{2} \\ -\sqrt{2} \end{pmatrix} = 2 \begin{pmatrix} \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \\ -\frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{2} \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}$$

from which we can solve the translation parameters as $(t_x, t_y) = (\sqrt{2}, 0)$.

3

3D Cartesian coordinate systems

Three–dimensional (3D) coordinates systems are used to describe the position of objects in 3D–space. In this chapter we discuss 3D Cartesian coordinate systems, before discussing spherical and ellipsoidal coordinate systems in Chapter 4. 3D Cartesian coordinate systems can be considered a straightforward extension of 2D Cartesian coordinate systems by adding a third axis.

3.1. Introduction

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 Chapter 2, 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 Z-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 and covered in Section 4.4. 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 or total station, but the third 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.

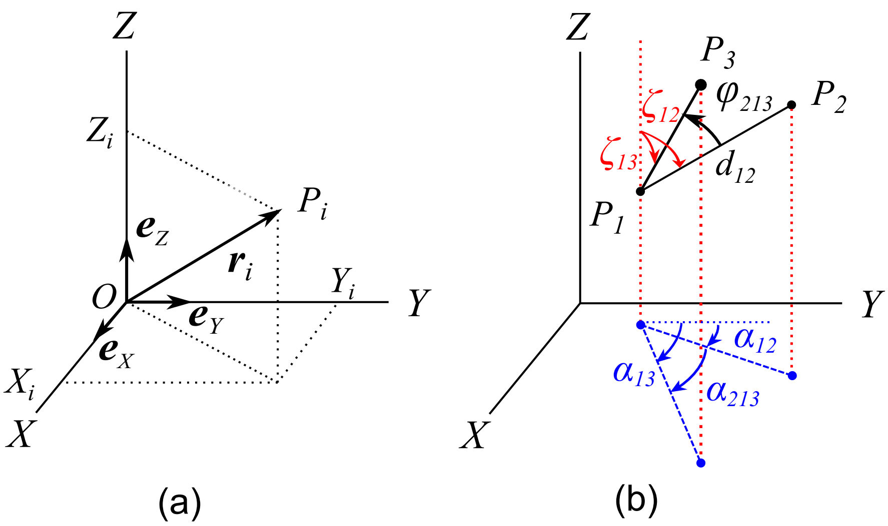


Figure 3.1: 3D Cartesian coordinate system (a) and definition of azimuth α_{12}, horizontal angle α_{213}, vertical angle ζ_{12}, angle φ_{213} and distance d_{12} (b).

3.2. 3D Cartesian coordinates

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

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

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

$$\varphi_{213} = \angle P_2 P_1 P_3 = \arccos \frac{< \mathbf{r}_2 - \mathbf{r}_1, \mathbf{r}_3 - \mathbf{r}_1 >}{\|\mathbf{r}_2 - \mathbf{r}_1\| \ \|\mathbf{r}_3 - \mathbf{r}_1\|} \tag{3.3}$$

with $< \mathbf{u}, \mathbf{v} >$ the dot (inner) product of two vectors. For a *topocentric* system, with the Z-axis in the up direction and Y-axis to the North, 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. For practical computations the $\arctan$ in Eq. (3.4) should be replaced with the $\operatorname{atan2}(X_2 - X_1, Y_2 - Y_1)$ function in order to obtain the right quadrant for the azimuth α_{12}. 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 $\alpha_{213} = \alpha_{13} - \alpha_{12}$.

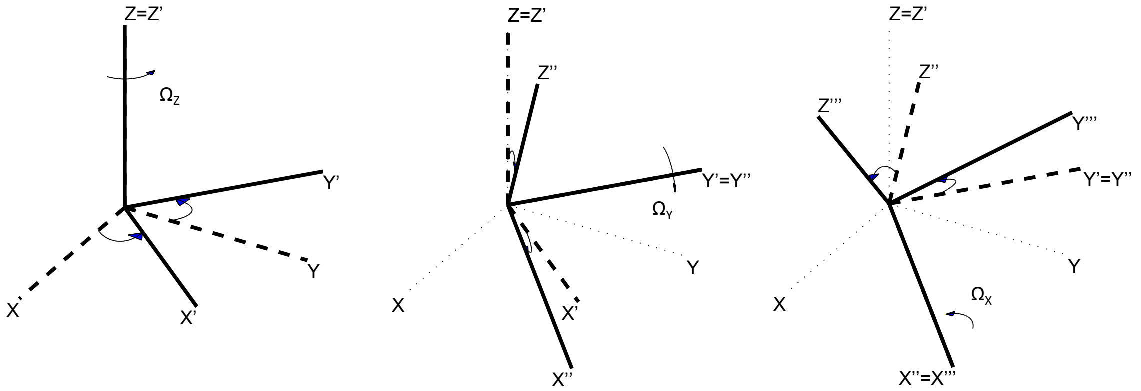


Figure 3.2: Definition of rotation angles for the 7-parameter similarity transformation. The figure on the left shows a rotation Ω_z about the Z-axis, the middle figure a rotation Ω_y about the newly obtained Y'-axis, and the figure on the right a rotation Ω_x about the final X''-axis. The angles are positive for a counter-clockwise rotation when viewed along the axis towards the origin (right-handed rotation is positive) and defined to turn the source coordinate system axes into the target system axes.

3.3. 3D similarity transformations

We start this section with a brief overview of 3D coordinate transformations. Of the 3D transformations, the 3D similarity transformation, that preserves shape, is by far the most often used coordinate transformation for 3D coordinates, and the remainder of this section is devoted to this important type of transformation.

3.3.1. Overview 3D coordinate transformations

The *affine transformation* is the most general transformation which can be represented in terms of linear algebra. The 3-by-3 matrix $\mathbf{R}$ has nine different elements, implying rotation, scaling and so-called shearing, the latter meaning that a square is turned into a parallelogram (or, actually a cube into a parallelepiped).

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \underbrace{\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}}_{\mathbf{R}} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} \tag{3.5}$$

The 3D affine transformation is specified by a total of 12 parameters.

The *similarity transformation* preserves the shape of objects. The 3-by-3 matrix $\mathbf{R}$ now implies only a rotation (or actually series of rotations). Matrix $\mathbf{R}$ has 9 elements, but needs to satisfy 3 orthogonality conditions and 3 orthonormality conditions (the rows are orthogonal, and they are all of unit length), and thereby only 3 degrees of freedom remain (3 rotation angles).

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \lambda \mathbf{R}(\Omega_x, \Omega_y, \Omega_z) \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} \tag{3.6}$$

The 3D similarity transformation is specified by a total of 7 parameters. In the sequel the scale parameter λ, which often close to one, will be replaced by $1 + \mu$.

The *congruence transformation* preserves the shape and size of objects. It is the so-called 'rigid body' transformation. It is a special case of the similarity transformation, with the scale parameter fixed to one $\lambda = 1$ (or equivalently, $\mu = 0$).

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \mathbf{R}(\Omega_x, \Omega_y, \Omega_z) \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} \tag{3.7}$$

The 3D congruence transformation is specified by a total of 6 parameters (3 for rotation, and 3 for translation).

Of the three transformations, the 3D similarity transformation is by far the most often used 3D coordinate transformation. It will be covered in more detail in the next subsections.

3.3.2. 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,

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = (1 + \mu) \cdot \mathbf{R}(\Omega_x, \Omega_y, \Omega_z) \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} \tag{3.8}$$

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 a sequence of so–called Euler rotations. A 3-2-1 series of Euler rotations gives the rotation matrix

$$\mathbf{R}_{321}(\Omega_x, \Omega_y, \Omega_z) = \mathbf{R}_1(\Omega_x) \cdot \mathbf{R}_2(\Omega_y) \cdot \mathbf{R}_3(\Omega_z) =$$
$$\begin{pmatrix} \cos\Omega_z \cos\Omega_y & \sin\Omega_z \cos\Omega_y & -\sin\Omega_y \\ \cos\Omega_z \sin\Omega_y \sin\Omega_x - \sin\Omega_z \cos\Omega_x & \sin\Omega_z \sin\Omega_y \sin\Omega_x + \cos\Omega_z \cos\Omega_x & \cos\Omega_y \sin\Omega_x \\ \cos\Omega_z \sin\Omega_y \cos\Omega_x + \sin\Omega_z \sin\Omega_x & \sin\Omega_z \sin\Omega_y \cos\Omega_x - \cos\Omega_z \sin\Omega_x & \cos\Omega_y \cos\Omega_x \end{pmatrix} \tag{3.9}$$

with Ω_z, Ω_y and Ω_x the rotation angles around – and in that order – the z-, y- and x-axis respectively. The corresponding Euler rotation matrices, $\mathbf{R}_i(\Omega_i)$, describing rotations around the coordinate axis, are

$$\begin{aligned} \mathbf{R}_3(\Omega_z) &= \begin{pmatrix} \cos\Omega_z & \sin\Omega_z & 0 \\ -\sin\Omega_z & \cos\Omega_z & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{R}_2(\Omega_y) = \begin{pmatrix} \cos\Omega_y & 0 & -\sin\Omega_y \\ 0 & 1 & 0 \\ \sin\Omega_y & 0 & \cos\Omega_y \end{pmatrix}, \\ \mathbf{R}_1(\Omega_x) &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\Omega_x & \sin\Omega_x \\ 0 & -\sin\Omega_x & \cos\Omega_x \end{pmatrix} \end{aligned} \tag{3.10}$$

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 $\mathbf{R}_{321}(\Omega_x, \Omega_y, \Omega_z)$ 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.

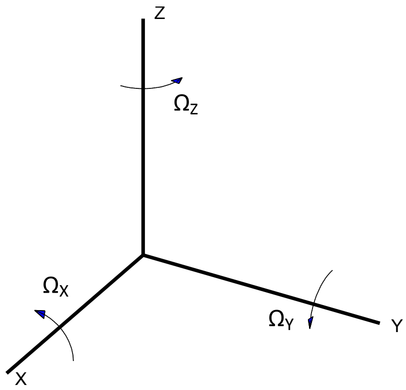


Figure 3.3: Definition of infinitesimal small rotation angles Ω_x, Ω_y and Ω_z for the *Helmert* transformation.

Changing the order of the Euler rotations in Eq. (3.9) will result in a different equation for the rotation matrix with different rotation angles. This is typical for Euler rotations. Reverting the order of rotations in Eq. 3.9, gives a 1-2-3 sequence of Euler rotations with rotation matrix,

$$\mathbf{R}_{123}(\Omega'_x, \Omega'_y, \Omega'_z) = \mathbf{R}_3(\Omega'_z) \cdot \mathbf{R}_2(\Omega'_y) \cdot \mathbf{R}_1(\Omega'_x) =$$
$$\begin{pmatrix} \cos\Omega'_z \cos\Omega'_y & \cos\Omega'_z \sin\Omega'_y \sin\Omega'_x + \sin\Omega'_z \cos\Omega'_x & -\cos\Omega'_z \sin\Omega'_y \cos\Omega'_x + \sin\Omega'_z \sin\Omega'_x \\ -\sin\Omega'_z \cos\Omega'_y & -\sin\Omega'_z \sin\Omega'_y \sin\Omega'_x + \cos\Omega'_z \cos\Omega'_x & \sin\Omega'_z \sin\Omega'_y \cos\Omega'_x + \cos\Omega'_z \sin\Omega'_x \\ \sin\Omega'_y & -\cos\Omega'_y \sin\Omega'_x & \cos\Omega'_y \cos\Omega'_x \end{pmatrix} \tag{3.11}$$

with Ω'_x, Ω'_y and Ω'_z the rotation angles around – and in that order – the x-, y- and x-axis respectively. The rotation matrix $\mathbf{R}_{123}(\Omega_x, \Omega_y, \Omega_z)$ is thus the product of first a rotation around the x-axis, followed by a rotation around the new y-axis, and finally a rotation around the then current z-axis. The rotation angles Ω'_x, Ω'_y and Ω'_z are different from the rotation angles Ω_x, Ω_y and Ω_z used in Eq. 3.9.

The rotation cannot be inverted by just changing the sign of the parameters (except for very small angles). The inverse of the rotation matrix is $\mathbf{R}_{321}(\Omega_x, \Omega_y, \Omega_z)^{-1} = (\mathbf{R}_1(\Omega_x) \cdot \mathbf{R}_2(\Omega_y) \cdot \mathbf{R}_3(\Omega_z))^{-1} = \mathbf{R}_3(-\Omega_z) \cdot \mathbf{R}_2(-\Omega_y) \cdot \mathbf{R}_1(-\Omega_x) = \mathbf{R}_{123}(-\Omega_x, -\Omega_y, -\Omega_z)$. This is *not* the same as changing the sign of the angles in Eq. 3.9, it also means changing the order of the rotations. As the rotation matrices are orthogonal matrices, the inverse of the rotation matrix is equal to the transpose of the matrix. Thus we can write $\mathbf{R}_{321}(\Omega_x, \Omega_y, \Omega_z)^T = (\mathbf{R}_1(\Omega_x) \cdot \mathbf{R}_2(\Omega_y) \cdot \mathbf{R}_3(\Omega_z))^T = \mathbf{R}_3(\Omega_z)^T \cdot \mathbf{R}_2(\Omega_y)^T \cdot \mathbf{R}_1(\Omega_x)^T = \mathbf{R}_3(-\Omega_z) \cdot \mathbf{R}_2(-\Omega_y) \cdot \mathbf{R}_1(-\Omega_x) = \mathbf{R}_{123}(-\Omega_x, -\Omega_y, -\Omega_z)$, or in short, $\mathbf{R}_{321}(\Omega_x, \Omega_y, \Omega_z)^T = \mathbf{R}_{123}(-\Omega_x, -\Omega_y, -\Omega_z)$. Compare the terms in Eqs. 3.9 and 3.11, and you will see it is true. It means that to do the reverse rotation, we not only have to change the sign of the rotation angles, but also need to revert the order of the rotations.

3.3.3. 7–parameter Helmert (small angle) transformation

In the case when the rotation angles are very 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

$$\mathbf{R}(\Omega_x, \Omega_y, \Omega_z) \simeq \begin{pmatrix} 1 & \Omega_z & -\Omega_y \\ -\Omega_z & 1 & \Omega_x \\ \Omega_y & -\Omega_x & 1 \end{pmatrix} \tag{3.12}$$

with the rotation angles as defined in Figure 3.3. The 7-parameter similarity transformation of Eq. (3.8) in it's simplified form is

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} + \begin{pmatrix} \mu & \Omega_z & -\Omega_y \\ -\Omega_z & \mu & \Omega_x \\ \Omega_y & -\Omega_x & \mu \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{3.13}$$

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.

3.3.4. 10–parameter Molodensky–Badekas transformation [*]

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 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 (e.g. in the middle of the area of interest on the Earth's surface). 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

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = (1+\mu) \cdot \mathbf{R}(\Omega'_x, \Omega'_y, \Omega'_z) \cdot \begin{pmatrix} X - X_0 \\ Y - Y_0 \\ Z - Z_0 \end{pmatrix} + \begin{pmatrix} t'_x \\ t'_y \\ t'_z \end{pmatrix} + \begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \end{pmatrix} \tag{3.14}$$

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. (3.14) uses the same sign convention as Eq. (3.8). Note that, whereas many publications use for Eq. (3.8) the same sign convention as this reader, most publications use for Eq. (3.14) the opposite convention. You are warned.

3.4. Realization of 3D coordinate systems

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 similarity transformation, consisting of three rotations, three translations and a scale factor, see Section 3.3. This transformation is again a conformal (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.

Using 7 coordinates for 3 points is just enough to define a 3D-coordinate reference system. When using more than 7 coordinates and 3 points to define the coordinates there is a serious risk of introducing distortions in the coordinates. For example, suppose we have 3 points, each with 3 coordinates. Suppose we use the coordinates of the first two points and the Z-coordinate of the third point to define the coordinate system, then in general the X and Y coordinates of the third point will not match the given coordinates, unless by coincidence. The same is true if four or more points are 'given'. The only proper way to handle such a situation is to set up a system of equations with a 3D-similarity transformation and then minimize in least–squares sense the differences between the given and computed coordinates. In more technical terms this is known as an **S-transformation**. In this way, using more than 7 coordinates for 3 points, has the important advantage of added redundancy in practical computations and becoming less sensitive to outliers, especially in combination with statistical testing, without introducing distortions in the coordinate network. This is called a **free–network**.

It is also possible that we want to do an *over-determined* connection to given coordinates. In this case the resulting coordinates for the connection points with be the same as the given values, but the original network of coordinates will be distorted. This can also be done in a weighted sense, whereby weights or a co–variance matrix is assigned to the given coordinates, and the network of coordinates is fitted in least–squares sense to the given coordinates. This results in an **over–determined network** of coordinates.

3.5. Problems and exercises

Question 1 Coordinate system I is related to coordinate system II through a rotation (counter-clockwise) about the Z-axis over 90 degrees. Both systems are three-dimensional Cartesian coordinate systems. Compute the rotation matrix for transforming coordinates given with

respect to system I, into coordinates with respect to system II.
Answer 1 The 3x3 rotation matrix is given by Eq. (3.10). The angle $\Omega_z = 90$ degrees. Hence, the matrix becomes

$$\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

So, a point on the Y-axis in source system I (e.g. with coordinates $(0, 1, 0)$), becomes a point on the X-axis in targete system II (e.g. with coordinates $(1, 0, 0)$).

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.

4.1. Spherical coordinates (geocentric 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).

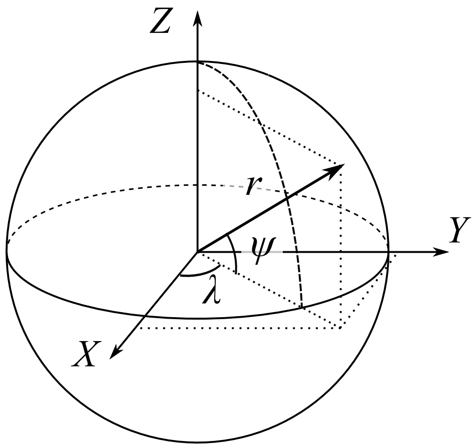


Figure 4.1: Spherical coordinates ψ, λ, r and Cartesian coordinates X, Y, Z.

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,

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

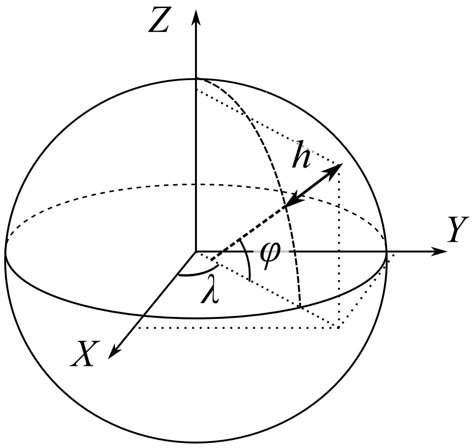


Figure 4.2: Ellipsoidal and Cartesian coordinates. The ellipsoidal latitude φ is also known as *geodetic* or *geographic* latitude. The ellipsoidal coordinates φ and λ are also called geographic coordinates.

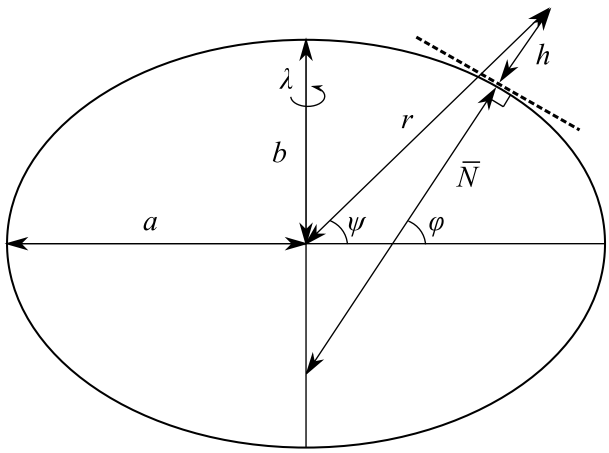


Figure 4.3: Ellipsoidal, geodetic or geographic latitude φ, geocentric (or spherical) latitude ψ, radius of curvature $\bar{N} = \bar{N}(\varphi)$, radius r, ellipsoidal height h, semi-major axis a and semi-minor axis b of the ellipsoid.

The inverse relationship is given by,

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

$\bar{N}$ 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 East-West and North-South direction.They are called respectively radius of curvature in the prime vertical, $\bar{N} = \bar{N}(\varphi)$, and the radius of curvature in the meridian, $\bar{M} = \bar{M}(\varphi)$, These two radii are not the same as the physical radius, the distance from the center of the Earth to the ellipsoid. This is different from a sphere, where all three radii are the same, and have a single value R. The radius of curvature in the prime vertical, $\bar{N} = \bar{N}(\varphi)$, and the radius of curvature in the meridian, $\bar{M} = \bar{M}(\varphi)$, for an ellipsoid 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 $\bar{N}$ normal to $\bar{M}$. On the equator the radius of curvature in East-West is equal to the semi-major axis a, with $\bar{N}(0°) = a$, while the radius of curvature in North-South is smaller than the semi-minor axis, with $\bar{M}(0°) = a(1-e^2) = b(1-f) = b^2/a$. On the poles the radius of curvature $\bar{N}(\pm 90°) = \bar{M}(\pm 90°) = a/\sqrt{(1 - e^2)} = a^2/b$ is larger than the semi-major axis a, see Figure 4.4.

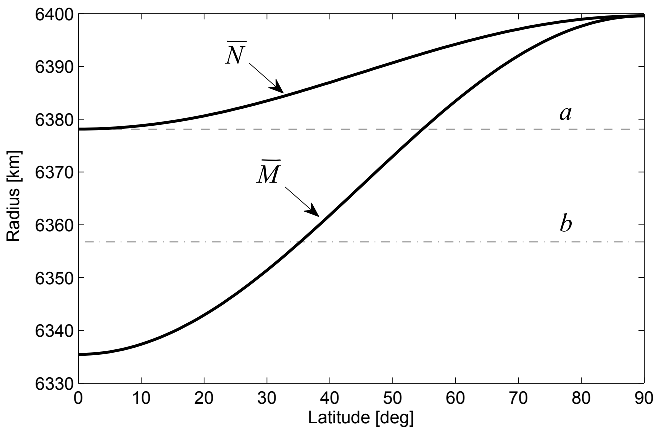


Figure 4.4: Radius of curvature $\bar{N}(\varphi)$ and $\bar{M}(\varphi)$ as function of latitude φ. The dashed lines represent the semi-major axis a and semi-minor axis b.

The radii of curvature play a role in the conversion of small differences in latitude and longitude into linear distances on the surface of the Earth. If $d\varphi$ is the differential latitude in radians, and $d\lambda$ the differential longitude in radians, then

$$\begin{aligned} dN &= (\bar{M}(\varphi) + h)\ d\varphi \\ dE &= (\bar{N}(\varphi) + h) \cos\varphi\ d\lambda \end{aligned} \tag{4.7}$$

with dN the differential distance in North-South (latitude) direction, with positive direction to the North, and dE the differential distance in East-West (longitude) direction, with $\bar{M}(\varphi)$ and $\bar{N}(\varphi)$ the meridian radius of curvature and radius of curvature in the prime vertical as given by Eq. (4.6) and Figure 4.4. Both dN and dE are in units of meters and are often referred to as *Northing* and *Easting*. The relations in Eq. (4.7) come in very handy if you wish to express small differences in latitude and longitude in units of meters. This happens for instance when you have latitude and longitude for two nearby points, but instead of a latitude and longitude differences in angular units, you are more interested to have the difference in meters. It is also very useful to convert for instance standard deviations in angular units to standard deviations in meters. For a first approximation, e..g. when differences in latitude and longitude are small or when accuracy does not matter, $\bar{M}(\varphi) + h$ and $\bar{N}(\varphi) + h$ in Eq. (4.7) can be replaced simply by the radius R of the spherical Earth.

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

$$N'_0 = Z$$

for $i = 1, 2, \ldots$

$$\begin{aligned} \varphi_i &= \arctan\left(\frac{Z + e^2 N'_{i-1}}{\sqrt{X^2+Y^2}}\right) \\ \bar{N}_i &= \frac{a}{\sqrt{1 - e^2 \sin^2 \varphi_i}} \\ N'_i &= \bar{N}_i \sin\varphi_i \end{aligned} \tag{4.8}$$

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),

$$\varphi = \arctan\left(\frac{Z + e'^2 b \sin^3\mu}{\sqrt{X^2+Y^2} - e^2 a \cos^3\mu}\right) \tag{4.9}$$

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

$$\begin{aligned} e'^2 &= \frac{e^2}{1-e^2} = \frac{a^2 - b^2}{b^2} \\ \mu &= \arctan\left(\frac{aZ}{b\sqrt{X^2+Y^2}}\right) \end{aligned} \tag{4.10}$$

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

$$\psi(\varphi) = \arctan\left(\frac{\bar{N}(1-e^2)+h}{\bar{N}+h}\tan\varphi\right) \simeq \arctan((1-e^2)\tan\varphi) \mid h \ll \bar{N} \tag{4.11}$$

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[1] at a geodetic latitude of $45°5'$. 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.

4.3. Astronomical latitude and longitude [*]

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 (φ, λ) is given by the unit direction vector $\bar{\mathbf{n}}$

$$\bar{\mathbf{n}} = \begin{pmatrix} \cos\varphi\cos\lambda \\ \cos\varphi\sin\lambda \\ \sin\varphi \end{pmatrix} \tag{4.12}$$

[1] 1 minute of arc is $1/60$ of a degree; so 11.5 minutes of arc is equal to $11.5/60 \simeq 0.19°$.

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* (in Dutch: *schietlood*) given by the direction of the local gravity field, $\mathbf{g}$, at that point. Gravity is the resultant of the gravitational acceleration and the centrifugal acceleration at that point, see also Chapter 7. The direction of the true vertical $\mathbf{n}$ is given by the *astronomical* latitude ϕ and longitude Λ,

$$\mathbf{n} = -\frac{\mathbf{g}}{g} = \begin{pmatrix} \cos\phi\cos\Lambda \\ \cos\phi\sin\Lambda \\ \sin\phi \end{pmatrix} \tag{4.13}$$

with $g = ||\mathbf{g}||$. The astronomical latitude and longitude can be determined through (zenith) measurements to the stars. 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 longitude Λ and ellipsoidal longitude λ. The ellipsoid is a purely geometric shape, but astronomical latitude and longitude are driven by physics, namely the direction of gravity.

The angle between the directions of the ellipsoidal normal and true vertical at a point is called the *deflection of the vertical*. The deflection of the vertical is divided in two components, defined as,

$$\begin{aligned} \xi &= \phi - \varphi \\ \eta &= (\Lambda - \lambda)\cos\varphi \end{aligned} \tag{4.14}$$

Astronomical latitude ϕ and longitude Λ are obtained from astronomical observations to stars whose positions (declination δ and right ascension α) in a Celestial reference system are accurately known, or from gravity observations using gravimeters[2]. 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.

4.4. Topocentric coordinates, azimuth and zenith angle [*]

It is not always convenient to use Cartesian coordinates in a global reference system with the origin in the center of mass of the Earth. Sometimes it is more convenient to choose the origin in a point on or near the surface of the Earth, and define the coordinate axis with respect to the local vertical and geographic North. This type of 3D Cartesian coordinate system is called a local *topocentric* coordinate system and it's coordinates are called topocentric coordinates. In Figure 4.5 the origin of the local topocentric system is the (observation) point A with geographic coordinates (φ, λ, h). The vectors $\bar{\mathbf{e}}_E$, $\bar{\mathbf{e}}_N$ and $\bar{\mathbf{n}}$ form the three axis of a right-handed local topocentric system centered at A, with z-axis along the normal of the ellipsoid $\bar{\mathbf{n}}$, x-axis $\bar{\mathbf{e}}_E$ orthogonal to the plane of the meridian and positive to the East, and y-axis $\bar{\mathbf{e}}_N = \bar{\mathbf{n}} \times \bar{\mathbf{e}}_E$ in the plane of the meridian completing the topocentric system. The coordinates in the local topocentric system are denoted by E (East), N (North) and U (Up). The system itself is sometimes also called a East-North-Up (ENU) coordinate system.

[2]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.

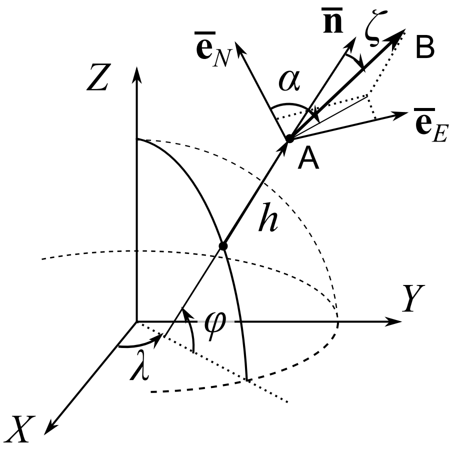


Figure 4.5: Local right-handed topocentric system in point A, with ellipsoidal coordinates (φ, λ, h), with the local azimuth α and zenith angle ζ for the direction $\vec{AB}$. The vertical (ellipsoidal normal vector) $\bar{\mathbf{n}}$ is the z-axis of the local right-handed system , $\bar{\mathbf{e}}_E$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

The relation between the differential coordinates $(\Delta X, \Delta Y, \Delta Z)$ and $(\Delta E, \Delta N, \Delta U)$ is,

$$\begin{aligned} \begin{pmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{pmatrix} &= \begin{pmatrix} -\sin\lambda & -\sin\varphi\cos\lambda & \cos\varphi\cos\lambda \\ \cos\lambda & -\sin\varphi\sin\lambda & \cos\varphi\sin\lambda \\ 0 & \cos\varphi & \sin\varphi \end{pmatrix} \begin{pmatrix} \Delta E \\ \Delta N \\ \Delta U \end{pmatrix} \\ &= \begin{pmatrix} \bar{\mathbf{e}}_E & \bar{\mathbf{e}}_N & \bar{\mathbf{n}} \end{pmatrix} \begin{pmatrix} \Delta E \\ \Delta N \\ \Delta U \end{pmatrix} \end{aligned} \tag{4.15}$$

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

$$\begin{pmatrix} \Delta E \\ \Delta N \\ \Delta U \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{e}}_E & \bar{\mathbf{e}}_N & \bar{\mathbf{n}} \end{pmatrix}^{-1} \begin{pmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{e}}_E & \bar{\mathbf{e}}_N & \bar{\mathbf{n}} \end{pmatrix}^{T} \begin{pmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{pmatrix} \tag{4.16}$$

where we used the property that the inverse of a rotation matrix is the transpose of the matrix. The vectors $\bar{\mathbf{e}}_E$, $\bar{\mathbf{e}}_N$ and $\bar{\mathbf{n}}$ form the three axis of a right-handed local topocentric system centered at the observation point (φ, λ, h), with z-axis along the normal of the ellipsoid $\bar{\mathbf{n}}$, x-axis $\bar{\mathbf{e}}_E$ orthogonal to the plane of the meridian and positive to the East, and y-axis $\bar{\mathbf{e}}_N = \bar{\mathbf{n}} \times \bar{\mathbf{e}}_E$ in the plane of the meridian completing the topocentric system. The 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°, and a point to the East $+90^\circ$, see Figure 4.5. The azimuth α and zenith angle ζ are defined by,

$$\begin{pmatrix} \Delta E \\ \Delta N \\ \Delta U \end{pmatrix} = s \begin{pmatrix} \sin\alpha\sin\zeta \\ \cos\alpha\sin\zeta \\ \cos\zeta \end{pmatrix} \tag{4.17}$$

with s the slant range $\sqrt{\Delta E^2 + \Delta N^2 + \Delta U^2}$. The inverse relation is,

$$\begin{aligned}
\alpha &= \arctan(\frac{\Delta E}{\Delta N}) = \arctan(\frac{<\mathbf{e}_E, \mathbf{s}>}{<\mathbf{e}_N, \mathbf{s}>}) \\
\zeta &= \arctan(\frac{\Delta U}{\sqrt{\Delta E^2 + \Delta N^2}}) = \arccos(\frac{\Delta U}{s}) = \arccos(\frac{<\bar{\mathbf{n}}, \mathbf{s}>}{s}) \\
s &= \sqrt{\Delta E^2 + \Delta N^2 + \Delta U^2} = \sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2} = \sqrt{<\mathbf{s}, \mathbf{s}>}
\end{aligned} \tag{4.18}$$

with $\mathbf{s} = (\Delta X, \Delta Y, \Delta Z)^T$. Compare this to Eq. (3.4), of Chapter 3, where the azimuth and zenith angles were defined in general terms of X, Y and Z coordinates for a topocentric system, whereas in this section the coordinates for the topocentric system have been named E (East), N (North) and U (Up) to emphasize the topocentric nature of the system. If the azimuth α and zenith angle ζ is computed in point A, the origin of the topocentric system, as shown in Figure 4.5, then $(\Delta E, \Delta N, \Delta U)$ in Eqs. (4.15), (4.16), (4.17) and (4.18) can be replaced by (E, N, U), and $(\Delta X, \Delta Y, \Delta Z)$ equals $(X - X_A, Y - Y_A, Z - Z_A)$, with (X_A, Y_A, Z_A) the 3D global Cartesian coordinates of point A computed by Eq. (4.4).

The order of the coordinates in Eqs. (4.15) and (4.16) is sometimes changed, with the N (North) coordinates given before the E (East) coordinate. This forms a left–handed coordinates system and is called a North–East–Up (NEU) system. It follows the common practive for geographic coordinates of giving the latitude before the longitude.

The coordinates N and E (ΔN and ΔE) are sometimes also referred to as *Northing* and *Easting*, however, Northing and Easting have been defined before in Eq. (4.7) of Section 4.2, and the two are not exactly the same.

In Section 4.2, differential ellipsoidal coordinates $d\varphi$ and $d\lambda$), given in radians, where expressed in Northing dN and Easting dE in meters, using the relation of Eq. (4.7)

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

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.

The difference with ΔN and ΔE is that dN and dE are curvilinear coordinates, whereas N and E are Cartesian coordinates. For small values of ΔN and ΔE we have $dN \simeq \Delta N$ and $dE \simeq \Delta E$, but although $dh \simeq \Delta U$, the surface dh = const represents a curved surface, whereas ΔU = const is a plane tangent to the curved Earth. While we may get away with the approximation $dN \simeq \Delta N$ or $dE \simeq \Delta E$, it is a bad idea to mix U and h.

The algorithms that involve $(\Delta X, \Delta Y, \Delta Z)$, $(\Delta N, \Delta E, \Delta U)$, or, the azimuth α and zenith angle ζ can be used with very large values, provided $(\Delta N, \Delta E, \Delta U)$ 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.

The algorithms that involve $d\varphi$, $d\lambda$ and dh, or dN, dE and dh, are only valid for small values or over the surface of the Earth. They are useful mainly for 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.

4.5. Practical aspects of using latitude and longitude

The ellipsoidal height differs by not more than 100 m from an 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 latitude 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 8 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 (See Chapter 7). 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).

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° at the Prime Meridian to 180° Eastward (written 180° E or $+180^\circ$) and 180° Westward (written 180° W or -180°) 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 is called the equator. A *great circle* is the intersection of a sphere and a plane which passes through the center point of the sphere, otherwise the intersection is called *small circle*. Planes parallel to the equatorial plane intersect the surface in circles of constant latitude; these are the parallels. Parallels are small circles. The equator has a latitude of 0°, the North pole has a latitude of 90° North (written 90° N or $+90^\circ$), and the South pole has a latitude of 90° South (written 90° S or -90°). 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

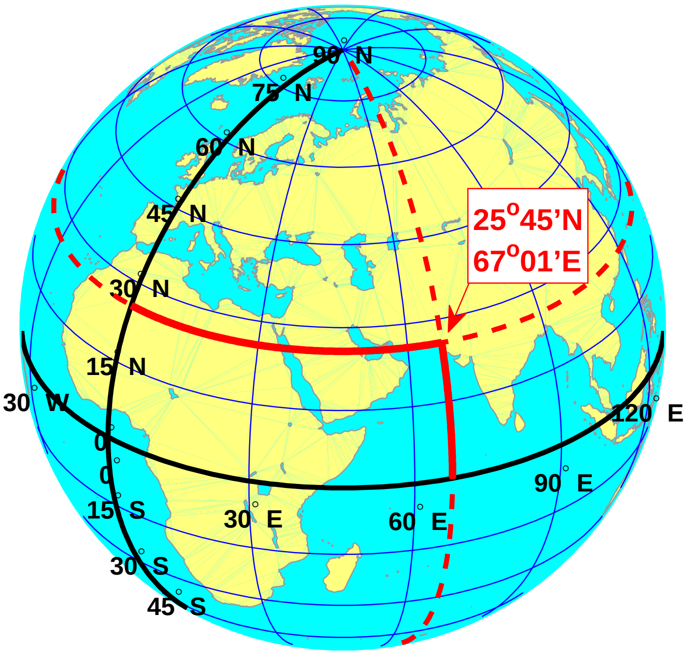


Figure 4.6: Latitude and longitude grid as seen from outer space (orthographic azimuthal projection). The prime meridian (through Greenwich) and equator are in black with latitude and longitude labels. The meridian and parallel through Karachi, Pakistan, $25°45'$N $67°01'$E, are the dotted lines in red. 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).

notation as $23°27'30"$ [EWNS]. The seconds can include a decimal fraction. An alternative representation uses decimal degrees, whereby degrees are expressed as a decimal fraction: $23.45833°$ [EWNS]. Another option is to express minutes with a fraction: $23°27.5'$ [EWNS]. The [EWNS] [3] 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 $\pm 180°$ meridian must be handled with care in calculations, for example when subtracting or adding two longitudes.

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. (4.7) (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. (4.7) 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[4]. Thus, at

[3] In Dutch we use the terms O.L. (Oosterlengte) for E, W.L. (Westerlengte) for W, N.B. (Noorderbreedte) for N and Z.B. (Zuiderbreedte) for S.

[4] In surveying and geodesy a circle is not divided in $360°$ but in 400 gon or grad. This has the added advantage that one gon (grad) measured along the meridian corresponds to 100 km, and one milli-gon (milli-grad) to 100 m, a decimilligon (decimilligrad, 10^{-4} grad) to 10 m and 10^{-7} gon (grad) corresponds to 1 cm. However, this "decimal" system for angular measurement never gained a big following outside surveying. But, be aware,

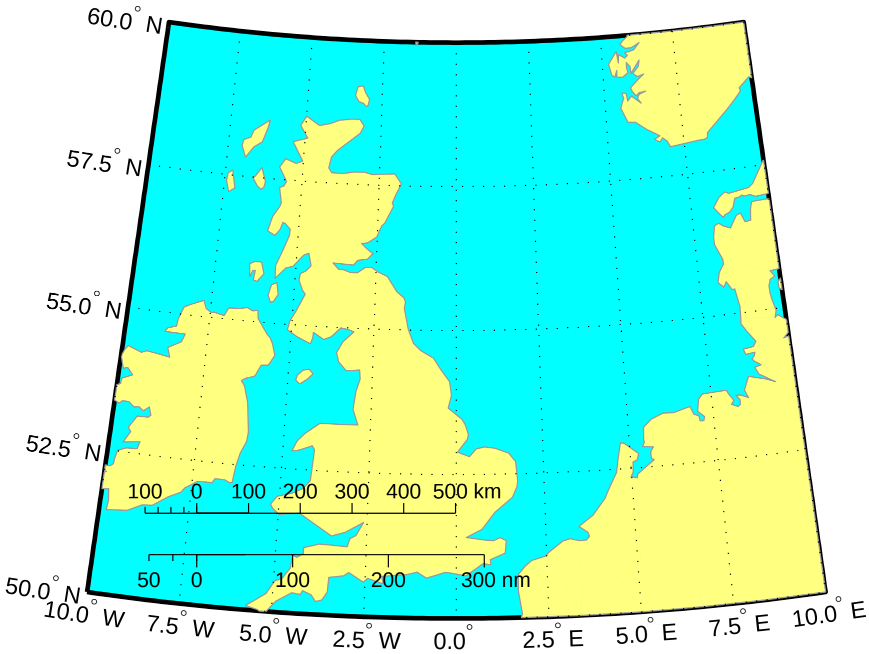


Figure 4.7: The latitude and longitude grid over the North-Sea in an equidistant conic projection of uniform scale. One degree of latitude is about 60 nm (Nautical miles), or more precisely 111.2 km at 50° and 111.4 km at 60° latitude. However, one degree of longitude is much shorter, it varies between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

52° latitude, one arc-second ($1''$) along the meridian corresponds to roughly 30.9 m and one arc-second along the 52° 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 4.6, map projections are discussed in Chapter 5.

4.6. 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 geom-

quite often surveying equipment uses *gon* or *grad* to measure arcs instead of degrees.

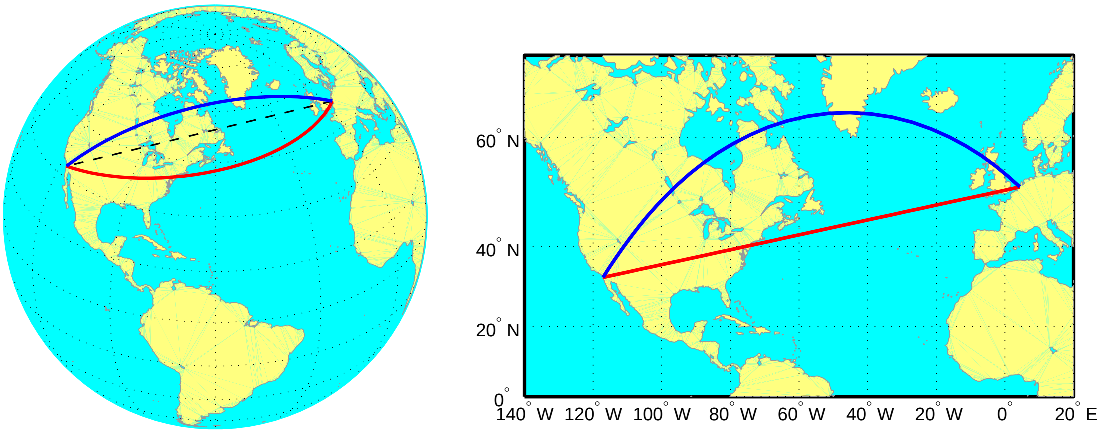


Figure 4.8: Great circle (blue), rhumb line or loxodrome (red) and straight line (black dashed) distance between Delft, NL, $52°$N $4.37°$E and San Diego, CA, USA, $32.8°$N $117.1°$W. The plot on the left uses an orthographic azimuthal projection, with the Earth as seen from outer space, while the plot 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, $61.0°$N $71.7°$W. 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 ($313.5°$) when leaving Delft and SSW ($212.1°$) 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 or loxodrome (red) from Delft to San Diego would have to steer a constant WSW course ($257.8°$). Rhumb lines become straight lines in a Mercator projection.

etry is the shortest path between points on a sphere or ellipsoid, which is called *geodesic* (in Dutch: *de geodetische lijn*). On a sphere geodesics are great circles[5]. 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. The starting point is writing the geodesic as a differential equation relating an elementary

[5]A great circle is the intersection of a sphere and a plane which passes through the center point of the sphere, otherwise the intersection is called small circle.

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. (4.7) 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 $\lambda_{12} = \lambda_2 - \lambda_1$. [6] 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 of the direct geodetic problem. Given the initial course α_1 and distance s_{12} along the great circle, the latitude and longitude of P2

[6] Please note that in this equation the ϕ is used for the latitude, but strictly, since this is a computation on the sphere, we should have used the geocentric latitude ψ. However, as these formules are often used as an approximation to the more difficult problem on the ellipsoid, you find them often expressed in ϕ instead of ψ.

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*.

4.7. Problems and exercises

Question 1 The geographic position coordinates of a geodetic marker in Vlissingen are given as $51°26'34.3501''$ North, $3°35'50.3686''$ East (in ETRS89). Express the geographic position coordinates (latitude and longitude) in decimal degrees.
Answer 1 Going from an angle expressed in degrees, minutes and seconds to decimal degrees, means taking the amount of degrees, adding the number of minutes divided by 60, and adding the number of seconds divided by 3600. This yields $\varphi = 51.442875°$ North, $\lambda = 3.597325°$ East.

Question 2 The geographic position coordinates of a geodetic marker on Terschelling are given as $53.362736°$ North, and $5.219386°$ East (in ETRS89). Express the geographic position coordinates (latitude and longitude) in degrees, arcminutes and arcseconds.
Answer 2 Going from an angle expressed in decimal degrees to degrees, minutes and seconds of arc, means taking the decimal part and multiplying it by 60 and the integer part yields the number of minutes; next taking the original decimal part again, and subtracting the integer number of minutes divided by 60, and multiplying this by 3600. This yields $\varphi = 53°21'45.8496''$ North, $\lambda = 5°13'9.7896''$ East.

Question 3 For the WGS84-ellipsoid, the semi-major axis is given as $a = 6378137.000$ m. And the flattening is $f = 1/298.257223563$. Compute the length of the semi-minor axis b, and also the eccentricity e.
Answer 3 The eccentricity e and semi-minor axis follow from Eq. (4.3), this results in $e = 0.081819191$ and $b = 6356752.314$ m, hence the distance from the pole to the Earth's center is about 21 km shorter than the distance from the equator to the Earth's center.

Question 4 The geographic position coordinates of a geodetic marker on Terschelling are given as $\varphi = 53.362736°$ North, $\lambda = 5.219386°$ East (in ETRS89), and $h = 56.098$ m. Express the position coordinates in Cartesian coordinates. The ellipsoidal parameters of the WGS84 ellipsoid can be found in Table 6.1.
Answer 4 Converting geographic coordinates into Cartesian coordinates is done through Eq. (4.4), with the expression for the radius of curvature in the prime vertical in Eq. (4.6). At the given latitude, the radius is $\bar{N}(\varphi) = 6391928$ m. The Cartesian coordinates are $X = 3798580.857$ m, $Y = 346993.872$ m, $Z = 5094780.835$ m.

Question 5 The Cartesian coordinates of a location (in the Atlantic Ocean) are given as $X = 6378137.000$ m, $Y = 0.000$ m, $Z = 0.000$ m. Compute, using the WGS84-ellipsoid (see Table 6.1), the geographic coordinates of this location.
Answer 5 The formal computation goes through Eqs. (4.5) and (4.6), and requires an iteration, see Eq. (4.8). However, in this special case, as we note that $Z = 0.000$ m, we can immediately conclude that this location lies in the equatorial plane, and latitude $\varphi = 0°$. The longitude follows easily as $\lambda = 0°$. And eventually the ellipsoidal height $h = 0.000$ m, as the radius of curvature in the prime vertical equals $\bar{N} = a = 6378137.000$ m, see also Figure 4.4.

Question 6 The position of a GPS receiver on the Delft campus is computed in 2015 using two different processing services: NETPOS and NRCAN. The result from the NETPOS processing service, given in ETRS89, is $\varphi_1 = 51°59'50.80858''$ North, $\lambda_1 = 4°22'33.0427''$ East and $h_1 = 43.5579$ m. The result from the NRCAN processing service, given in ITRF2008, is $\varphi_2 = 51°59'50.82510''$ North, $\lambda_2 = 4°22'33.0659''$ East and $h_2 = 43.5490$ m. After conversion to ETRS89 the coordinates from the NRCAN processing are $\varphi_3 = 51°59'50.80910''$ North, $\lambda_3 = 4°22'33.0433''$ East and $h_3 = 43.5513$ m. Compute the differences in meters between the NETPOS and NRCAN processing, both in ETRS89, and compute the differences in meters between the ITRF2008 and ETRS89 solutions for NRCAN.
Answer 6 It is clear that the differences are very small, only a fraction of a second of arc. The difference between the NETPOS and NRCAN solution, both in ETRS89, is $\Delta\varphi = \varphi_3 - \varphi_1 = 0.00052''$, $\Delta\lambda = \lambda_3 - \lambda_1 = 0.0006''$ and $\Delta h = h_3 - h_1 = -0.0266$ m. To convert the differences into units of meters Eq. (4.7) is used. At $52°$ latitude we have

$$\Delta N[\text{m}] = \frac{\pi}{180 * 3600} * 6391000 * \Delta\varphi[''] \simeq 31.0 * \Delta\varphi['']$$
$$\Delta E[\text{m}] = \frac{\pi}{180 * 3600} * 6376000 * \cos(\varphi) * \Delta\lambda[''] \simeq 19.0 * \Delta\lambda['']$$

whereby we obtained $\bar{N}(\varphi) \simeq 6391$ km and $\bar{M}(\varphi) \simeq 6376$ km from Figure 4.4 or Eq. (4.6). Note that $R = 6371$ km instead of $\bar{N}(\varphi)$ and $\bar{M}(\varphi)$ would have given a more or less similar result. The difference between the NETPOS and NRCAN solution is thus $\Delta N = 31 * 0.00052'' = 0.0161$ m , $\Delta E = 19 * 0.0006'' = 0.0114$ m and $\Delta h = -0.0266$ m. The differences between the two solutions are in the order of centimeters.

The difference between the ETRS89 and ITRF2008 solution is $\Delta\varphi = \varphi_2 - \varphi_3 = 0.01600''$, $\Delta\lambda = \lambda_2 - \lambda_3 = 0.0226''$ and $\Delta h = h_2 - h_3 = -0.0177$ m. To convert the differences into units of meters again Eq. (4.7) is used, which results in $\Delta N = 31 * 0.016'' = 0.496$ m, $\Delta E = 19.088 * 0.0226'' = 0.429$ m and $\Delta h = -0.0177$ m.

Please note that the horizontal differences between the ITRF2008 and ETRS89 solutions of NRCAN, at decimeter level, are much larger than the differences between NRCAN and NETPOS solutions in the same reference frame. This is due to ETRS89 moving along with the European plate, with station velocities in Europe close to zero, whereas in ITRF2008 Delft is moving yearly 2.3 cm to the North-East. Over a period of 26 years (the epoch of observation, 2015, minus 1989, the year ETRS89 and ITRF2008 coincided), this corresponds to about 0.60 m. See also Chapter 9 for more information. It also shows the importance of datum transformations, which we used to convert ITRF2008 coordinates to ETRS89, and is the topic of Chapter 6.

5
Map projections

Geographic latitude and longitude are convenient for expressing positions on the Earth, but computations on the sphere, let alone the ellipsoid, are quite complicated as we have seen in the previous Chapter. 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. This is known as a *map projection*. From then on simple 2D Eucledian metric can be used.

5.1. Introduction

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 (x, y) are called *map* or *grid* coordinates. The grid coordinates are often referred to as Northing (y) and Easting (x).

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 incorrect results when using coordinates from a different type of projection.

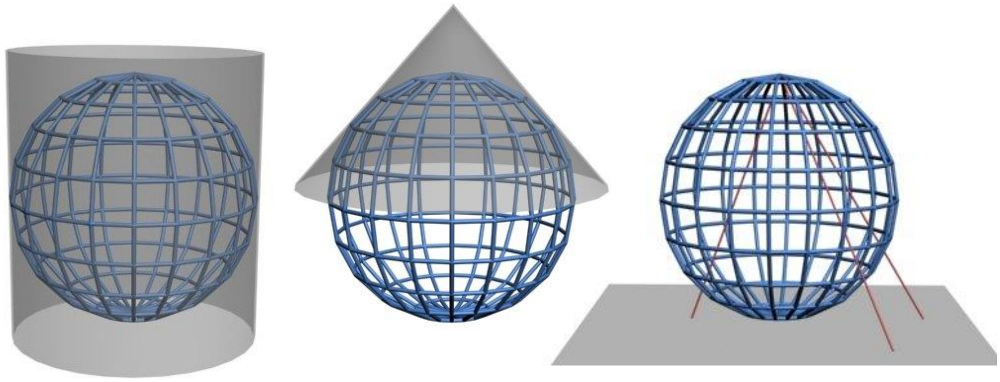


Figure 5.1: Cylindrical, Conic and Azimuthal map projection types (Source: Wikimedia Commons).

5.2. Map projection types and properties

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. Figure 4.8 (right part) of the previous chapter is a 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. An example is Figure 4.7 of the previous chapter.

Azimuthal map projections The plane of projection is a plane tangent to the Earth. Two well known examples are the stereographic projection, which is used for instance by the Dutch RD system, see Chapter 10, and the orthographic azimuthal projection used in Figure 4.6 of the previous chapter.

Miscellaneous projections Mostly used for cartographic purposes.

Any projection can be applied in the *normal*, *oblique* and *transverse* position of the cylinder, cone or tangent 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 of a cylindrical map projection is the *Mercator* projection, with the equator as the line of contact of the cylinder, see Section 5.4.2 and Figure 4.8. 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.

Map projections also differ in the *point of perspective* that is used. For instance, the point of perspective for the azimuthal *stereographic projection* is a point on the Earth opposite to the tangent plane, as is depicted in Figure 5.1 on the right. On the other hand, for the *orthographic azimuthal projection*, which was used for Figure 4.6 and in Figure 4.8 (left part), the point of perspective is at infinite distance. The orthographic azimuthal projection depicts a hemisphere of the globe as it appears from outer space, which results in shapes and areas distorted particularly near the edges.

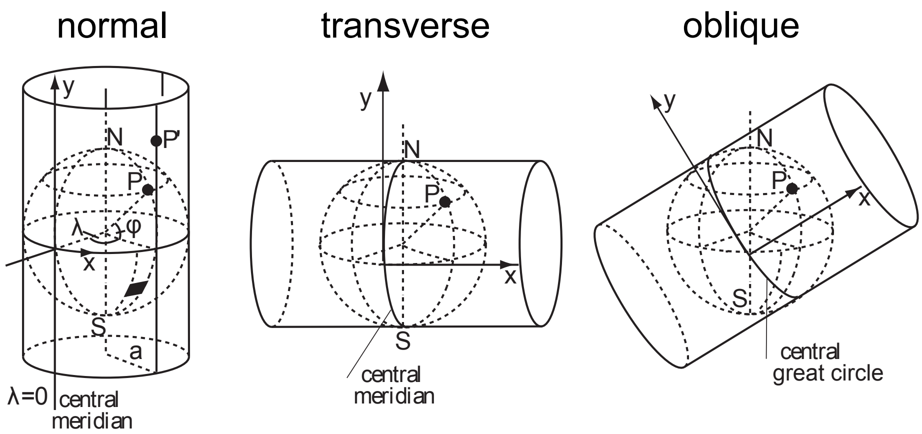


Figure 5.2: Normal, transverse and oblique projection for a cylinder (Source: Wikimedia Commons).

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 (see Section 2.2) in an infinitesimally 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 from the Bessel ellipsoid 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 (see Section 6.4)
- 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.

5.3. Practical aspects of map projections

Working with planar grid coordinates to compute distances, angles and areas 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 areas. 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, see Section 5.4.5, which uses between $80°$S and $84°$N latitude 60 zones, each of $6°$ 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).

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 geodetic 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 (> 30 km) 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 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.

5.4. Cylindrical Map projection examples

In this section several examples of cylindrical projections are presented. Cylindrical projections have been chosen because the mathematics are less complicated than those of other map projections, and thus serve well to illustrate some principles of map projections. Some of the cylindrical projections that are discussed are only for illustration, but others, like the Mercator, Web Mercator and UTM projections, are used (almost) on an every day basis.

With the cylindrical projection the Earth's surface is projected onto a cylinder tangent to the equator, as shown in the left part of Figure 5.3. The map projection turns (spherical)

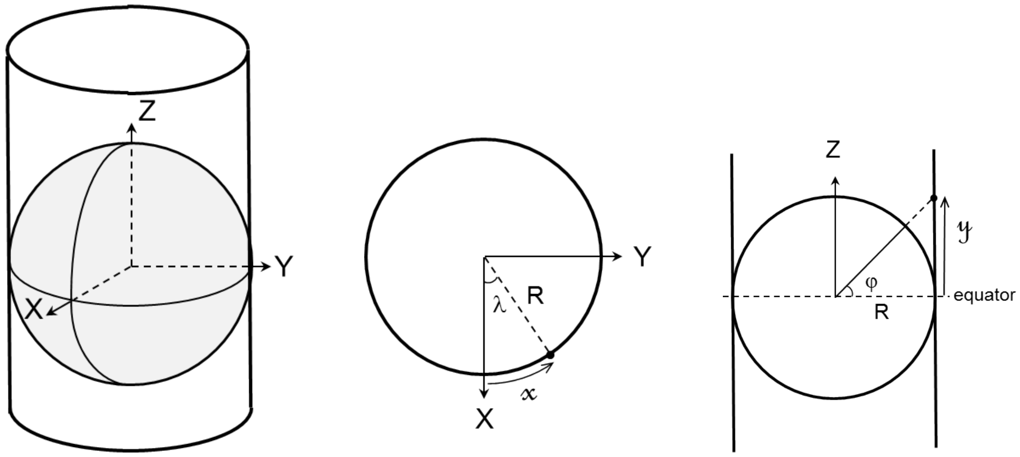


Figure 5.3: For the cylindrical projection, the mapping plane is wrapped around the Earth like a cylinder (left), longitude λ turns into map-coordinate x (middle), and latitude φ turns into map-coordinate y (right).

coordinates (φ, λ) [1] of points on the Earth's surface into map or grid coordinates (x, y). The map origin $(x = 0, y = 0)$ is at the intersection of the equator and the Greenwich meridan $(\varphi = 0, \lambda = 0)$. In the sequel latitude φ and longitude λ are expressed in radians. The Earth's surface is approximated by a sphere with radius R. The middle part of Figure 5.3 shows a top view of the equatorial plane. The distance from $(\varphi = 0, \lambda = 0)$ *along* the equator to an object at longitude λ equals $R\lambda$, hence we simply have: $x = R\lambda$ for all normal cylindrical projections. This is a property of all normal cylindrical projections: points on the meridian have a constant x value.

The function $y = y(\varphi)$ to project latitude φ onto y values is still open, it can be any one from an unlimited number of functions. In Figure 5.3, on the right, one such function is illustrated: the central cylindrical projection.

5.4.1. Central cylindrical projection

In case of the central cylindrical projection points on the Earth are projected, from the origin at the middle of the Earth, onto a cylinder tangential to the Earth at the equator. The right part of Figure 5.3 shows a meridianal cross section of the Earth at longitude λ. The object point, projected onto the cylinder, has a distance $R \tan \varphi$ from the equator, hence we have: $y = R \tan \varphi$. The map-projection equations for the central cylindrical projection are thus

$$\begin{aligned} x &= \mu R(\lambda - \lambda_0) \\ y &= \mu R \tan \varphi \end{aligned} \tag{5.2}$$

with μ a scaling factor and λ_0 the central meridial (e.g. Greenwhich meridian with $\lambda_0 = 0$), To represent coordinates on an actual paper map, or graphical display, a very small scaling factor μ is applied, e.g. to obtain a paper map with map-scale $1 : 10000000$ you would select $\mu = \frac{1}{10000000}$ rather than 1.

The true scale on the Equator is unity for $\mu = 1$. Everywhere else the linear scale is stretched by a factor of $1/\cos \varphi = \sec \varphi$ in the x-axis direction, and $1/\tan \varphi = \cot \varphi$ in the y-axis direction. The central cylindrical projection is neither conformal or equal area. Distortion

[1]The notation used here is the one for geographical coordinates. The distortions that are inherent to this projection make that the use of spherical or geographic coordinates doesn't matter for a graphical representation. However, formallly, geographic coordinates should first be projected onto spherical coordinates, before the map projection is applied.

increases so rapidly away from the equator, see Figure 5.4, that the central cylindrical is seldomly used for practical maps. Its vertical, latitudinal, stretching is even greater than that of the Mercator projection, which we discuss next.

5.4.2. Mercator projection

The Mercator projection is a cylindrical map projection, presented by the Flemish geographer and cartographer Gerardus Mercator, in 1569. It became the standard map projection for nautical navigation, as a line of constant course, known as rhumb line, see the red-line in Figure 4.8, is shown as a straight line, that conserves the angle with the meridians.

As in all cylindrical projections, parallels and meridians are straight and perpendicular to each other. The Mercator map-projection is a conformal map projection, meaning that angle between any two straight lines or curves is preserved. To this end the East-West stretching of the map (to 'undo' the meridian-convergence), which increases as distance away from the equator increases, is accompanied by a corresponding North-South stretching. The distance between the parallels gets larger and larger, the further one gets away from the equator, like in any cylindrical projection, but the amount by which is chosen carefully as to preserve angles.

As the radius of a parallel, or circle of latitude, is $R\cos\varphi$, the corresponding parallel on the map, a line with with a constant y coordinates has been stretched by a factor of $1/\cos\varphi = \sec\varphi$ in the x-coordinate direction. To preserve angles the same amount of stretching needs to be applied in the y-coordinate direction. This implies that the derivative of the map-coordinates function $y(\varphi)$ must be $y'(\varphi) = R\sec\varphi$. [2] Integrating this equation gives

$$y(\varphi) = R\ln\left[\tan(\frac{\pi}{4}+\frac{\varphi}{2})\right] \tag{5.3}$$

This function is illustrated in Figure 5.4. The map projection formulae for a basic normal Mercator projection are thus

$$\begin{aligned} x &= \mu R(\lambda-\lambda_0) \\ y &= \mu R\ln\left[\tan(\tfrac{\pi}{4}+\tfrac{\varphi}{2})\right] \end{aligned} \tag{5.4}$$

with μ a scaling factor and λ_0 the central meridian. The true scale on the Equator is unity for $\mu = 1$. Everywhere else the linear scale is stretched by a factor of $1/\cos\varphi = \sec\varphi$. This distorts the size of geographical objects far from the equator; objects like Greenland and Antartica appear to be much larger than they in reality are, see Figure 4.8, and also Figure 5.4. The Mercator projection is conformal, it preserves angles, but it is definitely not an equal area projection. By choosing a value of μ slightly smaller than one (effectively decreasing the radius of the cylinder) we can create a Mercator projection with the unity scale for two parallels, but this does not solve the problem of distortions. At higher latitude the Mercator projection becomes unusable, and even becomes singular at the poles (the North and South pole become lines at $y = \infty$).

In the previous equations it was assumed that the Earth was modelled by a sphere, or, more precisely, we should have used sperical coordinates (λ,ψ) instead of geographical coordinates (λ,φ). To use geographic coordinates instead of sperical coordinates is only a minor approximation for global small scale maps.

When the Earth is modelled by an ellipsoid, with (λ,φ) is the geographic longitude and latitude, the Mercator projection must be modified to remain conformal. The map projection

[2] The derivative for the central cylindrical projection is $y'(\varphi) = R\sec^2\varphi$.

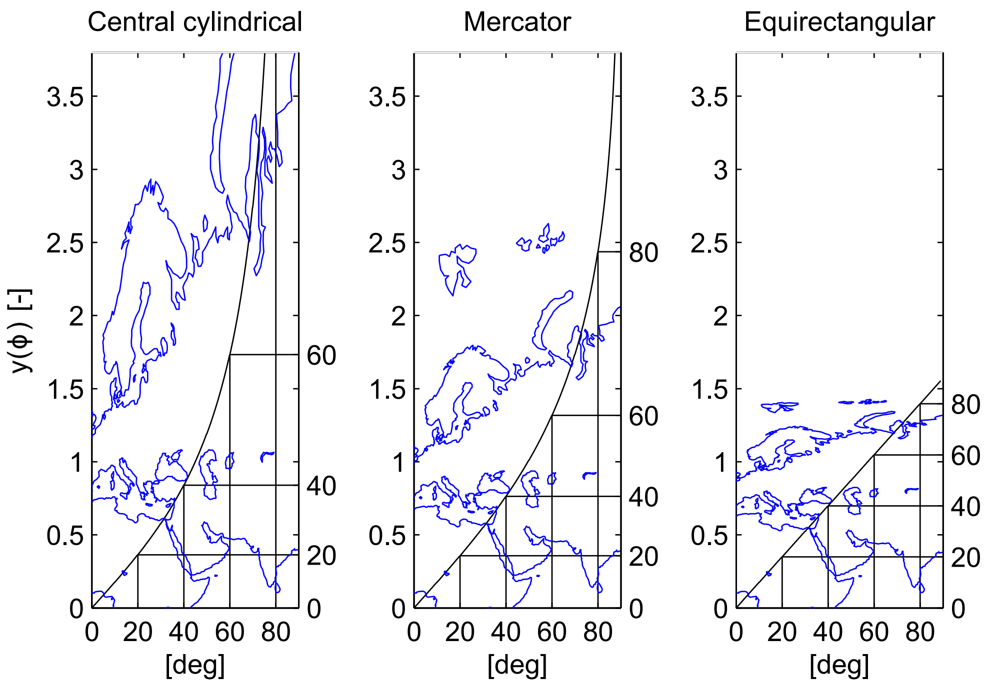


Figure 5.4: Mapping function $y = y(\varphi)$ for the Central cylindrical, Mercator and Equirectangular (Plate Carrée) projections. The function $y(\varphi)$ with, $R = 1$, is the black line. The x-axis is φ in degrees, the y-axis on the left of each plot gives the map-coordinate $y = y(\varphi)$, the y-axis on the right of each plot gives the latitude (in degrees) that corresponds to $y(\varphi)$. The blue lines are coast lines for part of the Earth plotted with the function $y(\varphi)$ on the y-axis, with longitude (in degrees) on the x-axis.

formula in case of the ellipsoidal model are

$$\begin{aligned} x &= \mu R(\lambda - \lambda_0) \\ y &= \mu R \ln \left[\tan(\frac{\pi}{4} + \frac{\varphi}{2}) \left(\frac{1 - e \sin\varphi}{1 + e \sin\varphi} \right)^{\frac{e}{2}} \right] \end{aligned} \tag{5.5}$$

with e the excentricity of the ellipsoid.

5.4.3. Plate carrée and equirectangular projections

A simple longitude-latitude presentation is obtained when the x- and y-coordinates are scaled by R in the same way. This is called the *plate carrée* projection. The map-projection equations for this simple cylindrical map-projection are

$$\begin{aligned} x &= R(\lambda - \lambda_0) \\ y &= R\varphi \end{aligned} \tag{5.6}$$

The parallels and meridians are being equidistant in the map and form a square grid[3], as can be seen in Figure 5.4. The scale in the latitude (N-S) direction is uniform, at least for a spherical Earth. However, the scale for the longitude (E-W) direction is not uniform and decreases with the latitude. The plate carrée projection is a special case of the equirectangular projection.

The map projection equations for the *equirectangular* projection, with standard parallels

[3]With the default Mercator projection the parallels get further and further apart the more to the North (South) you go, and with the central cylindrical projection this will be even more the case.

at φ_1 North and South of the equator, are

$$x = R(\lambda - \lambda_0)\cos\varphi_1$$
$$y = R(\varphi - \varphi_1) \tag{5.7}$$

The projection maps meridians to vertical straight lines of constant spacing, and circles of latitude to horizontal straight lines of constant spacing, to form a rectangular grid. The scale of the projection is true at both standard parallels φ_1. The projection is neither equal area nor conformal.

Because of the distortions introduced by this projection it has little use in navigation or cadastral mapping. However, it is an easy to use projection for mapping small areas, and it does a much better job than simply plotting longitude and latitude values in an xy-plot, what the Plate Carré projection basically does.

5.4.4. Web Mercator

The Web Mercator projection is a variant of the Mercator projection that is used by many Web mapping applications, including Google Maps, Bing Maps, OpenStreetMap and others. Its official EPSG identifier is EPSG:3857. It uses the same spherical formulas of Eq. 5.4 as the standard Mercator, however, the Web Mercator uses the spherical formulas with the geographical coordinates (λ, φ) in the WGS 84 ellipsoidal datum. The discrepancy is imperceptible at the global scale, but causes maps of local areas to deviate slightly from true ellipsoidal Mercator maps. This discrepancy also causes the projection to be slightly non-conformal. For these reasons, several agency have declared this map projection to be unacceptable for any official use.

5.4.5. Universal Transverse Mercator (UTM)

The normal Mercator projection works quite well in a small band around the Equator, but performs very poorly at higher latitudes. Switching from a normal projection, to a transverse projection, as in Figure 5.2, results in a projection that works quite well in a small band around the central meridian. This approach is used by the popular Universal Transverse Mercator (UTM) projection for latitudes between $80°$S and $84°$N. The UTM projection uses 60 zones, each of $6°$ width in longitude (up to 668 km), centered around a central meridian. Each zone is numbered. For instance, the Netherlands falls in two zones, 31N and 32N. Zone 31N covers longitude $0°$ to $6°$E, zone 32N covers longitude $6°$E to $12°$E.

The scale factor along the central meridian is not 1, but 0.9996, so that the inevitable distortion is spread more uniformly over the zone. The amount of distortion is less than $1/1000$.

In each zone the scale factor of the central meridian reduces the diameter of the transverse cylinder to produce a secant projection with two standard lines, or lines of true scale, about 180 km on each side of, and about parallel to, the central meridian (Arc cos 0.9996 = 1.62° at the Equator). The scale is less than unity inside the standard lines and greater than unity outside them, but the overall distortion is minimized

The polar regions South of $80°$S and North of $84°$N are excluded.

5.5. Problems and exercises

Question 1 We do have a geographic database available, with position coordinates in a three-dimensional Cartesian Earth Centered, Earth Fixed (ECEF) reference system. We would like to create a map of the Northern hemisphere, using an orthographic azimuthal projection (with the mapping plane being parallel with the equatorial plane, and lying/touching the North pole).

Set up the 3-by-3 projection matrix to perform the mapping operation on the three dimensional coordinates in the database.

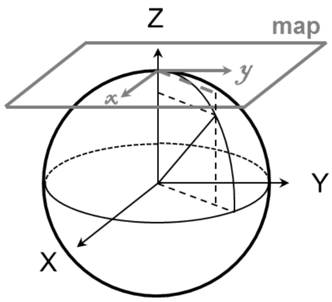


Figure 5.5: Orthographic azimuthal map projection. The point of tangency of the mapping plane is the North Pole.

Answer 1 The mapping plane is $Z = b$, with b the semi-minor axis of the ellipsoid (or the radius of the sphere). Next, orthographic means that the projection lines are all perpendicular to the mapping plane, and in this case parallel to the Z-axis. Hence the projection matrix is

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and the Z coordinate is eventually to be translated to $Z = b$. So mapping, or grid coordinates (x, y) become $x = X$ and $y = Y$. This is shown in Figure 5.5. Eventually you may want to apply a scale factor μ, so that $x = \mu X$ and $y = \mu Y$.

6

Datum transformations and coordinate conversions

Geospatial projects often involve spatial coordinates from different sources, each using their own coordinates representation and reference system. In order to establish the correct spatial relationships, first the coordinates have to be transformed into the same reference system and representation. Transformations between reference systems are also called *geodetic datum transformations*. In this chapter we discuss geodetic datum transformations and coordinate conversions.

6.1. Geodetic datum

In the previous chapters several types of spatial 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. But, somehow, spatial coordinates need to be linked to the Earth, the so–called *geodetic datum*. For instance, take the example of Cartesian and ellipsoidal coordinates first. For 3D Cartesian coordinates we need to define 7 parameters: three for the origin, three for the orientation of the axis, and one for scale (see Section 3.4). For ellipsoidal coordinates, i.e. geographic latitude, longitude and ellipsoidal height, we need to define

- Shape of the ellipsoid: length of semi-major axis and flattening (or length of semi-minor axis or inverse flattening) of the chosen ellipsoid (2 parameters)
- Position of the ellipsoid with respect to the Earth: origin, orientation and scale of the ellipsoid (7 parameters)

When countries developed their national coordinate systems at the end of the 19th and beginning of the 20th century, each country choose an ellipsoid of revolution that best fitted their country based on astronomical observations. This resulted not only in different choices for the shape of the ellipsoid, but also in different positions of the ellipsoid on the Earth. Table 6.1 gives the parameters for three commonly used ellipsoids in the Netherlands. Because of the limited accuracy of astronomical observations at the time, the position of the ellipsoids also differed.

Therefore, each spatial coordinate systems has 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

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.

scale is defined. However, this can only be done in a relative sense, using a *geodetic datum transformation* to link one reference system to the other. The parameters that are involved, usually origin, orientation and scale changes, are the so-called *geodetic datum transformation parameters*.

6.2. Coordinate operations

It is common for spatial reference systems to use different geodetic datums, reference ellipsoids and map projections. Therefore, coordinate transformations between two different systems not only involve a 7-parameter similarity transformation, the datum transformation, but often also a change in the reference ellipsoid, type of map projection and projection parameters. Two types of coordinate operations have to be distinguished

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.

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.

A diagram showing the relations between datum transformations and coordinate conversions is presented in Figure 6.1.

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 of both systems are also possible, but not shown in Figure 6.1. These affine or polynomial transformations are mostly course approximations.

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.

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,

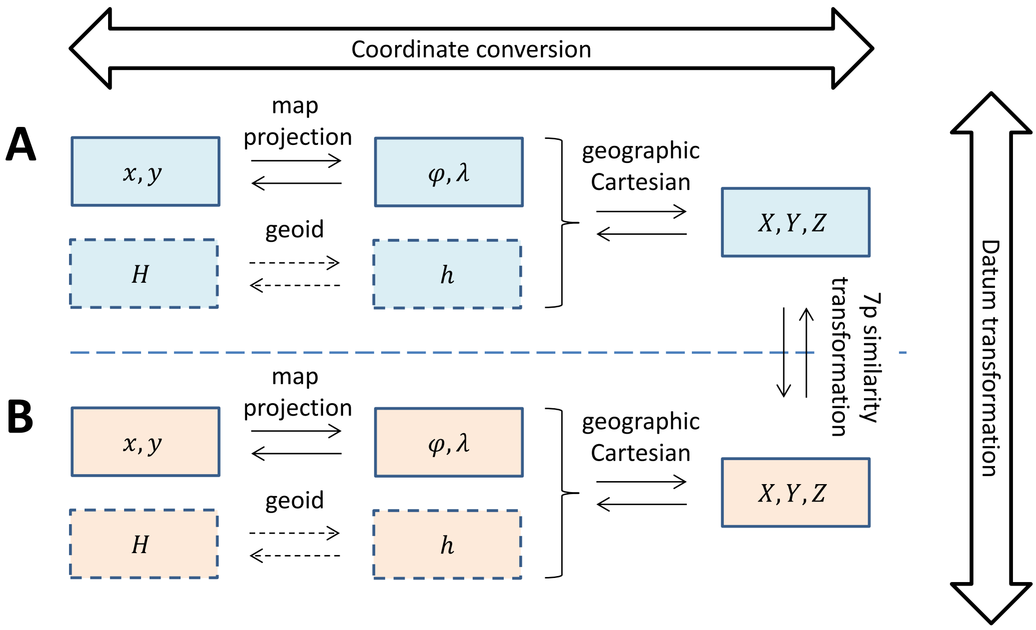


Figure 6.1: Coordinate conversions and datum transformations. Horizontal operations represent coordinate conversions. The vertical operations are datum transformations from system A to B. Not shown in this diagram are polynomial transformations (approximations) directly between map coordinates or geographic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

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.

A big difference between datum transformations and coordinate conversions is that the parameters for the datum transformation are often empirically determined and thus subject to measurement errors, whereas coordinate conversions are fully deterministic. More specific, three possibilities need to be distinguished for the datum transformation parameters

1. 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).

2. 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.

3. 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.

6.3. A brief history of geodetic datums

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 whether 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.

6.4. EPSG dataset and coordinate conversion software

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 a 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 EPSG database[1]. The EPSG database, although extremely useful, has no official status, and sometimes contains only approximate parameters. For high precision applications, that require an accuracy better than 1 meter, the user should be very careful. For instance, the EPSG code for the Dutch RD coordinate system is EPSG:28992 (Amersfoort / RD New – Netherlands) `https://epsg.io/28992`, but the accuracy is just a few decimeters. This is fine for visualizations on a map or GIS system, but it by no means a substitute for the official coordinate transformation that is discussed in Chapter 10.

Software for map projections, coordinate conversions and datum transformations is provided for instance by the open source PROJ package (`https://proj.org/`) used by several Geographic Information System (GIS) packages (e.g. the open source QGIS package). PROJ started purely as a cartography application, but over the years support for datum shifts and more precise coordinate transformations were added to PROJ. In their own words: "Today PROJ supports more than a hundred different map projections and can transform coordinates between datums using all but the most obscure geodetic techniques".

PROJ includes command line applications for conversion of coordinates from text files or user input, and an application programming interface. Coordinate transformations are defined by string that holds the parameters of a given coordinate transformation, e.g. the example string `+proj=merc +lat_ts=56.5 +ellps=GRS80` specifies a Mercator projection with the latitude of true scale at $56.5°$N on the GRS80 ellipsoid. The command `proj +proj ...` converts the geographic (geodetic) coordinates, read from standard input, to map coordinates. The `cs2cs` command line utility is used to transform from one coordinate reference system to another, using two `+proj` strings to specify the source and destination system. If you know the EPSG identifiers these can be used to specify the source and destination.

PROJ supports the Dutch RD coordinate system, see Chapter 10, through EPSG code EPSG:28992 for the approximate transformation, but recent versions of PROJ can also be used to implement the RDNAPTRANS™2018 procedure using a daisy chain of `+proj` strings.

6.5. Problems and exercises

Question 1 Compute the difference in semi-major axis length between the WGS84 ellipsoid and the Bessel-1841 ellipsoid.

Answer 1 The length of the semi-major axis of the WGS84 ellipsoid is $a_{\mathrm{WGS84}} = 6378137$ m, and of the Bessel-1841 ellipsoid $a_{\mathrm{Bessel}} = 6377397.155$ m (see Table 6.1). Hence, the difference is 739.845 m. That is a nearly a kilometer difference at the equator!

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

7

Gravity and gravity potential

In this chapter we introduce, as a preparation for the next chapter, on vertical reference systems, the concepts of gravity and gravity potential. These concepts are illustrated by means of the very simple example of the Earth being a perfect sphere. Eventually we introduce the geoid.

7.1. Introduction

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) $\ddot{\mathbf{x}}$

$$\mathbf{F} = m\ddot{\mathbf{x}} \tag{7.1}$$

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].

7.2. 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).

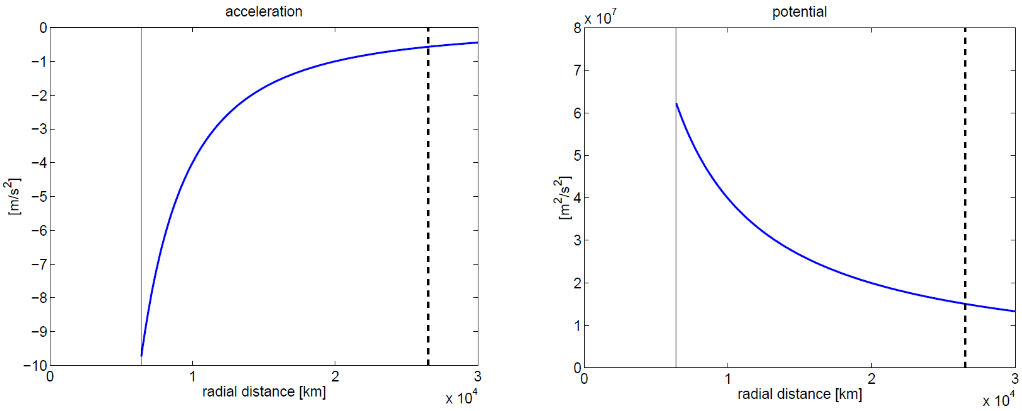


Figure 7.1: The (size of) acceleration $-g$ (left) and the potential W (right) both approach zero as radial distance r goes to infinity. A radial distance of zero corresponds to the Earth's center. The curves start at the Earth's surface (the thin vertical line at $r = 6378$ km). The dashed line indicates the radial distance of the orbit of a GPS satellite; the acceleration (across track) is less than 1 m/s^2, at a speed of almost 4 km/s. The negative sign for the acceleration, $-g$, is used to match the radial direction for an object in free fall.

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.

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

$$\mathbf{F} = -\frac{GMm}{r^2}\frac{\mathbf{r}}{r} \tag{7.2}$$

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

$$\mathbf{g} = -\frac{GM}{r^2}\frac{\mathbf{r}}{r} \tag{7.3}$$

The force in Eq. (7.2) has been divided by mass m to obtain the acceleration in Eq. (7.3), which consequently could be interpreted as the force per unit mass. The magnitude of the gravitational acceleration $g = ||\mathbf{g}||$ in Eq. (7.3) decreases with increasing height above the Earth's surface, and reduces to zero at infinite distance, see Figure 7.1 (at left). The acceleration (vector) $\mathbf{g}$ also points toward the Earth's center.

7.3. 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

$$W = \int_A^B \mathbf{F} \cdot d\mathbf{s} \tag{7.4}$$

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 Eq. (7.4), which is actually the acceleration in Eq. (7.3), and interpreting the work in Eq. (7.4) as a difference in energy $\Delta W = W_B - W_A$ after and before the force carrying the object from A to B, causes W in Eq. (7.4) 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 in Eqs. (7.2)) and (7.4), the potential of gravitation becomes

$$W = \frac{GM}{r} \tag{7.5}$$

for a pointmass or spherical Earth. In this (geodetic) sense, the potential is expressed in [Nm/kg], which equals [m^2/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 7.1 at right. Substituting Eq. (7.5) in Eq. (7.3), gives the relation between gravitational acceleration and gravitational potential

$$\mathbf{g} = -\frac{W}{r}\frac{\mathbf{r}}{r} \tag{7.6}$$

with for the magnitude $g = \|\mathbf{g}\| = \frac{W}{r}$. Also the derivative of the potential of gravitation (with respect to position) is roughly equal to the gravitational acceleration, i.e. $\frac{dW}{dr} \simeq g$ (for a homogeneous sphere the relation is exact), in other words, the slope of the curve at right in Figure 7.1 equals the acceleration shown at left.

From a physics perspective W in Eq. (7.5) 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 Eq. (7.5)).

7.4. Geoid

According to Eq. (7.5), 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 $\mathbf{g}$ 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 (MSL).

A geoid is shown in Figure 7.2. Obviously, good knowledge of the geoid is crucial for coastal engineering and construction of canals. On a global scale, the Earth Gravitational Models (EGMs) are the most commonly used geopotential models of the Earth. They consist of spherical harmonic coefficients published by the US National Geospatial-Intelligence Agency (NGA) with reference to the GRS80–ellipsoid that is also used by WGS84 and ITRS (`https://earth-info.nga.mil/GandG/update/index.php?dir=wgs84&action=wgs84#tab_egm2008`). Three versions of EGM are published: EGM84 with degree and order of harmonic coefficients 180, EGM96 with degree and order 360, and EGM2008 with degree and order 2160. The higher the degree and order of harmonic coefficient, the more parameters the models have, and the more precise they are. Also provided by NGA is a 2.5-minute worldwide geoid height file, precomputed from the EGM2008. The first EGM, EGM84, was defined as a part of WGS84, and is still used by many GPS devices to convert ellipsoidal height into

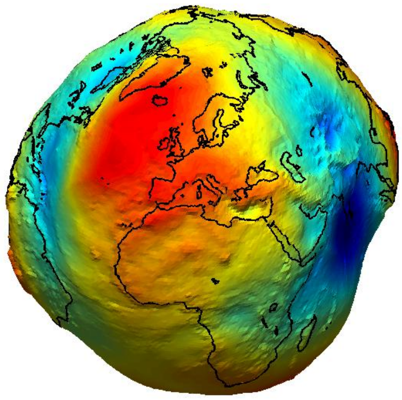


Figure 7.2: The height of the geoid with respect to the best fitting Earth ellipsoid (GRS80). The color-scale ranges from about -100 m (blue) to +70 m (red). This geoid is based on GRACE data. Image from ESA (http://www.esa.int/spaceinimages/Images/2004/10/The_Earth_s_gravity_field_geoid_as_it_will_be_seen_by_GOCE).

height above mean sea-level. The resolution and precision of global models is not sufficient for applications on a local scale. Therefore, many countries, including the Netherlands (see Section 10.3), have computed more precise geoids over a smaller region of interest.

As an introduction, the (shape of the) Earth and its gravity field have been treated here as being (perfectly) spherical, just like in Section 4.1. 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) as in Section 4.2, and subsequently the inhomogeneous distribution of mass within (and on) the Earth, and the presence of heavenly bodies are considered. Hence, the shape of the geoid, in particular departures from being a sphere or an ellipsoid, is determined by the actual mass distribution of the Earth, the outside surface shape, and also inside. The shape of the geoid may vary over time, think for instance of mass loss in polar regions due to ice and snow melt, sea-level rise and ground water level changes.

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 rotations (this effect is largest at the equator, and absent at the poles). 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.

7.5. 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

sphere	ellipsoid	geoid	Earth's surface
< 25 km	reference	< 150 m	< 10 km

Table 7.1: Deviations of different (best fitting) models of the Earth, and also the actual Earth's surface (topography), all referenced to the shape of the ellipsoid.

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).

At the Earth's surface the magnitude of gravity changes by $3 \cdot 10^{-6}$ m/s^2 over a 1 meter height difference ($\frac{dg}{dr}$). This is the slope of the curve in the graph at left in Figure 7.1,

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 bend as it travels through a (strong) gravity field.

7.6. Conclusion

In this chapter we learned that surfaces where the gravity potential W is constant are equipotential surfaces. The gravity vector $\mathbf{g}$ is everywhere orthogonal to them, dictating the local level, and hence water flow. The equipotential surface at mean sea level (MSL), the *geoid*, is thererfore the ideal surface of *reference* in a vertical sense.

A rotational ellipsoid (oblateness of the Earth), as in Section 4.2, is a reasonable approximation to the Earth's geoid. This approximation is popular when not specifically dealing with physical heights and the flow of water. The deviations between the geoid and rotational ellipsoid are smaller than 150 meters, as shown in Figure 7.2 and Table 7.1. Table 7.1 also includes the deviation with topography and a spherical approximation of the Earth.

The shape of the geoid, in particular departures from being a sphere or an ellipsoid, is determined by the actual mass distribution of the Earth. The shape of the geoid may vary over time, think for instance of mass loss in polar regions due to ice and snow melt, sea-level rise and ground water level changes.

7.7. Problems and exercises

Question 1 Compute the magnitude of the acceleration due to attraction by the Earth's mass, at the equator, and at a pole, assuming the Earth is a perfect ellipsoid (WGS84), and all mass is concentrated in the Earth's center.

Answer 1 The acceleration due to attraction by the Earth is given by Eq. (7.3), which holds for a spherical Earth with its mass homogeneously distributed, or all mass concentrated in the Earth's center. $\|g\| = \frac{GM}{r^2}$, this is the magnitude of the gravitational acceleration at radius r away from the Earth's center. The Earth's gravitational constant is $GM = 3986004.418 \cdot 10^8$m^3/s^2 (Table 6.1). At the equator the distance to the Earth's center equals

$a = 6378137.0$ m (Table 6.1, semi-major axis of WGS84 ellipsoid), and at a pole $b = 6356752.314$ m, see question 3 in Section 4.7. Hence the acceleration at the equator (with $r = a$) is $||g|| = 9.798\text{m/s}^2$, and at a pole (with $r = b$) is $||g|| = 9.864\text{m/s}^2$.

Question 2 As a follow-up on question 1, compute the (magnitude of the) centrifugal acceleration at the equator.
Answer 2 The magnitude of the centrifugal acceleration follows from the velocity and the radius: $a = \frac{v^2}{r}$ (uniform circular motion). Hence, we need the velocity. The Earth makes a full turn in (one solar day) of $T = 23\text{h}56\text{m} = 86160$ seconds. At the equator the circumference is $2\pi a$ (with the radius set equal to the length of the semi-major axis a, not to be confused with the symbol for acceleration which is used later), and hence velocity v is $v = 2\pi a/T = 465.1\text{m/s}$. The acceleration becomes $a = 0.034\text{m/s}^2$. The centrifugal acceleration is pointing outward. The acceleration due to the attraction by the Earth's mass is pointing inward to the center of the Earth. At a pole, the centrifugal acceleration is zero.

Question 3 Suppose again that the Earth is a perfect ellipsoid, and that all mass is concentrated in the Earth's center. Would water flow from the equator to the poles, in case the Earth would be not rotating?
Answer 3 Water flow is dictated by potential. The Earth is not rotating, hence we need to consider only the gravitational potential, due to the attraction by the Earth's mass. The equation for potential is simply $W = \frac{GM}{r}$ (7.5), at a location r away from the Earth's center. At the equator we have $r = a$ (the length of the semi-major axis of the ellipsoid), and at a pole $r = b$ (the length of the semi-minor axis). As $b < a$, we have $W_{\text{pole}} > W_{\text{equator}}$. The potential is zero at $r = \infty$, and the potential is larger at the pole (than at the equator), hence in this case water would flow from the equator to the poles.

8

Vertical reference systems

Until now the focus has been on the geometry 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.

8.1. Ellipsoidal heights

The elevation of a point can only be expressed with respect to 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.

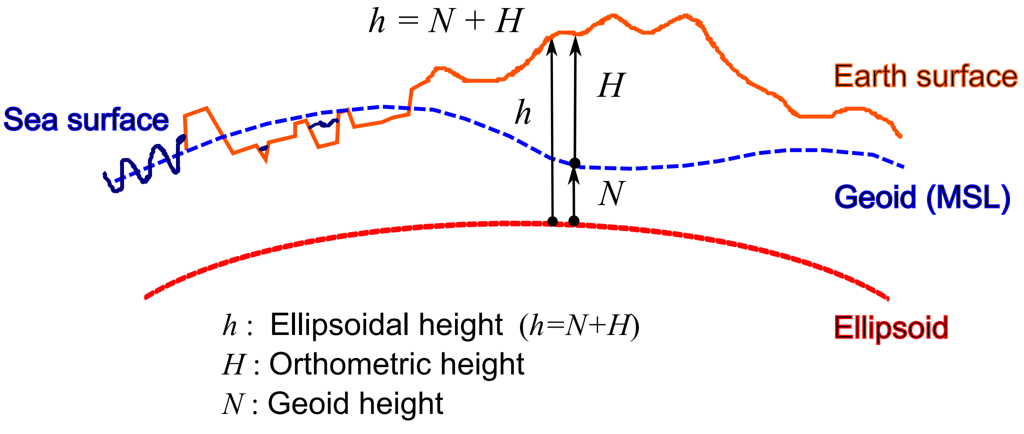


Figure 8.1: Relation between ellipsoidal height h, orthometric height H and geoid height N.

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 and hydraulic engineering a necessary requirement is that water always flows from a point with a higher height to points with lower height.

8.2. Orthometric and normal heights

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 a discussion of gravity and potential numbers the reader is referred to Chapter 7. Here it suffices to recapitulate from Chapter 7 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^2]. From Eq. (7.3) and (7.5) follows that the potential difference ΔW and height difference ΔH are related,

$$\Delta W = -g\ \Delta H \tag{8.1}$$

with g the gravitational acceleration in [m/s^2]. The gravitational acceleration is a positive number: the minus sign in Eq. (8.1) is because the gravitational potential W decreases with increasing height, see the right of Figure 7.1, and therefore ΔW and ΔH have opposite sign. Note that the gravitational acceleration g is not constant but depends on the location on Earth and height. *Orthometric heights* are defined by the inverse of Eq. (8.1),

$$H_{\text{orthometric}} = -\frac{1}{g}(W - W_0) \ . \tag{8.2}$$

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,

$$H_{\text{normal}} = -\frac{1}{\gamma}(W - W_0) \tag{8.3}$$

with γ the *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

$$h = N + H \tag{8.4}$$

with N the height of the geoid above the ellipsoid. This is illustrated in Figure 8.1. 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.

Instead of the word 'height', the word 'elevation' is often used in the US, to refer to the height of a point on the Earth's surface above a geoid.

8.3. Height datums

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 (MSL). In Figure 8.2 the reference tide-gauges used for different European countries are shown,

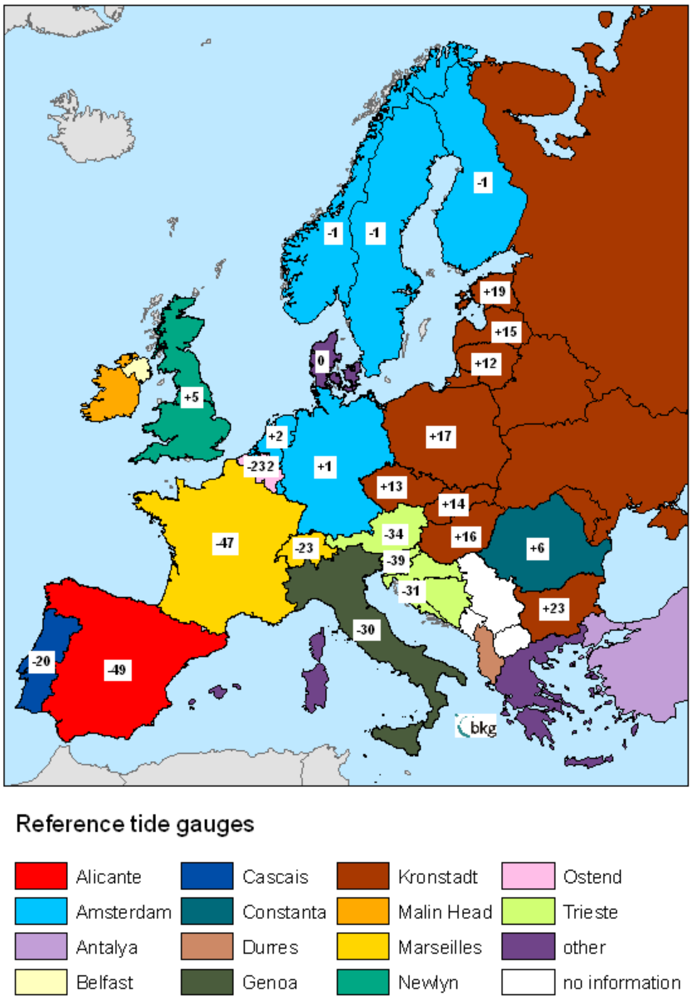


Figure 8.2: Differences between national height datums for Europe and reference tide-gauges in centimeters (Source: BKG http://www.bkg.bund.de).

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.

Lowest Astronomical Tide (LAT) is the lowest predicted tide level that can occur under any combination of astronomical conditions assuming average meteorological conditions. The advantage for Hydrographic chart datums is that all predicted tidal heights must then be positive, although in practice lower tides may occur due to e.g. meteorological effects. In the Netherlands, UK and many other countries charted depths and drying heights on nautical charts are given relative to LAT, and tide tables give the height of the tide above LAT. The depth of water, at a given point and at a given time, is then calculated by adding the charted depth to the height of the tide, or by subtracting the drying height from the height of the tide, with all heights and depths given with respect to LAT.

8.4. Problems and exercises

Question 1 Modeling the Earth as a sphere with radius equal to the semi-major axis of the WGS84-ellipsoid (see Table 6.1), and assuming that all mass is concentrated in the Earth's center, compute the gravitational acceleration.

Answer 1 The semi-major axis of the WGS84-ellipsoid is $a = 6378137$ m. The Earth's gravitational constant is $GM = 3986004.418 \cdot 10^8$ m^3/s^2. The magnitude of the gravitional acceleration at radius r away from the Earth's center is simply $a = GM/r^2$, see Chapter 7, Eq. (7.2). This yields $a = 9.798$ m/s^2. The acceleration vector points downwards to the Earth's center.

Question 2 The ellipsoidal height of a geodetic marker on Terschelling is 56.098 m. The ellipsoidal height of a geodetic marker in Eijsden, near Maastricht is 103.797 m. These coordinates are given in ETRS89 (and hence based on the WGS84-ellipsoid). The geoid-height-difference between these two locations is 4.601 m (that is, the geoid height near Maastricht is larger than in Terschelling; NLGEO2004 geoid with respect to the GRS80/WGS84 ellipsoid). Compute the orthometric (level) height difference between Terschelling and Eijsden.

Answer 2 The relation between ellipsoidal height h and orthometric (levelled) height H is $h = H + N$, with N the geoid height. This relation can also be exploited in a heigh-difference $h_{TE} = H_{TE} + N_{TE}$, with $h_{TE} = h_E - h_T$, with T for Terschelling, and E for Eijsden. The ellipsoidal height difference between Terschelling and Eijsden is $h_{TE} = h_E - h_T = 47.699$ m. The geoid-height difference was given as $N_{TE} = 4.601$ m. Hence, $H_{TE} = 43.098$ m. Hence, the levelled height difference is about 4.5 m smaller than the ellipsoidal height difference. With the levelled height of Terschelling being $H_T = 14.695$ m, the levelled height of Eijsden becomes $H_E = 57.793$ m.

9

International reference systems and frames

In this chapter a number of common international reference systems and frames is discussed. We start with the well know world wide WGS84 system used by GPS, but quickly shift forcus to the more important International Terrestrial Reference System (ITRS), which is realized through the International Terrestrial Reference Frames (ITRF). Then the focus is shifted to regional reference systems and frames, with the European Terrestrial Reference System ETRS89 as our prime example.

9.1. World Geodetic System 1984 (WGS84)

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

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 satellite orbits as civilian users have no direct access to tracking sites or tracking data from the US military. 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, discussed in the next sections.

In fact there are different WGS84 realizations. Until GPS week G730[1], WGS84 was based on the US Navy Doppler Transit Satellite System. Newer realizations of WGS84 are coincident with International Terrestrial Refernce Frame (ITRF), see Section 9.2, at about the 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).

In general WGS84 is identical to the ITRS and its realizations ITRFyy at the 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) are sometimes simply referred to as "WGS84".

[1]The GPS week number is the number of weeks counted since January 6, 1980.

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. For surveying and geoscience applications the more accurate ITRF, or ETRS89 in Europe, should be used.

9.2. 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 (International Union of Geodesy and Geophysics) resolution No 2 adopted in Vienna, 1991. As a consequence, the ITRS is

- a geocentric co-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,
- 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). Realizing a global terestrial reference system is not trivial as the Earth is not a rigid body. Even the outer layer, the Earth's crust, is flexible and changes under the influence of solid Earth tides, loading by the oceans and atmosphere, and tectonics. From a global perpective points are not stationary, but moving. Therefore each individual ITRF contains station positions and velocities, often together with full variance matrices, computed using observations from space geodesy techniques[2]. The stations are located on sites covering every continent and tectonic plate on Earth. To date there are thirteen realizations of the ITRS [3]: ITRF2008 and ITRF2014 are the latest two realizations. ITRF2000 will be the next realization, using more recent data, reprocessing of old data, improved models and processing softwares.

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 $\mathbf{r}(t_0)$ and velocities $\dot{\mathbf{r}}$ due to the availability of a longer time span of observations, which is in particular important for the velocities,

[2]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).

[3]ITRF88, ITRF89, ITRF90, ITRF91, ITRF92, ITRF93, ITRF94, ITRF96, ITRF97, ITRF2000, ITRF2005, ITRF2008 and ITRF2014. 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.

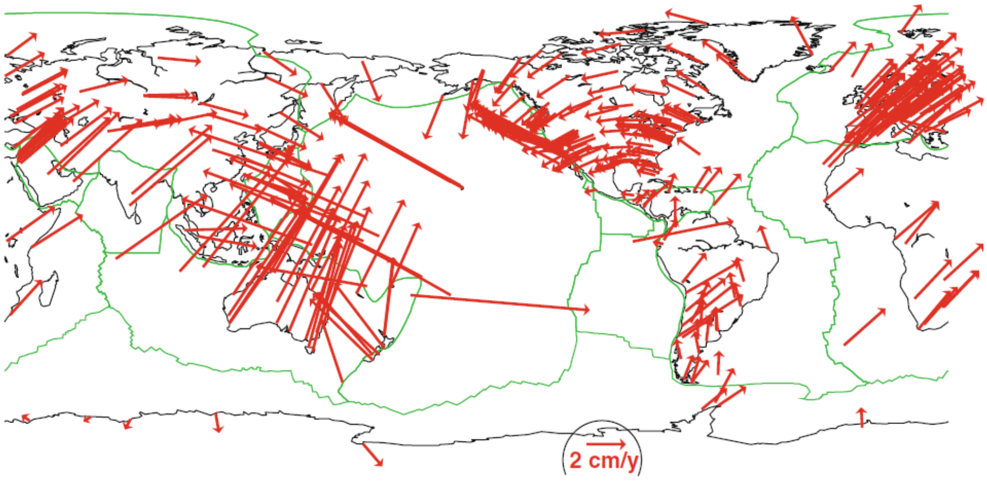


Figure 9.1: ITRF2008 velocity field with major plate boundaries shown in green (Figure from http://itrf.ensg.ign.fr/).

- 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 t_0.

All ITRF model the (secular) changes in the Earth's crust. The position $\mathbf{r}(t)$ at a specific epoch t is given by

$$\mathbf{r}(t) = \mathbf{r}(t_0) + \dot{\mathbf{r}} \cdot (t - t_0) \tag{9.1}$$

The ITRF2008 velocities are given in Figure 9.1. The velocities, in the vector $\dot{\mathbf{r}}$, 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 with sub–daily periods, can be computed using models specified in the IERS conventions, chapter 7 (http://www.iers.org/IERS/EN/DataProducts/Conventions/conventions.html).

In Figure 9.2 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 9.2, 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 9.3 shows the same time series, but in the European ETRF2000 reference frame, which is discussed in Section 9.3.

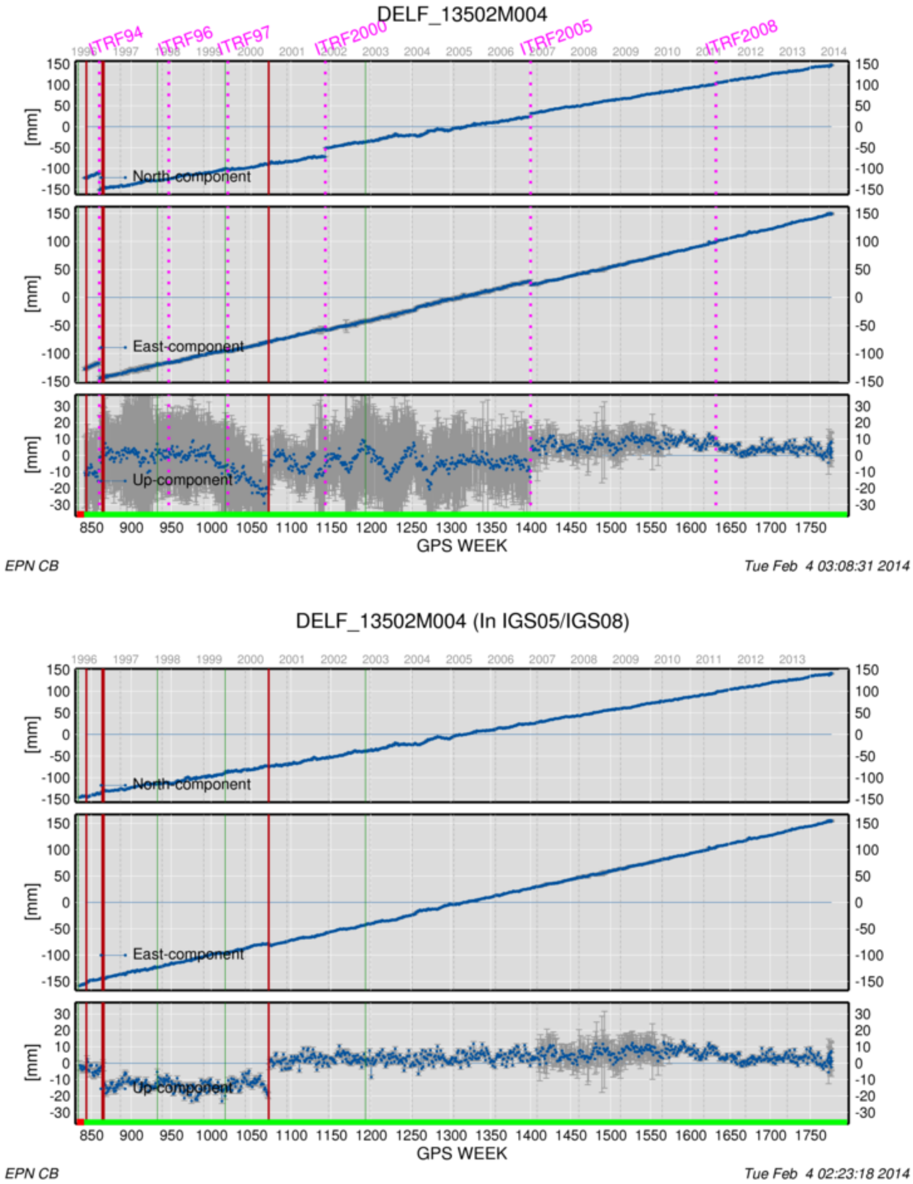


Figure 9.2: Time series of station positions of a permanent GPS receiver in Delft from 1996-2014. The top figure shows the time series in the ITRFyy reference frame that was current at the time the data was collected. The bottom figure shows the data after re-processing in the IGS05/IGS08 reference frame 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 [4]. The ITRF origin and rates are

[4]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.

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 published on ITRS website http://itrf.ensg.ign.fr/doc_ITRF/Transfo-ITRF2014_ITRFs.txt. The transformation formula is essentially Eq. (3.13), except for a different sign of the transformation parameters.

Web-based Precise Point Positioning (PPP) services, which utilize satellite orbits and clocks from the International GNSS Service (IGS), 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, see Table 6.1. This is the same ellipsoid as used for instance by WGS84.

9.3. 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[5]. 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 a 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 9.1 and 9.2.

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[6]. ETRS89 is accessed through the EUREF Permanent GNSS Network (EPN) [7], 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.

[5] 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).

[6] Sometimes people think the coordinates should be given at epoch 1989.0, but this is not necessary, as coordinates can be given at any epoch. The best practice is to give the coordinates at the epoch of observation.

[7] 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.

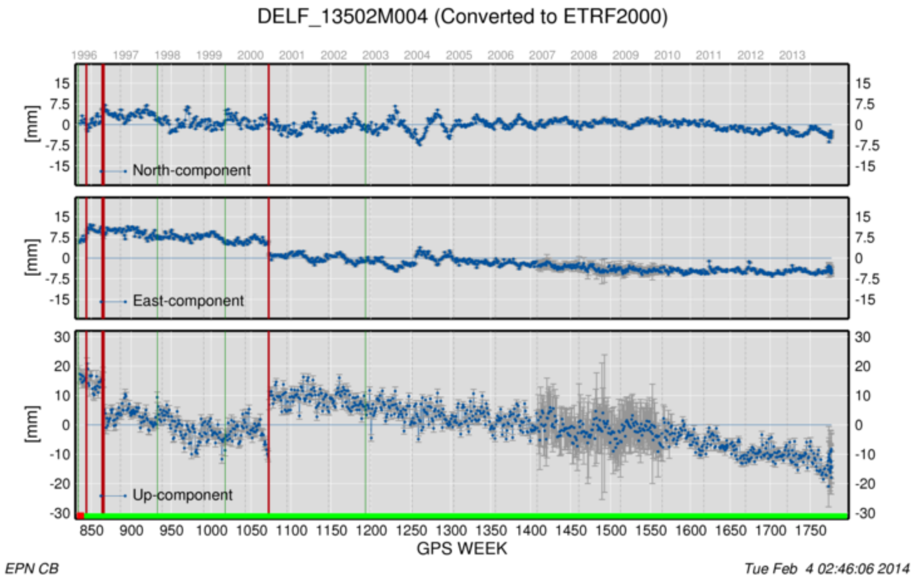


Figure 9.3: Time series of station positions of a GPS receiver in Delft from 1996-2014 in the European reference frame ETRF2000. The horizontal station velocity In ETRS89 is at the few mm level. The vertical red lines indicate equipment changes which cause jumps of a few mm in the time series (Figure from `http://www.epncb.oma.be/`).

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 9.3 for the GPS station in Delft that was also used for Figure 9.2. 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 Information System (GIS) applications. Also, ETRS89 is well suited for the exchange of geographic data sets between European national and international institutions and companies. On other continents solutions similar to ETRS89 have been adopted.

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 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. Each realization has a new set of improved positions and velocities. The three most recent realizations of ETRS89 are ETRF2000, ETRF2005 and ETRF2014[8]. Since each realization also reflects improvements in the datum definition of ITRF, which results in small jumps in the coordinate time series, the EUREF Governing Board (formally Technical Working Group) recommends not to use the ETRF2005 for practical applications, and instead to adopt ETRF2000 as a conventional frame of the ETRS89 system. However, considering the diverse needs of individual countries', it is the countries' decision to adopt their preferred ETRS89 realization. Most countries adopted the recommended ETRF2000, but not every European country has, and considering the improved accuracy and stability of the ITRF2014, some could switch to ETRF2014.

Another way to realize ETRS89 is by using GNSS campaign measurements or a network

[8]There is no ETRF2008.

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, as well as several commercial providers, operate real-time network RTK services (NETPOS, 06-GPS, LNRNET, and others) that provide GPS data and corrections in real-time which allows instantaneous GPS positioning in ETRS89 at the centimeter level. The 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. However, EuroGeographics does recommend to use one of three selected projection: *Lambert Azimuthal Equal Area* (ETRS89/LAEA, EPSG:3035) for statistical mapping at all scales and other purposes where true area representation is required, *Lambert Conformal Conic 2SP* (ETRS89/LCC, EPSG:3034) for conformal mapping at 1:500,000 scale or smaller, or *Universal Transverse Mercator* (UTM) for conformal mapping at scales larger than 1:500,000. In several countries, including the Netherlands, a conventional transformation from ETRS89 to grid (map) coordinates that resemble the old national systems is provided, see Chapter 10. 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/pub/EUREF-TN-1.pdf`.

9.4. Problems and exercises

Question 1 The position coordinates of a geodetic marker in Westerbork are given in the ITRF2008 at epoch 2015.0 as $X = 3828735.710$ m, $Y = 443305.117$ m, $Z = 5064884.808$ m, with velocities $V_X = -0.0153$ m/y, $V_Y = 0.0160$ m/y, $V_Z = 0.0096$ m/y. Compute the position coordinates of this marker, in ITRF2008, for January 1st, 2016.

Answer 1 The position coordinates are given in the International Terrestrial Reference Frame 2008, a realization of the ITRS. Generally positions are subject to small movements within a global reference system (due to Earth's dynamics). In this question the coordinates are given for January 1st, 2015. And also the velocities are given (in meter per year) to compute the position at any other instant in time. We can propagate the position over 1 year to January 1st, 2016. The resulting coordinates become: $X = 3828735.695$ m, $Y = 443305.133$, $Z = 5064884.818$.

10

Dutch national reference systems

In this chapter the Dutch triangulation system RD and height system NAP, and their relation to the ETRS89, are presented. The focus in this chapter is on the the Netherlands, with the remark, that many other countries have undergone similar developments and adopted similar approaches.

10.1. 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.

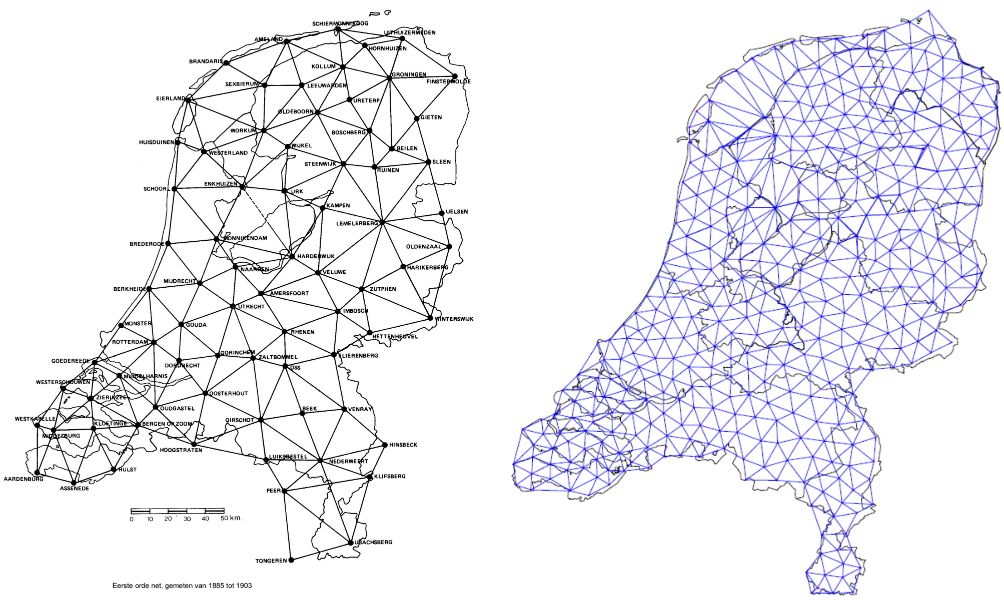


Figure 10.1: First order triangulation network for the Netherlands of 1903 (left) and GPS base network of 1997 (right). Figure from de Bruijne et al., 2005.

10.1.1. RD1918

The first-order triangulation grid was measured in the years between 1885 and 1904 (Figure 10.1). The church tower of Amerfoort was selected as the 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 10.2.

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° 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.

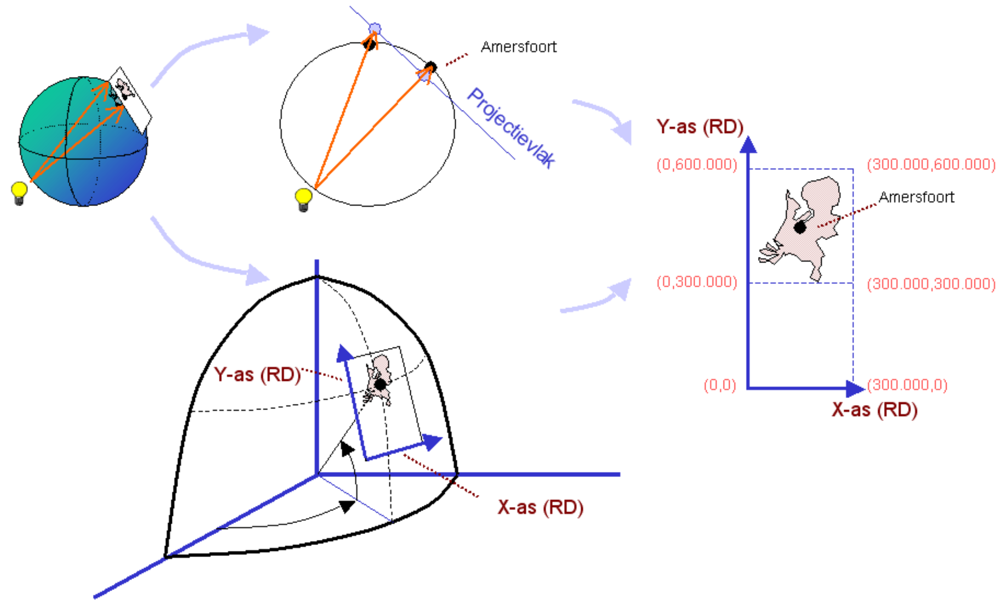


Figure 10.2: RD double projection (Bessel ellipsoid → Sphere → Plane) and definition of RD coordinates (Figure T.Nijeholt at nl.wikibooks).

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 the x-coordinates and y-coordinates, and to obtain always positive coordinates, the origin of the coordinates was shifted 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. It also avoids confusion between the old and new 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 10.1. 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 10.3 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.

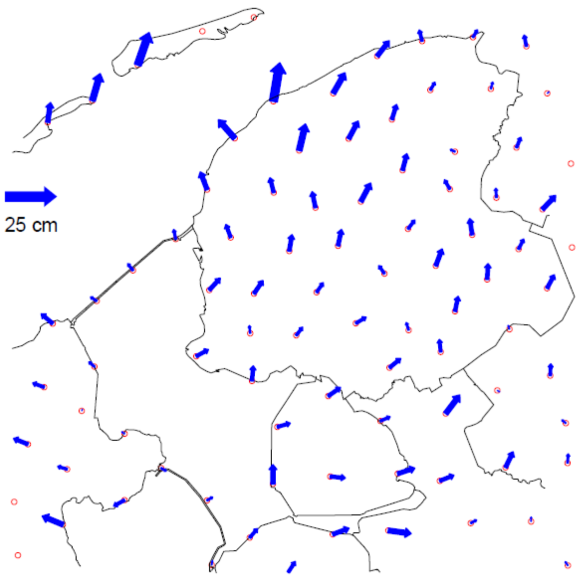


Figure 10.3: Differences between RD and ETRS89 based coordinates for the GPS Kernnet in the province of Friesland, showing significant regional correlation between the vectors (Figure from de Bruijne et al., 2005).

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.

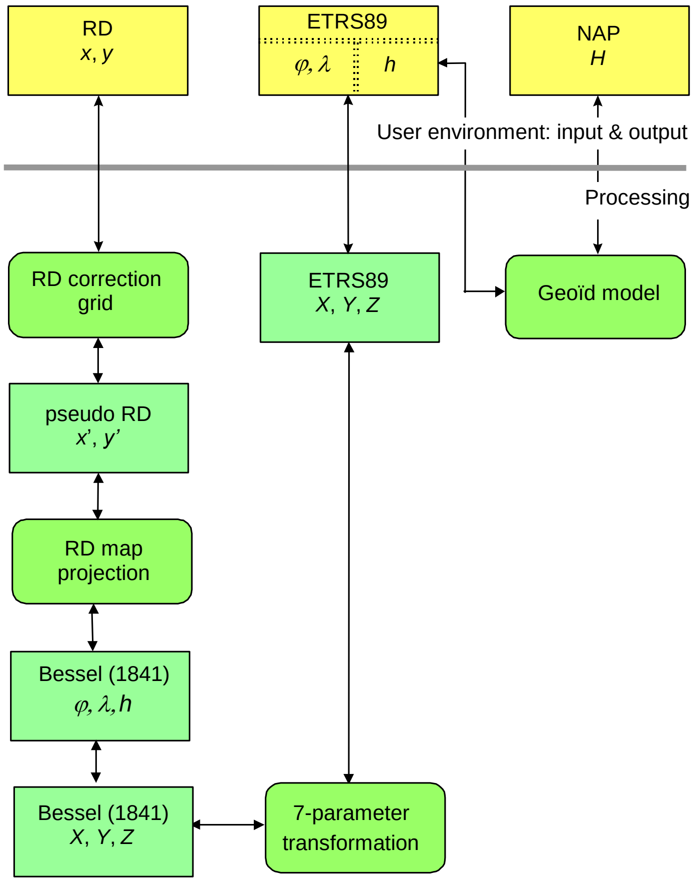


Figure 10.4: RDNAPTRANS transformation procedure until RDNAPTRANS™2018. The figure outlines the relationships and transformations between ETRS89, RD2000 and NAP (Figure after de Bruijne et al., 2005). The coordinates below the line, with the exception of Cartesian coordinates in ETRS89, are used only for computational purposes and should never be published or distributed to other users. In the new RDNAPTRANS™2018 procedure a correction grid is applied to the ETRS89 latitude and longitude instead, and pseudo RD coordinates become the final RD coordinates. In the new RDNAPTRANS™2018 there is also a version in which the 7-parameter transformation is included in the correction grid, see Figure 10.5.

10.1.2. 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 to be used in practice: (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 10.4 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, including 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. quasi–geoid for the conversion between NAP heights and height on the GRS80 ellipsoid, which will be discussed next in Section 10.2.

The transformation procedure works in both directions, and, both for 2D and 3D coordinates. In case no heights are available the Kadaster recommends to use an approximate height, e.g. by using a digital terrain model, or, when that is not possible, to use $h = 43$ m (which is close to $NAP = 0$) so that one gets the same result after transforming back and forth. 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 in 2004 and 2008, and one major revision in 2019. The minor revisions were related to changes in the European reference frame, which affected the 7-parameter transformation, and the introduction of an improved NLGEO2004 geoid in 2004. The original 2000 version used the 'De Min' geoid and older transformation parameters. The modified transformation procedures are referred to as RDNAPTRANS™2004 and RDNAPTRANS™2008. The original transformation of 2000 is since then also referred to as RDNAPTRANS™2000. In 2018 work started on a major revision of RDNAPTRANS, resulting in RDNAPTRANS™2018, which was published in 2019. In RDNAPTRANS™2018 the correction grid is aplied to the latitude and longitude coordinates instead of the pseudo RD coordinates, the NLGEO2018 quasi–geoid is used instead of NLGEO2004 geoid, and the interpolation and correction grids are based on international standards.

More 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.

10.1.3. RDNAPTRANS™2018

Although the RDNAPTRANS transformation procedure is well documented and example source code in C and Matlab is available free of charge, the transformation procedure was only supported by a few Geographic Information System (GIS) packages. The correction grid that was used in older versions of RDNAPTRANS was often not directly supported by software and the map projection chosen for RD was 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. At the heart of the discussion is that many users find it difficult to work directly with geographic coordinates

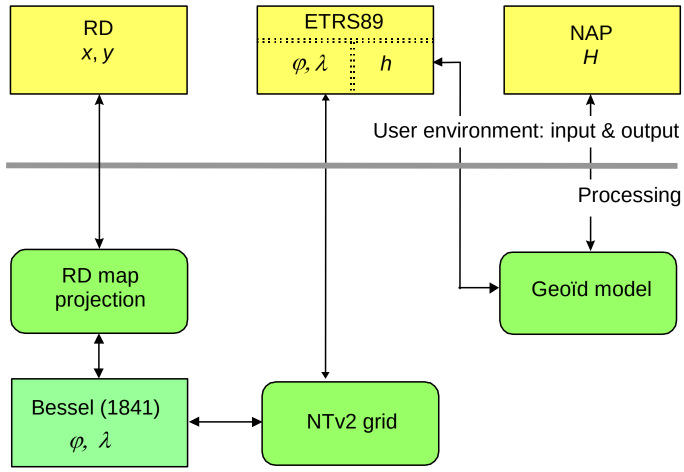


Figure 10.5: NTv2 transformation procedure used by RDNAPTRANS™2018. The figure outlines the relationships and transformations between ETRS89, RD2000 and NAP using the proposed NTv2 procedure, in variant 2 of RDNAPTRANS™2018 where the datum transformation is included in the correction grid. The coordinates below the line are used only for computational purposes and should never be published or distributed to other users.

(latitude and longitude) and prefer working with rectangular 2D grid coordinates, but lack the expertise and software to do the conversion to RD.

As the result of this discussion the RDNAPTRANS procedure has been modified and is based on the Canadian NTv2 correction procedure (National Transformation version 2) that is better supported by existing softwares. As shown in Figure 10.5, the NTv2 procedure employs a correction grid to convert latitude and longitude in ETRS89 directly to latitude and longitude on the Dutch Bessel 1841 ellipsoid, which are the input for the RD map projection. The procedure shown in Figure 10.5 includes the datum transformation into the correction grid. There is also a variant whereby the datum transformation is still implemented as a separate step. Other technical changes to the RDNAPTRANS procedure were the introduction of an easier to use and more standard bi–linear interpolation method and extension of the domain over which the procedure is valid. These technical changes, apart from the introduction of the new and improved NLGEO2018 quasi–geoid and improved transformation parameters, were carried out in such an way as to maintain consistency at the milimeter level with previous RDNAPTRANS versions. The new NLGEO2018 quasi–geoid represents a real improvement for the height, but even so, consistency with the previous RDNAPTRANS for the heights is still at the centimeter level.

The new version of the RDNAPTRANS procedure, called RDNAPTRANS™2018, is much easier to implement than the RDNAPTRANS procedure of Figure 10.4. Also, the NTv2 procedure is a (relatively new) standard that is now supported by many coordinate transformation and GIS software packages, including the PROJ generic coordinate transformation software. This makes it possible to fully implement RDNAPTRANS™2018 in the PROJ transformation software, that is also used by many GIS softwares. Although RDNAPTRANS™2018 can be implemented in the PROJ transformation software, using a daisy chain of smaller transformation steps, each with its own `+proj` string (see Section 6.4), it doesn't have an EPSG code yet. EPSG code EPSG:28992 only implements an approximate transformation, which is good enough for visualizations, but should *never* be used to facilitate the exchange of coordinates.

Latitude and longitude in the ETRS89 reference frame is the default for the exchange of geo-information in *Europe*, but in the Netherlands, users have the choice between latitude and longitude in the ETRS89 reference frame and RD coordinates. However, for the transformation between ETRS89 and RD, only software that support the official RDNAPTRANS should be used.

10.2. 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).

10.2.1. Precise first order levellings

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.

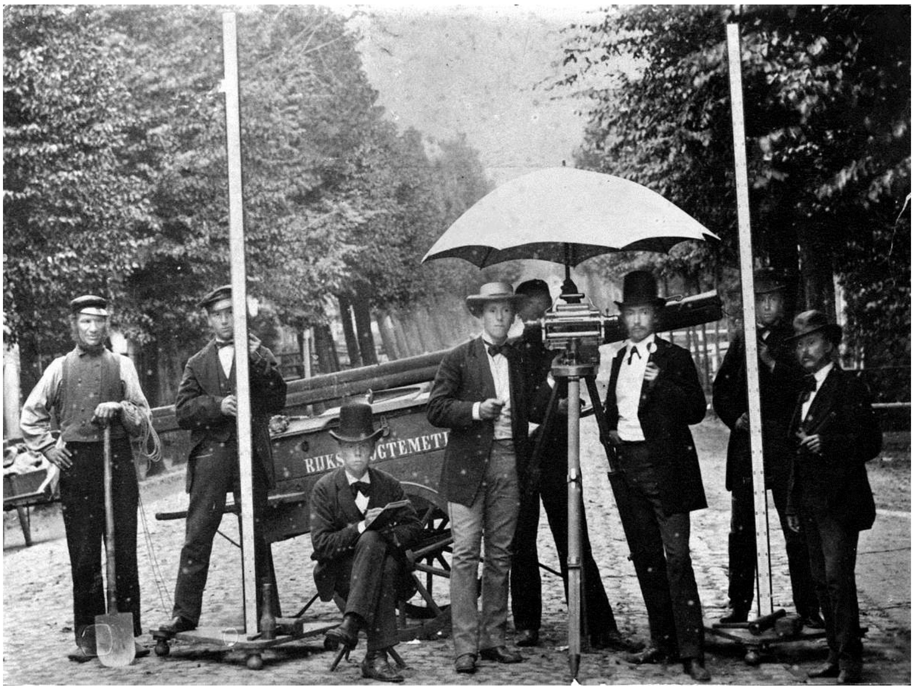


Figure 10.6: Team of surveyors posing for the camera during the first precise levelling, in Dutch *Eerste Nauwkeurigheids-waterpassing*, or *Rijks-Hoogtemeting*, probably in 1875 or 1876. The person with the white hat is Cornelis Lely (1854-1929), who just graduated as civil engineer in Delft (1875). Later, as minister of infrastructure, he introduced the bill that resulted in the *Zuiderzee werken*, which comprised the construction of a 30 km dike forming the Ijselmeer lake, and the creation of the two western polders in the former Zuiderzee. Figure from H.W. Lintsen (red.), Geschiedenis van de techniek in Nederland, De wording van een moderne samenleving 1800-1890. Deel VI. Techniek en samenleving. Walburg Pers, Zutphen 1995, http://www.dbnl.org/tekst/lint011gesc06_01/lint011gesc06_01_0012.php).

A series of five first order levelling campaigns has been carried out to date. The 1st national precise levelling dates from the period 1875-1885, including 410 already existing points and 2100 km of continuous levelling lines. See also Figure 10.6. 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. During later periods, until the 1980's, three more precise first order levellings were carried out. It saw the installation of new underground reference points in - presumably - stable geological strata throughout the Netherland, including several posts (nulpalen) in the vicinity of "tide" gauges (water level gauges), the introduction of hydrostatic levelling, and new routes, e.g. over the Afsluitdijk. On the other hand, many existing points were lost, including all stone datum points in the Amsterdam locks. During the 3rd levelling 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.

Figure 10.7: Levelling lines of the 5th precise levelling in the Netherlands, 1996-1999 (Figure from de Bruijne et al., 2005).

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 10.7. 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.

In 1998 a NAP monument was created at the Amsterdam Stopera. This monument, designed and created by Louis van Gasteren and Kees van der Veer, consists of a NAP pillar rising through the building with on top a bronze bolt a precisely the zero NAP level, two water columns showing the current tide levels at Ijmuiden and Vlissingen, and a third water column showing the water level at the time of the 1953 Zeeland flood disaster. See the front cover.

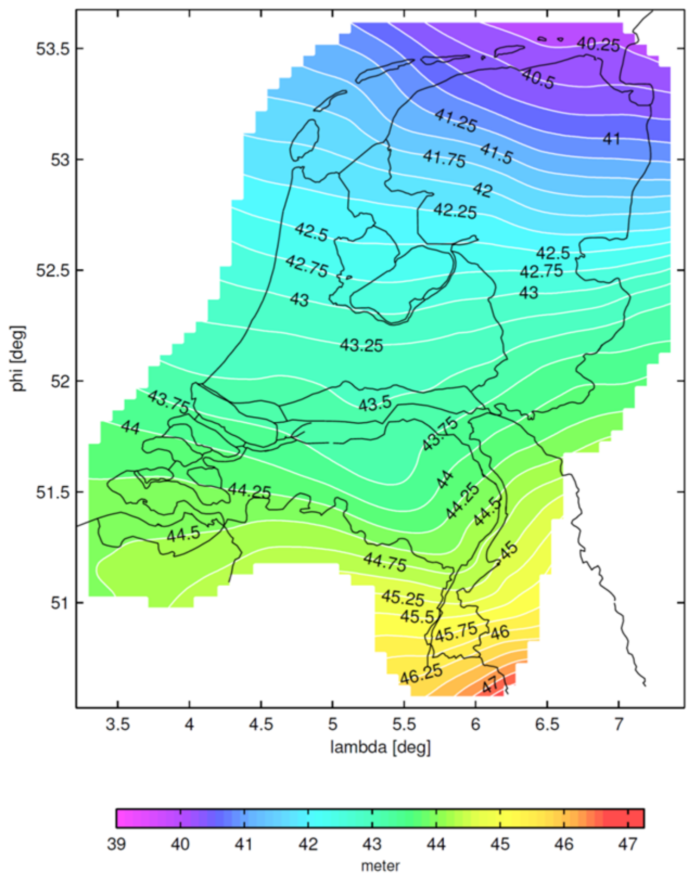


Figure 10.8: NLGEO2004 geoid for the Netherlands, with geoid height in [m] with respect to the GRS80 ellipsoid in ETRS89 (Figure from de Bruijne et al., 2005).

10.2.2. NAP Benchmarks

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 (https://www.rijkswaterstaat.nl/zakelijk/open-data/normaal-amsterdams-peil). These markers serve as the basis for height determination by consulting engineers, water boards, municipalities, provinces, state, and other authorities, 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 8 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.

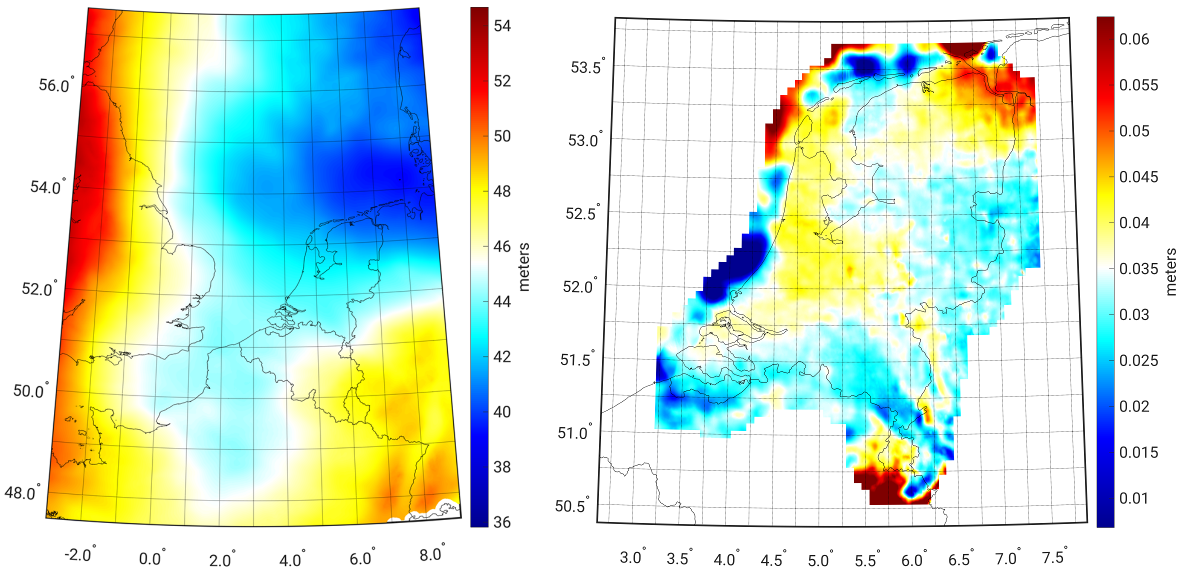


Figure 10.9: NLGEO2018 quasi–geoid for the Netherlands on the left, with geoid height in [m] with respect to the GRS80 ellipsoid in ETRS89, with on the right the differences with NLGEO2004. Clearly visible the much larger domain over which the NLGEO2018 quasi–geoid is computed (Figure courtesy Cornelis Slobbe).

10.3. Geoid models – NLGEO2004 and NLGEO2018

Although it is unlikely that GPS will replace levelling alltogether, GPS can be used to obtain heights with an accuracy of about $1 - 2$ cm using the RDNAPTRANS procedure, as outlined in Figure 10.4. 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, called the 'De Min' geoid, became available in 1996 and had a precision of one to a few centimeters. This was the first accurate geoid model of the Netherlands and was used by the original RDNAPTRANS procedure.

The geoid model was improved in 2004, resulting in the NLGEO2004 model, that is used by RDNAPTRANS™2004 and RDNAPTRANS™2008, see Figure 10.8. 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 define a correction surface to the gravimetric geoid. 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 km 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.

In 2018 a new (quasi–)geoid, called NLGEO2018, was computed. Contrary to NLGEO2004, it is based on a least-squares approach using a parametrization of spherical radial basis functions. This approach allowed to account for systematic errors in the gravity datasets, enables proper error propagation, and the computation of the full variance-covariance matrix of the resulting quasi-geoid model. The model itself was computed over a much larger domain than NLGEO2004 (it now includes the Dutch Exclusive Zone (EEZ) in the North–Sea) and based on re-processed datasets. Also new datasets have been used, including datasets in Limburg, Belgium, Germany and shipboard and airborne gravimetry data over the North Sea. Moreover, along-track geometric height anomaly differences from various satellite radar altimeters were used. Since the data area was much larger than before it became necessary to apply so called terrain corrections, which aim to remove the high-frequency signals in the data. Another improvement is that the remove-compute-restore procedure relied on a satellite-only geopotential model obtained from GRACE and GOCE data. Over the land area of the Netherlands, the precision of the NLGEO2018 gravimetric quasi–geoid is 0.7 cm standard deviation. After application of the innovation function (which aims to reduce the differences between the quasi-geoid and height reference surface) the standard deviation reduces to 0.5 cm. For the NLGEO2004 gravimetric geoid, the precision was 1.3 cm. After application of the so-called correction surface this number was 0.7 cm.

Figure 10.9 shows the NLGEO2018 quasi–geoid and the differences with the NLGEO2004 geoid. Differences are in the range of 1–6 cm, with a systematic difference of about 3.5 cm. These differences are to be expected because the innovation function and corrector surface are based on different GNSS and levelling datasets and the permanent tide is handled differently. Also, when the NLGEO2018 is used in the RDNAPTRANS™2018 procedure part of the differences will be resolved in the transformation parameters. Therefore, differences in the height resulting from RDNAPTRANS versions 2004 and 2018 are much smaller, with the maximum height difference of about 2.5 cm.

10.4. Lowest Astronomical Tide (LAT) model – NLLAT2018

The vertical datum for nautical maps in the Netherlands, and other countries around the North–Sea, is Lowest Astronomical Tide (LAT) . Lowest Astronomical Tide (LAT) is the lowest predicted tide level that can occur under any combination of astronomical conditions assuming average meteorological conditions, which implies that all predicted tidal heights must be positive (although in practice lower tides may occur due to e.g. meteorological effects). Tide

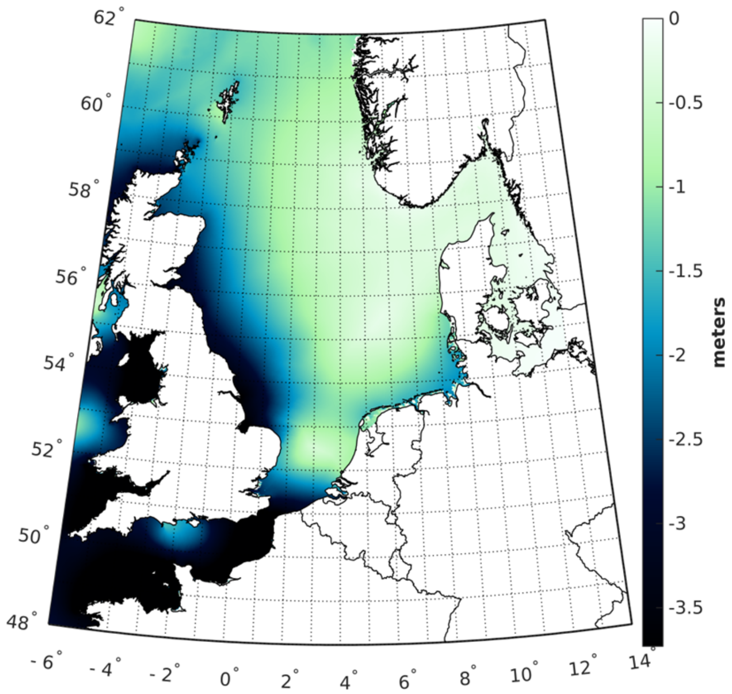


Figure 10.10: NLLAT2018 with respect to the NLGEO2018 quasi–geoid (Figure courtesy Cornelis Slobbe).

tables, as well as charted depths and drying heights on nautical charts, are given relative to LAT. The depth of water, at a given point and at a given time, is then calculated by adding the charted depth to the height of the tide, or by subtracting the drying height from the height of the tide, with all heights and depths given with respect to LAT.

The Dutch LAT model is called NLLAT2018. The LAT surface is always below the Dutch NLGEO2018 quasi–geoid, but the separation between the two is not constant and depends on the location. NLLAT2018 has been computed using hydrological and meteorological models, tidal water levels from 31 tide–gauge, and the NLGEO2018 quasi–geoid. The separation between NLLAT2018 and NLGEO2018 is shown in Figure 10.10. The LAT reference surface in NLLAT2018 itself is given with respect to the GRS–80 ellipsoid. The accuracy of the LAT reference surface is about 1 decimeter.

11

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`.

PROJ contributor. PROJ coordinate transformation software library. Open Source Geospatial Foundation, 2020, `https://proj.org/`.

A

Quantity, dimension, unit

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 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 $360°$, 1 minute of arc is $1/60$ of a degree, and 1 second of arc is $1/60$ of a minute or $1/3600$ of a degree.

designation	symbol	power
Exa	E	10^{18}
Peta	P	10^{15}
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}
femto	n	10^{-15}
atto	a	10^{-18}

Table A.2: Most common powers of ten and their designation, also known as SI-prefixes.

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. 'ppm' stands for parts per million and implies a 10^{-6} effect, and 'ppb' stands for parts per billion, a 10^{-9} effect.

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].

Colophon

This reader has been produced initially for the 3rd year BSc course CTB3310 *Surveying and Mapping* in the Civil Engineering curiculum at the Delft University of Technology, and later adapted to be used also for several other courses in the bachelor and master programs.

Title: Reference Systems for Surveying and Mapping

Author: Hans van der Marel

Contributors: Hans van der Marel, Christian Tiberius (Section 2.4 2D examples, Chapter 7 gravity, Appendix A units and the exercises at the end of each chapter)

Layout: Hans van der Marel

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
1.2a	28 Aug 2014	Interim release for CIE4606 Geodesy and Remote Sensing course
2.1	7 Feb 2016	Release for 2016 CTB3310 course
2.1a	22 Feb 2016	Updated release for 2016 CTB3310 course
2.1b	20 Apr 2016	Updated release for CTB3425, CIE4606 and CIE4614 courses.
3.1	15 Feb 2020	Third main release. Changes: - Section 2.4 2D examples (Courtesy Christian Tiberius) - Section 3.3 3D similarity transformation revised - Section 4.4 Topocentric coordinates revised - Section 5.4 Map projection examples (Initial version Christian Tiberius) - Chapter 10 RDNAPTRANS updated - Chapter 10 NLGEO2018 and NLLAT2018 (Courtesy Cornelis Slobbe) - Many minor improvements by the author, and from Christian Tiberius and Jochem Lesparre - Licensed under CC BY-NC-SA

File name: `TRS_3-1_xelatex.tex`