# Surveying and Mapping

Christian Tiberius, Hans van der Marel, René Reudink & Freek van Leijen

TU Delft OPEN

# Surveying and Mapping

Christian Tiberius - Hans van der Marel - René Reudink - Freek van Leijen

Delft University of Technology

The front cover shows a 'true color' remote sensing image of the South-Western part of the Netherlands. Data from ESA Sentinel-2 satellite (Copernicus Sentinel data 2019) obtained through Sentinel-Hub [1], CC BY-NC 4.0, image taken on August 24th, 2019 (10:56:43 UTC), see Figure 25.5.

Surveying and Mapping
Christian Tiberius, Hans van der Marel, René Reudink, Freek van Leijen
November 2021 - November 2022

Publisher:
TU Delft OPEN TU Delft Open Textbook
Delft University of Technology — The Netherlands

keywords:
land surveying, mathematical geodesy, GPS, remote sensing, reference system, mapping

ISBN (softback/paperback): 978-94-6366-490-5
ISBN (e-book): 978-94-6366-489-9
DOI: https://doi.org/10.5074/T.2021.007




The text has been typeset using the MikTex 2.9 implementation of LaTeX. Graphs and statistical tables have been created with Matlab, and drawings and diagrams with a variety of programs, such as Adobe Illustrator, LibreOffice, Mayura Draw, Inkscape, and MS Powerpoint. Maps have been created with QGIS.

# Contents

# Preface

## Introduction

This textbook provides an introduction, at academic bachelor level, into the field of surveying and mapping. This book has grown from a series of readers and hand-outs which had been compiled for the third year bachelor course on Surveying and Mapping, in the program of Civil Engineering at Delft University of Technology, on the run, while developing and teaching this course, over the years, from academic year 2013-2014, when this course was offered for the first time, to date.

## Overview and structure

This textbook consists of six parts, which are, to a large extent, self-contained. Each part can be used separately, and there is no real need to go through the book from the beginning to the end. These six parts cover underlying subjects, such as land surveying, remote sensing and mapping, which together constitute the field of surveying and mapping. The order of the six parts in this book follows from the order in which the subjects are taught in the third year bachelor course. A simple overview of the different subjects and their relations is shown in Figure 1.

In this book we cover a wide range of measurement techniques to collect geospatial data. A selection of measurement techniques is illustrated in Figure 1. In Part I we cover classic land surveying techniques such as leveling and tachymetry, in Part III we cover GPS positioning, and in Part IV various remote sensing techniques are presented. Next, measurements are to be processed, indicated by the cloud with the formula for least-squares parameter estimation, addressed in Part II, in order to obtain, primarily geometric, information of interest, for instance position coordinates of topography on the Earth's surface. Topography refers to the forms and features of the Earth's surface, where features can be natural, like rivers and mountains, and man-made, like roads, railways, canals, bridges and buildings. The word has its roots in Greek, with *topos* (meaning 'place') and *graphia* (meaning 'writing'), hence meaning a description of a place in terms of its physical forms and features. The aforementioned positioning activities require linking in a coordinate reference system, shown at left, and covered in Part V. Eventually the goal is to store, archive and describe geospatial information in a Geographic Information System (GIS) and visualize this information in a map, a three-dimensional map or a Digital Terrain Model, available for further consultation and dissemination, as shown at bottom. This subject is covered in Part VI.

## Content

We start in Part I with covering classical land surveying techniques, specifically leveling and tachymetry. An introduction is given of these measurement techniques, as well as the underlying principles of the leveling instrument and the total station. The aim is to provide the reader with sufficient knowledge, insight and instructions in order to be able to acquire actual surveying skills in practice, be it at a basic level.

Mathematical geodesy, or the calculus of observations, covered in Part II, provides us with the 'data processing engine' for land surveying in particular, and geoscience in general. Basically it addresses the question how to go from measurements to information, and inter-

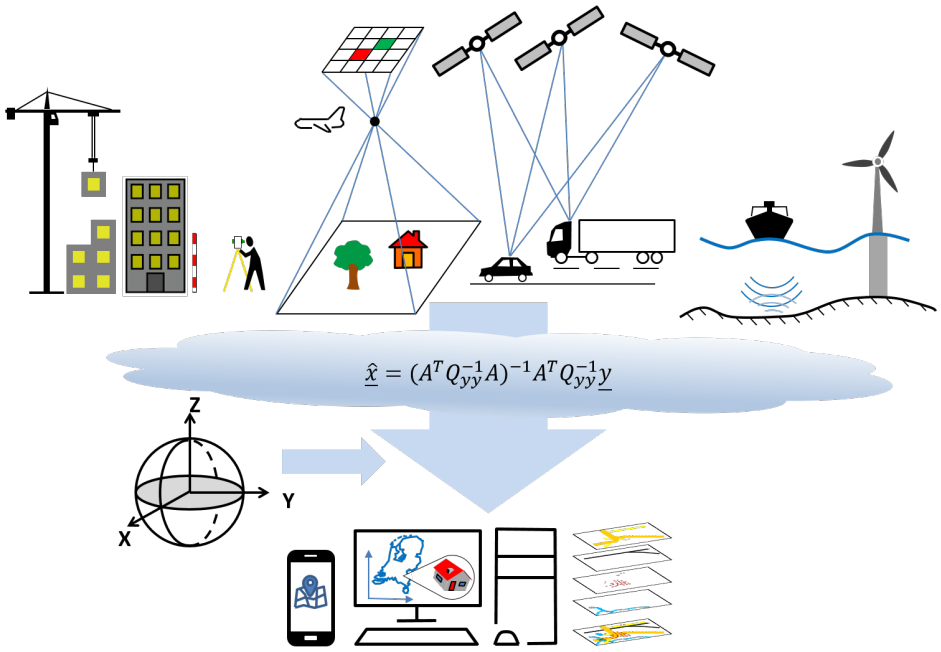



Figure 1: Overview of the subjects in this textbook on surveying and mapping, and their relationships.

pretation. In practice, surveying equipment is used, for instance, to measure distances. But typically one is not interested in knowing these distances. Instead, one is interested in the position of a certain object, which can be determined using these measured distances. And, being aware that in reality measurements are never perfect, one would like to know and quantify the quality of the resulting position, for instance, what is the uncertainty in the outcome one can expect under ordinary circumstances, or, to phrase it differently, how close can one expect the obtained outcome to be to the actual true position? The approach to 'handling measurement data' in this book is widely applicable in engineering, and as such contributes to observational data science.

GPS positioning has a vast range of applications in society, and is an extremely useful measurement technique also in surveying. In a relative set-up, millimeter-to-centimeter level positioning results are obtained in real-time, on the spot. In Part III the concept of standalone positioning is presented, as well as high-accuracy Real-Time Kinematic (RTK) positioning used in surveying, civil engineering and geoscience.

Part IV provides an introduction to the physics principles underlying measurements for surveying and remote sensing. Remote sensing means acquiring information about an object or phenomenon without 'going there' or 'touching it', in contrast to taking measurements in-situ or on site. Remote sensing is about *taking measurements from a distance*.

Remote sensing mostly refers to the use of electromagnetic signals (light, radar, laser) and sensors on board satellites and aircraft, in order to detect, measure and classify objects on Earth, including the surface, atmosphere and oceans. Also acoustic signals are used, with sensors on board vessels for measurements in oceans and waterways, and for sub-surface measurements.

Remote sensing with electromagnetic radiation in general serves a multitude of applications, for example creating topographic maps and three-dimensional terrain and elevation models, mapping coastlines and wetlands, mapping land-use (think of deforestation and urbanization), monitoring vegetation health, monitoring ice and snow coverage for climate studies, measuring ocean temperature and monitoring ocean circulation, hazard assessment (such

as flooding, erosion and landslide), measuring structural and surface deformations and subsidence, and observing atmospheric parameters such as cloud coverage for weather forecasting and monitoring air pollution.

Part V provides an introduction to coordinate systems and map projections. The latter allow us to visualize the Earth's curved surface on a flat sheet of paper or computer screen. The geometric infrastructure for coordinated surveying relies on reference systems, and as there are many in existence in practice, transformation from one reference system to another is an important subject to cover. Particular attention is given to heights and vertical reference. Part V on reference systems concludes with an overview of commonly used reference systems, both on- and offshore, at a worldwide, European and national Dutch scale.

In observing and describing the Earth's surface, as well as working on it while carrying out civil engineering projects, maps are indispensable tools for effectively communicating spatial information to colleagues, to the customer and to the public. Part VI of this book provides a brief introduction to the subjects of cartography and Geographic Information Systems (GIS). Working with visual variables and presenting spatial data is a skill, which is widely applicable in engineering.

## Educational considerations

This textbook keeps an informal and introductive style. Subjects are often described, explained and illustrated by means of examples. This textbook, after all, has been written with an *educational* goal in mind. The idea is to give students in civil engineering, and hopefully many others as well, a first *insight* into commonly used measurement techniques, the calculus of observations, the concepts and relevance of geometric infrastructure, and working with and visualizing spatial geodata, i.e. the art of mapping. Knowledge of high school physics and geometry should be sufficient to follow this textbook.

For Part II on mathematical geodesy, it is assumed that students have completed a course in linear algebra as well as an introductory course in statistics (in Dutch: kansrekening en statistiek), the latter for instance by using textbook [2].

Chapters and sections marked with a [*] contain material which is optional for the course Surveying and Mapping (CTB3310). Part of this material is covered in the course Monitoring and Stability of Dikes and Embankments (CTB3425).

## Acknowledgements

Our students are acknowledged in the first place, as they are the target audience for this material, yet at the same time teaching gives rise to interaction and feedback, either explicitly by raising questions and comments, or implicitly by showing questioning faces during a lecture. Your interaction provided valuable feedback, please keep it up!

Many colleagues from the department of Geoscience and Remote Sensing have contributed directly or indirectly to this textbook. We would like to mention a few of them in particular. Colleague Roderik Lindenbergh provided the original outline for the chapter on interpolation (Chapter 11), and colleague Marc Schleiss carefully reviewed this chapter. Roderik Lindenbergh is also acknowledged for his assistance with the chapter on laser scanning (Chapter 22).

Former colleague Peter de Bakker compiled an initial version of a reader on GPS positioning back in 2017 [3], which eventually evolved into the present part on GPS (Part III).

Former colleague Alijafar Mousivand wrote the first version of Chapter 25 on radiometric sensing, back in 2015, currently making up primarily Sections 25.1, 25.3 - 25.5 in Part IV.

Jochem Lesparre of the Netherlands' Cadastre, Land Registry and Mapping Agency (Kadaster) is credited with many improvements to the chapters on reference systems in Part V. Colleague Cornelis Slobbe is acknowledged for contributing the material on the 2018 Dutch quasi-geoid

and LAT chart datum in Sections 35.3 and 35.4.

Former MSc-student Jigme Klein assisted, back in 2019, with writing and compiling a first version of Chapters 37 and 38 in Part VI. For the subjects of mapping and Geographic Information Systems (GIS), earlier offerings of the course CTB3310 relied on the e-book by Michael Schmandt [4], which provided inspiration while writing the earlier reader on mapping. The on-line tutorial on the free, open-source GIS software QGIS [5] is also acknowledged. TU Delft colleague Edward Verbree is acknowledged for a first review of the part on mapping.

Roland Klees (TU Delft), Maarten Vergauwen (KU Leuven) and Menno-Jan Kraak (University of Twente) are acknowledged for feedback on the first edition of the book, in their respective domains of expertise.

The authors welcome corrections and suggestions for improvement, as well as feedback in general.

Delft, November 2021
(minor editing November 2022)

# I

# Land surveying

# 1

# Introduction

The goal of surveying is to gather information about the Earth, the Earth's surface and its topography. In order to gather information about phenomena and processes on Earth, surveying is about taking *measurements*. From these measurements we extract the information needed to model the Earth's surface and processes taking place on Earth, and mostly we focus on *geometric* information. As a simple example: we can identify the three vertices (corners) of the small plot of grass in Figure 1.1, determine that this plot is approximately locally level and that it can be represented by a triangle in the local horizontal plane. Next, we could measure the lengths of the three edges (sides). Thus, a real-world entity, the plot of grass, gets mapped as a two-dimensional horizontal area, namely a triangle, in Euclidean geometry, as shown at right.

For a civil engineer, land surveying is an essential skill. Surveying is used in all kinds of construction works like building bridges and infrastructure. Buildings and infrastructure need to be erected at the correct location, respect property boundaries, and have the correct dimensions, corresponding to the design. In particular big infrastructural projects cannot become a success without land surveying. Especially with big projects, improper surveying can lead to loss of the construction and/or big insurance claims.

## Overview of this part

In this part we cover leveling and tachymetry, which are considered classical land surveying techniques. These techniques are typically used on a local scale, for topographic mapping,

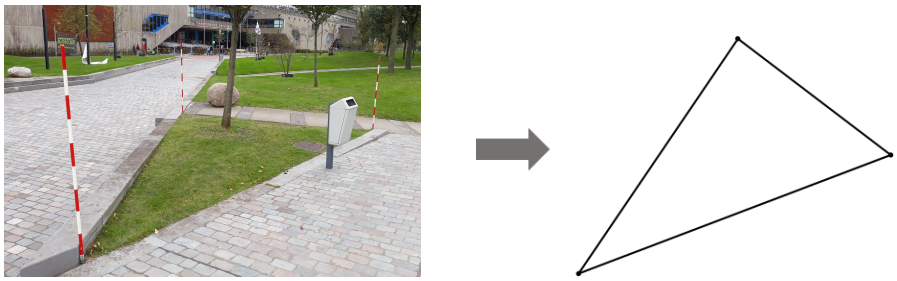
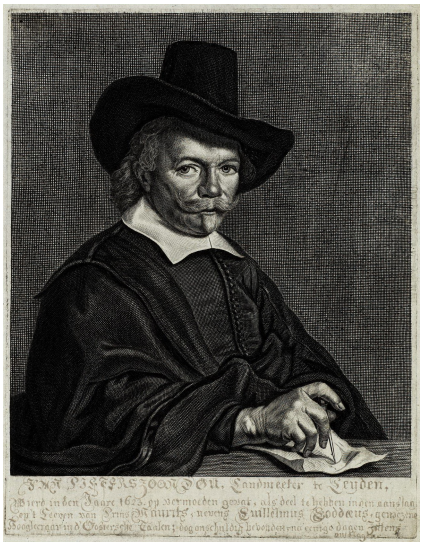


Figure 1.1: As a simple example, a small plot of grass is surveyed and mapped. This plot is geometrically modelled as a triangle and the three vertices (corners) are marked by red-white range poles (in Dutch: jalons) in the photo at left. When measurements of distance are taken along the three edges (sides), the simple map at right can be created. Surveying and mapping is about 'reducing the world around us to points, lines, polygons and volumes'.



TRACTAET
Vant maken ende Ge-
bruycken eens nieu gheor-
donneerden Mathematischen
Instrumens.
In welcke verscheyden konstighe stucken (de
Geometrie betreffende) vervatet ende
begrepen zijn.
Beschreven ende in druck uytgegeven
DOOR
IAN PIETERSZOON DOV
Der stadt Leyden Landtmeter ende Wijnroeyer.
PRÆSTAT
TOT AMSTERDAM
By Willem Janssen op't Water/ inde ver-
gulde Sonnewyser. Anno 1620.

Figure 1.2: At left: etching of Jan Pieterszoon Dou (1573-1635) by Reinier van Persijn (undated), taken from Wikimedia Commons [9] Public Domain. At right: title-page of the 'Tractaet vant maken ende Gebruycken eens nieu gheordonneerden Mathematischen Instruments', by Jan Pieterszoon Dou [8]. Public Domain.

and in support of construction works. Leveling is about measuring height differences, and tachymetry about measuring angles and distances.

There is much more to say about land surveying, as it also involves subjects like triangulation, surveying polygons (traversing), underground surveying, stake-out, and measurement and control for construction. For these subjects the reader is referred to the classical land surveying textbook [6] (in Dutch).

Before discussing leveling and tachymetry, in Chapter 3 and 4 respectively, we give a brief account of the history of land surveying in Chapter 2. Land surveying has a long and rich history, going back to ancient Greece, Egypt and Mesopotamia.

## Dutch historical perspective on land-surveying

In the 17th century, an important contribution to land surveying was made in Leiden, by Jan Pieterszoon Dou [7], see Figure 1.2 at left. Around 1610 he invented the Holland Circle (in Dutch: Hollandse Cirkel) and described this instrument in his book 'Tractaet vant maken ende Gebruycken eens nieu gheordonneerden Mathematischen Instruments', published in 1612 and reprinted in 1620 [8], see the title page in Figure 1.2 at right.

The instrument consists of two fixed sightlines, perpendicular to each other and attached to a circle, and one additional sightline (by a pair of sights, a so-called alidade) that can be rotated about the center of the circle, see Figure 2.5. This circle is graduated (marked) with a resolution of one tenth of a degree ($0.1^\circ$ of 360 degrees), an extraordinary achievement at that time. This instrument made it possible to precisely measure angles.

Later, by the end of the 18th century, theodolites replaced the Holland Circle. The optical telescope was invented by Hans Lipperhey in 1608 [10], a master lens grinder and spectacle maker, born in Wesel, Germany, living in Middelburg. He described his invention as a 'seecker instrument om verre te sien'.

Since the sixties and seventies of last century, as discussed in the next chapter, the theodolite has developed into the total station, the instrument commonly used in today's surveying.

# 2

# History of land surveying

The history of land surveying goes back at least some 3000 years. The river Nile in Egypt flooded parts of the country every year. To re-outline the farming land, surveyors re-measured the land. It was important to return the land properly to the farmers, as the amount of tax the farmers had to pay every year was proportional to the size (area) of their land. In those days the surveyors were not so well equipped as today. Their tool was a calibrated rope. That is why the Egyptian expression for a land surveyor was a 'rope stretcher', see Figure 2.1, and surveying was known as 'stretching a rope'.

Next in history, the Romans were known for their road system, their aqueducts and other big infrastructural civil engineering constructions. A well known surveying instrument of them was the groma. With this instrument they could stake-out straight and perpendicular lines, see Figure 2.2. It is a vertical staff with a horizontal cross at right angles on top and plumb lines hanging vertically from all four ends. Actually the groma was not a Roman invention. The instrument originated from Mesopotamia about 1100 BC.

The earliest preserved writing on land surveying is by Heron of Alexandria, a Greek who lived in Alexandria, Egypt, around 10 - 70 AD. His writing includes a treatise 'Dioptra' (surveyor's transit), see Figure 2.3, and a geometry book 'Metrica', see [11] and [12].

In the Middle Ages the art of surveying got almost forgotten until the beginning of the Renaissance. Around 1590 the plain table came in use, see Figure 2.4 — and it is occasionally still in use today. This was the starting point for todays surveying.

A well known instrument for surveying is the Holland Circle (Hollandse Cirkel in Dutch), see Figure 2.5. With this instrument angles could be measured, 360 degrees around, to distant

Figure 2.1: An ancient Egyptian surveyor at work in the farming fields. Harvest scenes from the tomb of Menna ca. 1400 - 1352 BC. Image by Charles K. Wilkinson, Metropolitan Museum of Art, New York, taken from Wikimedia Commons [9] under a CC0 1.0 Public Domain license.

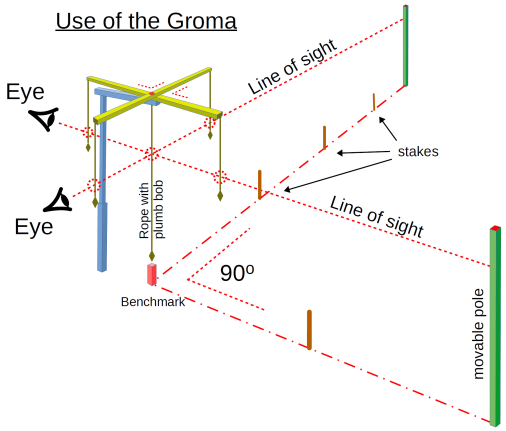



Figure 2.2: The groma, as used by Roman surveyors, allows to stake-out straight and perpendicular lines.

Figure 2.3: The dioptra described by Heron of Alexandria. Image courtesy of Kotsanas Museum of Ancient Greek Technology, Katakolo, Ilia, Greece, [13].

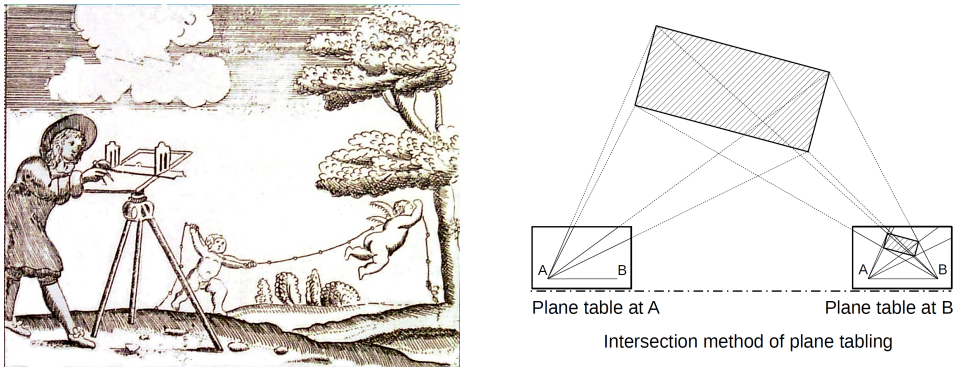



Figure 2.4: Example of the use of the plain table. Image at left taken from 'The compleat surveyor: or, the whole art of surveying of land, by a new instrument lately invented', by William Leybourn, 1722 [14], Public Domain.

Figure 2.5: The Holland Circle (in Dutch: Hollandse cirkel), an instrument for measuring directions to, and angles between distant objects. Photo courtesy of Rijksmuseum Boerhaave in Leiden, the Netherlands [15].

objects (as for instance church towers). This was a development from the 17th century, and an important step forward in land surveying.

With the invention of optical lenses and the telescope in the 17th century, surveying instruments for more precise measurements came within reach. From hereon two main directions of instrument developments can be distinguished. One is the *leveling* instrument. Two first examples of such an instrument — one without (at left), and the other with optics (at right) — can be seen in Figure 2.6. These are the so called dumpy levels. Dumpy levels are still in use today as one can see in Figure 2.7 at left, available in lower price categories. More often in use is the 'automatic' leveling instrument or automatic level; 'automatic' in the sense that the instrument has a built-in correction system for leveling, instead of a spirit level. Leveling instruments are used for measuring height differences between two points. A more detailed explanation follows later, in Chapter 3.

The other instrument is the *theodolite*. This instrument can measure horizontal and vertical angles. These angles were used together with a measurement of length between two points, and next all the points and objects could be positioned. This surveying technique is known as triangulation. Some early examples of theodolites are shown in Figure 2.8.

A new development in the 1950's was the Electronic Distance Measurement (EDM). In the beginning these were huge instruments capable of just measuring distance. In the seventies this technique got integrated in a theodolite, and the total station (or tachymeter) was born, see Figure 2.9 at left. This is an instrument which can measure distance and directions all in one go. Today the *total station* is the main workhorse in surveying, though classical theodolites are still sometimes used for special projects.

In this part you will learn about the use of the leveling instrument and the tachymeter (and total station). Principles and measurement techniques will be explained. We will first start with leveling in Chapter 3, followed by tachymetry in Chapter 4.

Figure 2.6: Leveling instruments from the early days of surveying. The construction with the spirit level attached to a telescope is also known as a dumpy level (at right). Photo at left courtesy of Museum of Lands, Mapping and Surveying, Queensland, Australia [16]. The leveling instrument at right was manufactured by B. Holsboer in Arnhem (1875-1900).


Figure 2.7: A modern version of a dumpy level, Johnson 22X Builder's Level (left), and an automatic level, Wild NA20 (right). Photo at left courtesy of Johnson Level and Tool Mfg Company [17].

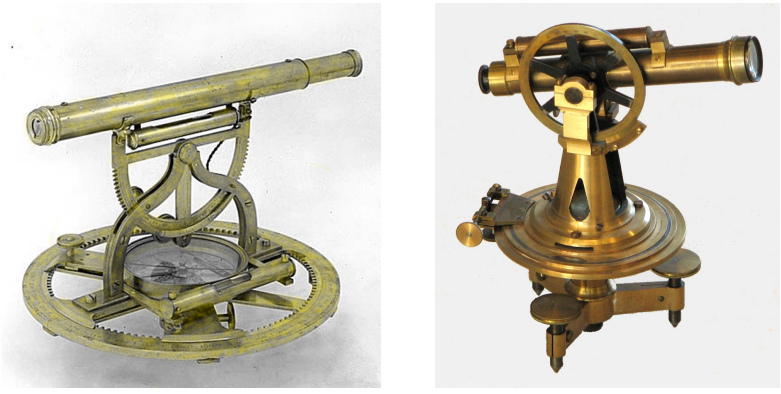


Figure 2.8: Theodolites from the early days of surveying. Photo at left, theodolite by Gilbert, London, taken from Smithsonian National Museum of American History [18] for educational and non-commercial use.


Figure 2.9: At left an example of a modern total station (South 6N+), and at right a modern theodolite (South DT-02L). The theodolite at right looks more lean, and has a display showing just vertical and horizontal angle. With the total station at left much more electronics is involved, in particular for the built-in Electronic Distance Meter (EDM). Photos courtesy of South Surveying and Mapping Technology Co., Ltd. [19].

# 3

# Leveling

## 3.1. Principle of leveling

Leveling is measuring the height difference between two points A and B with respect to an imaginary stationary open water surface — an equipotential surface, see Figure 3.1. The measured height difference, as it results from leveling, is $h_{AB}$, which equals $h_{AB} = H_B - H_A$ the difference of the height of point B and the height of point A.

There is a direct way of carrying a certain height to another location, namely by using a hose, filled with water. The water levels at both ends of the hose represent the same equipotential surface. This direct way of leveling is known as hydrostatic leveling (see Figures 3.2 and 3.3). The Survey-department of Rijkswaterstaat in the Netherlands used this method in the past for determining the height of the 'Wadden' islands with respect to the mainland. For this purpose an 11 kilometer length hose was used.

In this chapter we adopt a purely, and simply geometric approach to measuring and determining height. This is an approximation of reality which holds over small regions, where we can approximate the Earth's equipotential surface by just a flat plane. In Chapters 32 and 33 the true nature of height is covered, outlining its origin in gravity. In this chapter, leveling is basically measuring a distance between two points along a vertical coordinate axis, and this axis is pointing up.

The leveling line in Figure 3.4 is materialized by a strain wire, and connects two points of equal height, using a spirit level in the middle. The air-bubble in the spirit level is driven

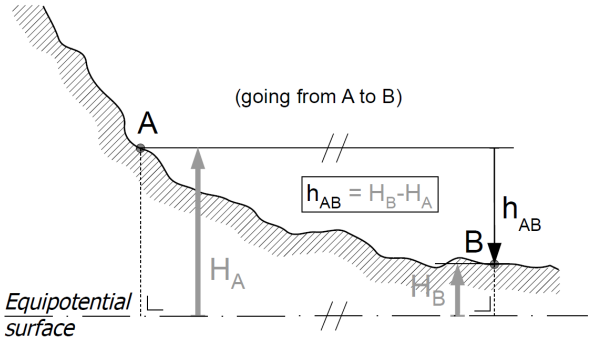



Figure 3.1: The principle of leveling. The height difference equals the height of point B minus the height of point A: $h_{AB} = H_B - H_A$. In this example the height difference is negative (B is lower than A), and hence the arrow is downward. It also holds that $h_{BA} = -h_{AB}$.

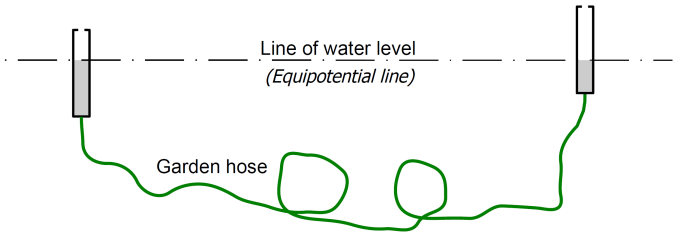



Figure 3.2: The hose method to create an equipotential surface/line — the principle of hydrostatic leveling.

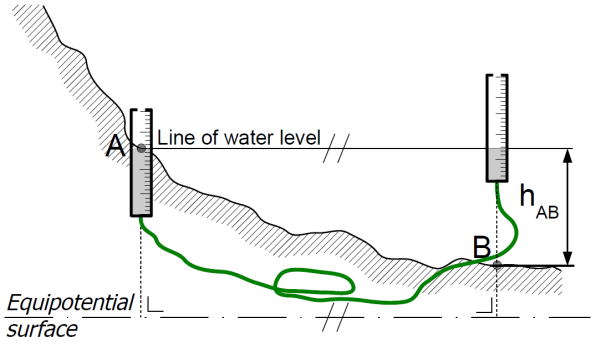



Figure 3.3: The hose method applied to measure the height difference $h_{AB}$ of Figure 3.1.

by gravity. The strain wire method is however not very practical, as you have to connect the two points physically (through the wire). And the method is not very accurate either. Still this method demonstrates the main principle of leveling. The physical wire has been replaced by an *optical* line, a so-called line of sight, see Figure 3.5.

To measure the height difference between points A and B, one installs a rod or staff on top of points A and B (and keeps it vertically, according to local gravity). This rod or staff is actually a sort of big ruler. The horizontal line (along the optical line set out by the instrument in the middle) hits the ruler at A and the one at B. One can read, through the telescope the two distances $l_A$ and $l_B$ on the rulers, which is the height of the line of sight over point A, and the height of the line of sight over point B, respectively. The difference of these two readings yields the height difference between points A and B, namely $h_{AB} = l_A - l_B$. Note that the staff reads zero at the bottom, and 'distances' $l_A$ and $l_B$ are measured upward (positive axis is pointing upward).

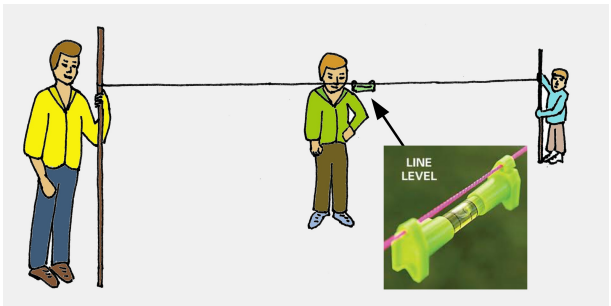



Figure 3.4: Leveling by using a strain wire between two rods or staffs, with a spirit level attached to the wire in the middle.

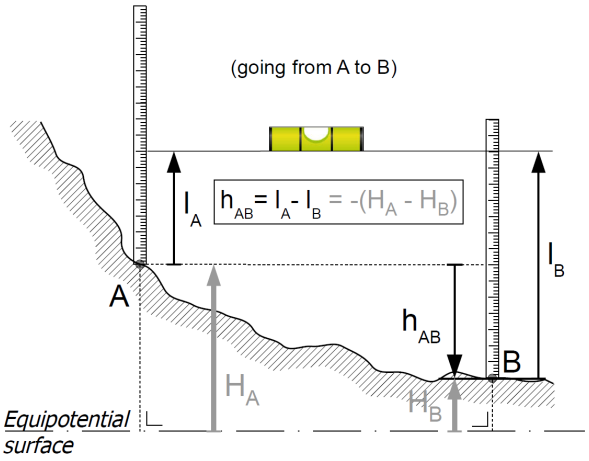



Figure 3.5: The main principle of leveling. A line of equal height is established optically, by means of a telescope in the middle (not shown here), which is set-up according to local gravity. The line of sight is perpendicular to the local plumb line. $h_{AB} = l_A - l_B$. Height difference $h_{AB}$ is negative here.

When the rod at point A is set up on a bolt with a given height, see Figure 35.9, the height of point B easily follows as $H_B = H_A + h_{AB}$.

A leveling instrument is shown in Figure 3.6. The instrument is a telescope with a spirit or automatic level, and the line of sight represents a certain, fixed height (indicated in gray). The instrument can be rotated only about its vertical axis (which is aligned with the local plumb line, indicated in red) and thereby you can see only objects which are at the same height, that is, objects which are at the same level.

To measure the height difference between the points A and B as shown in the Figures 3.1 and 3.4, the leveling instrument is installed — on a tripod — in the middle between the rods at A and B (Figure 3.7). Then a reading of rod A is taken, $l_A$, trough the telescope, followed by a reading, also through the telescope, of rod B, $l_B$. The height different between points A and B is then: $h_{AB} = l_A - l_B$.

In leveling, height differences are typically measured up to about 100 meters in distance in one step, or stretch. In practice, height differences can be determined over distances of several kilometers. That means that leveling is done step-by-step. Figures 3.8 and 3.9 shows two methods of how this can be done.

Figure 3.8 starts with one rod and one leveling instrument. First the rod is setup at point A. The leveling instrument is set up in the middle of points A and B. Then a reading of the rod at point A is taken. Then the rod is moved from point A to point B. Then turn the instrument from backsight direction to foresight direction and take a reading from the rod at B. While keeping the rod at B, take the leveling instrument and set it up halfway points B and C. Now follow the same routine as between the points A and B. And continue with this till the last and final point is reached.

Figure 3.9 shows a second method. It starts with a leveling instrument and *two* rods. One rod (the yellow one) is set up at point A, and the second one (in orange) is set up at point B. The leveling instrument is set up halfway the points A and B. Now first take a reading of the rod at point A (backsight). Turn the instrument to the foresight direction and take a reading of the rod at point B. Now move the yellow rod from point A to point C. Secondly, take the leveling instrument from its first setup (between A and B) and bring it to halfway inbetween points B and C. Measure first in backsight the rod at point B. Turn the instrument in the foresight direction and measure the rod at point C. Continue so, step-by-step, until the

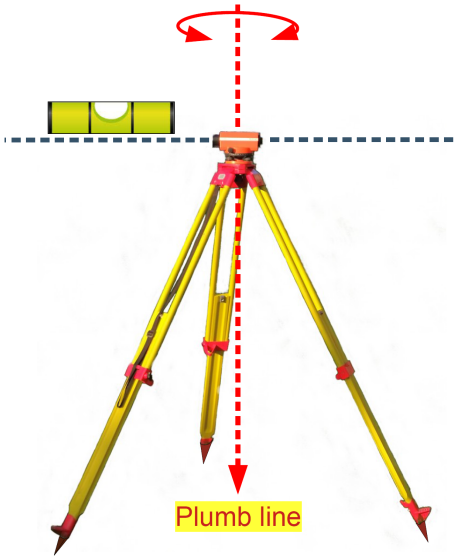



Figure 3.6: Leveling instrument in use. The vertical axis (in red) is set-up along the local plumb line. The line of sight (dashed line in gray), through the telescope, is in the local horizontal plane, and connecting — optically — points of the same height.

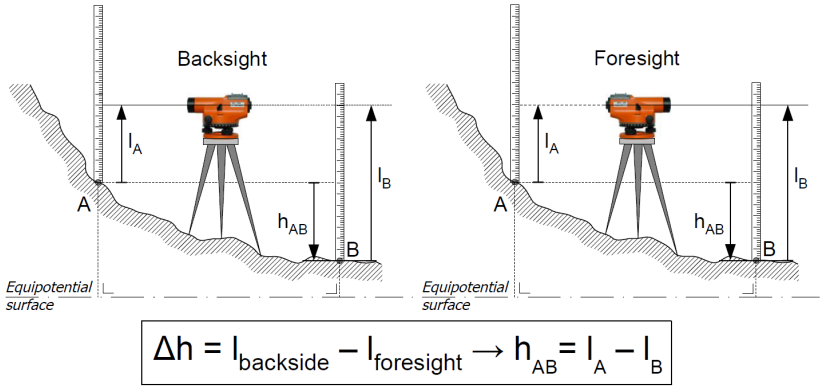



Figure 3.7: One step or stretch of leveling: first take a reading of rod A, then turn the instrument to rod B, and take a reading of rod B. The height difference is $h_{AB} = l_A - l_B$.

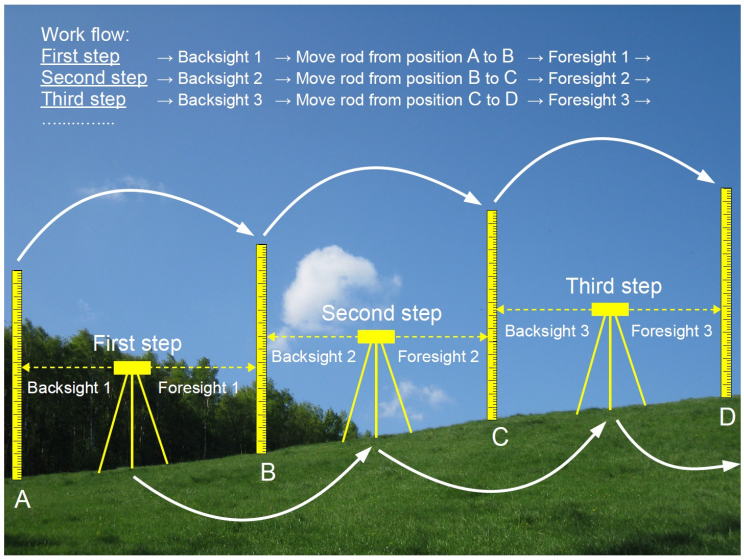



Figure 3.8: Leveling procedure: method 1, using a leveling instrument and a single rod.

final point is reached.

## 3.2. Leveling instrument: automatic level

In Figure 2.7 two types of leveling instruments were presented. One is the so called dumpy level, a leveling instrument with a spirit level (at left). This instrument was invented in the early 1700's, and is still in use today. It is not an easy instrument to handle, but it is very reliable. The surveyor has to work very accurately when setting up this instrument and taking measurements is rather time consuming.

A solution for the dumpy level came with the automatic level, or automatic leveling instrument. In this instrument the spirit level was replaced by an optical mechanical leveling system, the so called 'compensator'. Figure 3.12 shows the compensator inside the leveling instrument. Today most of the leveling instruments are automatic levels. They are available for different classes of accuracy, ranging from instruments for surveying (very accurate) to construction site purposes (built for measuring over short distances). Surveying grade automatic levels can even be equipped with a digital read-out and data logging, plus an electronic distance meter (see Figure 3.10).

In Figure 3.11 the outer parts of an automatic level are indicated. The figure shows, for example the Nikon AC-2SG, the part names, and these are generally the same for every leveling instrument. A step-by-step explanation about how to set-up a leveling instrument will follow later in Section 3.6. The only point of attention for the moment is the circular level (bottom left corner). When a leveling instrument is mounted on top of a tripod, one of the first things to do is to level the instrument by adjusting the leveling screws, so that the bubble in the circular level is right in the center.

A circular level is however not the most precise spirit level. To carry out accurate measurements with a leveling instrument, a much more precise way of leveling the instrument is needed. This is done inside the instrument — permanently — by a compensator (see Figure 3.12). The light beam — in red — from the objective is directed through a suspended prism (light blue triangle) in such a way that it automatically corrects the direction to the local horizontal, in case the leveling instrument has a small set-up error $\Delta\varphi$ with respect to the local

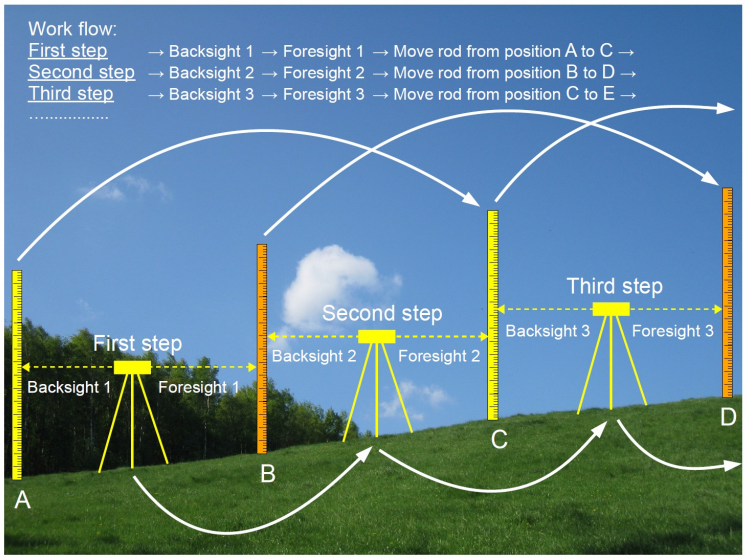



Figure 3.9: Leveling procedure: method 2, using a leveling instrument and two rods.


Figure 3.10: A modern electronic automatic level with digital read-out and data-logging functionality, Trimble DiNi 12. At right (a part of) the so-called bar-code rod or staff is shown, to be used with this digital leveling instrument.

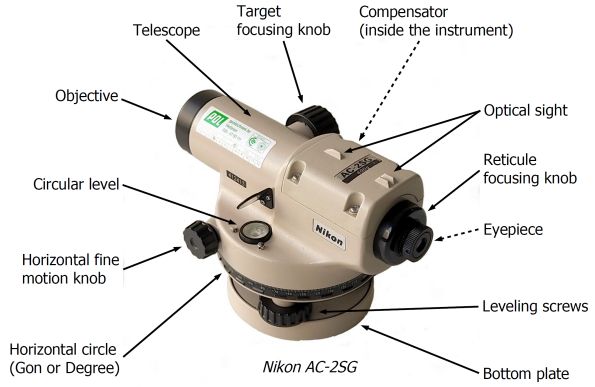



Figure 3.11: The outer parts of an automatic level instrument, in this example the Nikon AC-2SG.

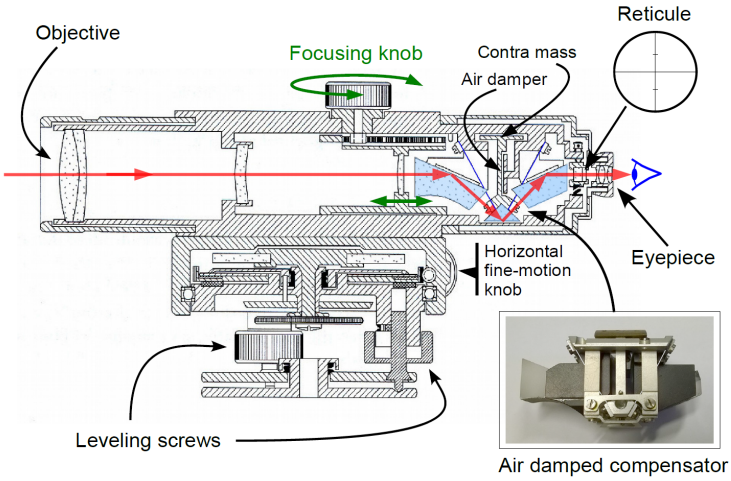



Figure 3.12: A cross section diagram of an automatic level instrument. The light beam is shown in red. The prisms in light blue constitute the compensator. The inset shows a photo of an actual compensator.

horizontal or 'leveling line' (see Figure 3.13).

Although the compensator corrects for small offsets in the instrument set-up, a minor error may still remain. This error depends on how well, and how long ago the instrument is calibrated. The remaining small offset in the horizontal line of sight is shown in Figure 3.14, and denoted by $\varphi$.

Though a small error $\varphi$ exists, it is still possible to obtain correct leveling measurements by setting up the instrument just in the *middle* between the two rods or staffs (see Figure 3.15), and the situation is perfectly symmetric to both sides of the instrument (lines of sight indicated in red). The same error occurs at A and B, $\xi$ (in red), and therefore perfectly cancels in the measured height difference, as $h_{AB} = l_A - l_B$.

When the instrument is set-up excentric, that is, away from the middle, the error in reading the rod is not the same at both rods, and therefore does *not* cancel in the measured height difference. This situation is shown in blue, in Figure 3.15: $L_A \neq L_B$ (in blue) and the same angular error in the line of sight translates in a much bigger error at A than at B, as $L_A > L_B$.

With the exercises and worked examples in Section 3.7, we elaborate on a practical procedure to determine this error $\varphi$ of the instrument in the field, even when the height difference

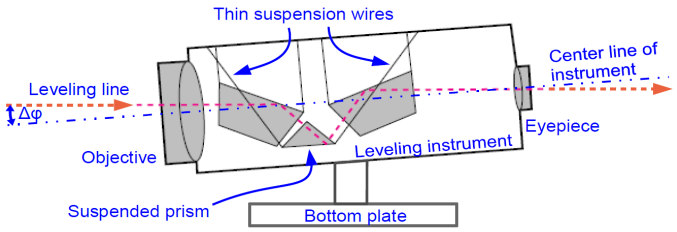



Figure 3.13: The automatic level instrument in action. The horizontal center line of the instrument (blue line) is intentionally set off-level. The prisms of the compensator nevertheless realize a horizontal line of sight (red line).

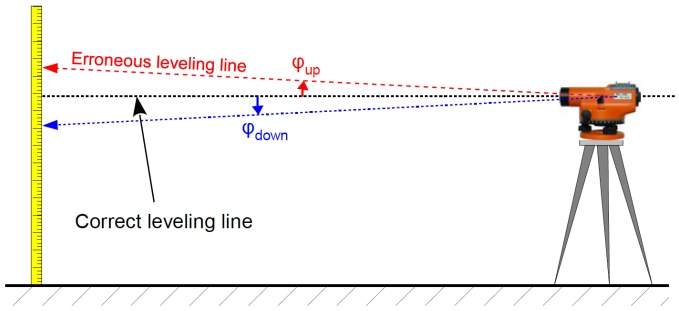



Figure 3.14: Even with a compensator, the line of sight, or leveling line (in red or blue) of the instrument may still be slightly off from the true local horizontal plane (indicated by the dashed line in black). This small remaining error is indicated by angle $\varphi$, and also referred to as a collimation error.

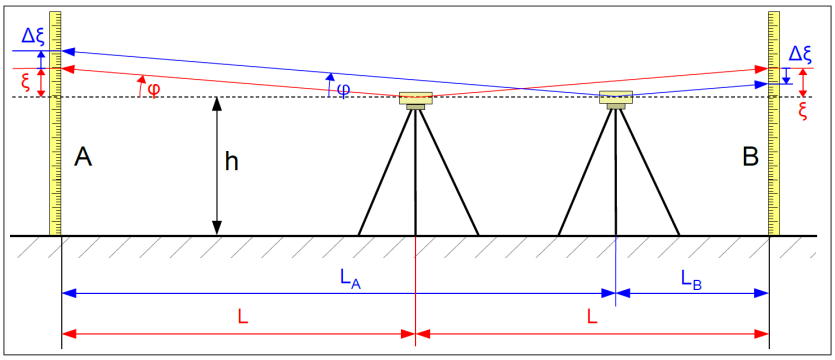



Figure 3.15: By setting-up the leveling instrument right in the middle between points A and B, a remaining small error in the line of sight (angle $\varphi$) has no impact on the measured height difference (situation indicated in red). The error is $\xi$ (in red) on both sides, and cancels in the height difference. When the instrument is *not* in the middle, the measured height difference will be in error (situation indicated in blue). At point A the error in reading the rod becomes $\xi + \Delta\xi$, at point B this is $\xi - \Delta\xi$. Hence, in the observed height difference $h_{AB}$, the error $2\Delta\xi$ remains. As the mis-alignment angle $\varphi$ is the *same* to both sides, $\Delta\xi$ (in blue) is the *same* on both sides (the lines of sight in red and blue run *parallel*).

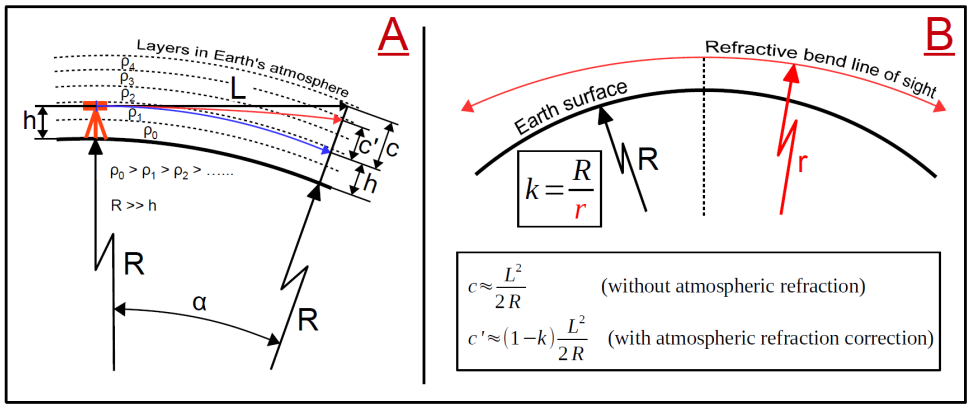



Figure 3.16: The effects of the Earth's curvature and atmospheric refraction on leveling. The reading of the rod will be too large, and consequently the determined height will be too low.

between the points A and B is not known in advance. As Figure 3.15 already suggests, the key lies in setting up the instrument twice: once in the middle, and once very much excentric.

To conclude this section on the automatic leveling instrument we mention that calibration of the instrument is important. During a calibration it is, for instance, also checked whether the reticle (cf. Figure 3.12 at right) is aligned with the line of sight of the instrument.

## 3.3. Curvature of the Earth

Early in Section 3.1, we made the assumption that the Earth — apart from topography — is flat. This assumption is a fair approximation over small regions. In this section we analyse the effect the curvature of the Earth has on leveling.

Figure 3.16 at left shows the instrument set-up, at height $h$ above the Earth's surface (here simply assumed to be a sphere with radius $R$), nicely leveled, and the leveling rod at some distance $L$, and at that location properly set up along the local plumb line (mind that at the instrument and at the rod, gravity is pulling in slightly different directions). In this analysis, the instrument height $h$ — typically 1.5 m — is neglected (or one can assume it to be included in $R$, the Earth's radius, which equals 6378 km); anyway we have $h \ll R$.

The line of sight of the instrument (the black line in Figure 3.16) not following the Earth's curvature, causes an error $c$ in the observed height difference; the true local horizontal is indicated by the blue line. Using Pythagoras in the triangle to the center of the Earth, with angle $\alpha$, we have: $R^2 + L^2 = (R + c)^2$. Realizing that $c \ll R$, so that the term $c^2$ can be neglected, we arrive at

$$c \approx \frac{L^2}{2R} \tag{3.1}$$

For a distance of $L = 100$ m, the effect is $c = 0.78$ mm, and for a distance of $L = 1000$ m, the effect is $c = 7.8$ cm. Hence, only over short distances, the effect of the Earth's curvature can be neglected.

Chapter 33 on vertical reference systems covers heights in further detail. Leveling — over short distances — yields (good approximations to) orthometric height differences.

## 3.4. Atmospheric refraction

In Figure 3.16 it is also indicated that, generally, the Earth's atmosphere gets more dense, the closer one gets to the surface ($\rho_0 > \rho_1 > \rho_2 \ldots$). Thereby also the refractive index of

the atmosphere gets larger, closer to the surface. By Snell's law of refraction and Fermat's principle of least travel time, light will bend towards the Earth's surface, and actually follow the red line in Figure 3.16 — light will take the fastest route. Section A.1 in the appendix provides a brief overview of the physical background on this subject.

As shown by the equations in Figure 3.16, the error due to the Earth's curvature gets actually reduced by the effect of atmospheric refraction, as we now have a factor of $(1-k)$ instead of just 1 in the equation, where $k$ is an empirically determined factor, often a default of $k$=0.13 is used. Similar to (3.1) for the Earth's curvature, the effect of the curvature of light waves due to the refraction is $\frac{L^2}{2r}$, where the curvature radius is $r = \frac{R}{k}$. Hence we obtain $c - c' \approx k\frac{L^2}{2R}$, leading to

$$c' \approx (1-k)\frac{L^2}{2R} \tag{3.2}$$

which represents the combined effect of the Earth's curvature and atmospheric refraction.

## 3.5. Good practice: recommendations

In order to avoid excessive errors due to atmospheric refraction it is recommended to generally keep a ground clearance of at least half a meter for the line of sight. One should also be aware of the fact that over a pool of cold water, a ditch or a canal, the refractive index may show strong gradients, leading to a strongly curved line of sight, and hence large leveling errors.

Good surveying practice is to limit the distance between instrument and rod to 50 meter (to avoid excessive errors), and in particular to set-up the instrument at equal distance from the two rods (i.e. halfway). The latter practice eliminates a possible tilt error of the line of sight (see Section 3.2), it also eliminates a big part of the Earth's curvature effect (Section 3.3), and — assuming that the Earth's atmosphere is nicely horizontally layered and one is working in a fairly flat area — it also eliminates most of the refraction error (Section 3.4).

## 3.6. Measurement procedure

In this section we outline the actual leveling procedure in the field. It consists of setting up the tripod on which we mount the instrument. Then we level the instrument itself, to get the line of sight horizontal, and we focus the reticle. Eventually we focus the telescope and take a reading of the leveling rod.

### 3.6.1. Set-up the tripod

The first step in carrying out leveling measurements is setting up the tripod. Normally the set-up height of the tripod corresponds to the length of the surveyor, so that he/she can look into the eyepiece of the instrument in a comfortable way. First extend the three legs of the tripod by the same amount. Spread out the legs at equal distance from each other as Figure 3.6 shows.

Figures 3.17 and 3.18 demonstrate how to fix the tripod legs on soft ground and on hard surface, respectively.

### 3.6.2. Mount leveling instrument on tripod

The next step is mounting the leveling instrument — carefully — on the tripod by fasten it with a screw (see Figure 3.19 left).

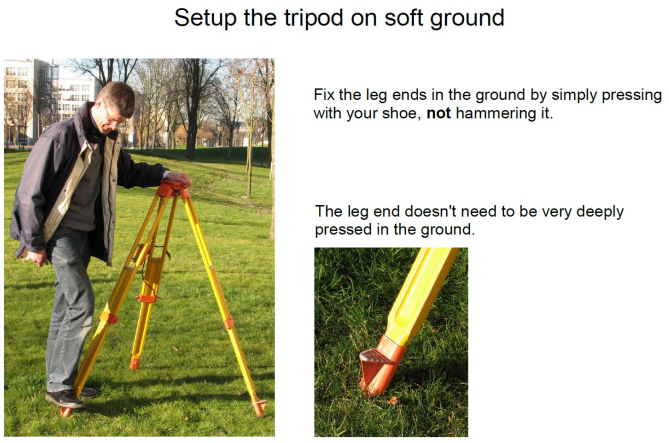



Figure 3.17: Fixing the tripod legs on soft ground.

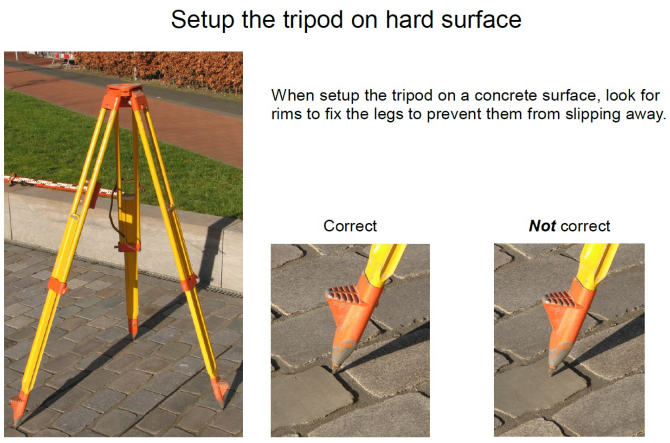



Figure 3.18: Fixing the tripod legs on hard surface.

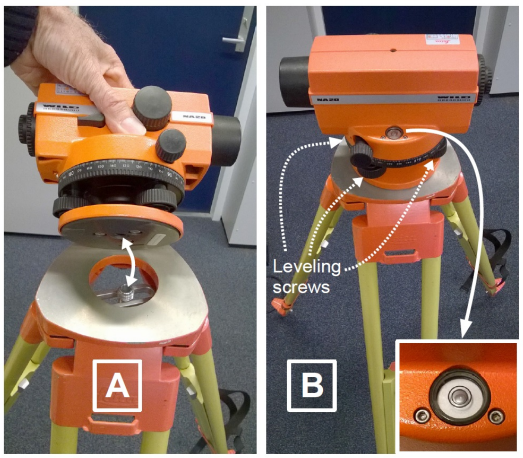



Figure 3.19: Mounting the leveling instrument on the tripod (left), and setting the instrument level (right).

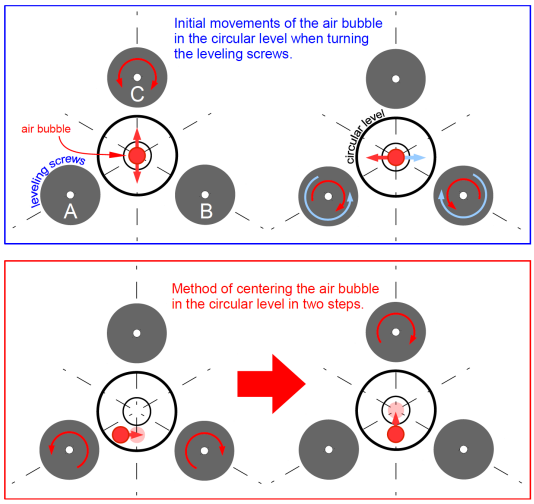



Figure 3.20: To move the air-bubble to the right from A to B (bottom-left panel of figure), turn the two screws A and B as indicated. Turn the screw on top (C) as indicated (bottom-right panel of figure) to make the air-bubble move upward to C, and to center the air-bubble in the bull's eye spirit level.

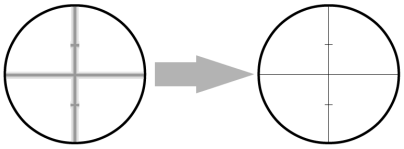


Figure 3.21: Focusing the reticle: the cross-hair should be seen ultimately clear and sharp in order to allow for parallax-free readings. You should end up with the image at right. When focusing the reticle you may turn the telescope to some white or light gray object, and leave the telescope unfocused.

### 3.6.3. Level the bull's eye spirit level

After mounting, the instrument needs to be leveled by adjusting the leveling screws in such a way that the air-bubble gets centered in the bull's eye spirit level (see Figure 3.19 right, in particular the inset). The instrument is suspended at three points. Correspondingly there are three leveling screws. How to adjust them is explained in the bottom panel of Figure 3.20.

Once you started measuring and reading 'backsight' and 'foresight', *never* re-adjust the leveling screws, as the line of sight of the instrument will change immediately (as well as the height of the instrument itself). When adjustment is needed, the set-up of the instrument should be re-done, as well as the measurements.

### 3.6.4. Focus the reticle

Adjust the eyepiece lens (Figure 3.22) to focus the reticle, see Figure 3.21.

### 3.6.5. Point instrument to leveling rod

Point the instrument to the leveling rod or staff for the 'backsight' measurement by using the indicated line of sight direction on the instrument housing (see Figure 3.22, dashed arrow in blue). This allows you to quickly point the instrument roughly in the right direction.

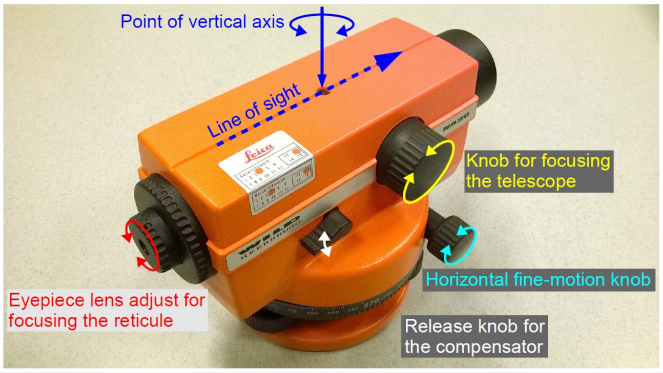



Figure 3.22: Automatic leveling instrument Wild NA20 with the eyepiece for focusing the reticle, the knob for focusing the telescope, and the line of sight direction indicated on top of the housing.

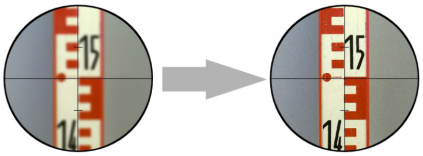



Figure 3.23: Focusing the telescope on the leveling rod.

### 3.6.6. Focus the telescope

Next, direct the leveling instrument straight onto the rod, i.e. look through the eyepiece and search for the leveling rod by using the horizontal fine-motion knob (cf. Figure 3.22). And, focus on the rod by adjusting the focusing knob of the telescope (Figure 3.23), so that you can see the rod and the markers very clearly. Eventually it should look like that the cross-hair sticks onto the leveling rod.

### 3.6.7. Reading the leveling rod

The leveling rod is a kind of ruler with a repeated decimeter pattern (Figure 3.24 at left). The indicated numbers ..., 13, 14, 15, ... represent a decimeter scale [dm]. The decimeter is divided, by an 'E'-pattern (red and white letters 'E'), into centimeters [cm] (this type of leveling rod is sometimes referred to as 'E-rod', or 'E-baak' in Dutch). Figure 3.24 at right shows the decimeter pattern with a [cm] and [mm] scale outlined.

Reading the leveling rod through the eyepiece of a leveling instrument, the horizontal center line (cross-hair) of the reticle directly gives the observation. In Figure 3.25 the observation reads: 14.22 dm. The last digit, here 0.02 dm, is an estimate, taking the scale of Figure 3.24 in mind (it is two millimeters).

To check for correctness of the observation, and in order to obtain an indication of the distance between the instrument and the rod, also readings from the upper and lower horizontal line in the reticle must be taken (these horizontal lines in the reticle are also known as stadia lines). In the example of Figure 3.25 the upper line reads 15.00 dm, and the lower line reads 13.44 dm. The average of these readings must equal, or be close to, the observation made with the (center) cross-hair. Here we have: (15.00 + 13.44)/2 = 14.22 dm, hence a perfect match.

The distance from the instrument to the rod or staff can be calculated too. Take the difference between the upper line and lower line reading and multiply this by 100. In the

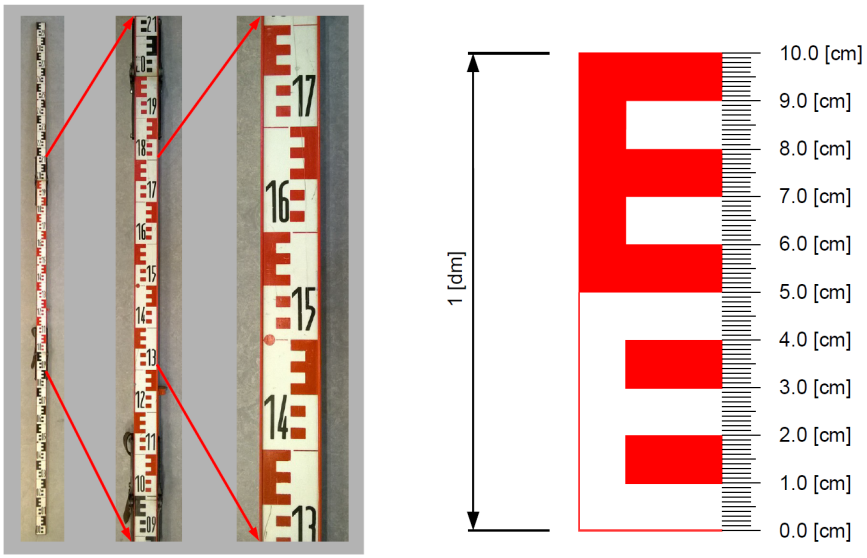



Figure 3.24: At left: leveling rod or staff, of so-called 'E-type', a white 'E' and red 'E' together cover one decimeter. The numbers indicated on the rod represent a decimeter scale. At right: A one decimeter piece of a leveling rod. When reading the rod, the surveyor can take a decimeter and centimeter reading, and should *estimate* the amount of millimeters.

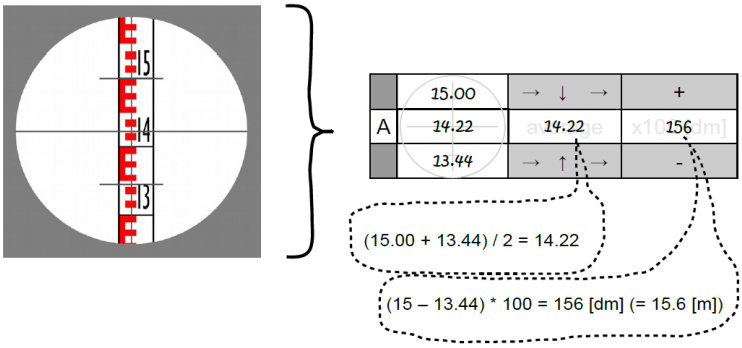



Figure 3.25: Taking an observation by reading the leveling rod through the eyepiece of the instrument (at left). And at right the registration of this observation.

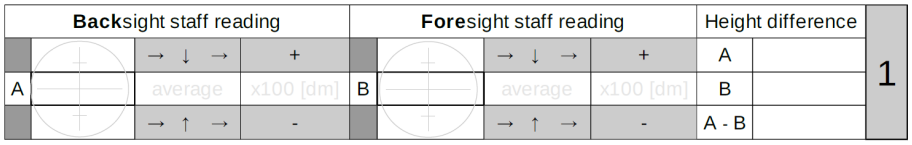



Figure 3.26: Example of a leveling observation registration form. In the left column of the 'backsight'-part, one supplies the readings of respectively the upper, center and lower line, next one computes the average and the distance. This is repeated for the 'foresight' staff in the middle column, and eventually the last column is completed, by carrying over the two center-line observations, and computing the difference.

example we have (15.00-13.44) * 100 = 156 dm, which is equal to 15.6 m. This measurement of distance is not very precise, but it is good enough to know how well the leveling instrument was centered in between the two leveling rods.

To conclude, Figure 3.26 shows an example of an observation registration form, in this case just a part is shown corresponding to leveling one step or stretch, with a backsight and a foresight reading (cf. Figure 3.7). For both the backsight and foresight one computes the average of upper and lower wire reading, and also the distance.

## 3.7. Exercises and worked examples

This section presents a couple of exercises and worked answers on leveling.

**Question 1** Starting from point 1, with given height (0.382 m), one has leveled from point 1 to point 2, from 2 to 3, and eventually from point 3 to point 4, in order to determine the height of point 4, see Figure 3.27. The backsight and foresight readings each time, are listed in Table 3.1; this is a simplified measurement form, without upper- and lowerline readings. Compute the height of point 4.

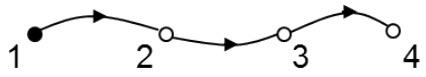



Figure 3.27: Leveling line from point 1 to point 4, observed in three runs.

| run | back | fore |
|---|---|---|
| 1 | 16.53 | 12.27 |
| 2 | 15.10 | 8.43 |
| 3 | 4.60 | 19.64 |

Table 3.1: Simplified measurement form with just center-line readings of back- and foresight. Readings are given in dm.

**Answer 1** The height of point 4 follows in this case straightforward as $H_4 = H_1 + h_{12} + h_{23} + h_{34}$. The height differences can be obtained from the back- and foresight readings, for instance $h_{12} = l_1 - l_2$, in this case $h_{12} = 16.53 - 12.27 = 4.26$ dm. We get $H_4 = 0.382 + 0.426 + 0.667 - 1.504 = -0.029$ m.

**Question 2** Due to imperfections in the optical part of a leveling instrument, the line-of-sight may not be perfectly horizontal, when the instrument has been properly set-up (that is, its vertical axis according to the local plumb line). The line of sight is slightly tilted, as shown

in the two set-ups in Figure 3.28; the dashed lines show the perfect horizontal line, and the solid lines show the actual lines of sight. Using a symmetric set-up (shown at left) and an a-symmetric set-up (shown at right) between the same two points, A and B, it is possible to determine the tilt error (that is, the small angle between the solid and the dashed line). With the following readings $z_A = 1.540$ m, $z_B = 1.268$ m (for the set-up at left), and $z'_A = 1.470$ m and $z'_B = 1.218$ m (for the set-up at right), and distance $d$ = 40 m, determine the tilt error.

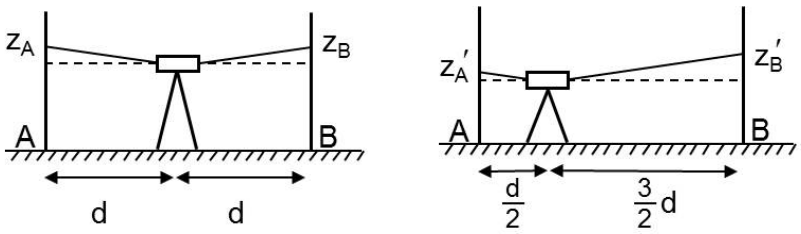



Figure 3.28: Two set-ups to measure the same height difference $h_{AB}$ between points A and B, in order to determine the mis-alignment error of the line of sight of the leveling instrument. This is referred to as the two peg test.

**Answer 2** This question refers to the situation shown in Figure 3.15. The situation at left in Figure 3.28 corresponds to the situation shown in red in Figure 3.15. In this (symmetric) situation, the readings at both rods are off by $\xi$. For the rod at point A the reading is $z_A = l_A + \xi = 1.540$ m, where $l_A$ is the reading as it should be (with a horizontal line of sight), and $\xi$ is the systematic error in the reading, due to tilt of the line of sight. For the rod at point B we have $z_B = l_B + \xi = 1.268$ m, and hence the height difference becomes $h_{AB} = z_A - z_B$, cf. Figure 3.5, in this case $h_{AB} = l_A + \xi - l_B - \xi = l_A - l_B = 0.272$ m, and the height difference is *not* affected by the mis-alignment of the line of sight in this *symmetric* set-up. In the a-symmetric set-up, shown at right, and indicated in blue in Figure 3.15, the mis-alignment error will *not* cancel. For the reading of the rod at A we have $z'_A = l'_A + \xi - \Delta\xi$, and for the reading of the rod at B we have $z'_B = l'_B + \xi + \Delta\xi$. So, the observed height difference now becomes $h'_{AB} = z'_A - z'_B = l'_A - l'_B - 2\Delta\xi = 0.252$ m; the height difference is off by $2\Delta\xi$. With the proper height difference $h_{AB} = l_A - l_B = 0.272$ m from the symmetric set-up (note that $l'_A - l'_B = l_A - l_B$), we arrive at $2\Delta\xi = 0.02$ m, or $\Delta\xi = 0.01$ m. Finally we consider the fact that the mis-alignment error is fixed all the time (*same* angle between the dashed and solid line everywhere). Hence, the two triangles in the figure at right have identical shapes. In the triangle at left, the distance to the rod is $\frac{d}{2}$ and the reading error is $\xi - \Delta\xi$. In the triangle at right, the distance to the rod is $\frac{3}{2}d$ and the reading error is $\xi + \Delta\xi$. Insisting on identical shapes for these triangles leads to

$$\frac{\xi - \Delta\xi}{\frac{d}{2}} = \frac{\xi + \Delta\xi}{\frac{3}{2}d}$$

from which follows that $\xi - \Delta\xi = \frac{1}{3}(\xi + \Delta\xi)$, hence $\frac{2}{3}\xi = \frac{4}{3}\Delta\xi$, which yields $\xi = 2\Delta\xi = 2 \cdot 0.01 = 0.02$ m. Considering the symmetric situation at left, the reading error is $\xi$ = 0.02 m, at a distance of $d$ = 40 m, hence the angle follows from $\tan\varphi = \frac{\xi}{d}$, as $\varphi = 5 \cdot 10^{-4}$ rad. Mind that it may look odd at first instance, that, going from the symmetric set-up to the a-symmetric set-up, both readings at A and B get less ($z'_A < z_A$ and $z'_B < z_B$). In practice the terrain may not be exactly flat, as suggested in the drawing, and/or the instrument height can be different for the two set-ups. The reading at A got however, less by a bigger amount than the reading at B.

# 4

# Tachymetry

Tachymetry is about measuring angles and distances, according to the principle of polar or spherical coordinates to determine the position of a target or object (in two or three dimensions respectively). A tachymeter or total station basically is a theodolite with a built-in distance meter. Therefore the tachymeter is explained in two parts in this chapter. First the theodolite is covered, and next, in the section on the total station, the distance meter is covered.

## 4.1. Theodolite

A theodolite is a surveying instrument meant to measure both horizontal and vertical angles. Therefore this instrument is equiped with two scales. A scale is a circular disc which is sub-divided with markers into very small equal parts (pies), see also Figure 18.1. Figure 4.1 shows the theodolite in its most basic form.

A diagram of the theodolite is shown in Figure 4.2. The zero direction on the horizontal scale is in an arbitrary direction; the zero direction on the vertical scale is by default exactly up, along the vertical axis.

The theodolite should be set up such that the vertical axis is aligned with the local plumb line, see Figure 4.3. And, the instrument has to be positioned such that the vertical axis, when extended downwards, hits the benchmark or survey marker on the ground. How to achieve this in practice is detailed later on.

Figure 4.4 shows how — at one set-up of the instrument — a horizontal and vertical angle are measured between two objects. The horizontal direction to object A is $\varphi_A$, and the horizontal direction to object B is $\varphi_B$. The horizontal angle between the two objects is $\varphi_{AB} = \varphi_B - \varphi_A$. The vertical scale is not shown in Figure 4.4. The zero direction is by default exactly up, along the vertical axis. The vertical direction (zenith angle) to object A is $z_A$, and the vertical direction to object B is $z_B$. The vertical angle is then $z_{AB} = z_B - z_A$.

The enclosed angle $\xi$ can be found, as stated in Figure 4.4 at bottom, using the cosine rule from spherical trigonometry. A unit sphere is taken, and a triangle on this sphere is considered with nodes A, B and the zenith.

In daily day life angles are measured in degrees, with a full turn (full circle) being 360 degrees, and with 1 degree sub-divided into 60 minutes, and 1 minute sub-divided into 60 seconds. This way of counting goes back to the ancient Babylonians (5000 BC). By the time of the French revolution (at end of the 18th century), Napoleon wanted to modernize this way of counting. He introduced the metric system, with a decimal way of counting, for instance with length expressed in meters, kilometers and centimeters, and also a *decimal* system for angles. The decimal degrees are known as gon [gon], grades or gradians [grad]. A full turn

Figure 4.1: The theodolite in its most basic form, actually Do-It-Yourself with paper and cardboard. There are two scales, one for measuring vertical angles, and the other for measuring horizontal angles. The historical theodolites of Figure 2.8 show similarly.

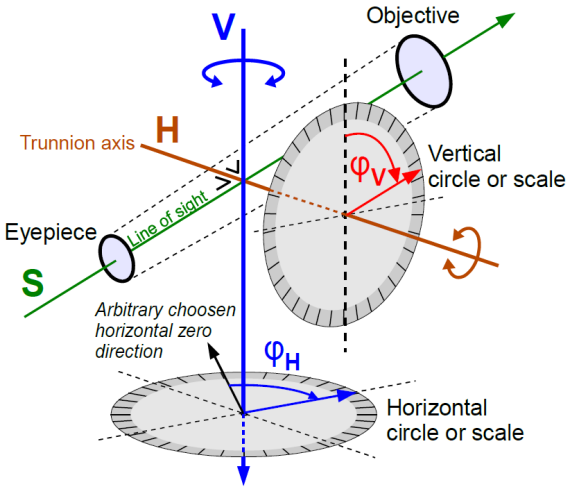



Figure 4.2: Diagram of a theodolite. The vertical (V) axis is set up along the local vertical (gravity, plumb line), the horizontal (H) axis is constructed orthogonal to the vertical axis, and the line of sight (in green) of the telescope is again orthogonal to the horizontal axis. An angle is the difference — looking from one set-up point — between the directions to two objects. At one set-up of the instrument the horizontal and vertical direction are measured to a first object, $\varphi_H$ and $\varphi_V$, and next, the horizontal and vertical direction are measured to a second object. And angles result as the difference of the two directions, both for horizontal and vertical.


Figure 4.3: The set-up of a theodolite, and identically, of a total station: the vertical axis should be set up along the local plumb line, and it points up to the local zenith. Unlike with leveling, it is *crucial* to set-up the instrument *right above* the benchmark or survey-marker in the ground. The vertical axis of the theodolite, when extended downward, should pass right through the survey-marker.

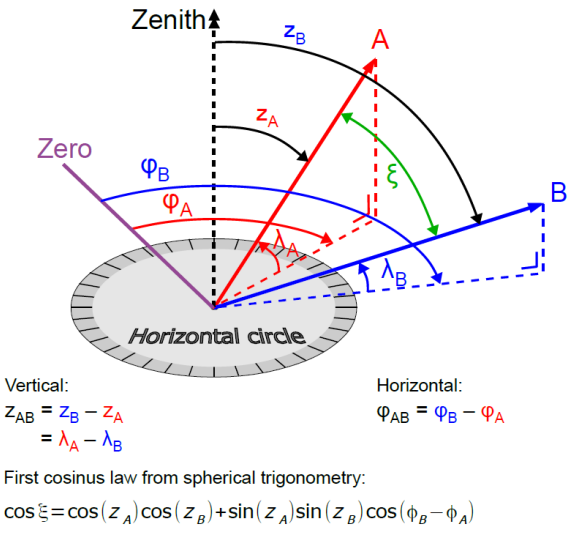



Figure 4.4: Measuring a horizontal and vertical angle between objects A and B with a theodolite.

is equivalent to 400 gon, instead of 360 degrees as the old Babylonian did. So a right angle is 100 gon. There is no sub-division into minutes or seconds, just decimal numbers. A right angle is 100.0000 gon.

This decimal system for measuring angles was adopted by surveyors in continental Europe Today decimal degrees ([gon]) is the way of measuring angles in theodolites and total stations in surveying. However in English speaking countries you will find equipment working with degrees, minutes and seconds, just like the mile, yard, foot and inches are in use there for distances.

Mind that generally, working with angles expressed in radians is mathematically more convenient, rather than angles in degrees or gon.

With a theodolite one can take a reading of a direction up to 0.0001 gon. Total stations may have a slightly poorer resolution for measuring directions (e.g. 0.001 gon), but they are equiped with a distance meter, and that is very practical. Total stations are very popular in surveying, and covered in the next section.

To conclude this section we review, like in the section on the automatic leveling instrument, the most relevant instrumental errors of the theodolite. The vertical axis should be set-up along the local plumb line. This is done using the (plate) bubble on the instrument. Therefore the instrument should have been properly calibrated. If, for some reason, the bubble is mis-aligned, the vertical axis may not be set-up perfectly vertical, and erroneous measurements will result.

Next, the horizontal axis should be constructed orthogonal to the vertical axis, and, the line of sight of the telescope again orthogonal to the horizontal axis (cf. Figure 4.2). The effects of small violations of these two construction requirements, as shown in Figure 4.5, can be eliminated by a special measurement procedure, refered to as measuring 'face left' and 'face right', and taking the average of the two measurements, see later Section 4.6. Directions to a series of objects are measured 'face left', and directions are measured to the same series of objects 'face right'. Then angles (or reduced directions) are formed, each time as the difference of two directions, from the 'face left' measurements, and from the 'face right' measurements as well. The average of a 'face left' and a 'face right' angle is then free from these two instrumental errors, as for example error $\beta$ in Figure 4.5 is involved once with a positive sign, and once with a negative sign.

## 4.2. Total station

A total station basically is an electronic theodolite with a built-in opto Electronic Distance Measurement device (EDM). The instrument is shown in Figure 4.6.

The EDM is integrated in the telescope of the instrument, see Figure 4.7. At the basis is an Infra-Red (IR) Light (laser) Emitting Diode (LED), which sends an amplitude modulated signal. Next, there is an Infra-Red (IR) photo-diode, to receive the reflected signal. The distance measurement then follows as the *phase* difference between the signal generated by the instrument (modulation oscillator) and the signal which is received back by the photo diode.

Figure 4.8 shows how the EDM measures a distance. A signal is sent by the total station, then reflected by a reflector, and received again at the total station, hence the signal travels in total a distance of $2d$ (two-way ranging), see also Section 18.2. Now, the distance traveled can be expressed in terms of the wavelength of the modulation of the signal:

$$2d = n\lambda_m + \Delta\lambda_m$$

The total distance $2d$ contains an integer number $n$ of modulation signal wavelengths $\lambda_m$, and a fractional phase difference $\Delta\lambda_m = \frac{\Phi_0 - \Phi_i}{2\pi}\lambda_m$, with the phase $\Phi$ expressed in radians, as

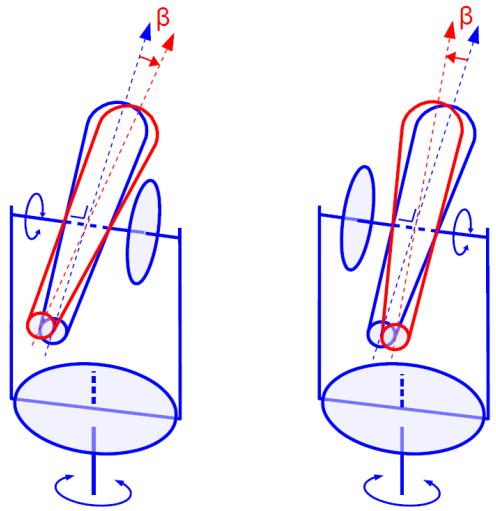



Figure 4.5: Instrumental error of a theodolite: the line of sight (in red) is not exactly orthogonal to the horizontal axis, it is off by a small angle $\beta$. Turning the telescope about the horizontal axis causes the line of sight to describe a cone, rather than a vertical plane. This will cause errors in observed horizontal directions. However, if one turns the telescope by 180 degrees about the horizontal axis, and then turn the theodolite by 180 degrees about its vertical axis, the same horizontal direction can be measured, but now mis-alignment error $\beta$ works the other way. The procedure is referred to as measuring both in face-left and face-right position.

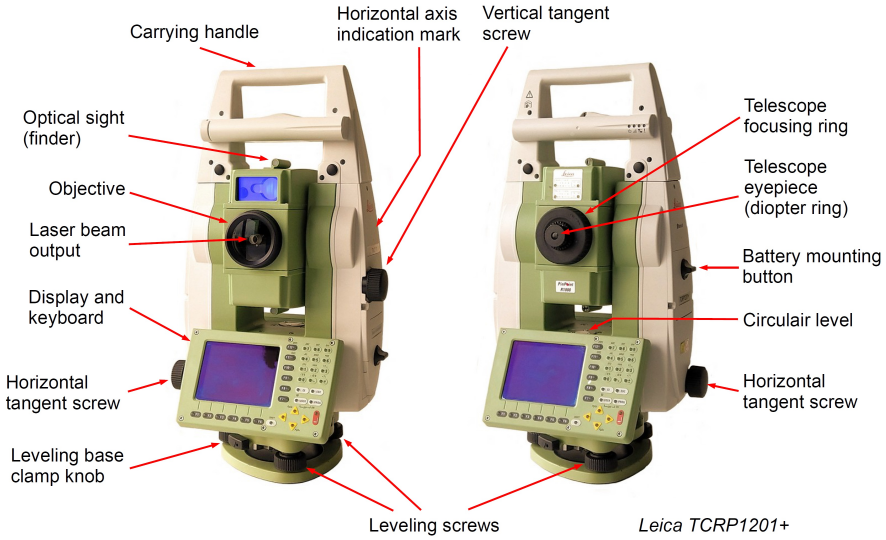



Figure 4.6: Overview of total station, Leica TCRP 1201+.

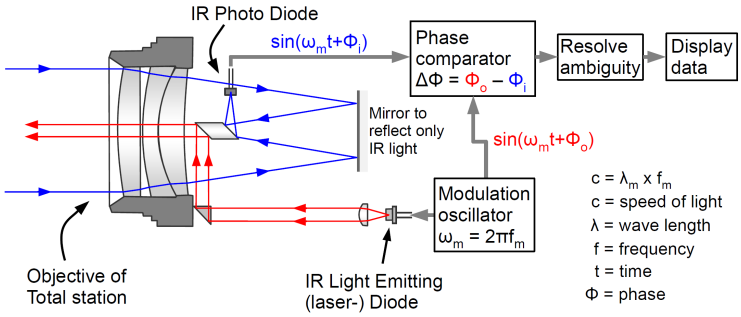



Figure 4.7: The Electronic Distance Measurement (EDM) instrument integrated in the telescope of a total station. The Infra-Red light is emitted (in red) and the received signal (in blue), which has been reflected by a reflector held at the point of interest, is captured by a photo diode. Essentially the signal travel-time is measured, and multiplied by the speed of light. This, divided by two, yields the distance between instrument and reflector.

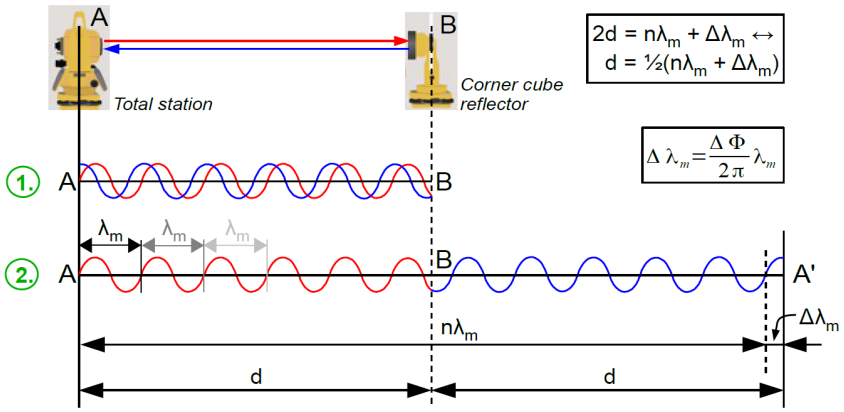



Figure 4.8: Phase comparison measurement technique of Electronic Distance Measurement (EDM) instrument. The distance to be measured between total station and reflector is $d$. The red wave is the transmitted signal, the blue wave the reflected signal. For clarity the blue wave has been 'unfolded' to the right, arriving at A' in the second line of the graph.

| frequency [Hz] | wavelength [m] | measured $\Delta\lambda[m]$ |
|---|---|---|
| $f_1$ 30 MHz | 10 | 2.845 |
| $f_2$ 3 MHz | 100 | 62.85 |
| $f_3$ 300 kHz | 1000 | 363.0 |
| $2d$ | | 362.845 |

Table 4.1: Measuring distance through phase comparison and using three different frequencies $f_1$, $f_2$ and $f_3$. The highest frequency yields 2.845 (un-truncated), for the middle frequency (with $f_2 = f_1/10$) we have 60, and from the lowest frequency, for which holds $f_3 = f_1/100$, the result of $\mathrm{trunc}(\Delta\lambda_3/100) * 100$ reads 300. Together this yields 362.845 m as a reading for distance $2d$ by the Electronic Distance Measurement (EDM) device. Mind that the same carrier (e.g. IR or laser light) can be used in all cases — the above mentioned frequencies refer to the *modulation* of the carrier (e.g. the IR/laser light is amplitude modulated).

shown in Figure 4.8.

The fractional phase difference $\Delta\Phi = \Phi_0 - \Phi_i$ can be measured by the phase comparator, as indicated in Figure 4.7, and this *fractional* phase difference is invariant to the number of full wavelengths, or cycles, of the signal. In this way one has no clue at all about how many full wavelengths $n\lambda_m$ are included in the measured distance $2d$; one wave looks just like another.

The solution lies in employing different frequencies for modulating the IR-light. For the example in Figure 4.9 we use three different frequencies, namely $f_1$ as the highest frequency, $f_2 = f_1/10$, and the lowest frequency $f_3 = f_2/10 = f_1/100$. The wavelengths are consequently related as $\lambda_1$, $\lambda_2 = 10\lambda_1$, and $\lambda_3 = 10\lambda_2 = 100\lambda_1$. The procedure is to measure $\Delta\lambda$ with all three frequencies, $\Delta\lambda_1$ for $f_1$, $\Delta\lambda_2$ for $f_2$, and $\Delta\lambda_3$ for $f_3$. The distance follows as

$$d = \frac{1}{2}(\Delta\lambda_1 + \mathrm{trunc}(\frac{\Delta\lambda_2}{10}) * 10 + \mathrm{trunc}(\frac{\Delta\lambda_3}{100}) * 100)$$

where $\mathrm{trunc}$ means rounding a positive number down to its nearest integer (i.e. cutting any decimal part). Table 4.1 provides an example. The result of the highest frequency, in this case 2.845 m, is carried un-truncated into the measured distance — the highest frequency yields the most precise phase comparison. The lowest frequency determines the working range of the instrument, in this example, 1000 m, divided by two. In practice, total stations employ four different modulating frequencies for distance measurements. The whole procedure described here is carried out automatically, upon a simple 'press the button' by the surveyor, and the resulting distance appears on the display. The distance measurement accuracy is in the order of one to a few millimeter, plus one or a few ppm (parts per million).

In this section the phase comparison method for measuring a distance was covered. Alternatively a distance can also be measured by determining the travel-time of a short *pulse*. The travel-time, multiplied by the speed of light, and divided by two, directly yields distance $d$, without ambiguity, see Section 20.1.1. The accuracy is however poorer than what can be achieved with the phase comparison method.

## 4.3. Atmospheric refraction

So far, we implicitly assumed that electromagnetic waves (e.g. Infra-Red or laser light to measure distance) travel with the speed of light in vacuum $c$. With surveying these measurements take place in the Earth's atmosphere, and one has to account for the effect of atmospheric refraction. The refractive index $n$ relates the actual propagation speed through the medium $v$, and the speed of light in vacuum $c$:

$$n = \frac{c}{v}$$

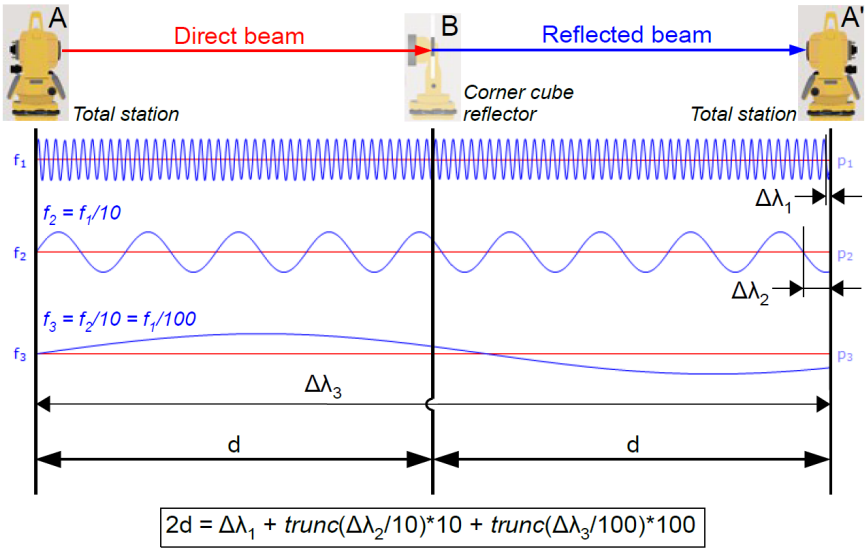



Figure 4.9: Using several frequencies in succession to measure distance $d$, through phase comparison.

see the section on propagation effects in Appendix G. As $n > 1$, we measure a *longer* signal *travel-time* and hence, experience a longer (than actual) distance.

In vacuum, frequency and wavelength are related through $c = \lambda f$, and in medium we obtain, with $c = nv$, $nv = \lambda f$, or $v = \frac{\lambda}{n} f$. Hence the apparent wavelength of a signal with frequency $f$ in the medium becomes $\frac{\lambda}{n}$, rather than just $\lambda$ in vacuum. This is shown in Figure 4.10.

As a side note we state that the refractive index is actually dependent on frequency $f$, an effect known as dispersion. This effect is not further dealt with here.

As outlined in the previous section, the distance is measured by an EDM through phase comparison, and the phase difference yields — apart from the ambiguity — the measured distance $\Delta\lambda_m = \frac{\Phi_0 - \Phi_i}{2\pi}\lambda_m$. If, in the multiplication at the right hand side, the wavelength in vacuum is used, whereas actually the (smaller) wavelength in the medium (atmosphere) should be used (as $n > 1$), the distance output by the instrument is too large.

For visible light, the refractive index of air (at $T$=15 degrees Celsius temperature and $p$=1013.25 mbar atmospheric pressure) is around $n = 1.0003$. The error due to atmospheric refraction is consequently a $3 \cdot 10^{-4}$ effect (300 ppm, i.e. 300 parts per million), hence on a 100 m distance, the effect is 3 cm. In practice, surveying equipment assumes standard or average atmospheric conditions (e.g. $p = 1013.25$ mbar, $T = 15^\circ$ C and a relative humidity of 50%), and only *deviations* from these conditions will cause errors. In general, these errors are small, or very small, in particular when only short distances are measured (up to 100 m), extreme weather conditions are avoided, and measurements are taken relatively close to sea level (not in mountainous areas).

Often, the remaining atmospheric refraction effect, in the order of a few ppm, can be corrected for, by entering the remaining scale factor in the instrument. The firmware in the instrument then applies this scale factor automatically to measured distances. Section A.1 in the appendix briefly elaborates on the physical background of atmospheric refraction. Finally it is wise, like with leveling, to keep sufficient clearance for the signal path from ground and obstacles like walls.

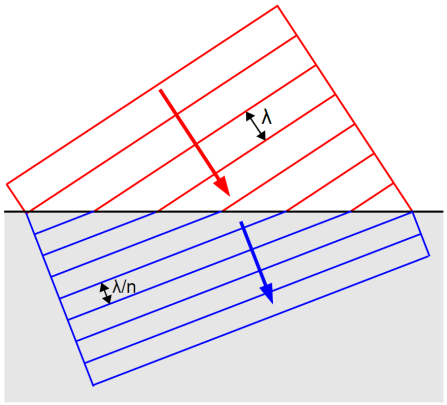



Figure 4.10: When an electromagnetic wave with frequency $f$ from vacuum (in white), enters a medium (in gray) with refractive index $n$, the wavelength apparent in the medium is $\frac{\lambda}{n}$, rather than $\lambda$ in vacuum. The waves are shown by lines of equal phase, see also Figure G.7.

Figure 4.11: Corner cube reflector for EDM measurements with a total station (Leica survey prism GPR121). The reflector can be mounted on a range pole and taken to survey points of interest, or by means of a carrier installed on a tribrach on a tripod.

## 4.4. Corner cube reflector and prism constant

When measuring a distance with the EDM of a total station, a special kind of reflector is usually needed to return the emitted IR-light, see Figure 4.11. This type of reflector is called a corner cube reflector based on its physical shape — a diagonally cut prism from a cube of clear glass (Figure 4.12).

The corner cube reflector has a very interesting property — it reflects an incident light beam back into exactly the same direction where it came from (Figure 4.13). The reflector does not need to be directed exactly to the light source for this.

Figure 4.14 explains the working of the corner cube reflector when it is used with a total station to measure a distance. A signal (e.g. IR-light) is sent by the EDM to the prism. In the reflector the light beam generally reflects three times before it travels back to the EDM. Two important issues should be noted now. First the signal travels a certain distance within the reflector, and second, the material of the reflector is glass and not air, and hence the signal experiences a lower propagation speed inside the reflector, than in air (as $n_{\text{glass}} > n_{\text{air}}$). For the refractive indices holds

$$n_{\text{air}} v_{\text{air}} = n_{\text{glass}} v_{\text{glass}}$$

Figure 4.14 shows a two-dimensional version of the set-up. As the two reflecting surfaces are under 45 degrees, the total path length through the reflector (indicated in blue arrows in

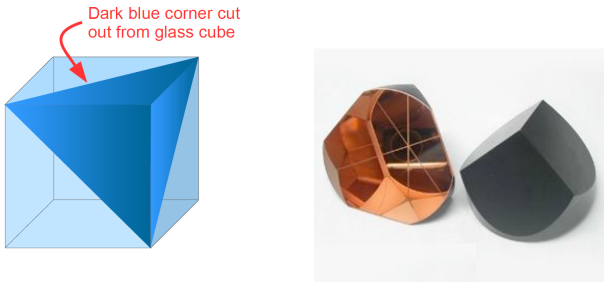



Figure 4.12: The corner cube prism constructed from a cube of clear glass.

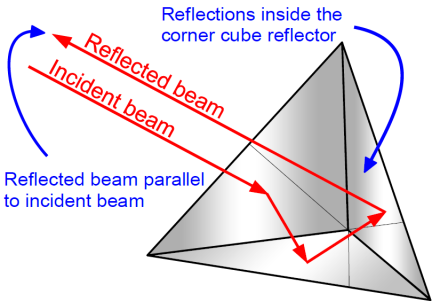



Figure 4.13: Corner cube reflector: incident light is reflected back to its source.

Figure 4.14) equals $d_{\text{prism,glass}} = 2h$, with $h$ being the (physical) depth of the reflector. The total path length through the reflector would be $d_{\text{prism,air}} = 2w$ in case the reflector would be made of air, see the right upper corner of the figure. And the relation reads

$$w = h\frac{n_{\text{glass}}}{n_{\text{air}}}$$

In the same amount of time, the signal travels (with $v_{\text{glass}}$) distance $h$ in glass, and could have traveled equivalently (with $v_{\text{air}}$) distance $w$. As shown in Figure 4.14, $w$ is the distance between the front-side of the reflector and the *apparent* signal reflection point So. If the surveyor is not aware of the fact that the reflector is made of glass, and assumes the signal is traveling all the way just through air, he/she would conclude — based on the obtained distance measurement — that the signal got reflected at point So. The measured distance is *longer*, as $n_{\text{glass}} > n_{\text{air}}$.

Hence, to retrieve correct geometric information from the distance measurement, a correction needs to be applied for using the reflector. We introduce yet another offset, the so-called *prism-constant*. In practice the reflector needs to be positioned at the point of interest, and as the reflector is not an infinitesimal small object, we need to define a point of reference on this object. Most logical would be to use the point So as the point of reference, see Figure 4.15 at left. Positioning So above the point of interest immediately yields the correct distance and we do not need to care about the prism being constructed of glass (instead of air). In practice however, often a so-called center of symmetry, Sc, is defined with the prism, which is a distance $C_{\text{prism}}$ away from So. The point Sc is used to connect the prism to a range pole, see Figure 4.15 at right. Since the range pole is positioned over the point of interest, one needs to correct the measured distance for the offset between Sc and So.

Figure 4.16 shows an example of a reflector with a zero prism constant; the vertical axis of reference coincides with the apparent reflection point So.

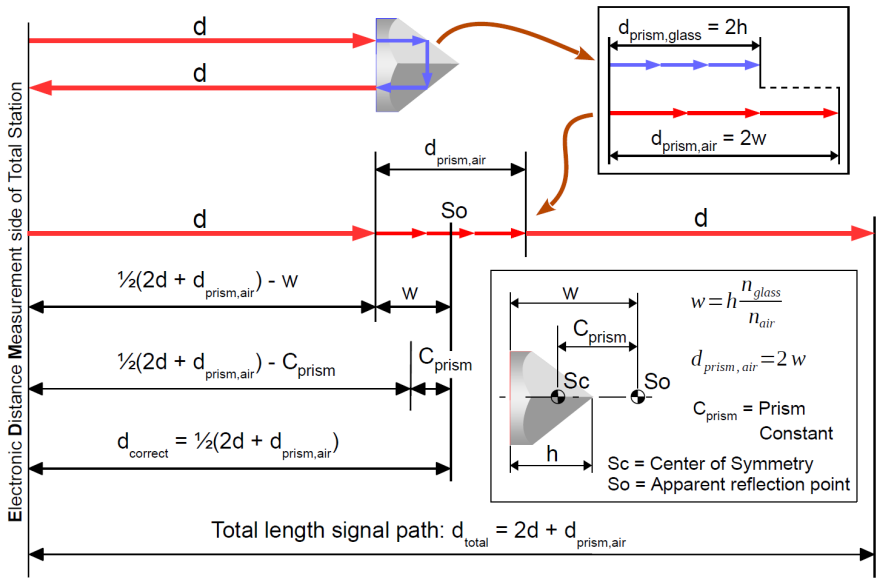



Figure 4.14: Measuring a distance using a corner cube reflector. Point So is the apparent reflection point in case the reflector would consist of air, and $d_{\text{correct}} = \frac{d_{\text{total}}}{2}$ refers to the distance to this point. Point Sc is the defined center of symmetry of the reflector. The so-called prism constant $C_{\text{prism}}$ is the distance between Sc and So.

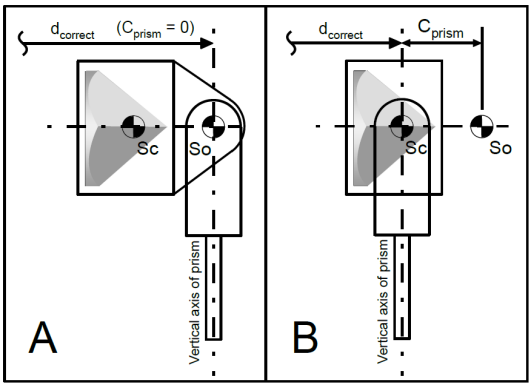



Figure 4.15: In practice, corner cube reflectors are either mounted on their apparent reflection point So (left), or on their center of symmetry Sc (right). For the case at left, no correction is needed to the measured distance (the one-way distance equals $d_{\text{correct}} = d + w$), as the distance — assuming an all-air signal path — directly refers to point So. For the case at right a correction through the so-called prism constant $C_{\text{prism}}$ is needed. The one-way distance equals $d + w$, and should be corrected by $C_{\text{prism}}$ in order to make this distance refer to point Sc. The prism constant should be subtracted from the measured distance: $d + w - C_{\text{prism}}$ in order to obtain the proper geometric distance between the total station and the point of reference of the reflector Sc.

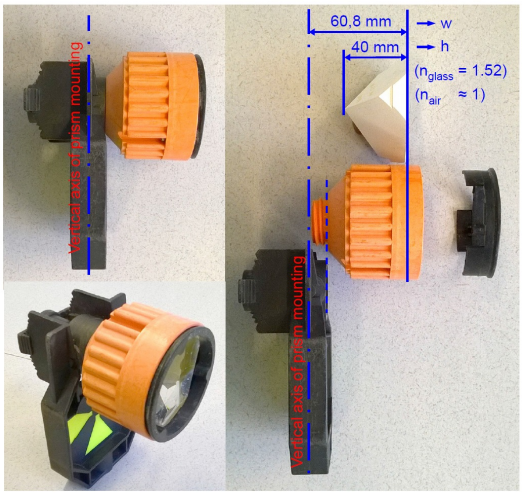



Figure 4.16: A corner cube reflector mounted on its apparent point of reflection So, and consequently $c_{\text{prism}} = 0$.

The center of symmetry Sc of a corner cube reflector is not its geometric center, nor its center of mass. It is the point where, if the prism is not exactly directed to the total station, a minimum line of sight error $\xi$ is made; Figure 4.17 illustrates this.

The value for the prism constant $c_{\text{prism}}$ is generally provided by the equipment manufacturer. And often this value for the prism constant is already implemented in the instrument's firmware, hence, the resulting measured distance presented by the instrument got already corrected for the prism-constant. A word of caution is in order of course, when the total station is used with a different prism. Further practical information with regard to the definition of the prism constant is given in Section A.2.

The measured distance in the instrument is corrected such that apparently the IR-light emitting diode and the photo diode of the EDM coincide with the vertical axis of the instrument, as this is the point of reference of the total station, see also Figure 4.20.

Figure 4.18 at left shows a so-called 360-degrees prism, which actually consists of six glass corner cubes neatly put together. This prism can return the signal from the EDM coming from any (horizontal) direction, which is particularly convenient when continuously tracking a moving object with a robotic total station.

Instead of using optical corner cube prisms, one can use, with laser ranging, simple self-adhesive retro-reflective targets, see Figure 4.18 at right, which can stay on the object of interest for (permanent) monitoring purposes, for instance during construction works.

Finally we mention that in practice, short distances (up to one hundred meters) can be measured even without a dedicated reflector. With these so-called reflector-less measurements, the objects of interest, for instance walls of buildings, reflect part of the laser-light back to the total station. Practically this is very convenient as the object of interest then does not need to be visited with a reflector, and the object can be measured remotely from the total station set-up. With *reflector-less* measurements typically a red laser is used, see also Figure 18.6.

## 4.5. Trigonometric leveling

Total stations are most often used for surveying in the local horizontal plane, for instance to produce a map with 'where objects, topography and infrastructure are'. By measuring directions and distances to points of interest, the coordinates of these points can be determined.

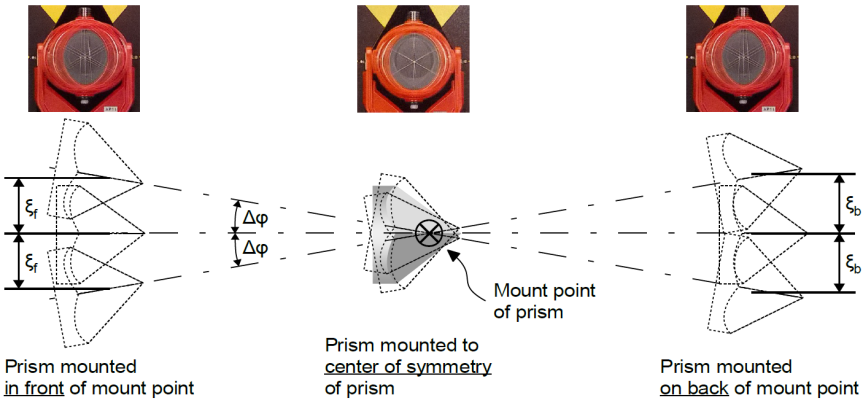



Figure 4.17: A surveyor aims the telescope of the total station at the center of the prism. If the prism is not directed exactly to the total station, an error may occur in this aiming, this is shown at left and at right with a line of sight error $\xi_f$ and $\xi_b$ respectively; in these cases when the center of rotation lies either (far) behind, or in front of the reflector, the reflector gets also displaced when not properly directed. When the prism is mounted at its center of symmetry, the error due to not exactly directing the prism to the total station is minimized.

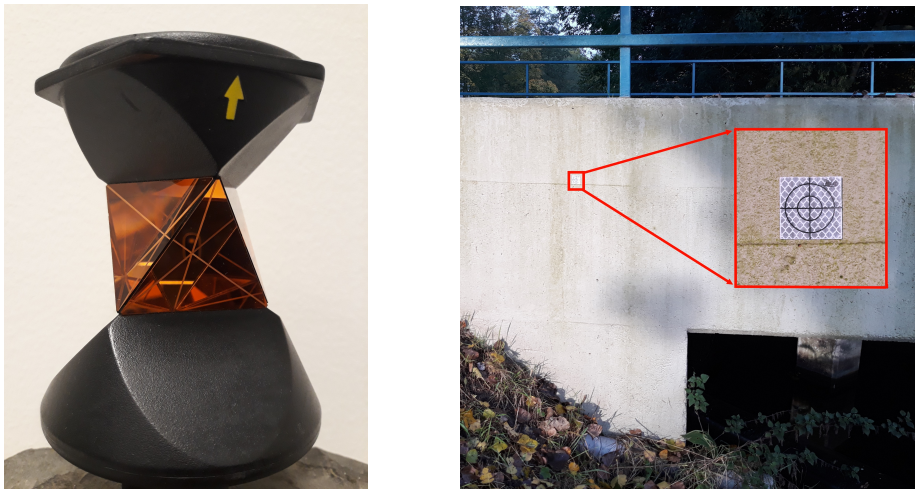


Figure 4.18: At left: 360-degrees prism (Leica type GRZ4) which returns the EDM signal coming from any horizontal direction. At right: self-adhesive retro-reflective target, measuring 4 cm x 4 cm, fixed on a concrete bridge.

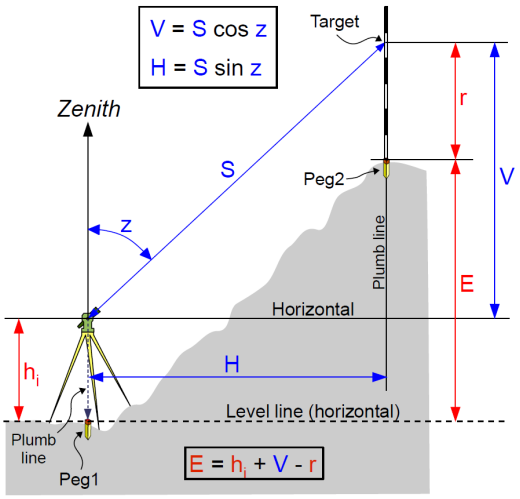



Figure 4.19: In a simple local situation, the height difference between the two points is easily obtained through trigonometric leveling.

With the observed vertical direction, one can also determine a height difference. In this section we briefly address trigonometric leveling, and we consider only the simple *local* situation, see Figure 4.19.

In Figure 4.19 the zenith angle $z$ to the target, point 2, is measured, as well as the (slant) distance $S$. These two measurements can be converted into the horizontal distance $H$ and vertical distance $V$. Accounting for the height of the total station above the marker (Peg1), $h_i$, and the height of the target reflector above the marker (Peg2), $r$, the geometric height difference $E$ between the two points follows as

$$E = h_i + V - r$$

In order to apply trigonometric leveling over larger distances, one has to account for the curvature of the Earth and atmospheric refraction, similar as with leveling in Chapter 3, and one should realize that trigonometric leveling is a geometric method, which is dissociated from gravity. Only over short distances (just like leveling), trigonometric leveling provides approximations to orthometric height differences.

## 4.6. Measurement procedure

In this section we outline the actual measurement procedure with a total station in the field. It consists of setting up the tripod on which we mount the instrument. Then we level the instrument and position it right above the marker in the ground, using a plummet, and we focus the reticle. Eventually we focus the telescope on the reflector and take the measurements (vertical, and horizontal direction and distance).

### 4.6.1. Set-up the tripod

The set-up of the tripod has been described, with leveling, in Section 3.6. One distinct difference is now that the theodolite or total station (usually) should be positioned exactly over the benchmark or survey marker in the ground. Hence, place the tripod approximately over the benchmark and make sure that the tripod top plate is approximately horizontal, and fix the legs, see Figures 3.17 and 3.18.

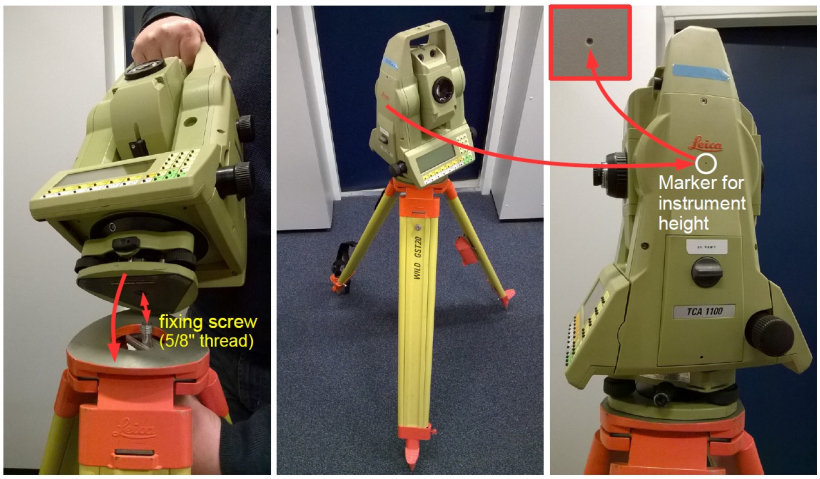



Figure 4.20: Mounting a total station on a tripod. Note, in the photograph at right, the marker on the instrument to measure the height of the instrument above the benchmark in the ground. The marker on the instrument coincides with the horizontal axis. The instrument height should be measured, with a tape, once the instrument set-up has been completed.

### 4.6.2. Mounting total station on tripod

Mounting a total station on a tripod goes in a very similar way as for a leveling instrument, see Figure 4.20. Fix the total station with the big screw to the tripod.

### 4.6.3. Level the instrument, and center it

As stated in Section 4.1, the theodolite or total station should be set up such that the vertical axis is aligned with the local plumb line, and, the instrument has to be positioned such that the vertical axis, when extended downwards, hits the benchmark or survey marker in the ground. The latter can be achieved in three ways, using an actual plummet, an optical plummet or a laser plummet, see Figure 4.21.

The optical plummet is most often used, and we present further details on its use, see Figure 4.22. The optical plummet is built in the tribrach on which the total station is mounted. It is a kind of (small) telescope viewing in a right angle to the ground. It is calibrated such that the viewing direction is downward and exactly perpendicular to the disc surface of the tribrach. Hence, using the circular level in the tribrach, the viewing direction of the optical plummet can be aligned with the local vertical (along the local plumb line). The steps of focusing the optical plummet telescope are the same as with leveling in Chapter 3. First adjust the reticle focusing-ring to get a clear and sharp view on the cross-wires. Then focus on the target point by adjusting the target focusing-ring, see Figure 4.22 at right.

Adjust the leveling screws of the tribrach so that the plummet gets aligned with the marker on the ground. Likely the instrument is not exactly level anymore after this adjustment. Now, adjust the lengths of the tripod legs (and *not* the leveling screws) in order to level again the instrument, using the circular level on the tribrach.

Next, the leveling of the instrument is fine tuned. This is done using a more sensitive level built-in the total station. This level can be a physical cylinder type spirit level or an electronic one — the latter one being accessible through the firmware of the total station. The fine tuning of the leveling is demonstrated in Figure 4.23.

By the fine tuning of the level of the instrument, the instrument may not be centered exactly over the survey marker in the ground. In order to repair this, release the central fixing screw a little bit so that you can move the total station — though within limited amount —

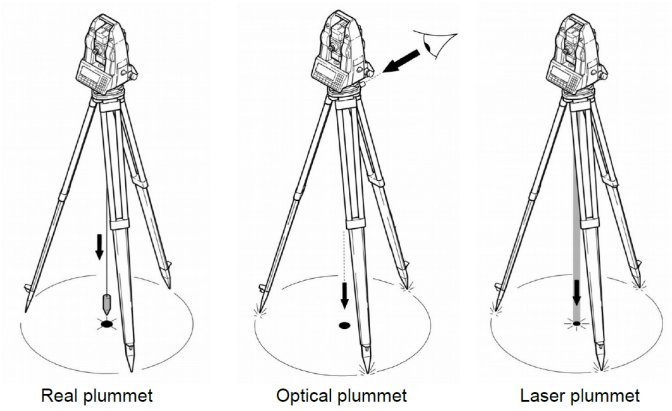



Figure 4.21: Three different implementations of a plummet, in order to get the total station positioned right above the benchmark or survey marker in the ground. Illustration courtesy of ©Leica Geosystems AG, Heerbrugg, 2021 [20].

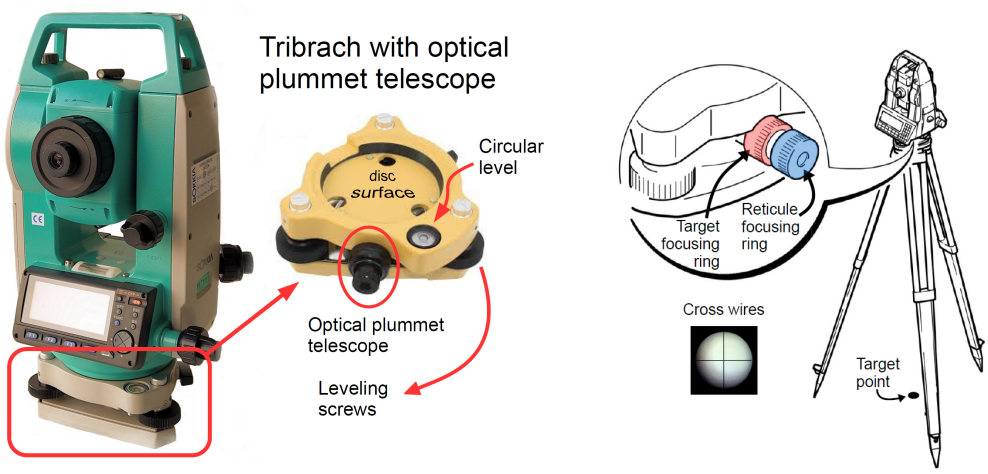



Figure 4.22: The tribrach with a built-in optical plummet. Image at right derived from illustration by ©Leica Geosystems AG, Heerbrugg, 2021 [21], with permission.

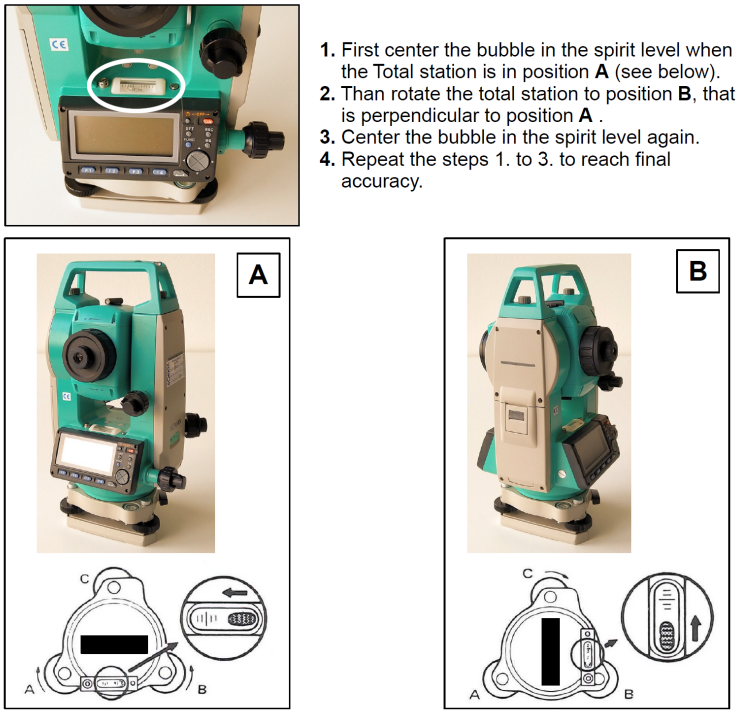



Figure 4.23: Fine tuning the level of the total station. Position the spirit level such that it is aligned with two of the leveling screws of the tribrach, screws A and B. Use the screws A and B in equal amount to get the bubble of the spirit level in the center. Next turn the instrument by 90 degrees and use leveling screw C to center the bubble. And repeat the procedure to check, or to refine the leveling of the instrument.

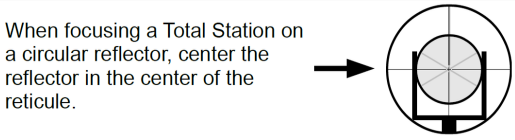



Figure 4.24: Aiming the telescope of the total station on the center of the reflector.

carefully over the top flat surface of the tripod. Shift the instrument such that the plummet indicates that the instrument is again centered exactly above the survey marker in the ground. And tighten the fixing screw again.

Finally, check the level as described above with fine tuning the level. And repeat this step, as well as shifting the instrument over the top flat surface of the tripod. If the spirit level is fine in any direction, and the instrument is centered exactly over the marker, one can measure the instrument height, cf. Figure 4.20 at right, and then proceed to the next step.

### 4.6.4. Focus the reticle

This is done exactly the same way as with the leveling instrument in Section 3.6, Figure 3.21.

### 4.6.5. Point instrument to reflector

This is done similarly as with the leveling instrument in Section 3.6. Make sure that the cross-wires in the telescope coincide with the center of the reflector, see Figure 4.24.

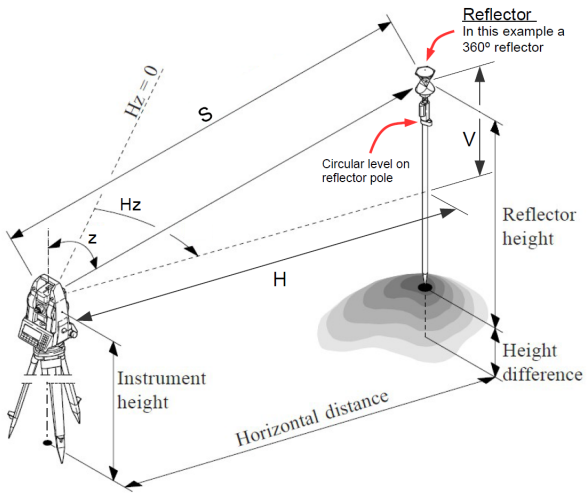



Figure 4.25: Measurement set-up with a total station. The vertical, or zenith angle 'z' is measured with respect to local zenith, horizontal direction 'Hz' is measured, with respect to an arbitrary zero-direction, and the slope distance 'S' is measured. The reflector is a so-called 360 degrees reflector, which returns the EDM measurement signal from any direction (Figure 4.18 at left). Image derived from illustration by ©Leica Geosystems AG, Heerbrugg, 2021 [20], with permission.

### 4.6.6. Focus the telescope

This is done similarly as with the leveling instrument in Section 3.6, Figure 3.23.

### 4.6.7. Take measurements

Figure 4.25 shows the general measurement setup of the total station. Two angles are measured (theodolite function), the horizontal angle 'Hz' and the vertical, or zenith angle 'z'. The slope or slant distance measurement 'S' results from the EDM function of the total station. A total station is an electronic instrument, and after pointing the telescope on the center of the corner cube reflector, the only remaining action is to push the measurement button to obtain the measurements. For surveying in the local horizontal plane, the vertical angle may not be needed. Typically it is used by the instrument's firmware to output also the *horizontal* distance $H$, which is computed internally, using 'z' and 'S'

Generally not a single point of interest is surveyed, but instead a whole series of points. The assistant of the surveyor will visit all the points of interest, and occupy them with the reflector to allow the surveyor to take measurements. The procedure is shown in Figure 4.26.

Optionally, the zero direction of the horizontal scale can be set, at the start of the survey. One can choose for instance a point left of all other points and set the reading of the horizontal angle to zero. In some cases the zero direction is set intentionally when the total station is pointed to a particular object of interest, or reference. When carrying out a survey with a tachymeter, for instance at a construction site, it is recommended to include measurements to a distant object — outside the construction site — for the purpose of verification and/or (later) reconstruction of the survey.

Then, the actual survey can start, by directing the telescope to the reflector at point 1, take measurements 'Hz', 'z' and 'S', repeat this for points 2, 3 and 4, in Figure 4.26.

Then, the telescope should be turned to the 'other face'. This means turning the telescope by 180 degrees about the horizontal axis, and then, turning the whole instrument by 180 degrees about its vertical axis, see Figure 4.27.

The reason for measuring with the telescope in 'face left' and in 'face right' position (or

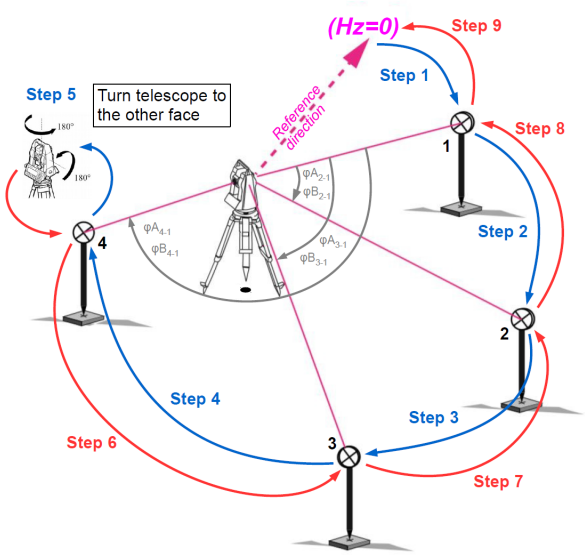



Figure 4.26: Example of taking a series of measurements with a total station at one set-up. Points 1 through 4 with reflectors installed on top of them, are surveyed, once in forward, and once in backward way, indicated by the steps 1 through 9. Image derived from illustration by ©Leica Geosystems AG, Heerbrugg, 2021 [22], with permission.

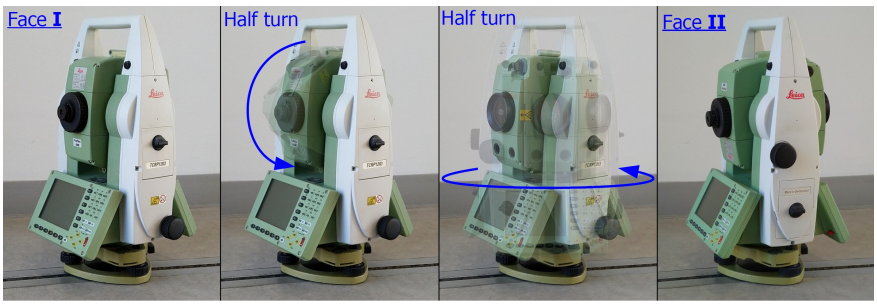
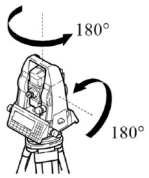



Figure 4.27: Turning the telescope to the 'other face'.

| **Face I** | Hz | z | S ( ) | H ( ) | V ( ) |
|---|---|---|---|---|---|
| | gon (grad) | gon (grad) | [m] | [m] | [m] |
| 1 | | | | | |
| 2 | | | | | |
| 3 | | | | | |
| 4 | | | | | |




| **Face II** | Hz | z | S ( ) | H ( ) | V ( ) |
|---|---|---|---|---|---|
| | gon (grad) | gon (grad) | [m] | [m] | [m] |
| 4 | | | | | |
| 3 | | | | | |
| 2 | | | | | |
| 1 | | | | | |

Figure 4.28: Total station measurement registration form. The first column Hz contains the horizontal directions, the second column z the vertical, or zenith angles, the third column S the slope distances, and columns 4 and 5 contain the horizontal H and vertical V distance. Parts A and B correspond to the two 'faces' of the instrument.

face I and face II) was explained in Section 4.1 with Figure 4.5.

Next, the same points are measured again, but in reverse order, hence points 4, 3, 2, and 1.

Figure 4.28 shows an example of a measurement registration form for the set up in Figure 4.26.

Theodolite procedure

| **Target** | **Direction** | **Reduced direction** | | | |
|---|---|---|---|---|---|
| ***Face I*** | gon (grad) | | gon (grad) | | |
| Hz.1 | | $\varphi FI_{1-1}$ = Hz.1 - Hz.1 | *0.000* | **Error** | gon (grad) |
| Hz.2 | | $\varphi FI_{2-1}$ = Hz.2 - Hz.1 | | $\Delta_{2-1} = \lvert\varphi FI_{2-1} - \varphi FII_{2-1}\rvert$ | |
| Hz.3 | | $\varphi FI_{3-1}$ = Hz.3 - Hz.1 | | $\Delta_{3-1} = \lvert\varphi FI_{3-1} - \varphi FII_{3-1}\rvert$ | |
| Hz.4 | | $\varphi FI_{4-1}$ = Hz.4 - Hz.1 | | $\Delta_{4-1} = \lvert\varphi FI_{4-1} - \varphi FII_{4-1}\rvert$ | |
| | | | | | |
| ***Face II*** | gon (grad) | **If φFII < 0 → φFII = φFII + 400** | gon (grad) | | |
| Hz.4 | | $\varphi FII_{4-1}$ = Hz.4 – Hz.1 | | **Average** | gon (grad) |
| Hz.3 | | $\varphi FII_{3-1}$ = Hz.3 – Hz.1 | | $\varphi_{2-1} = (\varphi FI_{2-1} + \varphi FII_{2-1})/2$ | |
| Hz.2 | | $\varphi FII_{2-1}$ = Hz.2 – Hz.1 | | $\varphi_{3-1} = (\varphi FI_{3-1} + \varphi FII_{3-1})/2$ | |
| Hz.1 | | $\varphi FII_{1-1}$ = Hz.1 – Hz.1 | *0.000* | $\varphi_{4-1} = (\varphi FI_{4-1} + \varphi FII_{4-1})/2$ | |

Figure 4.29: Total station measurement registration form — backside. The first column with directions is copied from the fore-side. Next, the reduced directions are computed, these are the horizontal directions with respect to the horizontal direction of the first point (hence, actually horizontal angles). Parts A and B correspond to the two 'faces' of the instrument; the angles have been measured twice. Differences between the two measurements are computed, and noted in the last column under 'Error', and the averages of the two measurements are computed, and stored under 'Average', these are the input for further data processing.

Typically on the back of the measurement registration form one performs the calculations for the so-called reduced horizontal directions, the errors and the averaged angles, which are eventually used in further processing. The back of the measurement registration form is shown in Figure 4.29.

## 4.7. Exercises and worked examples

This section presents two exercises, one on working with a theodolite, and one on the principle of the Electronic Distance Measurement (EDM) instrument.

**Question 1** With a theodolite two directions have been measured, to target points A and B, both in 'face left' and in 'face right' mode. The measurements are listed in Table 4.2. What is the value of symbol 'X'?

| face-left | target | direction |
|---|---|---|
| | A | 89.762 |
| | B | 139.468 |
| face-right | target | direction |
| | B | 339.475 |
| | A | X.760 |

Table 4.2: Simplified theodolite measurement form with just observed horizontal directions. Measurements are given in gon.

**Answer 1** Going from face left to face right implies a(n about) 200 gon difference in the horizontal directions, see Figure 4.27. We move up by 200 gon on the horizontal scale of the instrument. The horizontal direction to point A was 89.762 gon in face-left, and hence, in face-right, this becomes 89.762+200.000=289.762 gon, or, as indicated in Table 4.2, 289.760 gon (as target B has been measured again). Hence X means 289.

**Question 2** A distance can be observed by measuring the travel-time of a radio or optical signal. An Electronic Distance Measurement instrument transmits an infrared or laserlight, and this light travels forth and back to a reflector. The light has been amplitude modulated by a 15 MHz pulse signal. What is the distance between the EDM and the reflector, if, inside the EDM the phase difference between the modulation of the outgoing and received signal is $\frac{\pi}{2}$ rad, i.e. the received signal (pulse) arrives a quarter wavelength later, compared to the direct signal?

**Answer 2** This question concerns the working principle of the Electronic Distance Measurement (EDM) unit, as shown in Figure 4.7. It is given that the phase comparator yields a phase difference $\Delta\Phi$ of $\frac{\pi}{2}$ rad. This means that the detour of the signal to the reflector and back, causes the incoming blue signal to be late. The wavelength $\lambda_m$ is easily found through $c = \lambda_m f_m$, with $f_m$=15 MHz and $c = 3 \cdot 10^8$ m/s; the wavelength becomes $\lambda_m$=20 m (and we assume here that the refractive index in the Earth's atmosphere is $n \approx 1$). The detour, in terms of distance, is $\Delta\lambda_m = \frac{1}{4}\lambda_m$ = 5 m (hence a quarter wavelength). Next, we have that the two-way distance equals this quarter wavelength: $2d = \Delta\lambda_m$ (where so far, we assumed that there is no ambiguity involved, hence $n = 0$), and thus $d$=2.5 m; this is the single way detour of the signal to the reflector. Accounting for the ambiguity according to $d = \frac{n}{2}\lambda_m + \frac{1}{2}\Delta\lambda_m$ with integer $n$, the distance is 2.5 m, 12.5 m, or 22.5 m, or …. The blue signal is late compared to the red signal, in terms of time, by $\frac{1}{4}\frac{1}{f_m}$ = 16.7 ns (nanoseconds), in case $n = 0$, as the period $T$ of a periodic signal is $T = \frac{1}{f}$.

# II

# Mathematical geodesy

# 5

# Introduction

The development and application of mathematical theory needed to process, analyze, integrate and validate geodetic data, such as measurements for land surveying, is referred to as the discipline of *mathematical geodesy*. It is about the calculus of observation (in Dutch: waarnemingsrekening), the validation of measurements and mathematical models, and the analysis of spatial and temporal phenomena in geosciences, i.e. about parameter estimation, testing and reliability, and about interpolation and prediction.

An important step in this discipline was made with the discovery of the method of least-squares in the 18th century [23]. It occurred in the fields of astronomy and geodesy, as scientists and mathematicians sought to provide solutions to the challenges of navigating the Earth's oceans. The German Carl Friederich Gauss (1777-1855) proposed the method of least-squares, which allows to make an optimal combination of *redundant* measurements for the purpose of determining values for parameters of interest [23], and as such a solution to an *inconsistent* system of equations. The French mathematician Adrien-Marie Legendre (1752-1833) in 1805 published his work on the method of least-squares (a translation of the French term 'méthode des moindres carrés') in the context of the determination of the orbits of comets (Nouvelles méthodes pour la détermination des orbites des comètes). Later, in 1809, in a volume on celestial mechanics, Gauss published the least-squares method, and claimed he had been using this method already back in 1795, hence earlier than Legendre, although the latter published it first [23]. Many studies and investigations have been spent on this dispute between Legendre and Gauss, but generally it is believed that Gauss should indeed be regarded as the first inventor of the method of least-squares. Legendre came to the same method independently, and delivered a clear publication of it. Later, Gauss provided a probabilistic justification of the method of least-squares, and proposed the normal or Gaussian distribution. Today, more than 200 years later, the least-squares method is frequently used in a wide variety of scientific and engineering applications.

Further developments in mathematical geodesy followed on advances in the early 20th century in the field of statistical inference, which is about drawing conclusions from data which are subject to random variation, for example observational errors or sampling variation. Statisticians R.A. Fisher (1890-1962), J. Neyman (1894-1981) and E. Pearson (1895-1980) introduced concepts such as statistical confidence and hypothesis testing [24].

## Delft perspective on mathematical geodesy

Delft University of Technology professors J.M. Tienstra (1895-1951) and W. Baarda (1917-2005), see Figure 5.1, founded the 'Delft school' of mathematical geodesy. World-renowned

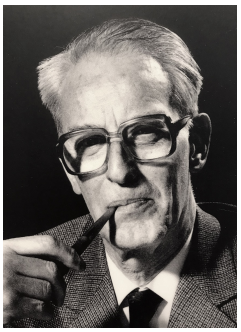



Figure 5.1: Delft University of Technology professors in mathematical geodesy Jacob Tienstra (1895-1951) at left and Willem Baarda (1917-2005) at right. Photo at left anonymous, from 'Persoonlijkheden in het Koninkrijk der Nederlanden in woord en beeld', Amsterdam, 1938, p. 1470; taken from Wikimedia Commons [9], Public domain. Photo at right by Axel Smits [25].

contributions have been made in Delft on the subject of statistical hypothesis testing, and theory was developed, for instance leading to the concept of reliability, and applied to land surveying networks, enabling the design and analysis of these networks. Later, professor P.J.G. Teunissen made extensions to the (automated) data quality control of dynamic measurement systems in navigation and satellite navigation (GPS) applications. In recent years substantial developments have taken place with regard to the theory for data processing and analysis of interferometric measurement techniques, such as high-precision GPS and radar interferometry, while in parallel a clear framework on prediction theory was set up, covering the subject of interpolation. Furthermore the reliability theory is being extended to risk evaluation for safety-critical applications, such as the navigation of unmanned vehicles, the monitoring of landslide and subsidence, and other (natural) hazards.

## Overview of this part

In the previous part it has been outlined how land surveying measurements are acquired by means of leveling and tachymetry. In Parts III and IV we continue the exposition of the acquisition of measurements by means of GPS and a wide range of remote sensing techniques. The present part, loosely speaking, is about extracting the information of interest from those measurements. The measurement process is cast in a mathematical model, and based on the measurements, values are determined for the parameters of interest, most often position coordinates; after all, surveying and mapping is about, again loosely speaking, 'what is where' and 'where is what'.

Chapters 6 and 7 present a refresher on basic probability and statistics, and Chapter 8 introduces the surveyor's workhorse: least-squares parameter estimation. Chapter 9 elaborates on its application in surveying, and Chapter 10 provides an approach of validating the measurements for the purpose of measurement quality control. Chapter 11, titled 'interpolation', is about taking measurements of a spatial phenomenon, for instance depth measurements by echo sounding to determine the seafloor bottom, with these measurements being taken at certain locations, and then wanting to know about the spatial phenomenon at another location.

# 6

# Random variable

## 6.1. Introduction

Suppose we would like to measure the inner width of a pre-fab tunnel element, just delivered to the construction site, in order to check whether it has been built according to requirements and specifications. To measure the distance in between the two concrete sidewalls, we use a decent laser distometer, see Figure 6.1. We take a measurement and the result reads 7.451 m. We ask a colleague to do the same, and when he/she does so, the outcome is 7.453 m. When a second colleague does, we get 7.454 m, and a third colleague gets 7.452 m, and so on. The point is that such a measurement is not perfect. If a series of measurements is carried out, even under unchanged circumstances, we will see a certain *spread* in the results; the measurements do not give all the same, exactly true, answer. This is for a number of reasons. First of all, the instrument is an electronic device, with (free) electrons moving around in its circuits (as we are not operating at zero Kelvin temperature); they will cause (small) errors in the reading of the distance.

Second, the concrete wall on one side to which we hold the instrument will not be perfectly smooth, and so will be the other wall, which has to reflect the laser pulse, used for measuring the distance. This may cause small differences between two measurements when we hold the instrument, take a measurement, remove the instrument, and put the instrument again etc. In addition, you may not hold the instrument completely still during the measurement. Finally, there are external circumstances which can cause errors in the measurement process, such as reduced visibility in the tunnel, due to dust and smoke, and motions and vibrations of the tunnel element itself. Generally, conditions at a construction site are sub-optimal … for

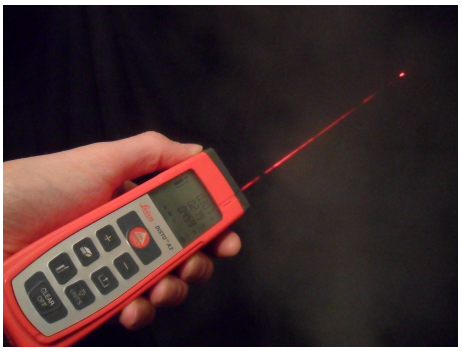


Figure 6.1: Laser distometer for measuring distances.

carrying out accurate measurements. A measurement does not give exactly the true answer; but, hopefully, its value is close though.

## 6.2. Random variable

The whole exercise of using a certain instrument, carrying out the measurement in a certain way (e.g. holding the instrument here, or there) and obtaining the result, is captured - in a mathematical sense - in a random variable. The bottom line is that the outcome of our measurement (generally) will be close to the desired, true value (which we do not know), but it will not be perfect - it will contain some amount of error. In the sequel, we will use the following notation: $y$ for the measurement value, $x$ for the unknown (true) distance, and $e$ for the (random) measurement error, so that we have

$$y = x + e \tag{6.1}$$

A random variable is a mathematical concept, it is denoted by a symbol with an underscore, such as $\underline{y}$. We can obtain a realization of this random variable, by actually taking a measurement, a sample $y$, or with an index $y_1$, where the index denotes that this is the first measurement. We can repeat the measurement a number of times to obtain $y_1$, $y_2$, ...., $y_N$. Later, we refer to $\underline{y}$ as the *observable*, the 'thing which *can* be observed', and $y_i$ is one of the *observations* ($y_i$ has a single, fixed value, namely the numerical outcome of that specific measurement, for example $y_3$=7.450 m).

Then (6.1) can be written as

$$\underline{y} = x + \underline{e} \tag{6.2}$$

The observable $\underline{y}$ equals the true, but unknown $x$ (distance, for instance), plus a *random measurement error* $\underline{e}$, and $e$ is referred to as an un-observable statistical error (we will never know this error in practice). We typically assume that the random measurement error $\underline{e}$ is — on average — equal to zero. Individual realizations $e_1, e_2, \ldots, e_N$ are not equal to zero, but the average over a large number of realizations of $\underline{e}$ will be close to zero.

The parameters we are trying to measure are of the continuous type. The unknown inner width of the tunnel element $x$ is a distance, which can have any real value, hence $x \in \mathbb{R}$. To allow for automated processing of the measured distance, the observation available to the user is presented and stored using a finite number of bits and hence decimals. The observation has been digitized and actually become discrete, though we will not address the subject of quantization here. By approximation, it still is a continuous quantity.

## 6.3. Histogram

When the measurement has been repeated a number of times, all measurements together can be presented in terms of a histogram. The range (or a part of it) of variable $y$ is divided into $k$ intervals (bins or classes) of equal length $h$, the bin width. With a chosen origin $y_o$, we

have the following intervals around $y_o$

$$\begin{array}{ll} j = 1 & [y_o - \frac{k}{2}h, y_o - (\frac{k}{2} - 1)h\rangle \\ \vdots & \vdots \\ \vdots & [y_o - 2h, y_o - h\rangle \\ j = \frac{k}{2} & [y_o - h, y_o\rangle \\ j = \frac{k}{2} + 1 & [y_o, y_o + h\rangle \\ \vdots & [y_o + h, y_o + 2h\rangle \\ \vdots & \vdots \\ j = k & [y_o + (\frac{k}{2} - 1)h, y_o + \frac{k}{2}h\rangle \end{array} \tag{6.3}$$

where we assumed $k$ to be even.

The $N$ samples, assumed to be all in $[y_o - \frac{k}{2}h, y_o + \frac{k}{2}h\rangle$, are divided over the bins. The observed (absolute) frequencies (or cell counts) in the $k$ bins are denoted by $f_j$, with $\sum_{j=1}^{k} f_j = N$.

The histogram, see also Chapter 17 in [2], is now created by plotting

$$\hat{f}(y) = \frac{f_j}{Nh} \text{ with } y \in \text{ interval } j \tag{6.4}$$

as a function of $y$; $\hat{f}(y) = 0$ outside the $k$ intervals.

The function $\hat{f}(y)$ is an interval-wise (width $h$) constant function. Figure 6.2 gives an example. The $\frac{f_j}{N}$ are the relative frequencies, and $h$ in the denominator assures that $\int_{-\infty}^{\infty} \hat{f}(y)dy = 1$, that is, the area under the histogram equals 1. This enables direct comparison of histograms of different data sets (with different bin widths $h$, and sample sizes $N$).

For the set of distance measurements, most of the measurements are close to the value of 7.452 m. The further we go away from 7.452 m, the fewer observed values we see. This is typical behaviour in practice. This behaviour is formalized in a *probability density function* (PDF) of the random variable. The PDF is a mathematical formula, describing the uncertain outcome of our distance measurement. It gives the probability density as a function of the value of the observed parameter $y$. The PDF describes the distribution of the random variable, and it is a mathematical model for the histogram.

The theoretical probability density function $f(y)$ can—for comparison—be directly imposed on the histogram $\hat{f}(y)$, as is done in Figure 6.2.

The bin width $h$ controls the amount of 'smoothing'. In practice one has to match the interval $[y_o - \frac{k}{2}h, y_o + \frac{k}{2}h\rangle$ with $y_{\min}$ and $y_{\max}$, and as a rule of thumb one often sets the number of bins to $k = \sqrt{N}$. Also in practice, in order to focus on the core of the distribution in case there are outlying measurements, one may want to match the interval with e.g. $q_{.01}$ and $q_{.99}$, the 1-st and 99-th (empirical) percentiles respectively, see (6.20), rather than $y_{\min}$ and $y_{\max}$.

The histogram is useful for presentation of the data. It gives a first impression of the probability density which lies at the basis of the samples. One might for instance visually judge — preliminary — whether normality is not unlikely. One must however, be very careful. The picture can be manipulated, certain features can be masked or over-emphasized, by the choice of the origin $y_o$ and the bin width $h$.

## 6.4. Probability Density Function

A histogram is useful to visualize a (big) set of repeated measurements. In practice, where 'time is money', one does not want to repeat measurements. Though one still would like to

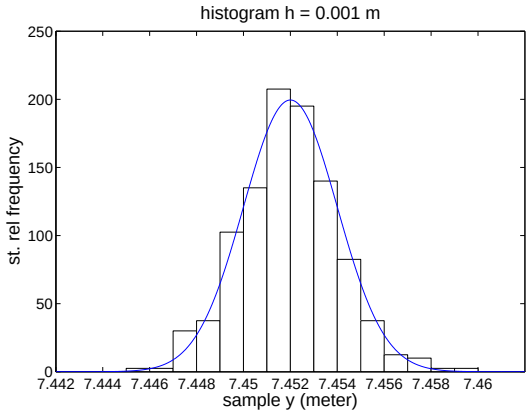



Figure 6.2: Histogram with binsize $h = 0.001$ m, $k = 20$ bins, and center $y_o = 7.452$ m. The vertical axis gives the standardized relative frequency. The data, with sample size $N = 400$, were generated here from a normal distribution with $x = 7.452$ m and $\sigma = 0.002$ m. The curve of this theoretical probability density function is imposed.

have some measure about the uncertainty that can be expected in the outcome, in particular, how big the chance is that the outcome is off from the truth, by more than a certain amount. Therefore, the histogram is captured by a Probability Density Function (PDF), denoted by $f(y)$, for the random variable $\underline{y}$. It gives a mathematical expression for the probability density, as a function of the value $y$ of the random variable.

A probability density function $f(y)$ has to satisfy two general requirements:

$$f(y) \geq 0 \; \forall \; y \text{ and } \int_{-\infty}^{\infty} f(y)dy = 1$$

The (cumulative) probability distribution function is denoted by $F(y)$ and can be found by

$$F(y) = \int_{-\infty}^{y} f(y)dy \tag{6.5}$$

or the other way around

$$f(y) = \frac{\partial F(y)}{\partial y} \tag{6.6}$$

$F(y)$ is a monotonic non-decreasing function, and provides a mapping from the $\mathbb{R}$ into $[0, 1]$. It holds that the probability $P[\underline{y} \leq k] = F(k)$.

### 6.4.1. Normal distribution

When $\underline{y}$ has a normal or Gaussian distribution, the probability density function reads

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-x)^2}{2\sigma^2}} \tag{6.7}$$

where $x$ is the mean, and $\sigma$ the standard deviation ($\sigma^2$ the variance); a proof can be found in Appendix B.1; the function is completely specified by these two parameters $x$ and $\sigma$. The density is a (unimodal) bell-shaped curve, see Figure 6.3 on the left. Often this distribution is denoted as $\underline{y} \sim N(x, \sigma^2)$, where $x$ represents the mean, and $\sigma^2$ the variance.

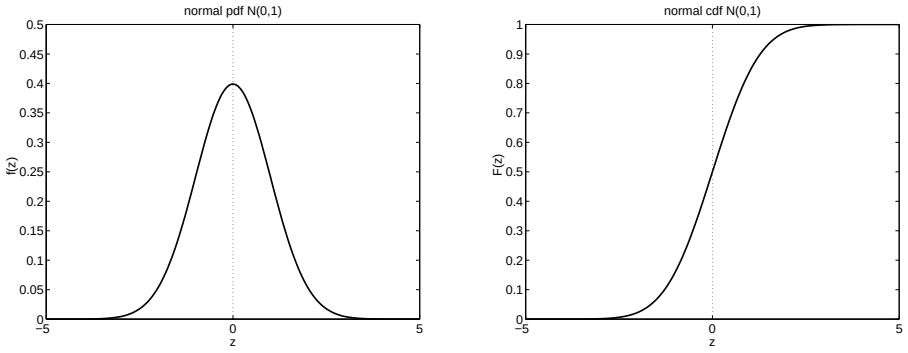



Figure 6.3: Standard normal, or Gaussian distribution, $\underline{z} \sim N(0,1)$ with $x = 0$ and $\sigma = 1$. On the left the probability density function $f(z)$. On the right the cumulative probability distribution function $F(z)$; as variable $z$ runs from $-\infty$ to $\infty$, the probability runs from 0 to 1.

In practice, the manufacturer of the equipment (e.g. of the laser distometer), has analyzed histograms of numerous repeated measurements, in the field, and in the lab, and provides users with information, stating that the error in the measurements will follow a normal distribution, with zero mean (hence on average, the measurement is correct), and a certain standard deviation, e.g. $\sigma$=0.002 m. With the random variables in (6.2) we have $\underline{y} = x + \underline{e}$ and hence, $\underline{e} \sim N(0, \sigma^2)$, and $\underline{y} \sim N(x, \sigma^2)$.

When random variable $\underline{z}$ has the following distribution $\underline{z} \sim N(0,1)$, it is said to be *standard* normally distributed, see Figure 6.3. The cumulative probability distribution function (CDF) $F$ of $\underline{z}$ is also denoted as $\Phi(z)$. Appendix C provides a table of the standard normal distribution. Given are the probabilities $\alpha$, as the right tail probabilities $\alpha = 1 - \Phi(r_\alpha)$, where $P[\underline{y} \leq r_\alpha] = \Phi(r_\alpha)$.

In practice not all observables are normally distributed. There are actually many different probability density functions — the normal one is certainly not the only one. Later, in Section 9.5, the Chi-squared distribution will be introduced.

## 6.5. Moments: mean and variance

In this section we present several characteristic measures of probability density functions. Most common are the mean and the variance. We will consider them from a theoretical (formal) point of view (Section 6.5.1), as well as from a practical (empirical) point of view (Section 6.5.2).

### 6.5.1. Formal moments

The expectation of $\underline{y}$ about some constant $\vartheta$ reads

$$E'(\underline{y}) = \int_{-\infty}^{+\infty} (y - \vartheta) f(y)\, dy \tag{6.8}$$

When we take, as usual, the expectation about zero ($\vartheta = 0$), we obtain the well known first moment, or mean,

$$E(\underline{y}) = \int_{-\infty}^{+\infty} y f(y)\, dy \tag{6.9}$$

and it holds that $E'(\underline{y}) = E(\underline{y}) - \vartheta$.

When $\underline{y}$ is distributed as $\underline{y} \sim N(x, \sigma^2)$, as in (6.7), it can be shown that $E(\underline{y}) = x$, see Appendix B.1. It gives the location (the center) of the normal curve.

The second central moment, or variance is

$$D(\underline{y}) = E((\underline{y} - E(\underline{y}))^2) = \int_{-\infty}^{+\infty} (y - E(\underline{y}))^2 f(y)\, dy \tag{6.10}$$

The word '*central*' refers to the fact that the moment is taken about the mean $E(\underline{y})$. We denote the variance by $D(.)$ (dispersion), rather than $Var(.)$ as done for instance in [2], and often the symbol $\sigma^2$ is used for the variance of a random variable (and $\sigma$ for standard deviation). Later, we will use the index $y$ to denote the variance of $\underline{y}$, hence $\sigma_y^2$.

When $\underline{y}$ is distributed as $\underline{y} \sim N(x, \sigma^2)$, as in (6.7), it can be shown that indeed the variance equals $D(\underline{y}) = \sigma^2$, see Appendix B.1. The standard deviation $\sigma$ describes the width of the normal curve. It presents the spread in the result; the standard deviation $\sigma$ is a numerical measure of the uncertainty, or conversely of *precision* ($\sigma$ small = little uncertainty = high precision; $\sigma$ large = large uncertainty = low or poor precision).

### 6.5.2. Empirical moments

We will now consider the empirical equivalents of the moments discussed above. Therefore we assume to have $N$ realizations of the random variable $\underline{y}$. The measurement has been repeated (under unchanged conditions), and as a result we have $N$ outcomes $y_1, y_2, \ldots, y_N$. These $N$ measurements are used to come up with estimates for the mean and variance, i.e to come up with the sample mean and sample variance.

The average deviation from some known constant $\vartheta$ reads

$$\hat{x}' = \frac{1}{N} \sum_{i=1}^{N} (y_i - \vartheta) \tag{6.11}$$

and all outcomes $y_1, y_2, \ldots, y_N$ are first 'corrected for' $\vartheta$ , and then the average is taken. The well known first *sample* moment is the (arithmetic) *mean* about zero ($\vartheta = 0$)

$$\hat{x} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{6.12}$$

which is denoted by $x$ with a 'hat'-symbol, meaning that it is an estimate for the unknown true mean $x$ of the random variable $\underline{y}$, and this estimate is based on data/measurements; (6.12) is the empirical counterpart of (6.9). It holds that $\hat{x}' = \hat{x} - \vartheta$.

The expectation (or mean) of $\underline{y}$, $x$, is unknown, and will remain unknown forever; we can only come up with an estimate $\hat{x}$ for this parameter.

The second central sample moment (hence second sample moment about the mean) reads

$$\hat{\sigma'}^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{x})^2 \tag{6.13}$$

which is an unbiased estimate for the variance, once the mean $x$ is known (a-priori), and substituted for $\hat{x}$ in the above equation. Unbiasedness is generally a desirable property for an estimator, meaning that on average the result is spot-on (see also Section 8.3). When the

mean is not known (a-priori), estimate (6.13), with $\hat{x}$ from (6.12) inserted, is *not* unbiased, meaning that on average it is 'somewhat off'. If the mean is unknown, we generally use the following estimate for the variance instead

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{x})^2 \tag{6.14}$$

which is unbiased. Generally (6.14) is referred to as the sample variance; it is the empirical counterpart of (6.10). The square root of estimate (6.14) will be our default estimate, the *sample standard deviation*.

The difference of $N$ and $N-1$ in the denominator in (6.13) and (6.14) is negligible in practice, when $N$ is large.

Note that in the second case with (hypothetically) only one observation $y_1$ ($N = 1$), the mean (6.12) becomes $\hat{x} = y_1$ and the variance (6.14) is undefined — from one sample it is not possible to estimate both mean and variance. Estimate $\hat{\sigma'}^2$ would give zero in this case.

The difference $y_i - \hat{x}$ in the above equations will later be denoted by $\hat{e}$, and referred to as the *measurement residual*

$$\hat{e}_i = y_i - \hat{x} \tag{6.15}$$

see Section 10.1, where $e$ is the unknown measurement error, cf. (6.1), and $\hat{e}$ is the *estimated* error. Residual $\hat{e}_i = y_i - \hat{x}$ equals the difference of the observation $y$ and the *estimated* mean $\hat{x}$ (or, the average fitted to the observations).

In the context of Section 10.1.1, the second sample moment about the mean (6.13) can be regarded as the mean of the squared residuals, see also Section 10.1.4, as $\sum_{i=1}^{N}(y_i - \hat{x})^2$ is the sum of squared residuals.

### 6.5.3. Empirical moments: precision [*]

The estimates for the mean and variance (6.12) and (6.13) can be shown to result as Maximum Likelihood (ML) estimates from the model $E(\underline{y}) = Ax$, with $\underline{y} = (\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_N)^T$, $A = (1, \ldots, 1)^T$, and $D(\underline{y}) = \sigma^2 I_N$, with $\underline{y}$ normally distributed, cf. Chapter 8 (8.2). Also, it can be shown that the estimator $\underline{\hat{\sigma}}^2$ (6.13) or (6.14) is not correlated with $\underline{\hat{x}}$ (6.12).

As the estimates $\hat{x}$ for the mean, and $\hat{\sigma}^2$ for the variance are based on the measurements (which are not perfect), we can expect these estimates to be not perfect either. The estimators are unbiased, meaning that $E(\underline{\hat{x}}) = x$ and $E(\underline{\hat{\sigma}}^2) = \sigma^2$. The variances of the estimators in (6.12) and (6.14) are given (without proof) by

$$\sigma^2_{\hat{x}} = \frac{\sigma^2}{N} \tag{6.16}$$

and

$$\sigma^2_{\hat{\sigma}^2} = \frac{2\sigma^4}{N-1} \tag{6.17}$$

The more measurements we take (bigger $N$) the smaller these variances get, which is intuitively appealing. The more data you use, the more precise these estimators (for mean and variance) get. The expression for $\sigma^2_{\hat{\sigma}^2}$ holds only for normally distributed observables.

Typically one is interested in the standard deviation rather than the variance, hence with $\underline{\hat{\sigma}} = G(\underline{\hat{\sigma}}^2) = \sqrt{\underline{\hat{\sigma}}^2}$ (the standard deviation is a non-linear function of the variance, through the

square-root) and through the first-order approximation $\frac{\partial G}{\partial \hat{\sigma}^2} = \frac{1}{2\sqrt{\hat{\sigma}^2}}$, one can use (7.12), and obtain $\sigma_{\hat{\sigma}} \approx \frac{\sigma_{\hat{\sigma}^2}}{2\sqrt{\hat{\sigma}^2}} = \frac{\sigma_{\hat{\sigma}^2}}{2\hat{\sigma}}$ (the standard deviation of the estimator for the standard deviation), and use in this case $\sigma_{\hat{\sigma}^2} = \frac{\sqrt{2}\sigma^2}{\sqrt{N-1}}$.

## 6.6. Mean square error: accuracy

In practice, a measurement may be biased. This means that on average it does *not* deliver (the value of) the parameter which we hope it delivers. The bias is another (typically undesired) parameter $\vartheta$, which enters the equation: $\underline{y} = x + \vartheta + \underline{e}$. Here, $x$ is the unknown parameter in which our interest lies, $\underline{y}$ is the observable (which is a random variable), $\underline{e}$ is the random measurement error (for which we assume that it has zero mean, hence, it will cause individual measurements (samples) to be off from the true value, but, taking the average over a large number of (repeated) measurements, will provide an outcome close to the true and wanted value), and $\vartheta$ represents the *bias*, or (constant) offset in the measurement, which is a *systematic* effect.

In terms of the laser distometer, one can think of an unwanted time delay of the signal in the electronic circuitry of the device, which translates into a certain fixed error (offset) in the travel-time of the laser pulse, and hence into the measured distance — the effect is there all the time (also in repeated measurements).

Systematic errors also include scaling effects, e.g. $\underline{y} = \lambda x + \underline{e}$, with $\lambda$ as a scale factor, but this is beyond the scope of this book.

In this section, we present a measure of the spread in the uncertain outcome of the measurement, which includes also the unwanted bias-part. In the end, we would like to have a measure of how close our observation is to the true (wanted) distance. Instead of the variance, we consider the mean squared error (MSE), cf. also Section 20.3 in [2].

In the previous section there was no bias, and the expectation of the observable was $E(\underline{y}) = x$, as $\underline{y} = x + \underline{e}$. Now, with $\underline{y} = x + \vartheta + \underline{e}$, we have instead $E(\underline{y}) = x + \vartheta$, where $\vartheta$ is a bias. The variance (6.10) is the second moment about the *mean* of the observable $E(\underline{y}) = x + \vartheta$.

But — as an all-in measure — we are now interested in the second moment about $E'(\underline{y}) = x$, (6.8), namely about the *true distance* $x$. Therefore one defines the Mean Squared Error (MSE)

$$D'(\underline{y}) = \int_{-\infty}^{+\infty} (y - E'(\underline{y}))^2 f(y)\, dy \quad = E((\underline{y} - E'(\underline{y}))^2)$$

instead of the variance (6.10), and the mean $E(\underline{y})$ (6.9) has been replaced by the *true* distance $E'(\underline{y}) = x$ (6.8).

We will show that the MSE can be written as $D'(\underline{y}) = \sigma^2 + \vartheta^2$. Therefore we develop the MSE into

$$D'(\underline{y}) = E((\underline{y} - E'(\underline{y}))^2) = E((\underline{y} - E(\underline{y}) + E(\underline{y}) - E'(\underline{y}))^2)$$

$$= E((\underline{y} - E(\underline{y}))^2 + (E(\underline{y}) - E'(\underline{y}))^2 + 2(\underline{y} - E(\underline{y}))(E(\underline{y}) - E'(\underline{y})))$$

$$= E((\underline{y} - E(\underline{y}))^2 + (E(\underline{y}) - E'(\underline{y}))^2)$$

as the factor $(E(\underline{y}) - E'(\underline{y}))$ is just a constant, and taking it out of the expectation operator in the cross-term, we are left with $2(E(\underline{y}) - E'(\underline{y}))E(\underline{y} - E(\underline{y}))$, which is just zero as $E(E(\underline{y})) =$

$E(\underline{y})$. Hence,

$$D'(\underline{y}) = E((\underline{y} - E(\underline{y}))^2 + (E(\underline{y}) - E'(\underline{y}))^2) = \sigma^2 + \vartheta^2$$

the MSE equals the variance plus the squared bias. The MSE accounts for the spread in the result, as well as a bias, when present. When there is no bias $\vartheta = 0$, the MSE simply equals the variance $\sigma^2$.

The name of Mean Squared Error (MSE) explains by recognizing that we take the error (in our case, by how much $\underline{y}$ deviates from the true value $E'(\underline{y}) = x$), square it, and eventually take the mean (expectation), as the observable $\underline{y}$ is a random variable (in general we have no knowledge or control about the random error $\underline{e}$ included in $\underline{y}$). The MSE provides us with a general, overall measure of the deviation we can expect in the outcome; the MSE measures *accuracy*.

According to [26] *accuracy* is 'the state of being exact or correct', with the specialized technical subsense as 'the degree to which the result of a measurement or calculation matches the correct value or a standard'; it represents the degree of closeness of a measurement of a certain quantity to the actual true, or reference value of that quantity.

The spread in the outcome of a repeated experiment is referred to as *repeatability*; the degree to which repeated measurements, under unchanged conditions, show same results. The formal notion of repeatability is *precision*. The standard deviation $\sigma$, or variance $\sigma^2$, is a measure of precision. For a common uni-modal Probability Density Function (PDF), the standard deviation measures the width of this formal function.

Loosely spoken, one could say that: 'accuracy equals precision plus bias'.

### 6.6.1. Empirical MSE

The Mean Squared Error (MSE) can be used in practice for instance in a calibration or verification campaign. With the laser distometer, the manufacturer may have a calibration test-range available, for which actual distances are known already (known with a much higher accuracy, better by orders of magnitude than what the laser distometer will deliver, for instance by using different equipment). Then, the laser distometer is employed on the test-range, and the spread of the (repeated) measurements $y_1, y_2, \ldots, y_N$ is quantified by the second moment about the true distance $x$, which is known in this case (and not about the mean). If — unexpectedly — a bias is present in the measurements, it will be reflected in the resulting mean squared error (MSE).

Also, one could correct the obtained measurements beforehand for the known distance, therefore be dealing with samples of $(\underline{y} - x) = \vartheta + \underline{e}$, which are assumed to present a zero mean *error*, as we are initially not aware of the bias $\vartheta$. Next, taking the second (sample) moment about zero yields the Mean Squared Error (MSE).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - x)^2 \tag{6.18}$$

In practice one often takes the square root of the MSE, leading to the Root Mean Squared Error (RMS) or (RMSE), which is in units of the observed quantity.

The above expression for the empirical MSE looks very much like the second sample moment about the mean (6.13), and (6.14), but carefully note that in (6.18) the true value $x$ is involved, whereas in (6.13), and (6.14), it is the *estimated* mean $\hat{x}$.

| case | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $\hat{x}$ | $\hat{\sigma}'$ | $RMSE$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 7.452 | 7.454 | 7.450 | 7.453 | 7.451 | 7.452 | 0.001 | 0.001 |
| 2 | 7.452 | 7.462 | 7.442 | 7.457 | 7.447 | 7.452 | 0.007 | 0.007 |
| 3 | 7.459 | 7.461 | 7.457 | 7.460 | 7.458 | 7.459 | 0.001 | 0.007 |

Table 6.1: Three different cases of testing a laser distometer on a calibrated distance. The true distance is $x$=7.452 m. Each time $N$=5 measurements have been taken. The sample mean (6.12), the (square root of the) second order sample moment (6.13), as an approximation to the sample standard deviation, and the Root Mean Squared Error $\sqrt{MSE}$ (6.18) have been computed (in the latter, $x$ has been replaced by $\hat{x}'$). All values are in [m].

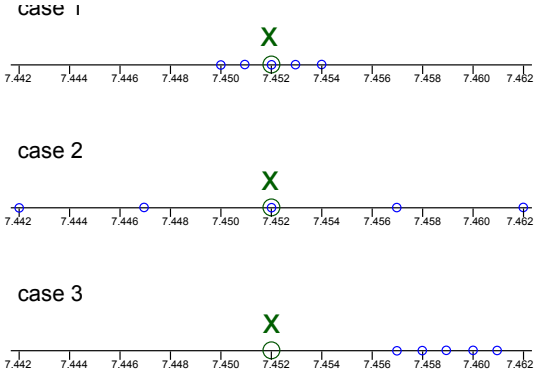



Figure 6.4: Three different cases of testing a laser distometer on a calibrated distance. The true distance is $x = 7.452$ m, indicated in green. Each time $N = 5$ measurements have been taken, shown by little blue circles.

### 6.6.2. Example on bias, precision and accuracy

With a large bias, the observable can be very precise (small random error), but it will not be accurate. With a big spread (large standard deviation), the observable is neither precise, nor accurate.

Suppose a laser distometer is being tested on an accurately known distance, with $x$=7.452 m. Five measurements are taken ($N$=5), and the MSE is computed, based on the known distance $x$, hence (6.18) can be rewritten (by subtracting and adding the term $\hat{x}$ in between the brackets) into

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{x})^2 + \frac{1}{N}\sum_{i=1}^{N}(\hat{x} - x)^2$$

in a way much similar to the variance plus bias decomposition of the MSE. The first term is the second sample moment $\hat{\sigma'}^2$ (6.13), which is approximately equal to the (estimated) variance (when $N$ is large), and the second term, with $\hat{x}$ in (6.12), equal to just $(\hat{x} - x)^2$, is the square of the estimated bias, as $\hat{x}$ is the (ordinary) mean of the observations, which now includes the bias $\vartheta$, and $x$ is the true distance.

We consider three different cases. They are listed in Table 6.1. In the first case the distance observable is obviously both precise (small standard deviation) and accurate (small MSE). In the second case the observable is not precise (big standard deviation) and not accurate (big MSE). In the first two cases there is no bias. In the third case, the observable is again precise (small standard deviation), but it is not accurate (big MSE), due to a bias in the measurements, the sample mean deviates a lot from the known distance (7.459 m versus 7.452 m).

Figure 6.4 displays the three cases, and Table 6.2 presents the summary in terms of the

| case | precise | accurate |
|---|---|---|
| 1 | yes | yes |
| 2 | no | no |
| 3 | yes | no |

Table 6.2: Summary on the precision and accuracy of the distance observable in three different cases of testing a laser distometer on a calibrated distance.

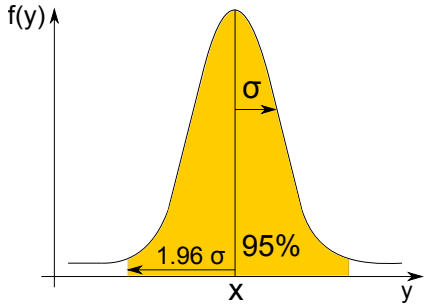



Figure 6.5: Normal probability density function (PDF) $f(y)$. The width of this function represents the uncertainty in measured outcomes of the observable $\underline{y}$. When the experiment could be repeated many times, the width of the PDF would reflect the spread in the outcome, for instance the observed distances. When the distance observable is normally distributed, the 1-sigma ($\sigma$) interval to both sides about the mean $x$ contains about 68% of the samples (it is referred to as the 68% interval), and 95% of the samples (in yellow) will lie in the interval [-1.96 $\sigma$, 1.96 $\sigma$] about the mean $x$.

distance observable being precise or not, and accurate or not. In some sources, you may find that case 2 is not precise, but, surprisingly, is accurate. The cause of this confusion may lie in the fact that in case 2 the distance observable is *not biased*, and thereby averaging can be used to reduce the uncertainty. According to (6.16), the standard deviation of the mean is smaller by a factor $\sqrt{N}$, when the mean is taken over $N$ samples. In case 2, a *single* distance observable is not precise, but the *mean* over all five together could be, and hence, all together they could provide a result which is close to the truth (and thereby — confusingly — rated accurate).

In case 3, the measurements are biased. This bias is preserved, when taking the average over all five measurements. The result will stay (far) off from the true value.

## 6.7. Probabilities and intervals

The Probability Density Function (and hence, also the Cumulative Distribution Function (CDF)) are mathematical concepts, which are very useful in practice, once one would like to compute probabilities, specifically, probabilities that a certain random variable lies in a certain interval.

$$P[a \leq \underline{y} \leq b] = \int_a^b f(y)\, dy = F(b) - F(a) \tag{6.19}$$

A normal probability density function $\underline{y} \sim N(x, \sigma^2)$ is shown in Figure 6.5.

Through the PDF (and CDF), probabilities and interval-bounds are intimately related. The $p$-th quantile (with $p$ a number between 0 and 1), or the $100p$-th percentile of the distribution of $\underline{y}$ is the smallest number $q_p$ such that

$$F(q_p) = P[\underline{y} \leq q_p] = p \tag{6.20}$$

Quantiles and percentiles can also be obtained empirically, from the observed data. Suppose that we have $N$ samples of a certain random variable, then we order the samples ascendingly, and each sample basically represents a probability of $\frac{1}{N}$. When a proportion $p$ is less than a certain number $k$ (i.e. $100p\%$ of the samples have all values less than $k$), and a proportion $1-p$ is greater than this number, this number $k$ is the $100p$-th empirical quantile, or the $100p$-th sample percentile.

For realizations of a random variable (error), which is (or can be) assumed to have zero mean, one is generally interested in just the *magnitude* of the error. Then typically the absolute value is taken of the samples, they are ordered ascendingly, and then the empirical percentile is determined. In that case, the corresponding formal $100p$-th percentile is defined as

$$P[|\underline{y}| \leq q_p] = p$$

## 6.8. Exercises and worked examples

This section presents a number of problems and worked answers.

**Question 1** A random variable is normally distributed, and has zero mean, and standard deviation equal to 1. You could think of this random variable as the measurement error $\underline{e}$, for instance of an observable with the laser distometer, and units in millimeters. Compute the probability that a single sample of this random variable will have a value of 1.27, or less. This is, compute the probability $P[\underline{e} < 1.27]$.

**Answer 1** The random variable has a standard normal distribution, $\underline{e} \sim N(0,1)$, which is tabulated in Appendix C. We need $P[\underline{e} < r_\alpha] = \Phi(r_\alpha)$, with $r_\alpha = 1.27$. The table gives probabilities $\alpha$, as the right tail probabilities $\alpha = 1-\Phi(r_\alpha)$, rather than left tail. With $r_\alpha = 1.27$, we obtain $\alpha = 0.1020$, hence the requested probability is $P[\underline{e} < r_\alpha] = \Phi(r_\alpha) = 1-\alpha = 0.8980$.

**Question 2** A random variable is normally distributed, and has zero mean, and standard deviation equal to 1, identical to question 1. Compute the probability that a single sample of this random variable will have a value of -1.27, or less. This is, compute the probability $P[\underline{e} < -1.27]$.

**Answer 2** The random variable has again a standard normal distribution, $\underline{e} \sim N(0,1)$, which is tabulated in Appendix C. We need $P[\underline{e} < -r_\alpha] = \Phi(-r_\alpha)$, with $r_\alpha = 1.27$. The table gives probabilities $\alpha$, as the right tail probabilities $\alpha = 1-\Phi(r_\alpha)$, but only for positive arguments. However, the normal distribution is symmetric about its mean. Hence, when the mean is zero, we have $\Phi(-r_\alpha) = P[\underline{e} < -r_\alpha] = P[\underline{e} > r_\alpha] = 1 - P[\underline{e} < r_\alpha] = 1 - \Phi(r_\alpha)$. With $r_\alpha = 1.27$, we obtain $\alpha = 0.1020$, hence the requested probability is $P[\underline{e} < -r_\alpha] = \alpha = 0.1020$.

**Question 3** A random variable is normally distributed, and has zero mean, and standard deviation equal to 1, identical to Question 1. Compute the probability that a single sample of this random variable will have a value of -1.23, or more. This is, compute the probability $P[\underline{e} > -1.23]$.

**Answer 3** The random variable has again a standard normal distribution, $\underline{e} \sim N(0,1)$, which is tabulated in Appendix C. We need $P[\underline{e} > -r_\alpha] = 1 - \Phi(-r_\alpha)$, which is $1 - \Phi(-r_\alpha) = 1 - (1 - \Phi(r_\alpha)) = \Phi(r_\alpha))$, with $r_\alpha = 1.23$. The table gives probabilities $\alpha$, as the right tail probabilities $\alpha = 1 - \Phi(r_\alpha)$. With $r_\alpha = 1.23$ we obtain $\alpha = 0.1093$, hence the requested probability is $\Phi(r_\alpha) = 1 - \alpha = 0.8907$.

**Question 4** A random variable is normally distributed, and has zero mean, and standard deviation equal to 1, identical to Question 1. Compute the probability that a single sample of this random variable will have a value lying in between -2.00 and 1.50. This is, compute the probability $P[-2.00 < \underline{e} < 1.50]$.

**Answer 4** The random variable has a standard normal distribution, $\underline{e} \sim N(0,1)$, which is tabulated in Appendix C. We need $P[-r_{\alpha,1} < \underline{e} < r_{\alpha,2}] = \Phi(r_{\alpha,2}) - \Phi(-r_{\alpha,1}) = \Phi(r_{\alpha,2}) - (1 - \Phi(r_{\alpha,1}))$, with $r_{\alpha,1} = 2.00$ and $r_{\alpha,2} = 1.50$. With the table we obtain $P[-2.00 < \underline{e} < 1.50] = (1 - 0.0668) - (1 - (1 - 0.0228)) = 0.9104$.

**Question 5** A random variable is normally distributed, and has zero mean, and standard deviation equal to 1, identical to Question 1. For what boundary value $r_\alpha$ holds that the probability that a single sample of this random variable will have a value of $r_\alpha$, or less, equals 0.975? This is, solve the probability statement $P[\underline{e} < r_\alpha] = 0.975$, for $r_\alpha$.

**Answer 5** The random variable has a standard normal distribution, $\underline{e} \sim N(0,1)$, which is tabulated in Appendix C. We need $P[\underline{e} < r_\alpha] = \Phi(r_\alpha) = 0.975$, or $\alpha = 0.025$, given in the table. This yields $r_\alpha = 1.96$.

**Question 6** A random variable is normally distributed, and has zero mean, and standard deviation equal to 1, identical to Question 1. For what boundary value $r_\alpha$ holds that the probability that a single sample of this random variable will have a value in between $-r_\alpha$ and $r_\alpha$, equals 0.95? This is, solve the probability statement $P[-r_\alpha < \underline{e} < r_\alpha] = 0.95$ for $r_\alpha$.

**Answer 6** The random variable has a standard normal distribution, $\underline{e} \sim N(0,1)$, which is tabulated in Appendix C. We need $P[-r_\alpha < \underline{e} < r_\alpha] = \Phi(r_\alpha) - \Phi(-r_\alpha) = \Phi(r_\alpha) - (1 - \Phi(r_\alpha)) = 0.95$. This equals $\Phi(r_\alpha) - (1 - \Phi(r_\alpha)) = 2\Phi(r_\alpha) - 1 = 0.95$, or $\Phi(r_\alpha) = \frac{1.95}{2}$. With $\alpha = 1 - \frac{1.95}{2}$, given in the table, this yields $r_\alpha = 1.96$.

**Question 7** A random variable is normally distributed, and has mean equal to 2 (unlike previous questions), and standard deviation equal to 1. Compute the probability that a single sample of this random variable will have a value in between 1.27 and 2.00. This is, compute the probability $P[1.27 < \underline{e} < 2.00]$.

**Answer 7** The random variable now does not have a standard normal distribution. Though, it can be turned into a standard normally distributed variable by subtracting the mean (this is a linear operation and the new random variable is normally distributed as well), $\underline{e} \sim N(2,1)$, and $(\underline{e} - 2) \sim N(0,1)$, which is tabulated in Appendix C. We need $P[r_{\alpha,1} < \underline{e} < r_{\alpha,2}] = P[(r_{\alpha,1} - 2) < (\underline{e} - 2) < (r_{\alpha,2} - 2)] = \Phi(r_{\alpha,2} - 2) - \Phi(r_{\alpha,1} - 2)$, with $r_{\alpha,1} = 1.27$ and $r_{\alpha,2} = 2.00$. With the table, we obtain $P[1.27 < \underline{e} < 2.00] = (1 - 0.5000) - (1 - (1 - 0.2327)) = 0.2673$.

**Question 8** Random variable $\underline{e}$ is normally distributed, and has mean equal to 1, and standard deviation equal to 2. Compute the probability that a single sample of this random variable will have a value in between -1.00 and 1.00. This is, compute the probability $P[-1.00 < \underline{e} < 1.00]$.

**Answer 8** The random variable now does not have a standard normal distribution. Though, it can be turned into a standard normally distributed variable by subtracting the mean, and dividing by the standard deviation (these are both linear operations and the new random variable is normally distributed as well), $\underline{e} \sim N(1,4)$, and $\frac{\underline{e}-1}{2} \sim N(0,1)$, which is tabulated in Appendix C. We need $P[-r_{\alpha,1} < \underline{e} < r_{\alpha,2}] = P[\frac{-r_{\alpha,1}-1}{2} < \frac{\underline{e}-1}{2} < \frac{r_{\alpha,2}-1}{2}] = \Phi(\frac{r_{\alpha,2}-1}{2}) - \Phi(\frac{-r_{\alpha,1}-1}{2})$, with $r_{\alpha,1} = 1.00$ and $r_{\alpha,2} = 1.00$. With the table, we obtain $P[-1.00 < \underline{e} < 1.00] = (1 - 0.5000) - (1 - (1 - 0.1587)) = 0.3413$.

**Question 9** For the error $\underline{e}$ in a distance observable $\underline{y}$, with $\underline{y} = x + \underline{e}$, and $x$ the unknown true distance, is given that it is distributed as $\underline{e} \sim N(0, \sigma^2)$, with standard deviation $\sigma$ = 3 mm. What is the probability that — when we take a distance measurement in practice — the *magnitude* of the measurement error is bigger than 6 mm?

**Answer 9** The required probability is $P[|\underline{e}| > 6]$. The absolute sign (because of the word

'magnitude') can be removed through: $P[|\underline{e}| > 6] = P[\underline{e} < -6] + P[\underline{e} > 6]$, which equals $2P[\underline{e} > 6]$, as the normal distribution is symmetric here about zero (given zero mean). Then we convert into a *standard* normal distribution through $2P[\frac{\underline{e}}{\sigma} > \frac{6}{\sigma}]$, which, with the table in Appendix C ($r_\alpha = 2.00$), yields $2P[\frac{\underline{e}}{\sigma} > \frac{6}{3}] = 2 \cdot 0.0228$ = 0.0456.

**Question 10** A laser distometer is deployed on an accurately calibrated test-range (hence the true distance $x$ is known). Suppose that the distance observable $\underline{y}$ has a standard deviation of 2 mm, and that the instrument has a bias $\vartheta$ of 1 cm (hence all measured distances are systematically too long by 1 cm). The distance observable is normally distributed. The manufacturer analyses, by measuring the known distance repeatedly, the *magnitude* of the error $\underline{y} - x$, where he presumes the error to be zero mean (as a-priori he is not aware of the presence of the bias). Can you give — based on the above given data — a good estimate for the 95th-percentile that the manufacturer is going to find?

**Answer 10** The deviation in the observable from the known distance $\underline{y} - x = \vartheta + \underline{e}$ is distributed as $\vartheta + \underline{e} \sim N(\vartheta, \sigma^2)$, in this case with $\vartheta$=10 mm, and $\sigma$=2 mm. We need to find the value for $q_p$ in $P[-q_p < \vartheta + \underline{e} < q_p] = p$, which is an interval symmetric about zero, with $p$=0.95. Transformation of the random variable yields $P[\frac{-q_p-\vartheta}{\sigma} < \frac{\underline{e}}{\sigma} < \frac{q_p-\vartheta}{\sigma}] = p$, or $\Phi(\frac{q_p-\vartheta}{\sigma}) - \Phi(\frac{-q_p-\vartheta}{\sigma}) = p$, as the random variable $\frac{\underline{e}}{\sigma}$ has a standard normal distribution. This has to be solved iteratively, by numerical search, i.e. trying different values for $q_p$ until we obtain the desired $p = 0.95$. Specifically, we start with value $q_p = 0$, and increase it each step by 0.01, until we reach the desired 95% probability, i.e. find the first occurrence where the above equation yields a probability of 95% or more. The result is $q_p$=13.29. In this case the left tail does actually not contribute to the exceedance probability, it is below $10^{-30}$; the 5% of the samples beyond the bounds $[-q_p, q_p]$ will typically all lie at the right hand side, that is, beyond $q_p$. Suppose the bias would be $\vartheta$=1 mm, then the bound is found to be $q_p$=4.37. In the left tail we have an exceedance probability of 0.0036, and in the right tail 0.0460. Please verify these figures yourself, with the table in Appendix C.

**Question 11** With a laser distometer four times the same distance has been measured on a calibration test-range. The distance is known, with very high accuracy, and equals 2.894 m. The four observed distances are: $y_1$ = 2.890 m, $y_2$ = 2.899 m, $y_3$ = 2.875 m, and $y_4$ = 2.886 m. Compute the (empirical) Mean Squared Error (MSE).

**Answer 11** The empirical MSE follows from Eq. (6.18). In this case $x$ = 2.894 m, there are four observations, hence $N = 4$, and the four observations $y_1$, $y_2$, $y_3$ and $y_4$ are given. Substituting this in Eq. (6.18) yields 116.5 $\mathrm{mm}^2$.

# 7

# Multi-variate: random vector

So far, we have been dealing with a single random variable. With two laser distometers in place, there are two random variables in parallel, denoted by $\underline{y}_1$ and $\underline{y}_2$ respectively, as we have two separate processes of taking measurements; a sample of $\underline{y}_1$ is denoted by $y_1$, and a sample of $\underline{y}_2$ by $y_2$. In this chapter we treat the multi-variate case.

From now on, the notion $\underline{y}$ will refer to a random vector. By default, we assume that this vector has $m$ elements, hence

$$\underline{y} = \begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_m \end{pmatrix} \tag{7.1}$$

and a sample of this vector is a set of one sample of each of the $m$ random variables in this vector

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

In the second part of this chapter we address the question of what happens to random variables upon mathematical manipulation. Suppose the two random variables $\underline{y}_1$ and $\underline{y}_2$ are independent, and have variances $\sigma^2_{y_1}$ and $\sigma^2_{y_2}$ respectively, what will be the variance of the sum of the two $\underline{y}_1 + \underline{y}_2$? Propagation laws are a crucial concept to be able to assess the uncertainty in a result, which is computed from a set of measurements.

## 7.1. Probability density function and moments

The mean of $\underline{y}$, the vector with random variables, is obviously a vector as well. Similar to (6.9) we have

$$E(\underline{y}) = \int_{-\infty}^{+\infty} y f(y)\, dy \tag{7.2}$$

which now is a multiple integral expression; for instance for the element $i$ of this vector we have

$$E(\underline{y}_i) = \int_{y_1=-\infty}^{+\infty} \int_{y_2=-\infty}^{+\infty} \dots \int_{y_i=-\infty}^{+\infty} \dots \int_{y_m=-\infty}^{+\infty} y_i f(y_1, y_2, \dots, y_i, \dots, y_m)\, dy_1 dy_2 \dots dy_i \dots dy_m$$

The mean, or expectation of vector $\underline{y}$ is

$$E(\underline{y}) = \begin{pmatrix} E(\underline{y}_1) \\ E(\underline{y}_2) \\ \vdots \\ E(\underline{y}_m) \end{pmatrix} \tag{7.3}$$

Instead of a single variance $\sigma^2$, we now get a full $m \times m$ variance matrix $Q_{yy}$. Formally the second central moment of the vector $\underline{y}$ reads

$$D(\underline{y}) = E((\underline{y} - E(\underline{y}))(\underline{y} - E(\underline{y}))^T) = \int_{-\infty}^{+\infty} (y - E(\underline{y}))(y - E(\underline{y}))^T f(y)\, dy \tag{7.4}$$

which is a multiple integral expression (the superscript $(\cdot)^T$ denotes the transpose of a vector or matrix). In case vector $\underline{y}$ consists of just a single random variable ($m$=1), the above form reduces again to (6.10). The variance matrix is

$$D(\underline{y}) = Q_{yy} = \begin{pmatrix} \sigma_{y_1}^2 & \sigma_{y_1 y_2} & \cdots & \sigma_{y_1 y_m} \\ \sigma_{y_2 y_1} & \sigma_{y_2}^2 & \cdots & \sigma_{y_2 y_m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{y_m y_1} & \sigma_{y_m y_2} & \cdots & \sigma_{y_m}^2 \end{pmatrix} \tag{7.5}$$

On the diagonal we find the variances for the individual random variables $\underline{y}_1$ through $\underline{y}_m$, and on the off-diagonals, we find the covariances, each time pertaining to a pair of random variables, hence $\sigma_{y_i y_j}$ is the covariance between $\underline{y}_i$ and $\underline{y}_j$, in textbooks on statistics often denoted as $Cov(\underline{y}_i, \underline{y}_j)$. The covariance is

$$\sigma_{y_i y_j} = E((\underline{y}_i - E(\underline{y}_i))(\underline{y}_j - E(\underline{y}_j)))$$

and the correlation coefficient is defined as

$$\rho_{y_i y_j} = \frac{\sigma_{y_i y_j}}{\sigma_{y_i} \sigma_{y_j}}$$

and can be regarded as a normalized covariance, as $|\rho_{y_i y_j}| \leq 1$, [2]. When $\sigma_{y_i y_j} = 0$, and thereby $\rho_{y_i y_j} = 0$, the two random variables $\underline{y}_i$ and $\underline{y}_j$ are said to be uncorrelated.

One could regard the variance $\sigma_{y_i}^2$ as the covariance of $\underline{y}_i$ with itself ($\sigma_{y_i y_i} = \sigma_{y_i}^2$); the square denotes the variance, in order to distinguish it from the standard deviation $\sigma_{y_i}$.

As $\sigma_{y_i y_j} = \sigma_{y_j y_i}$, matrix $Q_{yy}$ (7.5) is symmetric. The variance matrix $Q_{yy}$ is a positive (semi) definite matrix, just like the variance (6.10) by definition is a non-negative quantity, meaning that for any vector $u \in \mathbb{R}^m$, it holds that the quadratic form $u^T Q_{yy} u \geq 0$ in case it is positive semi-definite, and $u^T Q_{yy} u > 0$ in case it is positive definite.

### 7.1.1. Multi-variate normal distribution

In case all random variables in vector $\underline{y}$ are normally distributed, their joint distribution is a multi-dimensional, or multi-variate normal distribution, and the PDF $f(y) = f(y_1, y_2, \ldots, y_m)$ is given by

$$f(y) = \frac{1}{\sqrt{|2\pi Q_{yy}|}} e^{-\frac{1}{2}(y-x)^T Q_{yy}^{-1}(y-x)} \tag{7.6}$$

$$= \frac{1}{(2\pi)^{\frac{m}{2}}\sqrt{|Q_{yy}|}} e^{-\frac{1}{2}(y-x)^T Q_{yy}^{-1}(y-x)}$$

where $|Q|$ denotes the determinant of matrix $Q$, and vector $x$ denotes the mean of $\underline{y}$, hence $x = E(\underline{y})$.

## 7.2. Mean and variance propagation laws

We often transform one random vector into another one. When this transformation is linear, the mean and variance matrix of the new random vector can be fairly easily computed, once the mean and variance matrix of the original random vector are available; this takes place through the so-called propagation laws.

We consider the following linear transformation

$$\underline{v} = R\underline{y} + s \tag{7.7}$$

where vector $\underline{v}$ has $n$ elements, and consequently matrix $R$ has $n$ rows and $m$ columns, and vector $s$ is an $n$-vector.

The mean of $\underline{v}$ is easily obtained through

$$E(\underline{v}) = E(R\underline{y} + s) = RE(\underline{y}) + s \tag{7.8}$$

where $E(\underline{v})$ is an $n$-vector. The $n \times n$ variance matrix of $\underline{v}$ follows as

$$Q_{vv} = RQ_{yy}R^T \tag{7.9}$$

In practice, (7.9) is also referred to as the error propagation law.

Proofs of the above two propagation laws can be found in Appendix B.2.

Finally we state that when vector $\underline{y}$ is normally distributed, then through a linear transformation as (7.7), vector $\underline{v}$ is also normally distributed. Hence, if $\underline{y} \sim N$ then through (7.7) also $\underline{v} \sim N$.

## 7.3. Example

The height $x_1$ of a benchmark — monumented in the outer wall of a church tower — has been previously surveyed. The surveyed height of this point 1 is available and denoted as observable $\underline{y}_1$, with standard deviation $\sigma_{y_1} = \sigma$ (e.g. with $\sigma$ = 3 mm); the surveyed height will be a good estimate for the unknown height $x_1$, but not be perfect.

Next, we level from point 1 to 2, and eventually from point 2 to 3, see Figure 7.1. The measured height difference $y_{1,2}$ equals the difference of the height of point 2 and the height of point 1, hence $y_{1,2} = x_2 - x_1$, apart of course, from a measurement error; and $y_{2,3} = x_3 - x_2$. In order to properly account for random errors in these measurements, they are regarded as random variables: observables $\underline{y}_{1,2}$ and $\underline{y}_{2,3}$, with standard deviations $\sigma_{y_{1,2}} = \sigma_{y_{2,3}} = \sigma$ (and

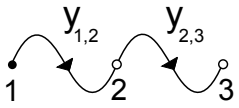



Figure 7.1: Levelling from point 1, via point 2 to point 3.

zero mean measurement error $\underline{y}_{1,2} = x_2 - x_1 + \underline{e}_{1,2}$ and $E(\underline{y}_{1,2}) = x_2 - x_1$, as $E(\underline{e}_{1,2}) = 0$ is assumed zero).

The surveyed height $\underline{y}_1$, and the height difference observables $\underline{y}_{1,2}$ and $\underline{y}_{2,3}$ are uncorrelated.

With the given surveyed height and the measured height differences we can determine the heights of points 2 and 3. The question now is, what uncertainty can be expected in these figures? So, what will be the variability in the determined height for point 3 for instance? Therefore, we need to compute the variance (or standard deviation) of the height (parameter) we determine for this point.

We handle this problem in a structured and systematic way. First the height of point 1 can be estimated; the estimator is trivial

$$\underline{\hat{x}}_1 = \underline{y}_1$$

From the second paragraph of text above we can deduce that

$$\underline{\hat{x}}_2 = \underline{y}_1 + \underline{y}_{1,2}$$

and that

$$\underline{\hat{x}}_3 = \underline{y}_1 + \underline{y}_{1,2} + \underline{y}_{2,3}$$

We see that there is only one straightforward way to get to know the heights of points 2 and 3. The above three equations can be cast in matrix-vector form

$$\begin{pmatrix} \underline{\hat{x}}_1 \\ \underline{\hat{x}}_2 \\ \underline{\hat{x}}_3 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}}_{M} \begin{pmatrix} \underline{y}_1 \\ \underline{y}_{1,2} \\ \underline{y}_{2,3} \end{pmatrix}$$

and now resembles (7.7), with the $R$-matrix as the above $3\times3$-matrix $M$ (which in this example is a square and full rank matrix). We basically have $\underline{\hat{x}} = M\underline{y}$. We are concerned here with a *linear* transformation.

In order to determine the variances (or standard deviations) of the height estimators (we want to compute $Q_{\hat{x}\hat{x}}$), we would like to apply (7.9). Therefore, we still need the variance matrix of the observables $Q_{yy}$. The third paragraph of text says that all $y$'s are uncorrelated, and we know that they all have standard deviation equal to $\sigma$. Hence variance matrix $Q_{yy}$ (7.5) is just an identity matrix, scaled by $\sigma^2$. Applying (7.9) yields

$$Q_{\hat{x}\hat{x}} = MQ_{yy}M^T = \sigma^2 \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$$

and the requested variances can be obtained from the diagonal, hence $\sigma^2_{\hat{x}_1} = \sigma^2$, $\sigma^2_{\hat{x}_2} = 2\sigma^2$, and $\sigma^2_{\hat{x}_3} = 3\sigma^2$.

What we see here is an *accumulation* of uncertainty, when you add together observables each with associated uncertainties. The estimator for the height of point 2 has a variance which is double the one for point 1, where we started. And for point 3, this is even a factor of three. We return to this phenomenon with Figure 9.14 on the so-called open levelling line. You can also see that the three height-estimators are correlated, the off-diagonal elements are not equal to zero; this also makes sense as they share part of the information they are based on; for instance, the surveyed height $\underline{y}_1$ appears in all three equations (for $\underline{\hat{x}}_1$, $\underline{\hat{x}}_2$, and $\underline{\hat{x}}_3$).

## 7.4. Non-linear mean and variance propagation laws

In Section 7.2 we considered a *linear* relation in Eq. (7.7). In practice, we may also face a *non-linear* relation:

$$\underline{v} = G(\underline{y}) \tag{7.10}$$

where $G$ is a mapping from $\mathbb{R}^m$ to $\mathbb{R}^n$; vector $\underline{v}$ has $n$ elements (random variables), and vector $\underline{y}$ has $m$ elements (random variables).

Theory is available to propagate the probability density function of $\underline{y}$, $f(y)$, into the one of $\underline{v}$, $f(v)$. In this section we restrict to just the expectation and the variance matrix of random vector $\underline{v}$.

An approximation for the expectation of $\underline{v}$ is given by

$$E(\underline{v}_i) \approx G_i(E(\underline{y})) + \frac{1}{2}\text{trace}\left(\frac{\partial^2 G_i}{\partial y y^T}\bigg|_{E(\underline{y})} Q_{yy}\right) \tag{7.11}$$

for $i = 1, \ldots, n$, and where trace means taking the sum of the diagonal elements, in this case of the matrix product $\frac{\partial^2 G_i}{\partial y y^T}\Big|_{E(\underline{y})} Q_{yy}$, which is an $m \times m$ matrix. Matrix $\frac{\partial^2 G_i(y)}{\partial y y^T}$ is the so-called Hessian matrix (in this case with dimensions $m \times m$), and contains the second order partial derivatives of the non-linear function $G_i(y)$. This equation shows that $E(\underline{v}_i) \neq G_i(E(\underline{y}))$; supplying the mean or expectation of $\underline{y}$ in the non-linear function $G$, does *not* yield the mean/expectation of $\underline{v}$! With approximation (7.11), there is already an extra term, which depends on the variance matrix of $\underline{y}$. The proof of (7.11) can be found in Appendix B.3.

An approximation for the variance matrix of $\underline{v}$ is given by

$$Q_{vv} \approx \frac{\partial G}{\partial y^T}\bigg|_{E(\underline{y})} Q_{yy} \left.\frac{\partial G}{\partial y^T}\right|^T_{E(\underline{y})} \tag{7.12}$$

where $\frac{\partial G(y)}{\partial y^T}$ is an $n \times m$ matrix, containing, as rows, the gradient vectors of non-linear functions $G_i(y)$, with $i = 1, \ldots, n$, all evaluated at $E(\underline{y})$. Defining the matrix $M = \frac{\partial G}{\partial y^T}\Big|_{E(\underline{y})}$, the above variance propagation law becomes $Q_{vv} \approx M Q_{yy} M^T$, which is then very similar to (7.9), though (7.12) is an approximation. The proof of (7.12) can be found in Appendix B.3.

In practice, the expectation of the observable vector $\underline{y}$ may not be known, hence the derivatives $\frac{\partial G(y)}{\partial y^T}$ and $\frac{\partial^2 G_i(y)}{\partial y y^T}$ are evaluated at a sample value instead.

## 7.5. Exercises and worked examples

This section presents three problems with worked answers on error propagation.

**Question 1** As shown in Figure 7.2, the height difference between point 1 and point 2 has been leveled, and the height difference between points 2 and 3 has been leveled. Both observables have a standard deviation of $\frac{3}{2}$ mm. The two observables are uncorrelated. Based on these two observed height differences, the height difference between point 1 and point 3 is determined. What is the standard deviation of this height difference?

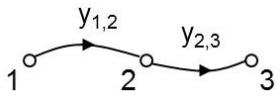



Figure 7.2: Levelling from point 1, via point 2 to point 3.

**Answer 1** The height difference between point 1 and point 3 follows as $y_{1,3} = y_{1,2} + y_{2,3}$. Or, in terms of a matrix and a vector (and random variables)

$$\underline{y}_{1,3} = \underbrace{\begin{pmatrix} 1 & 1 \end{pmatrix}}_{R} \begin{pmatrix} \underline{y}_{1,2} \\ \underline{y}_{2,3} \end{pmatrix}$$

With the variance matrix of $\underline{y}_{1,2}$ and $\underline{y}_{2,3}$ being

$$D\begin{pmatrix} \underline{y}_{1,2} \\ \underline{y}_{2,3} \end{pmatrix} = \begin{pmatrix} \frac{9}{4} & 0 \\ 0 & \frac{9}{4} \end{pmatrix}$$

application of Eq. (7.9) yields

$$\sigma^2_{y_{1,3}} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \frac{9}{4} & 0 \\ 0 & \frac{9}{4} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{18}{4}$$

In the absence of correlation, effectively the two *variances* are added: $\frac{9}{4} + \frac{9}{4} = \frac{18}{4}$. Hence, the requested *standard deviation* becomes $\sigma_{y_{1,3}} = \frac{3}{2}\sqrt{2}$ mm. By adding two equally precise observables, the result has a standard deviation which is worse by a factor of $\sqrt{2}$, not by a factor of 2.

**Question 2** The position coordinates of two points, 1 and 2, are available from an earlier survey. For simplicity we consider a one-dimensional coordinate system, shown in Figure 7.3. The coordinates of points 1 and 2 are $\underline{x}_1$ and $\underline{x}_2$ respectively, and their variance matrix is given as

$$D\begin{pmatrix} \underline{x}_1 \\ \underline{x}_2 \end{pmatrix} = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}$$

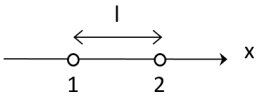



Figure 7.3: Computing distance $l$ from the coordinates $x_1$ and $x_2$.

hence, the variance of a single coordinate is $\sigma_{x_1}^2 = \sigma_{x_2}^2 = 4$, and the covariance is $\sigma_{x_1 x_2} = 2$. The correlation between $\underline{x}_1$ and $\underline{x}_2$ naturally results, as both coordinates together stem from the same, earlier survey. Compute the variance of distance $\underline{l}$ between the two points.

**Answer 2** Distance $l$ is defined as the difference between the two coordinates $\underline{l} = \underline{x}_2 - \underline{x}_1$ (assuming that $x_2 > x_1$), and this can be cast in the shape of (7.7) as

$$\underline{l} = \underbrace{\begin{pmatrix} -1 & 1 \end{pmatrix}}_{R} \begin{pmatrix} \underline{x}_1 \\ \underline{x}_2 \end{pmatrix}$$

and hence, the variance of the distance $\sigma_l^2$ follows with (7.9) as

$$\sigma_l^2 = \underbrace{\begin{pmatrix} -1 & 1 \end{pmatrix}}_{R} \underbrace{\begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}}_{Q_{xx}} \underbrace{\begin{pmatrix} -1 \\ 1 \end{pmatrix}}_{R^T} = 4$$

When the covariance between $\underline{x}_1$ and $\underline{x}_2$ would be zero, then $\sigma_l^2 = 8$, and the variance of the distance equals the sum of the two variances $\sigma_{x_1}^2$ and $\sigma_{x_2}^2$ (naturally as the uncertainty in the coordinates of *both* points contributes to the uncertainty in the distance). In practice, when there is positive correlation between $\underline{x}_1$ and $\underline{x}_2$ (practically meaning that chances are high that they share similar errors, which will then to some extent cancel in the difference), as in this exercise, the variance of the distance is actually *smaller* (than the sum). In this exercise we computed the precision of the distance, or actually the coordinate difference of two points, which is typically referred to as *relative precision*.

**Question 3** Given is a random variable $\underline{y}$ with mean equal to $E(\underline{y}) = 40$ mm, and a standard deviation of $\sigma_y = 3$ mm. Compute the expectation and standard deviation of $\underline{v}$, with $\underline{v} = \underline{y}^2$.

**Answer 3** The given transformation $\underline{v} = \underline{y}^2$ is *non-linear*, hence the propagation laws for a linear transformation can not be applied. We have to use (7.11) and (7.12) instead. With $G(y) = y^2$, we obtain $\frac{\partial G(y)}{\partial y} = 2y$ and $\frac{\partial^2 G(y)}{\partial y^2} = 2$ (which have to be evaluated at $E(\underline{y}) = 40$ mm), and this results into

$$E(\underline{v}) \approx 40^2 + \frac{1}{2} 2 \cdot 9 = 1609 \text{ mm}^2$$

$$\sigma_v^2 \approx 80 \cdot 9 \cdot 80 = 57600 \text{ mm}^4$$

hence $\sigma_v \approx 240 \text{ mm}^2$. Mind that the expectation $E(\underline{v})$ deviates, though slightly, from $(E(\underline{y}))^2 = 1600 \text{ mm}^2$. The larger the uncertainty in observation $y$, the larger its variance, and hence the larger the effect of the second term of (7.11).

# 8

# Observation modeling and parameter estimation

## 8.1. Introduction

In this chapter we propose a structured way of dealing with measurements. For each measurement we will formulate an equation, which expresses the observed quantity in terms of parameters of interest. In Chapter 6, we used a very simple example: we had a distance observable $\underline{y}$, and related it to the unknown (true) distance $x$. The observation equation reads:

$$\underline{y} = x + \underline{e} \tag{8.1}$$

This equation says that distance observable $\underline{y}$ is equal to the unknown, true distance, plus a random error term $\underline{e}$ representing the measurement error. When the instrument is fine, and the measurement is carried out properly, we can expect that the measurement (sample) $y$ is not perfect, hence we will *not* have $y = x$, though $y$ should be close to $x$. The symbol $e$ accounts for the measurement error, which, due to uncontrollable effects, will be positive one time, negative a next time, small, somewhat bigger etc. But, on average, the measurement is expected to be spot-on, that is, on average the measurement error is zero, $E(\underline{e}) = 0$, and therefore $E(\underline{y}) = x$. The spread we can expect once we would take many repeated measurements, and the uncertainty present in a single measurement, is quantified by standard deviation $\sigma$. With a laser distometer for instance, a value of $\sigma$=0.002 m is a fair number. Then we have $\sigma_y = \sigma_e = \sigma$.

In this simple example, the unknown parameter of interest is the distance. In practice survey problems can be much more complicated. Measurements can be angles and distances, and the unknown parameters are (position) coordinates of certain points of interest. An observation equation is an equation, expressing the observed quantity in terms of the unknown parameters. When you express a distance in terms of position coordinates (in a two or three dimensional space), this equation may even be non-linear. Later we return to non-linear observation equations, but first we address the linear model of observation equations.

## 8.2. Observation modeling

The model of observation equations is given by

$$E(\underline{y}) = Ax \; ; \; D(\underline{y}) = Q_{yy} \tag{8.2}$$

where $y$ is the $m$-vector of observations, $x$ the $n$-vector of unknown parameters, $A$ the $m \times n$ design matrix (of full rank equal $n$) containing a total of $mn$ *known* coefficients, and $Q_{yy}$ the $m \times m$ variance-matrix (rank equal $m$). The observation equations are assumed to be linear here, so (8.2) is a linear system. The system can also be written with the measurement errors occuring explicitly

$$\underline{y} = Ax + \underline{e} \; ; \; D(\underline{y}) = D(\underline{e}) = Q_{yy} \tag{8.3}$$

with $E(\underline{e}) = 0$.

The linear regression model in Chapter 17 of [2], with offset $\alpha$ and slope $\beta$ is indeed a simple example of the above linear model of observation equations; these two parameters $\alpha$ and $\beta$ would together constitute vector $x$.

### 8.2.1. Example

In the example on leveling in the previous chapter (Figure 7.1), the unknown parameters of interest are the heights of certain points (as $x_2$ and $x_3$), and through leveling we measure height *differences*: e.g. $y_{2,3} = x_3 - x_2$. The model of observation equations reads

$$\underbrace{E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_{1,2} \\ \underline{y}_{2,3} \end{pmatrix}}_{E(\underline{y})} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{x}$$

In this special case, $m = n = 3$ (there are exactly as many unknown parameters, as there are observations), and matrix $A$ is a square and invertible matrix, and indeed we had $\underline{\hat{x}} = M\underline{y}$ which, with $M = A^{-1}$, equals $\underline{\hat{x}} = A^{-1}\underline{y}$. The above system of equations is solved for by simply an inversion of matrix $A$, but this can be done *only* in the case with a square and invertible matrix $A$.

### 8.2.2. Redundancy

When there are $n$ unknown parameters (in vector $x$) to determine, we need at least $m = n$ observations (in vector $y$). In practice, typically (some) more measurements are done than strictly necessary (for good reasons), and in this part we will always consider the case $m \geq n$. The excess of $m$ over $n$ is referred to as redundancy; it equals $m - n$.

## 8.3. Parameter estimation

Given a vector of observations $y$ we have to determine the values for the entries in vector $x$. In the simple example above (in Section 8.1), computing the estimator for the unknown distance is very simple: $\underline{\hat{x}} = \underline{y}$, once we have a sample for $y$, we simply equate the result for $x$ to this observation: $\hat{x} = y$. By the hat-symbol, we denote that this is an estimate for the unknown distance. We will never know the actual true distance, but we can make a guess, or estimate of it, based on our observation. The true distance $x$ is still unknown, but, the 'best to our knowledge' guess of it is $\hat{x} = y$.

When, with multiple measurements and unknown parameters, we have $m = n$, then $\hat{x} = A^{-1}y$, as $A$ was assumed a full rank matrix before (and with $m = n$ it is also a square matrix). This is still a fairly simple case to solve.

As said before, in practice we typically deal with the case $m > n$, hence there are 'too many' measurements, and for the reason of measurement errors, vector $e$ in (8.3), the system $y = Ax$ will not be a consistent system, i.e. $y \neq Ax$; given the measurements, there will not be a solution for $x$ which exactly satisfies all equations in the system $y = Ax$.

### 8.3.1. Example

Suppose we measure the same distance twice. One time, we get $y_1$=7.452 m, and next we get $y_2$=7.454 m. Though, these two measurements are close together, we cannot make them fit perfectly in the assumed measurement model. For both observables holds that they are related to the same, unknown distance, hence

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} x$$

but we can never find a value for $x$ which satisfies $y_1$=7.452 m $= x$, and at the same time $y_2$=7.454 m $= x$. Therefore $y \neq Ax$. Adding measurement errors $e_1$ and $e_2$ on the right hand side would complete the above expression, where typically $e_1 \neq e_2$.

### 8.3.2. Least-squares estimate

The least-squares principle provides a solution to a system of observation equations which is redundant ($m > n$) and inconsistent ($y \neq Ax$ due to measurement errors). The estimate for the unknown parameter vector $x$ shall be computed according to

$$\hat{x} = (A^T A)^{-1} A^T y \tag{8.4}$$

The name least-squares explains, as we are trying to make the system $y = Ax$ consistent, by using — instead of the vector of observations $y$ — a vector of estimated, 'slightly adapted', observation values $\hat{y}$, that is $\hat{y} = A\hat{x}$ (properly said, $\hat{y}$ is the estimate for the mean of the observable $y$). Of course, one could choose very weird values for $\hat{x}$ and consequently for $\hat{y}$, and arrive at a consistent system, but, we would like to have $\hat{y}$ *close* to $y$, as afterall, we expect the measured values to be close to the true values (the measurements do contain useful information). So, we are going to 'change', or 'adjust' the observed values not too much.

The underlying criterion is, given vector $y$, to find $x$ such that the length (norm) of the vector $(y - Ax)$ is smallest

$$\min_x \|y - Ax\|^2 \tag{8.5}$$

The solution $\hat{x}$ to this minimization problem yields the smallest length, hence the vector $(y - A\hat{x}) = (y - \hat{y})$ is shortest (the norm is obtained by squaring all entries of the vector and summing them, and this should be at minimum, hence the term least-squares). A proof that solution (8.4) results from this minimization is provided in Appendix B.4.

### 8.3.3. Example

When we observe the same distance $x$ twice, then the least-squares estimator for the unknown distance equals just the average of the two measurements:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ 1 \end{pmatrix}}_{A} x$$

and using (8.4) we find that $\hat{x} = \frac{1}{2}(y_1 + y_2)$.

In Section 6.5.2 we actually used a model $y = Ax$ with $y = (y_1, \ldots, y_N)^T$ and $A = (1, \ldots, 1)^T$ with $m = N$ and $n = 1$, just a single unknown parameter.

### 8.3.4. Minimum variance estimator

With least-squares we have not used yet all available information. We used just the functional relations between observations and unknown parameters cast in $y = Ax$. In (8.2) we have also available, the variance matrix of the observables $Q_{yy}$ (and it does not occur at all in (8.4)). The parameter estimation process should use this information, to be optimal. Observables with small variances should get more weight in the solution than observables with larger variances. The first ones are more precise than the latter, and the resulting estimator should reflect this. The information contained in the variance matrix $Q_{yy}$ is taken into account in an optimal way, by the following estimator, which we state without proof:

$$\underline{\hat{x}} = (A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1} \underline{y} \tag{8.6}$$

You can easily see that, when the variance matrix $Q_{yy}$ is a (scaled) identity matrix, we are back at (8.4). Note that the above estimator $\underline{\hat{x}}$ for the vector of unknown parameters is a random vector, as it is a function of the vector of observables $\underline{y}$. The least-squares principle is not concerned with statistical aspects, and therefore no 'underscores' are used in Eq. (8.4).

The above estimator (8.6) has three distinct properties. The first one is that $\underline{\hat{x}}$ is a *linear* function of the observables $\underline{y}$, through $n \times m$-matrix $(A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1}$. Second, the estimator is *unbiased*, that is, on average it delivers values which agree with the unknown true values; taking the expectation of (8.6) and using (7.8) we get $E(\underline{\hat{x}}) = (A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1} E(\underline{y})$, which, with $E(\underline{y}) = Ax$, yields $E(\underline{\hat{x}}) = (A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1} Ax = x$, thus indeed $E(\underline{\hat{x}}) = x$. And finally the estimators in $\underline{\hat{x}}$ have *minimum variance*. The vector $\underline{\hat{x}}$ is a (linear) function of vector $\underline{y}$, which is a vector with random variables, hence all entries are subject to (random) measurement errors. Therefore, the entries of vector $\underline{\hat{x}}$ will also be subject to uncertainty. By using variance propagation law (7.9) on

$$\underline{\hat{x}} = \underbrace{(A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1}}_{H} \underline{y}$$

we obtain the $n \times n$ variance matrix for the estimator $\underline{\hat{x}}$ as

$$Q_{\hat{x}\hat{x}} = H Q_{yy} H^T = (A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1} \; Q_{yy} \; Q_{yy}^{-1} A (A^T Q_{yy}^{-1} A)^{-1}$$

which simplifies into

$$Q_{\hat{x}\hat{x}} = (A^T Q_{yy}^{-1} A)^{-1} \tag{8.7}$$

It can be shown that this matrix — among the variance matrices of all possible linear and unbiased estimators — has minimum trace, that is, this estimator is best in the sense that the sum of all $n$ variances together is smallest. The least-squares solution (8.4) only shares the first two properties with the minimum variance solution (8.6), hence being linear and unbiased.

The estimator (8.6) is unbiased, and therefore minimum variance implies best accuracy, cf. Section 6.6. The estimator (8.6) is also known as the Best Linear Unbiased Estimator (BLUE). It provides a generalization of the least-squares estimation of offset $\alpha$ and slope $\beta$ in Chapter 22 of [2]. An example on regression, or line-fitting is presented in Chapter 10. With the above BLUE one can compute a properly weighted least-squares solution (for vector $x$) to any proper linear problem. As the *inverse* of matrix $Q_{yy}$ is involved in (8.6), precise observables (small variances) receive larger weights, and less precise observables (large variances) receive smaller weights.

### 8.3.5. Example

We repeat the example of observing the same distance twice. However in this case the first measurement is made with a better instrument, and the standard deviation of $\underline{y}_1$ is equal to $\frac{1}{2}$. The second measurement is carried out with the default instrument, as before, and the standard deviation of $\underline{y}_2$ equals 1. We have

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ 1 \end{pmatrix}}_{A} x \; ; \; D\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \end{pmatrix} = Q_{yy} = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & 1 \end{pmatrix}$$

and using (8.6) we find that $\hat{\underline{x}} = \frac{4}{5}\underline{y}_1 + \frac{1}{5}\underline{y}_2$, hence we arrive at the weighted mean, rather than the ordinary mean. Observable $\underline{y}_1$ has a variance which is four times smaller than the variance of observable $\underline{y}_2$, and therefore the coefficient in the final estimator is four times bigger, $\frac{4}{5}$ versus $\frac{1}{5}$, and because $\frac{4}{5} + \frac{1}{5} = 1$ all information is taken into account (total weight equals 1, and the estimator is unbiased).

Finally, with (8.7), one can see that the variance of the estimator is $\sigma_{\hat{x}}^2 = \frac{1}{5}$, which is better, and smaller than that of any of the two observables at the input!

Collecting redundant measurements (here we have $m = 2$ and $n = 1$) typically leads to inconsistent systems of equations, but eventually improves the precision of the estimator!

## 8.4. Non-linear observation equations

The chapter started with a *linear* model of observation equations (8.2), and proposed the least-squares estimate and the Best Linear Unbiased Estimator (BLUE). It provides a nice theoretical framework. But, in practice there are hardly any measurements which carry a linear relation with the unknown parameters. Indeed, a leveled height difference is linear in the (unknown) height-parameters of the two points. But a distance is clearly a non-linear function of the coordinate differences, see (9.4).

The approach to systems of *non-linear* observation equations will be to *approximate* them by linear equations. The originally non-linear equations will be linearized with respect to the unknown parameters and the resulting system of linear(ized) equations will be treated using (8.4) or (8.6) on (8.2). Of course, we need to make sure that the approximation we make, is sufficiently good.

The model of observation equations

$$E(\underline{y}) = F(x) \; ; \; D(\underline{y}) = Q_{yy} \tag{8.8}$$

where matrix-vector product $Ax$ has been replaced by $F(x)$, a non-linear function, or mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$; it is a collection of $m$ non-linear functions of $n$ parameters.

The mapping $F(x)$ can be detailed as

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_m \end{pmatrix} = \begin{pmatrix} f_1(x_1, x_2, \ldots, x_n) \\ f_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ f_m(x_1, x_2, \ldots, x_n) \end{pmatrix}$$

or in words, (the expectation of) each observable $\underline{y}_i$ is a (scalar) function $f_i$ of $n$ parameters, namely $x_1, x_2, \ldots, x_n$ ($f_i$ is a mapping from $\mathbb{R}^n$ to $\mathbb{R}$), and row-by-row we do have $m$ such functions, $i = 1, \ldots, m$. Examples of function $f_i$ are given by (9.2) and (9.4).

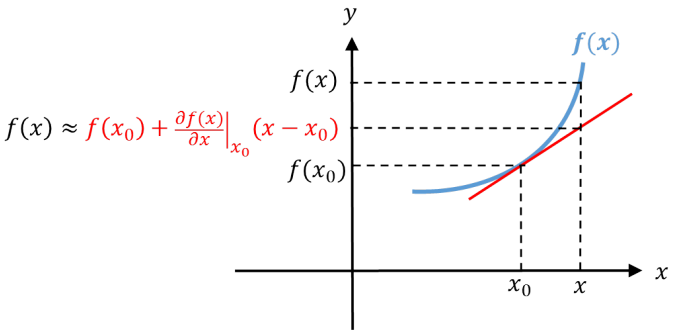



Figure 8.1: Linearization of the non-linear function $\underline{y} = f(x) + \underline{e}$ ($m = n = 1$). The red line represents the first degree Taylor polynomial of $f(x)$, shown by the blue curve, centered at $x_o$.

### 8.4.1. Linearization

The function $F(x)$ is approximated by the zero order and first order terms. Higher order terms are neglected. We rely on the *Taylor series*, using a point $x_o$, which is presumed to be reasonably close to the actual/true $x$; vector $x_o = (x_{1,o}, x_{2,o}, \dots, x_{n,o})^T$ contains approximate values for all $n$ unknown parameters. So,

$$F(x) \approx F(x_o) + \frac{\partial F(x)}{\partial x^T}\Big|_{x_o} (x - x_o)$$

where the zero order term and the first derivative are evaluated at $x_o$. To compute the first order derivative, all $m$ non-linear functions, one by one, are differentiated with respect to $x_1$, $x_2$, until $x_n$ (and these derivatives are organized in a row), hence this turns into an $m \times n$ matrix. The first row of this matrix reads $\frac{\partial f_1}{\partial x_1} \frac{\partial f_1}{\partial x_2} \dots \frac{\partial f_1}{\partial x_n}$, with the partial derivatives evaluated at $x_o$. The simple case of $m = n = 1$ is shown in Figure 8.1 - the slope of the red line is driven by $\frac{\partial f(x)}{\partial x}\Big|_{x_o}$.

Substituting the first-degree (linear) approximation of $F(x)$ into (8.8) yields

$$E(\underline{y}) \approx F(x_o) + \frac{\partial F(x)}{\partial x^T}\Big|_{x_o} (x - x_o) \; ; \; D(\underline{y}) = Q_{yy}$$

or

$$E(\underbrace{\underline{y} - F(x_o)}_{\Delta \underline{y}}) \approx \underbrace{\frac{\partial F(x)}{\partial x^T}\Big|_{x_o}}_{A} \underbrace{(x - x_o)}_{\Delta x} \; ; \; D(\underline{y} - F(x_o)) = Q_{yy} \tag{8.9}$$

Here, the first order derivative (of dimensions $m \times n$) takes the role of the design matrix $A$. The vector of observations $y$ is replaced by $y - F(x_o)$, that is, the observations minus the observations as they are computed based on just the approximate value $x_o$ for the unknown parameters: $y_o = F(x_o)$. And through least-squares we will not be estimating the vector of unknown parameters $x$, but $(x - x_o)$ instead, the differences of $x$ with respect to the approximate values $x_o$ that we already introduced.

### 8.4.2. Estimation

Accepting the approximation made in (8.9), the estimator for $\Delta x$ follows as

$$\underline{\Delta \hat{x}} = (A^T Q_{yy}^{-1} A)^{-1} A^T Q_{yy}^{-1} (\underline{y} - F(x_o))$$

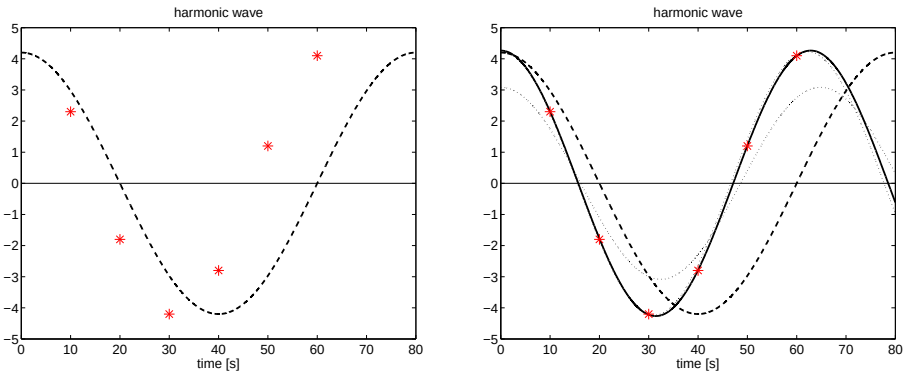



Figure 8.2: Example of non-linear estimation problem, fitting a cosine with unknown amplitude and frequency through a given set of data points, indicated by red asterisks (at left), and results of Gauss-Newton iterated estimation (at right).

with

$$Q_{\Delta\hat{x}\Delta\hat{x}} = (A^T Q_{yy}^{-1} A)^{-1}$$

and the eventual estimator for the vector of unknown parameter is obtained as

$$\underline{\hat{x}} = x_o + \underline{\Delta\hat{x}}$$

with

$$Q_{\hat{x}\hat{x}} = Q_{\Delta\hat{x}\Delta\hat{x}}$$

The model (8.9) is only an approximation of the actual non-linear model. For the approximation to be good and valid, the approximate value $x_o$ should be close to the true unknown value $x$. Therefore, the above procedure is repeated (iterated). One starts off with as good as possible guess for $x_o$, next one determines the estimate $\hat{x}$, and then takes this estimate as a new approximate value, as likely it is closer to the true, unknown $x$ than $x_o$ was, and repeats the above procedure (and on and on, if necessary). This iterative procedure is known as the Gauss-Newton method. A further discussion of this method, its properties and convergence behaviour is beyond the scope of this book (as well as alternatives to the Gauss-Newton method).

Concerns with regard to non linear estimation are briefly mentioned in Appendix B.5.

### 8.4.3. Example

A certain object moves harmonically, for instance a tall rise structure vibrating under the load of wind. The motion in one dimension, as a function of time, can be described by a cosine, with zero phase offset (our assumption/simplification in this example), but with unknown amplitude and frequency. The positions of the object are observed at times $t_1 = 10$, $t_2 = 20$, $t_3 = 30$, $t_4 = 40$, $t_5 = 50$, and $t_6 = 60$ seconds (timing is assumed to be perfect), and the measurements are denoted by $y_1$, $y_2$, $y_3$, $y_4$, $y_5$ and $y_6$ (and these position measurements are subject to errors). The observation model reads

$$E(\underline{y}_i) = x_1 \cos(2\pi x_2 t_i) \tag{8.10}$$

for $i = 1, \ldots, 6$. Unknown parameters are the amplitude $x_1$, and the frequency $x_2$ in Hertz.

| | iteration | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|
| $y_1 = 2.3$ | | 2.9698 | 1.7446 | 2.2474 | 2.3027 | 2.3046 | 2.3046 |
| $y_2$ =-1.8 | | 0.0000 | -1.1066 | -1.8431 | -1.7699 | -1.7729 | -1.7729 |
| $y_3$ =-4.2 | $F(x_o)$ | -2.9698 | -2.9975 | -4.2056 | -4.2163 | -4.2210 | -4.2210 |
| $y_4$ =-2.8 | | -4.2000 | -2.2871 | -2.6249 | -2.7887 | -2.7898 | -2.7898 |
| $y_5 = 1.2$ | | -2.9698 | 0.4080 | 1.4168 | 1.2011 | 1.2054 | 1.2054 |
| $y_6 = 4.1$ | | 0.0000 | 2.7491 | 4.1302 | 4.0874 | 4.0928 | 4.0928 |
| | | -0.6698 | 0.5554 | 0.0526 | -0.0027 | -0.0046 | -0.0046 |
| | | -1.8000 | -0.6934 | 0.0431 | -0.0301 | -0.0271 | -0.0271 |
| | $y - F(x_o)$ | -1.2302 | -1.2025 | 0.0056 | 0.0163 | 0.0210 | 0.0210 |
| | | 1.4000 | -0.5129 | -0.1751 | -0.0113 | -0.0102 | -0.0102 |
| | | 4.1698 | 0.7920 | -0.2168 | -0.0011 | -0.0054 | -0.0054 |
| | | 4.1000 | 1.3509 | -0.0302 | 0.0126 | 0.0072 | 0.0072 |
| $(x_1)_o$ =4.2000 | $\hat{x}_1$ | 3.0818 | 4.2308 | 4.2595 | 4.2641 | 4.2641 | 4.2641 |
| $(x_2)_o$ =0.0125 | $\hat{x}_2$ | 0.0154 | 0.0161 | 0.0159 | 0.0159 | 0.0159 | 0.0159 |

Table 8.1: Gauss-Newton iteration: the left column shows the observed values $y_1$, $y_2$, $y_3$, $y_4$, $y_5$ and $y_6$, and the two approximate values for the unknown parameters at bottom $(x_1)_o$ and $(x_2)_o$. The columns 0 through 5 show the results of the 5 iteration steps; on top the $F(x_o)$, $y - F(x_o)$ and the resulting estimate $\hat{x}$ at bottom.

The observation values are $y_1 = 2.3$, $y_2 = -1.8$, $y_3 = -4.2$, $y_4 = -2.8$, $y_5 = 1.2$ and $y_6 = 4.1$, and they are shown in Figure 8.2 by the red asterisks. The observation model is non-linear in the unknown paramaters $x_1$ and $x_2$, and will now be approximated by zero and first order terms. In order to do so, one needs approximate values for the unknown parameters. Our choice is to simply set the amplitude $(x_1)_o$ equal to the largest observed value (in absolute sense) $(x_1)_o = |y_4|$, and the data points in Figure 8.2 seems to show roughly one period of the harmonic motion in a time span of 80 seconds, hence the frequency is set to $(x_2)_o = \frac{1}{80}$ (Hz). The dashed line presents the harmonic wave defined by these approximate values, hence $y_o(t) = (x_1)_o \cos(2\pi(x_2)_o t)$.

The graph on the right in Figure 8.2 shows again the harmonic wave defined by the initial approximate values, and also the results (estimates) of the first and second iteration (in thin dotted lines), and eventually the result of the last step, the fifth iteration, as a solid line: $\hat{y}(t) = \hat{x}_1 \cos(2\pi\hat{x}_2 t)$. One can clearly see that with the initial approximate values the harmonic wave was a bit off, and that a few iterations quickly make the wave fit to the observated data points. Numerical results are reported in Table 8.1. The values $F(x_o)$ computed for the observations in each iteration, directly follow from using the values for $\hat{x} = (\hat{x}_1, \hat{x}_2)^T$ obtained in the step before.

It should be noted that the approximate values for the unknown amplitude and frequency should be carefully chosen — not any value will do. The problem is in fact highly non-linear — the cosinus is a highly curved function, in particular for high(er) frequencies. The approximate values should be sufficiently close already to the true, but unknown values for the parameters to be estimated.

## 8.5. Exercises and worked examples

In this section we present a sequence of questions on parameter estimation, all related to the same problem.

**Question 1** A one-dimensional positioning problem is considered, see Figure 8.3. Two

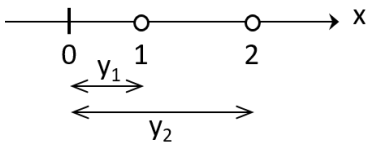



Figure 8.3: One-dimensional positioning problem: observing distances $y_1$ and $y_2$, and determining the position coordinates of points 1 and 2 (Question 1).

distances have been observed, $y_1$ and $y_2$ (each time from the origin of the coordinate-system, point 0 with $x_0 = 0$), and the purpose is to determine the position coordinates of the points 1 and 2, hence $x_1$ and $x_2$. The observations are given as $y_1$=21 and $y_2$=63.

**Answer 1** This is actually a trivial problem. There are $m = 2$ observations and $n = 2$ unknown parameters. There is a unique solution, and that is all there is.

$$\hat{x}_1 = y_1$$
$$\hat{x}_2 = y_2$$

Setting up the model of observation equations and though formally computing the least-squares parameter estimates yields an identical result. Applying (8.4) to

$$E\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

also yields $\hat{x}_1 = y_1$ and $\hat{x}_2 = y_2$, hence $\hat{x}_1$=21 and $\hat{x}_2 = 63$.

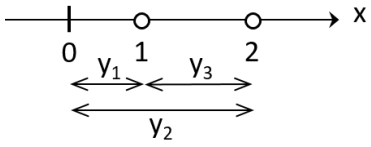



Figure 8.4: One dimensional positioning problem: observing distances $y_1$, $y_2$, and $y_3$, and determining the position coordinates of points 1 and 2 (Question 2).

**Question 2** Now, also the distance between point 1 and 2 has been observed, $y_3$=45, see Figure 8.4. The task is to compute least-squares estimates for the unknown coordinates $x_1$ and $x_2$.

**Answer 2** There are $m = 3$ observations and $n = 2$ unknown parameters, and they are cast in a model of observation equations.

$$E\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Applying (8.4) yields, with

$$A^T A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$$

$$\begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 2 & 1 & -1 \\ 1 & 2 & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

and supplying the numerical values, we obtain $\hat{x}_1$=20 and $\hat{x}_2$=64. These values are slightly different from the ones obtained with Answer 1. This requires further deliberations. With Question 1 we basically had two unknown parameters in two equations. The solution was trivial. With Question 2 there is actually one observation 'too much'; the redundancy equals $m - n$=1. The observations are *not consistent* with the assumed model. One expects that $E(\underline{y}_2) - E(\underline{y}_1) = E(\underline{y}_3)$, or that $E(\underline{y}_2) - E(\underline{y}_1) - E(\underline{y}_3) = 0$, but using the given observation values, we end up with $y_2 - y_1 - y_3 = -3$. It does not fit exactly. With least-squares parameter estimation we try — given the mathematical model — to find the solution best fitting *all* observations. The 'best' fitting implies that a misfit or discrepancy is distributed (in this exercise) equally over all observations. Using the least-squares parameter estimates $\hat{x}_1$ and $\hat{x}_2$, one could compute for diagnostic reasons, using the model $y = Ax$, estimates for the observables

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix}$$

and we obtain $\hat{y}_1$=20, $\hat{y}_2$=64, and $\hat{y}_3$=44, and the misfit has been resolved: $\hat{y}_2 - \hat{y}_1 - \hat{y}_3 = 0$. One can also observe that all three observation values have been 'adjusted' by the same amount $y_1 - \hat{y}_1 = 1$, $y_2 - \hat{y}_2 = -1$, and $y_3 - \hat{y}_3 = 1$. This is illustrated in Figure 8.5. We return to the difference $y - \hat{y}$ as the least-squares residuals in Section 10.1.

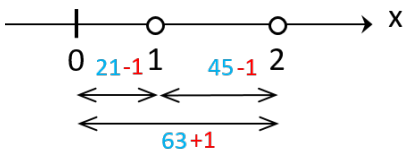



Figure 8.5: One dimensional positioning problem: observing distances $y_1$, $y_2$, and $y_3$, and determining the position coordinates of points 1 and 2, interpreting the results of least-squares parameter estimation.

**Question 3** When the observables' variance matrix is given, one can apply the minimum variance estimator (8.6) of Section 8.3.4. In this problem all observables are equally precise with $\sigma_{y_i} = 1$, with $i = 1, 2$ for Question 1, and with $i = 1, 2, 3$ for Question 2. The observables are uncorrelated. Compute the variance matrix of the minimum variance estimator for Question 1, and also for Question 2.

**Answer 3** The variance matrix of the minimum variance estimator is given by (8.7). For Question 1 $Q_{yy}$ is a $2 \times 2$ identity matrix and we obtain

$$Q_{\hat{x}\hat{x}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

For Question 2 $Q_{yy}$ is a $3 \times 3$ identity matrix, and we obtain

$$Q_{\hat{x}\hat{x}} = \frac{1}{3} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

For Question 1 the variance of the estimators for the coordinates is $\sigma^2_{\hat{x}_1} = \sigma^2_{\hat{x}_2} = 1$, and for Question 2 $\sigma^2_{\hat{x}_1} = \sigma^2_{\hat{x}_2} = \frac{2}{3}$, hence the precision of the coordinates in Question 2 is better than in Question 1. This is what redundancy brings on account of precision: using an extra observation pays off by obtaining *more precise* results!

# 9

# Land surveying

Land surveying is in support of civil engineering activities, as there is need for knowledge about shape, attitude and location (position) of objects, topography and the Earth's surface. Measurements are done to gather this knowledge, and the knowledge is eventually presented — most often — in terms of *position coordinates*. In this chapter we present the most common types of measurements in land surveying. The two basic measurement types are angle (or azimuth, or direction), and distance. And they are actually closely related, as the angle — expressed in radians — equals the quotient of the distance along the arc of a circle, and its radius.

As civil engineering activities mostly concern local and regional areas, we use a two dimensional (Euclidean) geometry in the (local) horizontal plane. The height is used separately as a third dimension. Azimuths, directions, angles, and distances are parametrized in terms of two-dimensional coordinates. The coordinates of a certain point i read $((x_1)_i, (x_2)_i)$. Commonly the coordinates of a point i are denoted as $(x_i, y_i)$, but we deviate here, as to avoid confusion with the model of observation equations $y = Ax$ in the previous chapter; therefore we uniquely use symbol $x$ for coordinates, and append appropriate indices to indicate the dimension.

Next, in Section 9.5 we introduce the theory necessary to evaluate the precision of the resulting position coordinate estimators. The last section of this chapter provides a review of elementary measurement set-ups, together with an analysis of precision of the resulting position coordinates.

## 9.1. Leveled height difference

A height difference $y_{ij}$ obtained through leveling is simply the difference of the heights of the two points involved:

$$E(\underline{y}_{ij}) = (x)_j - (x)_i \tag{9.1}$$

This is a linear equation in the unknown parameters $(x)_i$ and $(x)_j$, and has been dealt with before, see Section 7.3. With leveling, we deal with a *one-dimensional* coordinate system; we are only interested in the height of a point, and the height is represented by coordinate $x$.

## 9.2. Azimuth and angle measurements

The azimuth $a_{ij}$ is the argument of the line-segment from point i to (target) point j,

$$E(\underline{a}_{ij}) = \arctan \frac{(x_1)_j - (x_1)_i}{(x_2)_j - (x_2)_i} \tag{9.2}$$

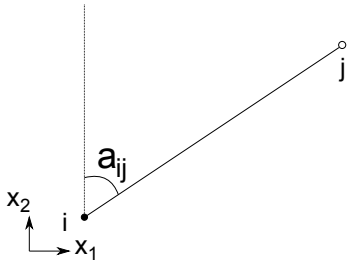



Figure 9.1: Azimuth measurement $a_{ij}$.

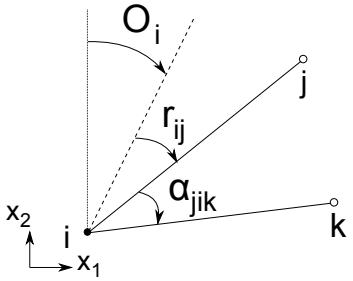



Figure 9.2: Measurement of direction $r_{ij}$, and measurement of angle $\alpha_{jik}$.

see Figure 9.1; we deal with *two-dimensional* geometry. The azimuth provides the angle of the line-segment from point i to j, with respect to a fixed reference direction, typically the $x_2$-axis. For measurements of azimuth we use the symbol $a$ rather than $y$; for distance we use $l$ in the next section, rather than $y$.

For an observation of direction $r_{ij}$, the zero reference — indicated by the dashed line in Figure 9.2 — has an arbitrary offset (angle) with respect to the second coordinate axis $x_2$. This unknown offset parameter enters the observation equation (9.2); it should be subtracted from the right-hand side. All observations of direction taken with a single set up of the instrument, share the same orientation offset.

$$E(\underline{r}_{ij}) = \arctan \frac{(x_1)_j - (x_1)_i}{(x_2)_j - (x_2)_i} - O_i$$

An angle observation $\alpha_{jik}$ is just the difference of two azimuths, hence the angle $\alpha_{jik}$ at point i from point k to point j, is just $a_{ik} - a_{ij}$, see also Figure 9.2.

With Section 8.4, linearization of (9.2) with azimuth $a_{ij}$ expressed in radians, yields

$$E(\underline{\Delta a}_{ij}) = -\frac{(x_2)_{ij,o}}{l^2_{ij,o}}(\Delta x_1)_i + \frac{(x_1)_{ij,o}}{l^2_{ij,o}}(\Delta x_2)_i + \frac{(x_2)_{ij,o}}{l^2_{ij,o}}(\Delta x_1)_j - \frac{(x_1)_{ij,o}}{l^2_{ij,o}}(\Delta x_2)_j \tag{9.3}$$

where $(x_1)_{ij,o} = (x_1)_{j,o} - (x_1)_{i,o}$, $(x_2)_{ij,o} = (x_2)_{j,o} - (x_2)_{i,o}$, and $l^2_{ij,o} = (x_1)^2_{ij,o} + (x_2)^2_{ij,o}$. The above equation follows by using that $d \arctan(x)/dx = 1/(1+x^2)$.

## 9.3. Distance measurements

The Euclidean distance between points i and j is the length of the line-segment from point i to point j,

$$E(\underline{l}_{ij}) = \sqrt{((x_1)_j - (x_1)_i)^2 + ((x_2)_j - (x_2)_i)^2} \tag{9.4}$$

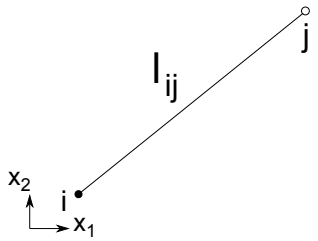



Figure 9.3: Measurement of distance in two dimensions between points i and j.

see Figure 9.3; we deal with *two-dimensional* geometry.

Linearization of (9.4) yields

$$E(\underline{\Delta l}_{ij}) = -\frac{(x_1)_{ij,o}}{l_{ij,o}}(\Delta x_1)_i - \frac{(x_2)_{ij,o}}{l_{ij,o}}(\Delta x_2)_i + \frac{(x_1)_{ij,o}}{l_{ij,o}}(\Delta x_1)_j + \frac{(x_2)_{ij,o}}{l_{ij,o}}(\Delta x_2)_j \tag{9.5}$$

where $(x_1)_{ij,o} = (x_1)_{j,o} - (x_1)_{i,o}$, $(x_2)_{ij,o} = (x_2)_{j,o} - (x_2)_{i,o}$, and the approximate value for the distance $l_{ij,o} = \sqrt{(x_1)^2_{ij,o} + (x_2)^2_{ij,o}}$.

In practice distance measurements can be subject to a scale factor and/or to an offset. These aspects are beyond the scope of this part.

## 9.4. Idealization

In the mathematical model of observation equations $\underline{y} = Ax + \underline{e}$ (Chapter 8), the parameters $x$, for instance representing position coordinates of a point, are deterministic quantities. Each parameter represents a single — though unknown — numerical value.

Next, we should realize that — for the purpose of surveying and making maps — we are *modeling* the Earth's surface and its topography and objects by means of basic geometric entities. The real world is reduced to points, lines and polygons/areas, see also Figure 1.1. For this reduction the surveyor relies on his insight and experience.

We should realize that there is a random component involved in this process. The corner of a brick-wall building can typically be identified quite well, though, looking at the millimeter scale, you will see that not all bricks in a wall are perfectly aligned, and some walls are poorer, see Figure 9.4 at left. For the center or border line of a ditch, the reduction will be much more difficult; how can we identify the (shape of the) ditch in rough terrain? Where does the ditch start and where does it end? The (additional) uncertainty may increase to the centimeter or even decimeter level in this case. The additional random component is referred to as the *idealization error*. Figure 9.4 shows two examples of reducing reality to a point (at left) and a line (at right).

Idealization is about the question how well we can *identify* what we actually measure. Upon concluding the surveying exercise, we can make for instance a statement about the distance between two objects on Earth and when stating the precision of this distance, we should account for the *measurement precision*, and also for the *idealization precision*, in order to present realistic figures. In this part however, we will *not* account for this additional random component in surveying. To say, our assumption is that we can identify the objects of interest infinitely precise.

## 9.5. Analysis of measurement set-up: confidence ellipse

In this section we present the confidence ellipse as a concept for evaluating the precision of a random vector, and this is instrumental to (performance) analysis of a survey measurement

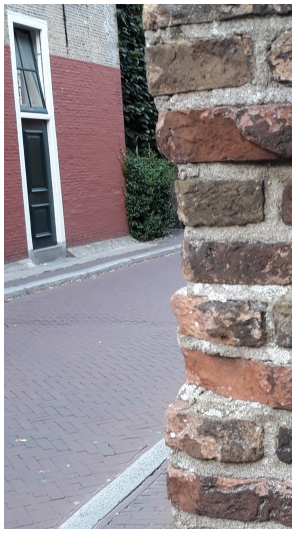


Figure 9.4: Example of idealization and generalization. The corner of the brick wall, at left, you find on a two-dimensional map as a point, and the country-road, at right, as a (series of) straight line-segment(s). In reality the corner is not just a single discrete point, and the edge of the road, shown in orange, which separates the asphalt-surface and the grass along the road, is really not a straight line. Geometric entities in maps, like points and lines have infinitesimally small dimensions. A line connecting two points for instance is infinitesimally thin, whereas objects in reality, like the edge of the road, have real dimensions. This reduction, or approximation of dimensions is referred to as idealization and generalization. Accurately surveying the actual edge with twists and turns every centimeter would be an unrealistic (and very costly) job.

set-up. The outcome of surveying (that is, taking measurements and carrying out their consequent processing) is generally a set of *position coordinates*. We take measurements, and we want to know where (a certain identified point on) an object is located. Observables are turned into estimators for position coordinates, and the quality of these estimators needs to meet requirements. An important aspect of quality is *precision*, and we present it in terms of a confidence ellipse (for the estimators of the two position coordinates of the point under consideration).

The confidence ellipse is an area (in the two-dimensional space) centered at the estimated position, which contains the true, but unknown position, with a certain stated probability. In case the position is described with just a single coordinate (for instance when we are only interested in the height) the confidence ellipse turns into an interval. And for a three-dimensional position coordinates vector, it is an ellipsoid. One can also use the general term confidence region, rather than confidence ellipse, to suit any dimension. In the sequel we work with two-dimensional position coordinate vectors. One can for instance think of the indicated center of a pillar to support a future bridge deck, and be concerned with the (horizontal) coordinates of this point of interest: $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. We are about to analyze the precision of the obtained position coordinate estimators $\underline{\hat{x}}$, which typically is a function of the observables $\underline{\hat{x}} = G(\underline{y})$ (non-linear or linear), with variance matrix $Q_{\hat{x}\hat{x}}$. The estimators are assumed to be *normally distributed* here (remember that normally distributed observables yield normally distributed estimators, with Section 7.2 and Eq. (8.6)).

We repeat Eq. (7.6), but now applied to $\underline{\hat{x}}$, rather than $\underline{y}$ in (7.6).

$$f(\hat{x}) = \frac{1}{\sqrt{|2\pi Q_{\hat{x}\hat{x}}|}} e^{-\frac{1}{2}(\hat{x}-E(\hat{x}))^T Q_{\hat{x}\hat{x}}^{-1}(\hat{x}-E(\hat{x}))} \tag{9.6}$$

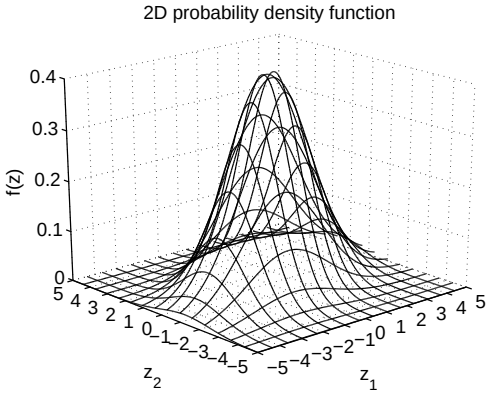



Figure 9.5: Joint probability density function (PDF) of $\underline{z}_1$ and $\underline{z}_2$.

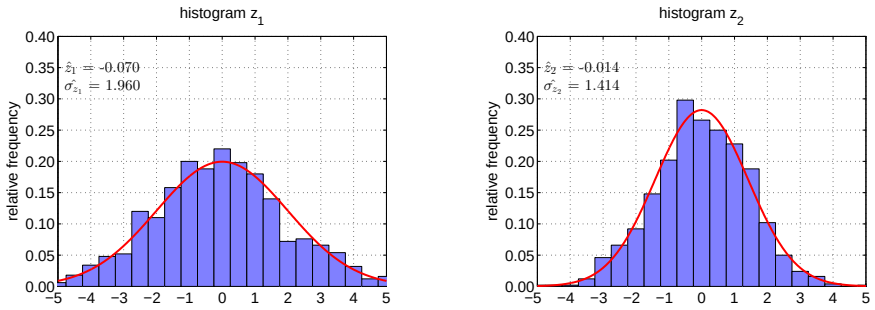



Figure 9.6: Histogram of $z_1$ on the left, and of $z_2$ on the right. The (theoretical) normal probability density function is imposed: for $\underline{z}_1$ with standard deviation $\sigma_{z_1} = 2$, and for $\underline{z}_2$ with standard deviation $\sigma_{z_2} = \sqrt{2}$, and both $\underline{z}_1$ and $\underline{z}_2$ with zero mean.

The peak of the PDF is centered at $E(\underline{\hat{x}}) = x$, the true value (as we assume an unbiased estimator here); $\underline{\hat{x}} \sim N(x, Q_{\hat{x}\hat{x}})$. Mind, that in practice, we generally do *not* know the true value …. What is actually of interest for precision analysis, is *how close* position estimates can expected to be to the true value, therefore we consider the PDF of the *difference* $\underline{z} = \underline{\hat{x}} - x$, which we temporarily denote by $z$; $\underline{z} \sim N(0, Q_{zz})$, which is easily obtained using Section 7.2 (subtracting a constant vector does not change the variance matrix, so $Q_{zz} = Q_{\hat{x}\hat{x}}$)). Such a PDF is shown in Figure 9.5. The PDF is specified by the mean vector and the variance matrix, in this two-dimensional example:

$$E\begin{pmatrix} \underline{z}_1 \\ \underline{z}_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad D\begin{pmatrix} \underline{z}_1 \\ \underline{z}_2 \end{pmatrix} = \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix}$$

Figure 9.6 shows the histograms for 1000 samples, separately of $\hat{x}_1$ and $\hat{x}_2$, and Figure 9.7 (on the left) shows the samples ($\hat{x}_1$,$\hat{x}_2$) in a two-dimensional scatter plot; in both cases they have been corrected for their known true values ($x_1$,$x_2$). So, we had to know the true values for the two coordinates, and we do so here for demonstration purpose.

In practice it is not very convenient to draw — for each pair of coordinate estimators — the PDF in a full three dimensional image, as done in Figure 9.5. Though one would like to present the main feature of this function. This is done by considering a cross-sectional cut of the bell-shaped curve, horizontally, at a certain height $k$; this cut provides an iso contour line.

$$f(z) = \frac{1}{\sqrt{|2\pi Q_{zz}|}} e^{-\frac{1}{2} z^T Q_{zz}^{-1} z} = k$$

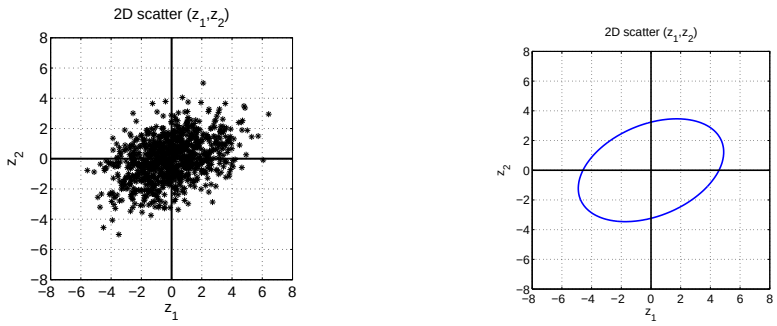



Figure 9.7: On the left: scatter plot of 1000 samples $(z_1,z_2)$; that is, all samples $(\hat{x}_1,\hat{x}_2)$ have been corrected for their known true value $(x_1,x_2)$. On the right: contour of PDF, ellipse of concentration, of position estimator, corrected for true position $(x_1,x_2)$, hence, centered at the origin; an estimated position $(\hat{x}_1,\hat{x}_2)$, also corrected for the true position, is inside the ellipse with probability $1-\alpha = 0.95$ and $k' = \chi^2_\alpha(2,0) = 5.9915$.

hence

$$z^T Q_{zz}^{-1} z = -2\ln(k\sqrt{|2\pi Q_{zz}|}) = k' \tag{9.7}$$

where $\ln$ is the natural logarithm, and vector $z \in \mathbb{R}^n$. By doing this, $f(z) = k$, we obtain an ellipse. The border of this ellipse represents all points $z$ (end-points of vectors $z$) with equal probability density. Optionally it is mentioned that the axes of the ellipse can be found through eigenvalue-decomposition of the variance matrix [1] .

The ellipse nicely captures the shape of the PDF, but still does not tell us much about how much probability is actually contained. For this, we need another type of probability density function, namely the Chi-squared distribution.

If random vector $\underline{z}$, consisting of $n$ random variables, is distributed as $\underline{z} \sim N(0, Q_{zz})$, then $\underline{z}^T Q_{zz}^{-1} \underline{z} \sim \chi^2(n,0)$. The PDF is shown in Figure 9.8, and a table of the Chi-squared distribution can be found in Appendix D. Hence we have for the quadratic form

$$\underline{T} = \underline{z}^T Q_{zz}^{-1} \underline{z} \sim \chi^2(n,0)$$

in practice typically with $n = 2$.

In terms of $\hat{x}$ we have (using $\underline{z} = \underline{\hat{x}} - x$)

$$(\underline{\hat{x}} - x)^T Q_{\hat{x}\hat{x}}^{-1} (\underline{\hat{x}} - x) \sim \chi^2(n,0)$$

Hence, the above constant $k'$ simply follows from the Chi-squared distribution. In the table in Appendix D values can be found for the one-minus-the Cumulative Distribution Function, or to say, the exceedence probability $\alpha$. The area $P[(\underline{\hat{x}} - x)^T Q_{\hat{x}\hat{x}}^{-1} (\underline{\hat{x}} - x) \leq \chi^2_\alpha(n,0)] = 1-\alpha$ contains probability $1-\alpha$, with the values for $\alpha$ on the top-row in the table, and the $\chi^2_\alpha(n,0)$-values tabulated. For the points in this area holds that $f(\hat{x} - x) \geq k$.

Above we considered the probability that the difference of the position estimator with the true position is inside an ellipse, which equals the probability that the true (and in practice,

[1] optional: the orientation (principal axes) of the ellipse, or ellipsoid, can be found through eigenvalue-decomposition of variance matrix $Q_{zz}$:

$$Q_{zz} = U \Lambda U^T$$

with orthogonal matrix $U = (u_1, \ldots, u_n)$ containing the orthonormal eigenvectors $u_i$, and diagonal matrix $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ with positive eigenvalues $\lambda_i$, with $i = 1, \ldots, n$, see also Chapter 7 on symmetric matrices and quadratic forms in [27]. The directions of the principal axes of the ellipse/ellipsoid are given by the vectors $u_i$ in matrix $U$, and the length of the i-th principal axis by $\sqrt{\lambda_i k'}$.

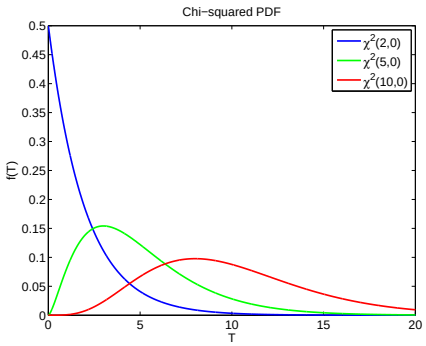



Figure 9.8: Probability density function of a Chi-squared distributed random variable $\underline{T} = \underline{z}^T Q_{zz}^{-1} \underline{z}$; shown are the central Chi-squared distributions for 2, 5 and 10 degrees of freedom respectively. The quadratic form $T$ ranges from zero to infinity.

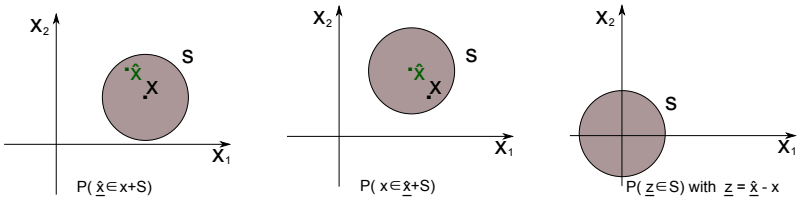



Figure 9.9: The diagram on the left shows, centered at the true position $x$, the area $S$ which contains the position estimator $\underline{\hat{x}}$ with a certain probability (contour of the PDF). The diagram in the middle shows the *confidence region* for the true position $x$, centered at the position estimate $\hat{x}$. The diagram on the right shows the area $S$ for the difference of the position estimator and the true position $\underline{z} = \underline{\hat{x}} - x$, it is centered at the origin (sometimes referred to as the error-region). In these diagrams the area $S$ has simply been shown as a circle.

unknown) position is inside the ellips, but centered at the estimate $\hat{x}$. The ellipse of identical shape as in Figure 9.7 on the right, but centered at the obtained estimate for the position coordinate vector, that is at $(\hat{x}_1, \hat{x}_2)$, is the *confidence ellipse*. It shows the area, which contains the true (but unknown) position, with a certain, specified probability. After all, the goal of surveying is to determine the position of a point of interest. We will never know the actual, true value, but instead we come up with an estimate for it, and then we would like to know, how close our estimate is to the actual, true position, see Figure 9.9.

In the sequel, in Section 9.7, we use the PDF contour ellipse to demonstrate the quality of position solutions using several different measurement set-ups, for instance using solely distance measurements, using solely azimuth measurements, and using a combination of them. Such an analysis is typically done during the design-phase of the survey.

Finally we note that with an estimator for a one-dimensional quantity, we are considering a single random variable, and the error region is just an interval, namely the '$\sqrt{k'}$-times-sigma' interval. In (9.7), the variance matrix then consists of just a single variance, $(\hat{x} - x)^T Q_{\hat{x}\hat{x}}^{-1} (\hat{x} - x) = k'$, and we have

$$\frac{(\hat{x} - x)^2}{\sigma_{\hat{x}}^2} = k'$$

or

$$|\hat{x} - x| = \sqrt{k'} \sigma_{\hat{x}}$$

The corresponding confidence interval is centered at $\hat{x}$, and extends to both sides by $\sqrt{k'} \sigma_{\hat{x}}$ and

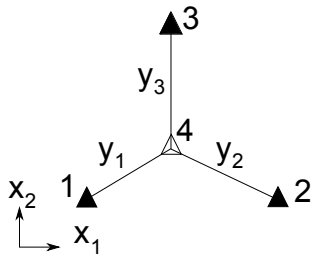



Figure 9.10: Geometric construction to determine position coordinates of new/unknown point 4 based on observed distances to three neighbouring known points.

contains the true $x$ with a probability $1-\alpha$, where values for $k'$ are tabulated in Appendix D, as a function of $\alpha$, with $n=1$.

## 9.6. Example: resection with distances

For a simple example of measuring three distances (from three known points) to determine the position of a fourth point, we will analyse the quality (precision) of the coordinate estimators of this fourth point. It is actually a 'strength'-analysis of the geometric construction (or surveying network in general).

As shown in Figure 9.10 distances are measured from points 1, 2 and 3, to the new/unknown point 4. The coordinates of the points 1, 2 and 3 are known and indicated in the figure: point 1=(0,0), point 2=($100\sqrt{3}$, 0), point 3=($50\sqrt{3}$, 150). Approximate values for the coordinates of point 4 are ($50\sqrt{3}$, 50). Units can be assumed to be in meters. The three distance observables are uncorrelated and all have variance $\sigma^2$ (for instance $\sigma$ = 5 m; quite a large value, but done for convenience here).

In the sequel we set up, and compute the design matrix as it occurs in the model of linearized observation equations (using the approach of Section 8.4). Next, we compute the variance matrix of the coordinate estimators for point 4.

The three non-linear observation equations read:

$$E(\underline{l}_{14}) = \sqrt{((x_1)_4-(x_1)_1)^2+((x_2)_4-(x_2)_1)^2}$$

$$E(\underline{l}_{24}) = \sqrt{((x_1)_4-(x_1)_2)^2+((x_2)_4-(x_2)_2)^2}$$

$$E(\underline{l}_{34}) = \sqrt{((x_1)_4-(x_1)_3)^2+((x_2)_4-(x_2)_3)^2}$$

There are $m=3$ observations, and only $n=2$ unknown parameters, namely $(x_1)_4$ and $(x_2)_4$; the other coordinates are known. With equation (9.5), the given coordinates of points 1, 2 and 3, and $(x_1)_{4,o}$ and $(x_2)_{4,o}$ as the approximate values for $(x_1)_4$ and $(x_2)_4$, the $3\times 2$ design matrix $A$ of the model of linearized observation equations (cf. Eq. (8.9)) becomes

$$A = \begin{pmatrix} \frac{1}{2}\sqrt{3} & \frac{1}{2} \\ -\frac{1}{2}\sqrt{3} & \frac{1}{2} \\ 0 & -1 \end{pmatrix}$$

The variance matrix is simply $Q_{yy}=\sigma^2 I_3$, a scaled identity matrix.

With (8.7) the variance matrix is obtained as

$$Q_{\hat{x}\hat{x}} = \sigma^2 \begin{pmatrix} \frac{2}{3} & 0 \\ 0 & \frac{2}{3} \end{pmatrix}$$

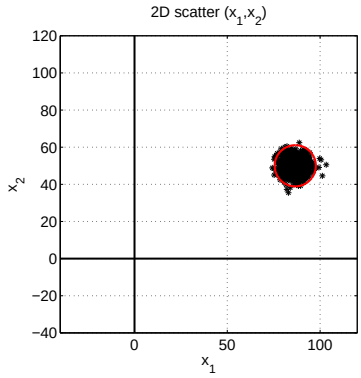



Figure 9.11: Coordinates estimator of point 4 based on distance observables to three neighbouring known points. Shown is the contour ellipse of the PDF, with 97.5% confidence, $k' = 7.3778$, as well as the outcome of carrying out this experiment $N = 1000$ times (each time measuring three distances and determining estimates for $(x_1)_4$ and $(x_2)_4$).

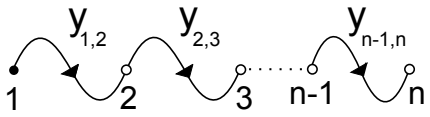



Figure 9.12: Open leveling line from (known) point 1 to point n.

With this particular geometry (a re-section of point 4 with three angles all nicely of 60 degrees), the two coordinate estimators are uncorrelated, and also happen to have equal precision $\sigma_{(\hat{x}_1)_4} = \sigma_{(\hat{x}_2)_4}$. The contour ellipse of the PDF turns into just a circle, see Figure 9.11. This figure also shows a scatter of $N = 1000$ trials of this experiment; 975 of these trials should lie inside the ellipse, and 25 should lie outside.

The above analysis of geometry and precision can be done once approximate values for the unknown parameters are available. Actual measurements are not (yet) needed. As there are no measuremements, we can not iterate as done in the example of Section 8.4, hence we obtain only a first approximation of the measures of precision.

## 9.7. Elementary measurement set-up

In this section we present six of the most elementary local measurement set-ups, three on leveling, which is about positioning in just one-dimension, and three on tachymetry (with angles and distances) in two-dimensional geometry. For these six set-ups we analyze the precision of the coordinate estimators for the unknown points.

### 9.7.1. Leveling

A height difference $y_{ij}$, observed with an optical leveling instrument in a local context, equals the difference of two (orthometric) heights $x_i$ and $x_j$ (apart from a measurement error). The observation equation $E(y_{ij}) = x_j - x_i$ is linear in the unknown parameters (coordinates) $x_i$ and $x_j$. The coordinate system is one dimensional, and the direction of the axis (up) is determined by gravity. The scale is provided by the marks on the levelling rod. With optical leveling, the measurement precision is typically in the order of a few millimeter.

Figure 9.12 presents a so-called *open leveling line*. The height of point 1 is known from an earlier survey, and one levels from 1 to 2, from 2 to 3 etc., until point n. The line consists of $(n-1)$ stretches, also called level-runs.

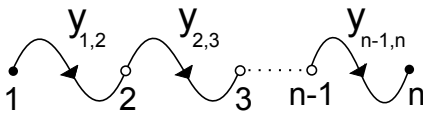



Figure 9.13: Leveling line which terminates at (known) points, point 1 and point n — connected leveling line.

In an open leveling line, the heights for the unknown points simply follow by adding the observed height differences to the height of the known point, just like we did in the example with three points in Chapter 7 (Section 7.3). This set-up does not allow for an internal consistency check. In case a gross error is present in one of the levelled height-differences, it will go unnoticed (and spoil part of the results). Hence, such a set-up is not recommended for critical applications!

When the height difference observables $\underline{y}_{ij}$ have standard deviation equal to $\sigma_{y_{ij}} = \sigma$, and also the given height of point 1 has standard deviation equal to $\sigma_{y_1} = \sigma$ (and all observables are uncorrelated), then the variances for the points along the line are

$$\sigma^2_{\hat{x}_i} = i\sigma^2 \quad \text{for } i = 1, \ldots, n \tag{9.8}$$

This behaviour is shown as a straight line in Figure 9.14. The variance of the height increases linearly with the number of the stretches (accumulation of errors).

What would be the variance of point 1, i.e. $\sigma^2_{\hat{x}_1}$ in case point n would be the known point (instead of point 1)?

Figure 9.13 shows a so-called *connected leveling line* (or, closed level traverse). The heights of both the first and last point are known from an earlier survey. The line connects known points 1 and n, through a series of level runs, visiting unknown points 2, 3, ..., n-1.

This set-up is more secure than an open leveling line, as there is now a check at the end (the line is connected). There is now one more measurement than strictly needed (redundancy equals one).

When the height difference observables $\underline{y}_{ij}$ have standard deviation equal to $\sigma_{y_{ij}} = \sigma$, and also the given heights of points 1 and n have standard deviation equal to $\sigma_{y_1} = \sigma_{y_n} = \sigma$ (and all observables are uncorrelated), then the variances for the points along the line are

$$\sigma^2_{\hat{x}_i} = i\sigma^2[1 - \frac{i\sigma^2}{(n+1)\sigma^2}] \quad \text{for } i = 1, \ldots, n \tag{9.9}$$

Compared with the variance expression for the open leveling line, an additional factor shows up, between the square brackets, and it is smaller than 1, hence the variances of the heights in the connected line are smaller than those in the open line. This behaviour is shown as a curved line in Figure 9.14. The curve (as a function of $i$) is actually determined by the parabola $-i^2 + (n+1)i$, which has its top at $i = \frac{n+1}{2}$. The curve starts at a (slightly) lower value for point 1 than the straight line for the open line, and it is symmetric about $i = 5\frac{1}{2}$.

The above expression for the variance can be checked for instance by considering an example with 4 points, with points 1 and 4 given, and unknown points 2 and 3, and we have three leveled height differences. The model of observation equations (8.2) reads

$$E\begin{pmatrix} \underline{y}_{-1} \\ \underline{y}_{-12} \\ \underline{y}_{-23} \\ \underline{y}_{-34} \\ \underline{y}_{-4} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

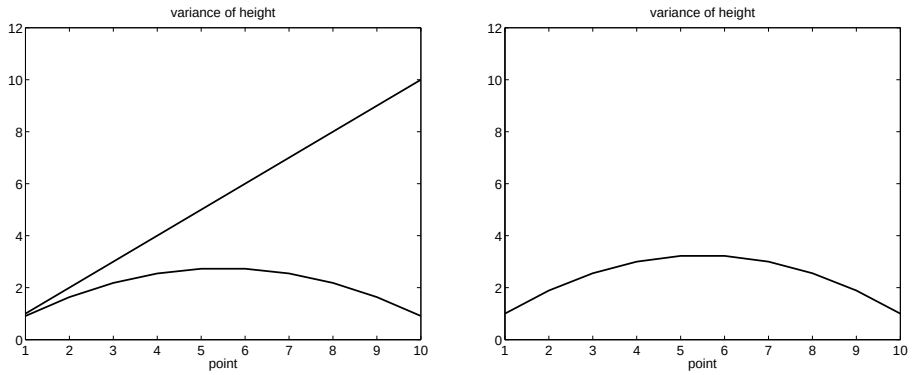



Figure 9.14: Variance of heights determined for points along open and connected leveling line of $n$=10 points (on the left), and along (closed) leveling loop of 9 points, where point 10 equals point 1 again, (on the right). Variance $\sigma^2$ was simply set to 1.

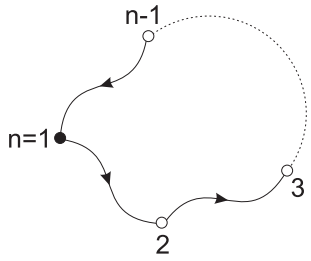



Figure 9.15: Leveling loop of $(n-1)$ points, where (known) point n equals point 1.

and the variance matrix is just a $5 \times 5$ identity matrix, scaled by $\sigma^2$. The expression for $\sigma^2_{\hat{x}_i}$ can be checked by computing the $Q_{\hat{x}\hat{x}}$ matrix, cf. (8.7).

Figure 9.15 shows a *leveling loop*, (or, closed loop level traverse). The leveling 'line' now returns to the starting point, and the height of the first point is known from an earlier survey.

Also this set-up is more secure than an open leveling line, as there is a check: all observed height differences added together need to yield just zero (return to the same point); this is the loop-closure condition. There is now one more measurement than strictly needed (redundancy equals one).

Suppose the leveling loop of Figure 9.15 consists of just four points, hence $n-1=4$. Then there are $m=5$ observations: $y_{12}, y_{23}, y_{34}$ and $y_{41}$, and the height of point 1 is given and fixed - this provides the fifth observation $y_1$. There are three points with unknown heights: $x_2, x_3, x_4$, they are to be determined, and $x_1$ is included in order to be able to relate 'observation' $y_1$ (the given height). The model of observation equations (8.2) reads:

$$E\begin{pmatrix} \underline{y}_{-1} \\ \underline{y}_{-12} \\ \underline{y}_{-23} \\ \underline{y}_{-34} \\ \underline{y}_{-41} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

With $m=5$ observations and $n=4$ unknown parameters, there is a redundancy of 1; to say, there is one measurement too much. This redundancy can be made explicit. The (expectation of the) five observables can be described by four unknown parameters. This implies that there

should hold a *condition* ($m - n = 1$, hence one condition in this case) on the observations.

We multiply both left and right side by the same matrix

$$\underbrace{\begin{pmatrix} 0 & 1 & 1 & 1 & 1 \end{pmatrix}}_{B^T} E \begin{pmatrix} \underline{y}_{-1} \\ \underline{y}_{-12} \\ \underline{y}_{-23} \\ \underline{y}_{-34} \\ \underline{y}_{-41} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

resulting in

$$E(\underline{y}_{-12} + \underline{y}_{-23} + \underline{y}_{-34} + \underline{y}_{-41}) = 0$$

and this is the above mentioned loop closure condition (the expectation operator is linear). It implies that the sum of the leveled height differences is (expected to be) equal to zero. We need to return to the same point, and this provides us with a check. In practice, the sum of the height differences $y_{12} + y_{23} + y_{34} + y_{41} \neq 0$ (though hopefully close to zero), and this is referred to as the *misclosure*. For the $(m - n)$x$m$ matrix $B^T$ describing the condition(s), holds that $B^T A = 0$.

When the height difference observables $\underline{y}_{-ij}$ have standard deviation equal to $\sigma_{y_{ij}} = \sigma$, and also the given height of point 1 has standard deviation equal to $\sigma_{y_1} = \sigma$ (and all observables are uncorrelated), then the variances for the points along the line are

$$\sigma^2_{\hat{x}_i} = \sigma^2 + \frac{\sigma^2}{n-1}(i-1)(n-i) \quad \text{for } i = 1, \ldots, n \tag{9.10}$$

where $n$ is indicating the number of points, as in Figure 9.15. The variance shows a behaviour similar as for the connected line, see Figure 9.14 on the right, driven by a similar parabola, but at a higher level than for the connected line (as there is only one known point involved in the loop, versus two with the connected line). Obviously the variance of point 10 equals the variance of point 1.

### 9.7.2. Intersection with azimuths

In Figure 9.16 the position of an unknown point is determined (in two dimensions) by observing the azimuths at two known points 1 and 2. For four possible locations of the unknown point (all on the line $x_1 = 5$) the PDF contour ellipse is given.

When the unknown point is close to the line connecting the points 1 and 2 (dashed line, which parallels the $x_1$-axis here), the precision is good in the $x_2$-direction (perpendicular to the connection line), but very poor in the $x_1$ direction (along the connection line). This situation is reversed when the point is located far away (to the top or bottom of the graph). An intersection of the two azimuth-lines at the unknown point, at a right angle, delivers a homogeneous precision (with the ellipse of concentration being close to a circle).

### 9.7.3. Polar coordinates

In two dimensions the position of an unknown point can be also be determined using one measurement of azimuth and one measurement of distance, taken at a single known point. This is the typical way of working with a tachymeter or a total station: measuring distances and azimuths (or directions) to points to be determined. In Figure 9.17 PDF contour ellipses are given again for the same four possible locations of the unknown point.

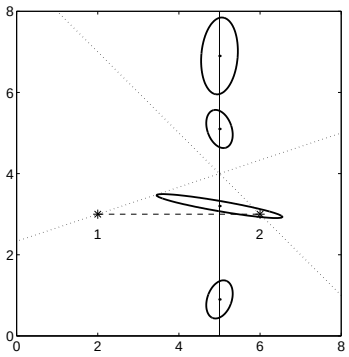



Figure 9.16: Precision of positioning in two dimensions using azimuths, measured at (known) points 1 and 2; PDF contour ellipses ($P$=0.3935) for four possible locations of the unknown point. For visualization of the ellipses, the standard deviation was set as $\sigma_a$ = 0.13 radian.

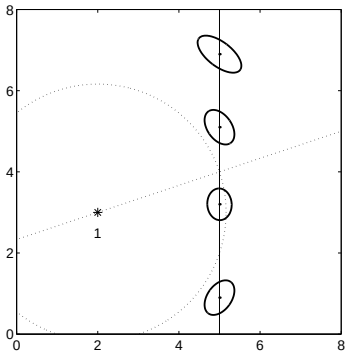



Figure 9.17: Precision of positioning in two dimensions using azimuths and distances (polar coordinates at point 1); PDF contour ellipses for four possible locations of the unknown point. For visualization of the ellipses, the standard deviations were set as $\sigma_a$ = 0.13 radian, and $\sigma_l$ = 0.30.

With the particular choice for the azimuth and distance observables' standard deviations as noted in Figure 9.17, a homogeneous precision is achieved for the two coordinates of the unknown point (ellipse close to a circle), provided that the unknown point is not too far away from the known station 1. For points far away, one has to realize that a fixed uncertainty in the measurement of angle translates — at a larger distance — into a larger uncertainty in the position, in the direction perpendicular to the line connecting the unknown point and the instrument.

### 9.7.4. Intersection with distances

Finally we determine the position of an unknown point using two measurements of distance, a set up which is pretty much similar to the example with azimuths in one of the previous sections. From each of the two known points a distance is measured to the unknown point. In Figure 9.18 PDF contour ellipses are given for the same four possible locations of the unknown point. The standard deviation of the distance observable was kept fixed, whereas in practice it may vary with distance (typically, larger standard deviation for larger distances).

When the unknown point is close to the line connecting the points 1 and 2 (dashed line, which parallels the $x_1$-axis), the precision is good in the $x_1$-direction (the direction along the connection line), but very poor in the other direction (perpendicular). This situation is reversed when the point is located far away (to the top or bottom of the graph). An intersection of the

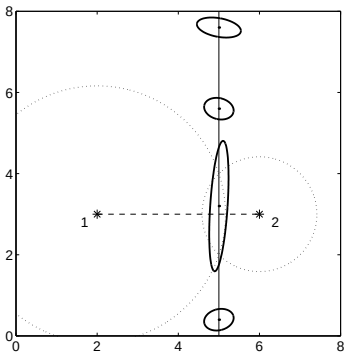



Figure 9.18: Precision of positioning in two dimensions using distance measurements at points 1 and 2; PDF contour ellipses for four possible locations of the unknown point. For visualization of the ellipses, the standard deviation was set as $\sigma_l = 0.30$.

two distance lines at the unknown point, at a right angle, delivers a homogeneous precision (with the ellipse of concentration being close to a circle).

# 10

# Validation

In Chapter 8 we introduced the mathematical model $\underline{y} = Ax + \underline{e}$, with $m \geq n$, where $m$ is the dimension of vector $y$, and $n$ the dimension of vector $x$. The system $y = Ax$ is generally not consistent, hence, $y \neq Ax$. The least-squares estimate (8.4), and the minimum variance estimator (8.6) were introduced, Sections 8.3.2 and 8.3.4. Once an estimate $\hat{x}$ for the unknown parameters is available, one can compute an estimate for the observations: $\hat{y} = A\hat{x}$, and this system is *consistent*. One could regard this parameter estimation as 'changing' (or adjusting) the observed values from $y$ to $\hat{y}$ in order to turn a non-consistent system into a consistent one. As outlined with the least-squares criterion (8.5), one keeps vector $y - \hat{y}$ as short as possible, the observed values should be 'changed only as little as possible'.

In this chapter we introduce the least-squares residuals, and show how they can be used in an overall *consistency check*, to answer the question whether the collected measurements $y$ and the assumed model $Ax$ can be deemed to be mutually consistent. Next we present a worked example of line-fitting (regression). Eventually we briefly introduce the more advanced (optional) subject of observation testing, and present the final results of our data processing and analysis.

## 10.1. Least-squares residuals

Once an estimator $\underline{\hat{y}}$ is available for the vector of observables, the least-squares residuals follow as

$$\underline{\hat{e}} = \underline{y} - \underline{\hat{y}} \tag{10.1}$$

The least-squares residuals measure the difference between the observations (as measured) $y$, and the estimated, or adapted ones $\hat{y}$ (see Section 8.3.2, $\underline{\hat{y}} = A\underline{\hat{x}}$). The least-squares residuals provide an estimate for the (unknown) measurement error $e$ (that is why also the residuals are denoted with a hat-symbol). They carry important diagnostic information about the parameter estimation process. When the residuals are small, the situation is looking good. One does not need to 'change' the observed values by much, in order to make them fit into the model $Ax$. However, when they are large, this might be a reason for reconsideration. It could be that there are large outliers or faults present in the measurements (e.g. entering an observed height difference as 0.413 m, instead of 0.143 m), or that the assumed model is not appropriate for the case at hand (in a dynamic system, with a moving object, we may assume that the object is moving with constant velocity, but this may turn out not to be the case). Small residuals tend to be OK — large ones are not. But what is small, and what is big? Fortunately, we can devise an objective criterion to judge on their size.

### 10.1.1. Overall model test - consistency check

When the data are normally distributed, $\underline{y} \sim N(Ax, Q_{yy})$, under the working model (the null-hypothesis), then also the residuals will be normally distributed (with zero mean). It can be shown — though the proof is omitted here — that $\underline{\hat{e}}^T Q_{yy}^{-1} \underline{\hat{e}}$ has — under the working model (8.2) — a central Chi-squared distribution with $m-n$ degrees of freedom, hence $\chi^2(m-n, 0)$ (from the $m$-vector of observations, $n$ unknown parameters in $x$ are estimated, and hence only $(m-n)$ degrees of freedom are 'left' for the residuals). The Chi-squared distribution was introduced in Section 9.5, and shown in Figure 9.8. The *squared norm of the residuals-vector* $\underline{\hat{e}}^T Q_{yy}^{-1} \underline{\hat{e}}$ is an *overall* measure of consistency[1] [2]. It provides an objective criterion on judging the amount by which we needed to 'change' the observed values, to fit the assumed or supposed model ($Ax$).

$$\underline{T} = \underline{\hat{e}}^T Q_{yy}^{-1} \underline{\hat{e}} \sim \chi^2(m-n, 0) \tag{10.2}$$

Now one can set a level of significance (probability) $\alpha$, see Chapter 26 in [2], e.g. 5%, and not accept the residual vector $\hat{e}$, once its squared norm is located in the upper 5% of the distribution (right tail); occurrence of such a value is deemed to be too unlikely to be true under the working model. When there are large outliers or faults present in the observations, or when the assumed model is not appropriate, the residuals tend to be larger, leading to a *larger* value for the test-statistic $T$, and hence we set a right-sided critical region.

In practice, one computes $T = \hat{e}^T Q_{yy}^{-1} \hat{e}$, and retrieves threshold $k' = \chi^2_\alpha(m-n, 0)$ from the table and concludes (decides):

- model and data are consistent when $T < k'$
- model and data are *not* consistent when $T > k'$

For example with $m$=5 and $n$=2, $m-n$=3, and with a 5% level of significance, the threshold (or critical) value $k'$ for the above squared norm of the least-squares residuals, $T$, is 7.8147, see the table in Appendix D ($\chi^2_\alpha(3, 0) = 7.8147$).

### 10.1.2. Simplification

In the simple case that the observables' variance matrix is a diagonal matrix $Q_{yy} = \mathrm{diag}(\sigma^2_{y_1}, \sigma^2_{y_2}, \ldots, \sigma^2_{y_m})$ (all observables are uncorrelated), the above overall model test statistic $\underline{T}$ can be given a simple and straightforward interpretation. Namely

$$\underline{T} = \underline{\hat{e}}^T Q_{yy}^{-1} \underline{\hat{e}} = \sum_{i=1}^{m} \frac{\underline{\hat{e}}_i^2}{\sigma^2_{y_i}} = \sum_{i=1}^{m} \left( \frac{\underline{\hat{e}}_i}{\sigma_{y_i}} \right)^2 \tag{10.3}$$

and it compares — per observation — the residual $\hat{e}_i$ with the standard deviation $\sigma_{y_i}$ of, i.e. the expected uncertainty in, the observable $\underline{y}_i$.

The overall model test statistic equals the sum of all those $m$ squared ratios. The overall model test, as the name says, aims to *detect*, in general sense, any inconsistency between observed data and the proposed or assumed model.

[1] actually a measure of in-consistency

[2] formally $\underline{\hat{e}}^T Q_{yy}^{-1} \underline{\hat{e}}$ is the square of the *weighted* norm (weighted because of $Q_{yy}^{-1}$) of the least-squares residuals; in statistics unweighted $\underline{\hat{e}}^T \underline{\hat{e}}$ is referred to as sum of squared residuals (SSR)

### 10.1.3. Discussion

As an example we consider the case in which the same unknown distance is measured twice, cf. Section 8.3.1. If one of these measurements is biased by a large amount (for instance 10 m), and the other is not, then the overall model test will likely detect that the two measurements are not consistent with the model, namely, according to the model (measuring the same distance twice) the numerical values of two measurements should be the same, or close together. Intuitively, as the measurements are not close together, this gives rise to suspicion (there might be something wrong with the measurements). Anomalies in the measurements which cause the data to be (still) consistent with the assumed model can*not* be detected by this test. In our example, if both observations are biased in the same way, let us say by 10 m, then the two observations are still consistent (with each other in the model; they are both in error). Intuitively, as their values are the same or close together, this does not raise any suspicion. Being able to detect all relevant anomaly scenarios is part of designing a good measurement set-up.

### 10.1.4. Example: repeated measurements [*]

Suppose $m$ measurements of the same unknown quantity are made. Then the general model of observation equations $E(\underline{y}) = Ax;\ D(\underline{y}) = Q_{yy}$ (8.2) reads:

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_m \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} x \ ; \ D\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_m \end{pmatrix} = \sigma^2 I_m$$

where we assumed that all observables have equal precision (all variances equal to $\sigma^2$), and there is no correlation. There are $m$ observations, and there is $n = 1$ unknown parameter.

In this case, the minimum variance estimate $\hat{x}$ for the unknown parameter $x$, equals just the mean of the observations, similar to (6.12).

In this simple example $\hat{y}_i = \hat{x}$, and hence, the least-squares residuals $\hat{e}_i = y_i - \hat{x}$. The squared norm of the residuals vector becomes

$$T = \hat{e}^T Q_{yy}^{-1} \hat{e} = \frac{1}{\sigma^2} \sum_{i=1}^{m} \hat{e}_i^2 = \frac{1}{\sigma^2} \sum_{i=1}^{m} (y_i - \hat{x})^2$$

which shows that, in this case, this overal model test-statistic is closely related to the sample variance (6.14), which you determine based on measurements taken. We have

$$\frac{T}{m-1} = \frac{1}{\sigma^2} \frac{1}{m-1} \sum_{i=1}^{m} (y_i - \hat{x})^2 = \frac{\hat{\sigma}^2}{\sigma^2}$$

where we use $m$ instead of $N$ as in (6.14). The overall model test statistic equals, in this case, the ratio of the sample variance and the formal, a-priori variance.

## 10.2. Example

In this example we consider the classical problem of *line fitting*. This problem is frequently encountered in science and engineering. One can think of measuring the extension of a certain object (e.g. of steel) due to temperature, for which one assumes a linear behaviour, so the length of the object is measured, at different temperatures, and next one would like to determine the length of the object at some reference temperature and the coefficient of extension,

i.e. by how much it extends for every increase of one degree in temperature. In literature, this subject is often referred to as regression analysis, see e.g. Chapter 6 on orthogonality and least-squares in [27], in particular Section 6 with applications to linear models. The subject can be seen much broader however, namely as *curve fitting*. The principle is not restricted to just straight lines, one can also use parabolas and higher degree polynomials for instance.

In this example we consider a vehicle which is driving along a straight line, and we are interested in the position along the road. Therefore, a laser tracker is used, and this device measures/reports the position of the vehicle every second. For convenience, the laser-tracker is at the origin of this one-dimensional coordinate system.

Over a period of four seconds, we take measurements: $y = (y_1, y_2, y_3, y_4)^T$, hence the vector of observations has dimension $m = 4$. Correspondingly at times $t_1$, $t_2$, $t_3$ and $t_4$, the unknown positions are $x(t_1)$, $x(t_2)$, $x(t_3)$ and $x(t_4)$. The measurements equal the unknown positions, apart from measurement errors, i.e. $y_i = x(t_i) + e_i$ for $i = 1, \ldots, 4$.

$$E\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} x(t_1) \\ x(t_2) \\ x(t_3) \\ x(t_4) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x(t_1) \\ x(t_2) \\ x(t_3) \\ x(t_4) \end{pmatrix}$$

In case we determine these $n = 4$ four unknown positions from the $m = 4$ four observed positions, estimation is pretty trivial, namely $\hat{x}(t_i) = y_i$ for $i = 1, \ldots, 4$.

We have reasons however, to assume that the vehicle is driving at constant speed, and thereby we can model the unknown motion of the vehicle, by just an unknown initial position $x(t_0)$, and its velocity $\dot{x}$.

$$\begin{pmatrix} x(t_1) \\ x(t_2) \\ x(t_3) \\ x(t_4) \end{pmatrix} = \begin{pmatrix} 1 & (t_1 - t_0) \\ 1 & (t_2 - t_0) \\ 1 & (t_3 - t_0) \\ 1 & (t_4 - t_0) \end{pmatrix} \begin{pmatrix} x(t_0) \\ \dot{x} \end{pmatrix}$$

which we can substitute in the above system of observation equations, and hence

$$E\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & (t_1 - t_0) \\ 1 & (t_2 - t_0) \\ 1 & (t_3 - t_0) \\ 1 & (t_4 - t_0) \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x(t_0) \\ \dot{x} \end{pmatrix}}_{x}$$

Now there are still $m = 4$ observations, but just $n = 2$ unknown parameters in vector $x$. The distance observables by the laser tracker are all uncorrelated, and all have the same variance $\sigma^2$. Hence $Q_{yy} = \sigma^2 I_4$. Standard deviation $\sigma$ can be taken here as $\sigma = 1$.

In the sequel we develop the example into a real numerical example. The observation times are $t_1 = 1$, $t_2 = 2$, $t_3 = 3$ and $t_4 = 4$ seconds, and $t_0 = 0$ (and timing is assumed here to be perfect — no errors; all coefficients of matrix $A$ are known — when this assumption cannot be made, refer to Appendix B.6).

As can be seen in Figure 10.1, we try to fit a straight line through the observed data points. Therefore we estimate the offset/intercept of the line $x(t_0)$, and its slope $\dot{x}$; in terms of regression, they are the (unknown) regression coefficients. In this example time $t$ is the explanatory or independent variable (also called the regressor); the (observed) position depends on time $t$. And, the observation $y$ is the dependent (or response) variable. Least-squares estimation will yield a best fit of the line with the observed data points, through minimizing (8.5).

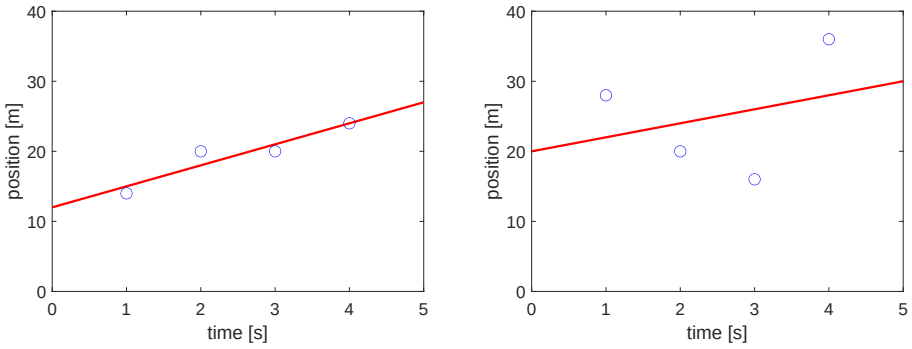



Figure 10.1: Line fitting on day 1 (left) and day 2 (right). Shown are the observed data points as blue circles, and the straight line in red, fitted by means of least-squares.

The experiment was repeated the next day, and in the sequel we consider — simultaneously — the outcomes of *both* experiments. The measurements on day 1 were $y = (14, 20, 20, 24)^T$, and on day 2 $y = (28, 20, 16, 36)^T$. In Figure 10.1, the observed distances (in meters) are shown, together with the fitted line, based on the least-squares estimates. Verify yourself that, with (8.6), one obtains

$$\hat{x} = \begin{pmatrix} 12 \\ 3 \end{pmatrix} \text{for day 1, and } \hat{x} = \begin{pmatrix} 20 \\ 2 \end{pmatrix} \text{for day 2}$$

with units in meters and meters per second respectively.

With $\hat{x}(t_0) = 12$ and $\hat{\dot{x}} = 3$ on day 1, one can determine $\hat{y}$ through $\hat{y} = A\hat{x}$, and eventually obtain $\hat{e} = y - \hat{y}$. And do this for day 2 as well.

$$\hat{e} = \begin{pmatrix} -1 \\ 2 \\ -1 \\ 0 \end{pmatrix} \text{ for day 1, and } \hat{e} = \begin{pmatrix} 6 \\ -4 \\ -10 \\ 8 \end{pmatrix} \text{ for day 2}$$

From these numbers, and also Figure 10.1, one can already conclude that for day 1 (on the left) we have a pretty good fit, while on day 2 (on the right) the fit is pretty poor, likely a second degree polynomial would do here much better (quadratic polynomial). It indicates that the motion of the vehicle on day 2 has not really been at constant speed. For an objective criterion in judging a good and poor fit, we use the squared norm of the residual vector (10.2).

$$T = \hat{e}^T Q_{yy}^{-1} \hat{e} = 6 \text{ for day 1, and } T = \hat{e}^T Q_{yy}^{-1} \hat{e} = 216 \text{ for day 2}$$

With the table in Appendix D, we find that with $m - n = 2$ and $\alpha = 0.01$ the threshold value equals $\chi^2_\alpha = 9.2103$, and hence the line fit of day 1 is not rejected ($6 < 9.2103$), but the line fit of day 2 is rejected ($216 > 9.2103$)!

At this stage we recall that least-squares estimation is driven, see (8.5), by minimization of the squared norm of the vector $y - Ax$, see Figure 10.2. Least-squares minimizes the objective function $\min_x \|y - Ax\|^2$. Vector $y - Ax$ contains the differences between the actual observations $y$ and the modeled observations by $Ax$. One should choose values for the elements in vector $x$, such that the squared norm of th vector $y - Ax$ is at minimum. Figure 10.2 shows at left, for the observations of day 1, along the vertical axis, the squared norm of the vector $y - Ax$, which is just a single number, as a function of $x(t_0)$ and $\dot{x}$ in the horizontal plane, with offset parameter $x(t_0)$ along the axis in front, and slope parameter $\dot{x}$ along the axis to the back.

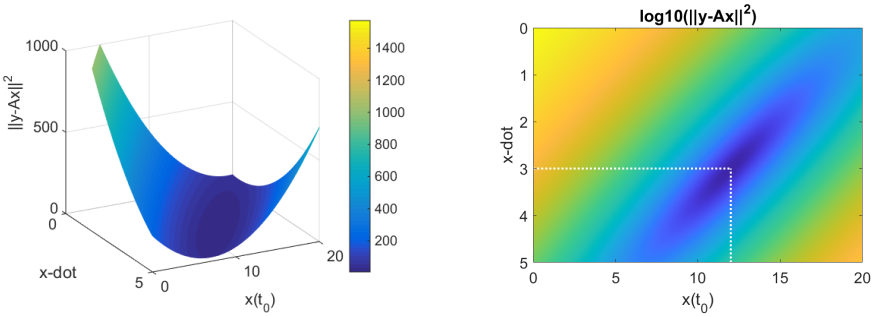



Figure 10.2: The sum of squared residuals as a function of $x(t_0)$ and $\dot{x}$. At right, actually the $\log 10$-value is shown. The minimum, which is equal to 6, is attained for $x(t_0)$=12 and $\dot{x}$=3, indicated by the dotted lines, hence $\min_x \|y - Ax\|^2 = 6$ (for day 1).

We have to keep in mind that the true motion of the vehicle is unknown. It is *our assumption* that we can mathematically describe the motion through a constant velocity model. For day 1 there is no indication that this is not working. But for day 2, a constant velocity model, most likely, is not providing a proper description of the actual motion. Maybe a quadratic model (constant acceleration) gives a better/acceptable fit.

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \underline{y}_3 \\ \underline{y}_4 \end{pmatrix} = \begin{pmatrix} 1 & (t_1 - t_0) & \frac{1}{2}(t_1 - t_0)^2 \\ 1 & (t_2 - t_0) & \frac{1}{2}(t_2 - t_0)^2 \\ 1 & (t_3 - t_0) & \frac{1}{2}(t_3 - t_0)^2 \\ 1 & (t_4 - t_0) & \frac{1}{2}(t_4 - t_0)^2 \end{pmatrix} \begin{pmatrix} x(t_0) \\ \dot{x}(t_0) \\ \ddot{x} \end{pmatrix}$$

There are now $n = 3$ unknown parameters, with still $m = 4$ observations.

Mind that if we go for a third degree polynomial (cubic polynomial), we will have $n = 4$ unknown parameters, and all information contained in the $m = 4$ observations is strictly needed to determine the unknown parameters, and 'nothing will be left' for the least-squares residuals, as $m - n = 0$. Without redundancy, the overall model test gets void. In that case it looks like the observations perfectly fit the assumed model, but you actually do not have any means to verify this.

## 10.3. Observation testing and outlook [*]

The squared norm of the residuals is an overall measure of consistency between observations and assumed mathematical model. The statistical test of $\hat{e}^T Q_{yy}^{-1} \hat{e}$ being smaller or larger than the critical value $k'$ from the Chi-squared distribution is referred to as the overall model test, see Section 10.1.1. In literature you may also encounter it as the F-test, then $\underline{\hat{e}}^T Q_{yy}^{-1} \underline{\hat{e}}/(m - n) \sim F(m - n, \infty, 0)$ under the working model (null-hypothesis), where $F$ represents the F-distribution.

The test can be derived from the principle of statistical hypothesis testing. More specific statistical hypothesis tests exist, for instance tests to identify outliers, blunders, faults and anomalies in single observations. A true coverage of this subject is beyond the scope of this book, but we will introduce — without any derivation, or proof of optimality — a simple test which aims to identify an *outlier* in a set of observations (then only *one* observation from this set is affected by the outlier). Typically this test is used multiple times, namely to test *each* of the $m$ observations in vector $y = (y_1, \ldots, y_m)^T$ separately. It is based again on the least-squares residuals $\underline{\hat{e}} = \underline{y} - \underline{\hat{y}}$. Using the error propagation law (7.9), with (8.6), and $\underline{\hat{y}} = A\underline{\hat{x}}$,

one can derive that

$$Q_{\hat{e}\hat{e}} = Q_{yy} - A(A^T Q_{yy}^{-1} A)^{-1} A^T = Q_{yy} - A Q_{\hat{x}\hat{x}} A^T$$

where $Q_{\hat{x}\hat{x}}$ was given by (8.7).

Under the working model the least-squares residuals have zero mean, hence $\underline{\hat{e}} \sim N(0, Q_{\hat{e}\hat{e}})$, and the vector $\underline{\hat{e}}$ is normally distributed as it is a linear function of $\underline{y}$, which is also taken to be normally distributed.

The least-squares residuals $\underline{\hat{e}}$ is a vector with $m$ random variables $\underline{\hat{e}} = (\underline{\hat{e}}_1, \underline{\hat{e}}_2, \ldots, \underline{\hat{e}}_m)^T$. For each of the residuals we have $\underline{\hat{e}}_i \sim N(0, \sigma_{\hat{e}_i}^2)$ with $i = 1, \ldots, m$. Usually, if (just) observation $y_i$ contains a large error, the corresponding residual $\hat{e}_i$ will deviate (substantially) from zero. We use this to propose — valid only for the case when the observables have a diagonal variance matrix $Q_{yy}$ — the w-test statistic as

$$\underline{w}_i = \frac{\underline{\hat{e}}_i}{\sigma_{\hat{e}_i}} \tag{10.4}$$

and check whether it deviates from zero.

Division of the residual by the standard deviation $\sigma_{\hat{e}_i}$ (you may use (7.9) again) causes the w-test statistic to be *standard normally distributed*: $\underline{w}_i \sim N(0, 1)$. In fact, it is the *normalized* residual. Setting a level of significance $\alpha$, one can find the critical value. Mind that deviations both in positive and negative direction may occur. Hence, the level of significance $\alpha$ is split into $\frac{\alpha}{2}$ for the right tail, and $\frac{\alpha}{2}$ for the left tail. The critical value $\tilde{k} = N_{\frac{\alpha}{2}}(0, 1)$ follows from the table in Appendix C. The hypothesis 'no outlier present in observation $y_i$' is rejected if $|w_i| > \tilde{k}$, i.e. if $w_i < -\tilde{k}$ or $w_i > \tilde{k}$. The test is typically done for all observations $i = 1, \ldots, m$.

In practice the largest of the w-teststatistics (in absolute sense) indicates the most likely faulty observation. This observation is removed from the data, and estimation and validation is repeated with one less observation, until no more measurements are rejected, or until redundancy runs low. This procedure of 'data checking' can be carried out automatedly, without human intervention, and allows for a robust processing of the measurements. This checking provides a safeguard against producing, unknowingly, largely incorrect results.

There are many alternative approaches to robust estimation, delivering robust estimators, which are insensitive to outliers, but these are beyond the scope of this book.

## 10.4. Example — continued

Eventually we wrap up, and present the final results of our data processing and analysis. We have computed the estimates for the parameters of interest in the example of Section 10.2, namely the initial position $x(t_0)$ and the velocity $\dot{x}$, and we have verified the consistency of the model and the data.

In Figure 10.3 we consider the data of day 1, hence giving a good fit, see Figure 10.1 at left. In Figure 10.3 at left, we present again the four measurements, and, we indicate with them the measurement uncertainty. The standard deviation of each measurement was $\sigma = 1$. The graph presents the so-called error-bars for a confidence level of 95%. When the observables are normally distributed, one can verify with the table in Appendix C, that the error-bar extends by $r_\alpha \sigma$ to either side of the observed value, with $r_\alpha$ = 1.96.

The graph at right presents again the fitted line, based on the least-squares estimate $\hat{x}$. As outlined in the introduction of this chapter, estimates for the observations can be computed as $\hat{y} = A\hat{x}$. With this, you can check yourself that while the observation $y_2$ = 20, $\hat{y}_2$ = 18 (and the latter is located on the red line).

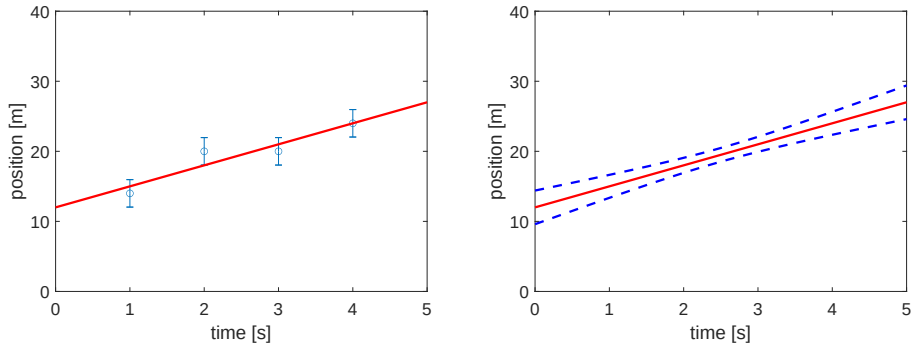



Figure 10.3: Line fitting on day 1: observed data points as small blue circles, with error-bars, also in blue, at the left, and the straight line fitted by means of least-squares with confidence interval at the right.

The corresponding variance matrix of $\underline{\hat{y}}$ can be found, using (7.9) on $\underline{\hat{y}} = A\underline{\hat{x}}$ as (7.7), as $Q_{\hat{y}\hat{y}} = AQ_{\hat{x}\hat{x}}A^T$, with $Q_{\hat{x}\hat{x}}$ from (8.7). So, for the estimated observation at $t_2$ holds that $\sigma_{\hat{y}_2} \approx 0.55$, whereas $\sigma_{y_2} = 1$ (hence, it is much smaller). Then the 95% confidence interval for the true $y_2$ is centered at $\hat{y}_2$, and extends to $r_\alpha \sigma_{\hat{y}_2} \approx 1.07$ on either side (and hence it is much smaller than with the individual observation, in the graph at left). Suppose the experiment is repeated, then 95% of the realizations of this (random) interval will actually contain the true position, see also Chapter 23 in [2]. In fact, the confidence interval can be presented for the position at any time $t$, see the dashed blue lines on either side of the fitted line, in the graph at right; it is the confidence interval for $y(t)$ based on $\hat{y}(t)$, for any time $t$.

## 10.5. Exercises and worked examples

This section presents several problems with worked answers on parameter estimation and validation.

**Question 1** To determine the sea level height in the year 2000, and also its rate of change over time, five observations of the sea level are available, for instance from a tide gauge station. The observations are:

- $y_1$ = 25 mm, in the year 1998
- $y_2$ = 24 mm, in the year 1999
- $y_3$ = 27 mm, in the year 2000
- $y_4$ = 26 mm, in the year 2001
- $y_5$ = 28 mm, in the year 2002

Based on these five observations, compute the *least-squares* estimates for the sea level in the year 2000 and the rate of change. All quantities are referenced to the start of the year. The rate of change can be considered constant over the time span considered.

**Answer 1** Similar to the example of Section 10.2, the model reads

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \underline{y}_3 \\ \underline{y}_4 \\ \underline{y}_5 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x(t_0) \\ \dot{x} \end{pmatrix}}_{x}$$

and basically we need to fit a straight line through five data points. In the above $y = Ax$ model, $x(t_0)$ is the *unknown* sea level height in the year 2000 (expressed in [mm]), and $\dot{x}$ is the (assumed constant) rate of change of the sea level (and also unknown, and expressed in [mm/year]). On the left hand side we have the five observations, of which the third, $y_3$, is the *observed* sea level height in the year 2000 (in [mm]). For example, the last observation, $y_5$, is related to the two unknown parameters as: $y_5$ equals (on average) the sum of the sea level in the year 2000 $x(t_0)$, plus twice the yearly change $\dot{x}$. These two coefficients, 1 and 2, show up, as the last row, in the A-matrix. There is no information given with regard to the precision of the observables (no matrix $Q_{yy}$), hence we use the basic least-squares equation (8.4) $\hat{x} = (A^T A)^{-1} A^T y$.

$$A^T A = \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix} \quad (A^T A)^{-1} = \begin{pmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{10} \end{pmatrix}$$

and $\hat{x} = (A^T A)^{-1} A^T y$ becomes

$$\hat{x} = \begin{pmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{10} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 25 \\ 24 \\ 27 \\ 26 \\ 28 \end{pmatrix} = \begin{pmatrix} 26 \\ \frac{8}{10} \end{pmatrix}$$

The least-squares estimate for the sea level height in the year 2000 is 26 mm, and the rate of change is 0.8 mm/year. Note that the least-squares estimate for the height in the year 2000 does *not* equal the observed height ($y_3$). The least-squares estimate is determined based on *all* available observations. In an era of climate change and sea-level rise, conscientious and responsible analysis and interpretation of sea-level height measurements over time is essential to design and maintenance of coastal defense infrastructure against flooding, see Figure 10.4.

Figure 10.4: The Oosterscheldekering, a series of dams and storm surge barriers to protect the province of Zeeland from flooding from the North-Sea. Photo by Rijkswaterstaat, 2007, taken from Beeldbank Rijkswaterstaat, under BY-NC license [28].

**Question 2** With the model and observations of the previous question, determine whether the overall model test is passed or not, when the level of significance is set to 10%. The observables can be assumed to be all uncorrelated, and have a standard deviation of 1 mm (which is not really a realistic value in practice, but it is fine for this exercise).

**Answer 2** The variance matrix of the observables reads $Q_{yy} = I_5$, and this does not change anything to the computed least-squares estimates (see Eq. (8.6)). The overall model test is

$T = \hat{e}^T Q_{yy}^{-1} \hat{e}$, hence we need the vector of least-squares residuals (10.1) $\hat{e} = y - \hat{y} = y - A\hat{x}$.

$$\begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \hat{e}_3 \\ \hat{e}_4 \\ \hat{e}_5 \end{pmatrix} = \begin{pmatrix} 25 \\ 24 \\ 27 \\ 26 \\ 28 \end{pmatrix} - \begin{pmatrix} 24.4 \\ 25.2 \\ 26.0 \\ 26.8 \\ 27.6 \end{pmatrix} = \begin{pmatrix} 0.6 \\ -1.2 \\ 1.0 \\ -0.8 \\ 0.4 \end{pmatrix}$$

Then the value for the overall model test statistic becomes $T = \hat{e}^T Q_{yy}^{-1} \hat{e}$ = 3.6. The threshold, with $\alpha = 0.1$, is $k' = \chi^2_\alpha(m-n, 0)$, and $m = 5$ and $n = 2$, hence $m - n = 3$. With the table in Appendix D we obtain $k' = \chi^2_{0.1}(3, 0)$ = 6.2514, and hence $T < k'$, and the overall model test is accepted. There is no reason to suspect that something is wrong; the assumed model and the made observations seem to be in agreement with each other, they seem to be consistent; the fit is good.

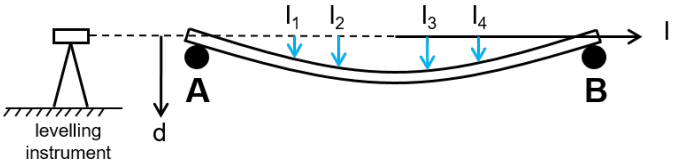



Figure 10.5: The deflection $d$ of a bridge deck is observed at four positions $l_1$, $l_2$, $l_3$, and $l_4$.

| $i$ | $l_i$ | $d(l_i)$ |
|---|---|---|
| 1 | -2 | 1 |
| 2 | -1 | 4 |
| 3 | 1 | 3 |
| 4 | 2 | 2 |

Table 10.1: Measurements of deflection $d(l_i)$ at four positions $l_1$, $l_2$, $l_3$, and $l_4$ at the bridge deck.

**Question 3** A bridge deck is supported by pillars at points A and B. By its own weight, the deck in between will bend (deflect) as shown in Figure 10.5. A vertical cross-section along the bridge deck center-line, i.e. coordinate-axis $l$, is shown. The deflection $d$ is modeled as a *quadratic* function of coordinate $l$: $d(l) = x_1 l^2 + x_2 l + x_3$, with unknown coefficients $x_1$, $x_2$, and $x_3$. Using leveling, the downward deflection $d$ (downward is positive) has been observed at four positions along the deck. The measurements are listed in Table 10.1. Set up the model of observation equations $E(\underline{y}) = Ax$, for the four observations, for the goal of estimating the unknown coefficients $x_1$, $x_2$, and $x_3$.

**Answer 3** We have four observation equations, according to the given quadratic function $d(l) = x_1 l^2 + x_2 l + x_3$. Hence,

$$E\begin{pmatrix} \underline{d}(l_1) \\ \underline{d}(l_2) \\ \underline{d}(l_3) \\ \underline{d}(l_4) \end{pmatrix} = \begin{pmatrix} l_1^2 & l_1 & 1 \\ l_2^2 & l_2 & 1 \\ l_3^2 & l_3 & 1 \\ l_4^2 & l_4 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 4 & -2 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

with observation vector $y$ as

$$\begin{pmatrix} 1 \\ 4 \\ 3 \\ 2 \end{pmatrix}$$

The horizontal coordinate $l$ is assumed to be known exactly in this exercise.

**Question 4** With the problem of the bridge deck in the previous question, one can assume that *maximum* deflection occurs (naturally) exactly in the middle, and we conveniently choose the origin of the $l$-coordinate axis in the middle ($l = 0$), in between the two support points A and B. Present the accordingly simplified model of observation equations.

**Answer 4** To find the extremum of the deflection function $d(l)$, we set the first derivative to zero: $\frac{\partial d(l)}{\partial l} = 2lx_1 + x_2 = 0$, which has to occur for $l = 0$, hence $x_2 = 0$. With the later found value for $x_1$ one can verify that this extremum indeed is a maximum (with positive deflection downward). Now, the second column of the A-matrix together with the given zero value for $x_2$ can be brought to the left-hand side of the equation. But, as $x_2 = 0$, this cancels all together. The resulting model of observation equations reads

$$E\begin{pmatrix} \underline{d}(l_1) \\ \underline{d}(l_2) \\ \underline{d}(l_3) \\ \underline{d}(l_4) \end{pmatrix} = \begin{pmatrix} 4 & 1 \\ 1 & 1 \\ 1 & 1 \\ 4 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_3 \end{pmatrix}$$

Matrix product $A^TA$ is found as

$$A^TA = \begin{pmatrix} 34 & 10 \\ 10 & 4 \end{pmatrix}$$

hence

$$(A^TA)^{-1} = \frac{1}{36}\begin{pmatrix} 4 & -10 \\ -10 & 34 \end{pmatrix}$$

and the least-squares estimates, according to (8.4), are found as

$$\begin{pmatrix} \hat{x}_1 \\ \hat{x}_3 \end{pmatrix} = \frac{1}{6}\begin{pmatrix} 1 & -1 & -1 & 1 \\ -1 & 4 & 4 & -1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}$$

and we obtain $\hat{x}_1 = -\frac{2}{3}$ and $\hat{x}_3 = \frac{25}{6}$.

# 11

# Interpolation

In this chapter we cover the subject of interpolation. After the introduction we cover deterministic interpolation, and next stochastic interpolation, thereby focussing in particular on Kriging.

## 11.1. Introduction

Interpolation is about determining the value of an attribute, like height or water depth, at a certain position within a spatial domain, from a set of observations of that attribute in that domain. For instance, with a hydrographic survey, by means of echo sounding, the attribute water depth is measured, and the measurements $y_1, \ldots, y_m$ are taken at specific discrete positions, for instance while the vessel is sailing along a regular pattern (forth and back) over the water-way. Next, one is interested to know the water depth $z$ at *another* position, where *no* specific *measurement* is available, see Figure 11.1.

One of the most common techniques is *linear* interpolation. The value of the attribute $z_0$ at position $p_0$ (position coordinate vector) is estimated as a linear combination

$$\hat{z}_0 = w_1 y_1 + \cdots + w_m y_m \tag{11.1}$$

of observations $y_1, \ldots, y_m$ at positions $p_1, \ldots, p_m$ in the neighborhood of $p_0$. The coefficients $w_1$ to $w_m$ indicate the weights given to each of the observations. By stacking the observations in vector $y = (y_1, \ldots, y_m)^T$, and the weights in vector $w = (w_1, \ldots, w_m)^T$, Eq. (11.1) can be summarized into

$$\hat{z}_0 = w^T y \tag{11.2}$$

Linear interpolation is not to be confused with the interpolated values being a linear function of time or position; this may be the case, cf. Figure 11.10, but generally it is not, cf. Figure 11.6 at right and Figure 11.7 at left.

Rudimentary interpolation could imply to copy just the nearest observation, in this case $w_i = 1$, with $p_i$ being the observation position closest to $p_0$, and all other $w_{j \neq i} = 0$. This is referred to as nearest neighbor interpolation.

Another simple alternative is to take the mean of all available observations in the domain. Then we have $w_1 = w_2 = \cdots = w_m = \frac{1}{m}$.

Figure 11.2 shows two interpolation configurations in a two-dimensional spatial domain. Observation positions are given in blue, and the desired interpolation position is indicated in red. In the configuration on the left (gridded data-set), many observations are available in all

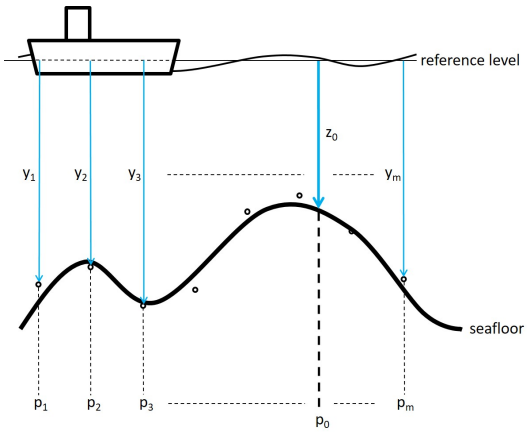



Figure 11.1: The water depth $z_0$ at position $p_0$ has *not* been measured. The water depth $z_0$ is interpolated, based on measurements of water depth $y_1, \ldots, y_m$ at positions $p_1$ through $p_m$. Note that the measurements — the small open dots — are *not* exactly on the sea floor, as they are subject to (small) measurement errors. Generally, the resulting estimate or prediction $\hat{z}_0$ (interpolated value) will not lie exactly on the sea floor either, though hopefully be close to it.

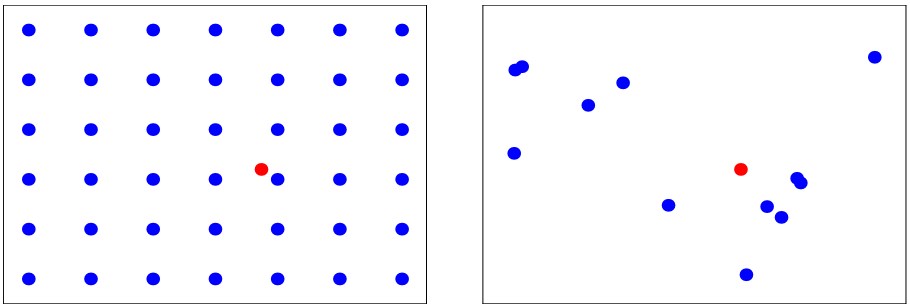


Figure 11.2: Observations of the attribute in a regular grid, in a two-dimensional domain, in blue (left), and at arbitrary positions (right). The intended interpolation position $p_0$ is indicated by the red dot.

directions in a regular way around the desired interpolation position, while in the configuration on the right, observations are sparse and irregular — in some directions observations are available, whereas not in other directions, while in addition, some observations have their positions almost coinciding. In the first case, a simple but fast method will probably work well, while in the latter case, one should be more careful on how the method handles clusters of data and directional variability.

If the desired position (red dot) is within the convex hull of surrounding observation positions, then one speaks of interpolation, otherwise of extrapolation. Two classes of interpolation methods are distinguished here. In deterministic methods, the uncertainty of, and correlation between observations is ignored, while stochastic methods do take this uncertainty and correlation into account. The next two sections present the basics of deterministic interpolation and stochastic interpolation.

## 11.2. Deterministic interpolation

Deterministic interpolation methods are distinguished by the way in which they distribute weights $w_i$ with $i = 1, \ldots, m$ over the available observations (with $w_i \geq 0$).

As mentioned before, two very basic ways of interpolation are nearest neighbor interpolation, in which all weight is given to the closest observation, and averaging all available

observations by distributing the weights equally over all observations. Most common, slightly more advanced, methods are inverse distance interpolation and triangular interpolation.

### 11.2.1. Inverse distance interpolation

It is intuitively appealing that observations close by (in the spatial domain) show large similarity with the attribute at the target interpolation position. The observations are weighted by their spatial distance to the interpolation position. Let $d_i$ be the distance between points $p_i$ and $p_0$, with $i = 1, \ldots, m$. The *inverse distance interpolation* of (for instance) the height at position $p_0$ is given by

$$\hat{z}_0 = \frac{1}{\sum_{i=1}^{m} \frac{1}{d_i}} \left( \frac{1}{d_1} y_1 + \cdots + \frac{1}{d_m} y_m \right) \tag{11.3}$$

based on the height observations $y_1, \ldots, y_m$. Therefore, the weight given to the $j$-th observation equals

$$w_j = \frac{\frac{1}{d_j}}{\sum_{i=1}^{m} \frac{1}{d_i}} \tag{11.4}$$

This approach to interpolation is also referred to as Inverse Distance Weighting (IDW). Note that the sum of all weights equals one.

When one interpolates specifically for one of the given data points, i.e. position $p_0$ coincides with observation point $p_j$: $p_0 = p_j$, one will get the attribute value of that data point. In this limiting case $d_j \downarrow 0$ and $w_j \uparrow 1$, and indeed $\hat{z}_0 = y_j$, and the observed height $y_j$ at position $p_j$ is returned as interpolated depth or height.

More or less weight to close by observations can be given by incorporating a power $p$ in Eq. (11.3), respectively by power $p > 1$ (more) and $p < 1$ (less). Inverse distance interpolation of power $p$ reads:

$$\hat{z}_0 = \frac{1}{\sum_{i=1}^{m} \frac{1}{d_i^p}} \left( \frac{1}{d_1^p} y_1 + \cdots + \frac{1}{d_m^p} y_m \right) \tag{11.5}$$

Inverse distance interpolation works well for dense and regularly space data.

Often only observations within some threshold distance $R$ to the interpolation location are incorporated ($d_i \leq R$); the neighborhood of position $p_0$ is restricted to those positions within a distance of $R$.

### 11.2.2. Triangular interpolation

Rather than using a single nearby observation for interpolation, as done with nearest neighbor interpolation, one may want to use the three observations, which are, in a two-dimensional spatial domain, directly surrounding the target interpolation position $p_0$. Triangular interpolation is often used for *dense* data, like airborne laser scanning data (for creating 3D terrain models for instance, cf. Chapter 22). This method consists of two steps. Assume $m$ height observations $y_1, \ldots, y_m$ are available at positions $p_1, \ldots, p_m$, with $p_1 = ((x_1)_1, (x_2)_1), \ldots, p_m = ((x_1)_m, (x_2)_m)$, and that we want to obtain a height estimate at a 2D position $p_0$.

The first step consists of determining the *Delaunay triangulation* of the observation positions $p_1, \ldots, p_m$, see Figure 11.3 on the left.

In the second step one finds the (smallest) triangle consisting of three nodes (positions $p_i$, $p_j$, and $p_k$) which contains $p_0$, see Figure 11.3 on the right. Only the three observations

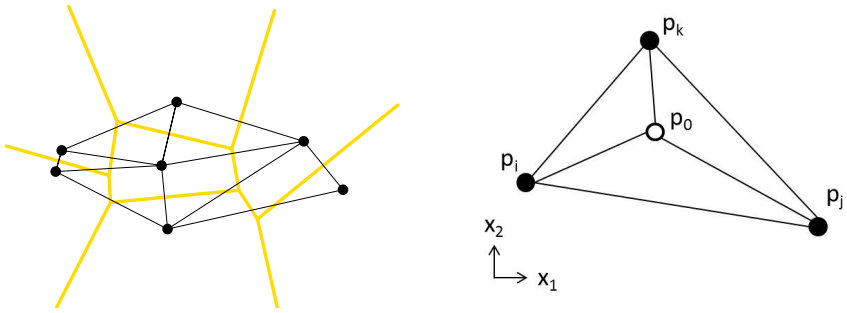



Figure 11.3: Delaunay triangulation in black of a set of observation points/positions (black dots) in two-dimensional $\mathbb{R}^2$, also referred to as Triangulated Irregular Network (TIN), and Voronoi diagram in yellow, on the left. In triangular interpolation the weights $w_i, w_j$ and $w_k$ of the attribute at positions $p_i$, $p_j$ and $p_k$ are proportional to their relative distance to the center point $p_0$.

$y_i, y_j, y_k$ at positions $p_i$, $p_j$, and $p_k$ are used. All other observations get a weight equal to zero for interpolation at $p_0$, hence $w_l = 0$ for $l \neq i, j, k$. Positive weights for $y_i$, $y_j$ and $y_k$ are obtained from the triangular weight equation

$$\hat{z}_0 = \frac{A_{0jk}}{A_{ijk}} y_i + \frac{A_{i0k}}{A_{ijk}} y_j + \frac{A_{ij0}}{A_{ijk}} y_k = w_i y_i + w_j y_j + w_k y_k \tag{11.6}$$

where $A_{ijk}$ denotes the area of the triangle with vertices $i$, $j$, $k$, and noting that $A_{0jk} + A_{i0k} + A_{ij0} = A_{ijk}$. So, the closer $p_0$ is to vertex $j$, the larger the triangle on the other side ($A_{0ik}$) and the more weight $y_j$ gets. Ultimately, triangular interpolation yields $\hat{z}_0 = y_j$, with $w_j = 1$ (and $w_i = w_k = 0$) when interpolation position $p_0$ coincides with one of the observation points, i.e. $p_0 = p_j$.

An example of a Delaunay triangulation is given in black in Figure 11.3 (on the left), while in yellow its dual structure, the Voronoi diagram, is given. To create the Voronoi diagram, one starts from the perpendicular bisectors between any two positions (black dots), and uses these to construct the smallest possible convex polygon around each position (practically spoken, the smallest possible area (cell) around each position). This yields the Voronoi cells in yellow, Figure 11.3 (on the left). Triangular interpolation is relatively fast, as efficient algorithms exist for creating a Delaunay triangulation, cf. [29].

Once the Delaunay triangulation has been done for the full set of observation positions, typically irregularly distributed, triangular interpolation is often used to create a regular interpolation grid; then the attribute is determined (interpolated) at each grid point.

## 11.3. Stochastic interpolation [*]

Though deterministic interpolation is intuitively appealing at first sight, there are a few problems associated.

In stochastic interpolation we consider the *spatial variation* of the attribute in a statistical way. The attribute to-be-interpolated, which is subject to variations, is modeled as a *random function* $\underline{z}(p)$, which depends on position $p$. As an example we can assume that the seafloor is flat (deterministic trend), but quite naturally, small variations may occur from place to place, see Figure 11.4. These random variations are referred to as the *signal* in the quantity of interest, in this case the seafloor depth; the seafloor naturally goes up and down, though smoothly. Considering the observation positions, the random function $\underline{z}(p)$ actually consists of a set of *random variables* $\underline{z}(p_1), \ldots, \underline{z}(p_m)$. We note that examples of *spatial* interpolation are shown here, but the approach presented in this section equally applies to *temporal* inter-

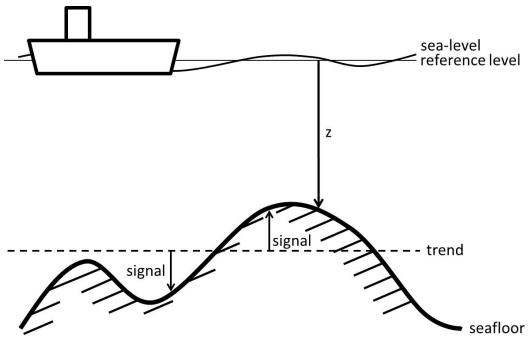



Figure 11.4: The to-be-interpolated attribute $z$ (e.g. the seafloor depth) is modeled (simply) as a trend, in this case the flat dashed horizontal line (average seafloor-level), supplemented by a small variation signal, the solid line. The signal behaves smoothly in the spatial domain, there are no sudden changes or jumps. The trend and signal together represent the actual seafloor depth $z$. By including a *random* signal in the model, the interpolation can accomodate small local variations of the attribute (e.g. the depth) with respect to the trend. Water depth is measured with respect to the thin line on top, which represents the reference or zero height/depth level (for instance average sea level).

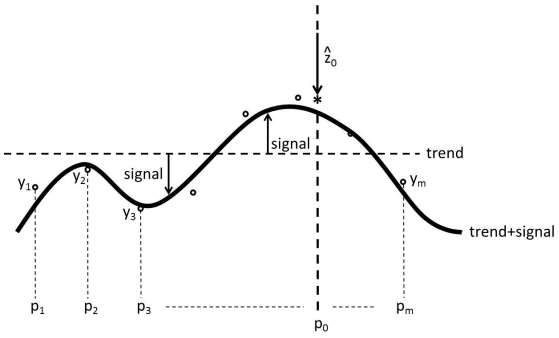



Figure 11.5: The result of interpolation: water depth $\hat{z}_0$. The small open dots indicate water depth observations $y_1, \ldots, y_m$. Because of measurement errors, they do not perfectly coincide with the actual seafloor. The actual seafloor is not known to us - we just got the observations at positions $p_1, \ldots, p_m$. Interpolation is about determining the water depth at another location $p_0$, and do this in a best way. The interpolated water depth $\hat{z}_0$, indicated by the asterisk, is hopefully close to the actual seafloor.

polation (with the attribute, for instance temperature, as a function of time). Mind that in this chapter position $p$ is deterministic; the position coordinates in $p$ are independent variables, just like time $t$ in the example of Section 10.2.

Secondly, as we know already from Chapter 8, observables $\underline{y}_1, \ldots, \underline{y}_m$ are *random* variables, and the *measurement* uncertainty, or noise, should be taken into account in the interpolation, like we did with parameter estimation. For *optimal* interpolation results, the measurement uncertainty should be reflected in the weights in (11.1).

Finally one should also propagate the quality of the observables into measures of quality of the interpolation result, taking into account also the natural variability of the signal, so that one is able, to not only present the result, but also in a realistic way to evaluate its quality (uncertainty). What counts in the end is, how far off the interpolated value is from the actual seafloor depth. The result of interpolation, estimate $\hat{z}_0$ at position $p_0$ is shown in Figure 11.5.

### 11.3.1. Kriging

A class of methods that takes care of the above issues is called Kriging. Kriging is a common technique in geology and environmental sciences, and fits in the theory of linear *prediction* developed in (geo-)statistics and geodesy.

Kriging takes as input, the observation values, and also takes into account stochastic properties of the attribute and the observable. In this section we restrict ourselves to working with the first two central moments of a random variables' distribution: the mean and the variance.

The spatial variation of the attribute with respect to the trend is captured by the signal, denoted by $\underline{s}$, and it depends on the position $p$, i.e. $\underline{s}(p)$. The signal is assumed here to have zero mean, i.e. on average, the variations of the seafloor with respect to the assumed flat plane equal zero. The variation is described by a *covariance function*, cf. Figure 11.7 on the right; this implies that signal values at two nearby positions are largely correlated (dependent), whereas signal values of two positions far away from each other will have little, or no 'dependence'. In this way physical smoothness of the variation is translated into statistical correlation. The covariance is assumed to depend on the separation vector between two positions, and in practice often only on its norm, hence the Euclidean *distance* between two positions, and not on the particular positions themselves.

Secondly we take into account the uncertainty in the observables $\underline{y}$. As one can see in Figure 11.5, random *deviations* in the observations *from the trend* consist of firstly, the signal variation (the solid line hoovers around the dashed line), and secondly, the measurement uncertainty (the open dots are not exactly on the solid line). With respect to model (8.3) in Chapter 8, the random signal $\underline{s}$ is added, and we have

$$\underline{y} = Ax + \underline{s} + \underline{e} \tag{11.7}$$

with vector $y$ holding the observations $y_1, \dots, y_m$ at positions $p_1, \dots, p_m$, with $Ax$ the so-called trend, vector $s$ the signal at positions $p_1, \dots, p_m$, i.e. $s = (s(p_1), \dots, s(p_m))^T$, and vector $e$ the measurement errors $e = (e_1, \dots, e_m)^T$. The full $m$x$m$ variance matrix $Q_{yy}$ of all observables, cf. (7.5), and Section 8.3.4, takes both effects into account (error in the measurement system and signal in the observed attribute). It starts from the variance matrix $Q_{yy}$ in Chapter 8, representing the measurement noise, and adding now the $m$x$m$ variance matrix $Q_{ss}$ of vector $\underline{s}$, constructed from the covariance function, with elements $\sigma_{s(p_i)s(p_j)}$ for $i = 1, \dots, m$ and $j = 1, \dots, m$. Positions $p_i$ and $p_j$ are separated by a certain distance $d_{ij}$, and given this distance, matrix entry $\sigma_{s(p_i)s(p_j)}$ is simply read from Figure 11.7, on the right. In short-hand notation we have $Q_{yy} := Q_{yy} + Q_{ss}$ (assuming that measurement noise and signal are not correlated). In many applications of Ordinary Kriging, the measurement noise — when compared to the signal variation — can often be, or is often neglected. In that case we simply have $Q_{yy} := Q_{ss}$.

In order to derive the *best* possible interpolator, the interpolation error is defined as $\underline{\hat{\epsilon}} = \underline{z}_0 - \underline{\hat{z}}_0$, the difference between the actual water-depth $z_0$ at position $p_0$ and the interpolated value $\hat{z}_0$. We start from the Mean Squared Error (MSE), cf. Section 6.6 and impose conditions on the solution, like that the interpolation is a *linear* combination of the observations, cf. Eq. (11.2), and that the interpolation is *unbiased*, thus $E(\underline{\hat{z}}) = E(\underline{z})$, and this leads to an expression for the error variance of the interpolator. Minimizing the error variance (implying minimum MSE) results in a solution that in practice is obtained by solving a system of linear equations. Kriging not only provides a numerical interpolation result that is optimal in the sense of minimum error variance $\sigma^2_{\hat{\epsilon}}$ (least uncertainty), but also provides this variance as a measure of the uncertainty of the interpolation result.

Finally we note that in this chapter we always work with a given covariance function. In practice, the covariance function will not be known, and has to be estimated. In particular

selecting an appropriate type or shape of covariance function is important. These subjects are beyond the scope of this book.

### 11.3.2. Ordinary Kriging

The most common way of Kriging is so-called Ordinary Kriging. Ordinary Kriging assumes, first, that the attribute under consideration has a constant mean over the entire spatial domain (i.e. the seafloor globally is assumed to be flat and level, as in Figure 11.4), but the mean water-depth is *unknown*. The attribute $\underline{z}$ at position $p_i$ equals the unknown trend $x$, plus the signal $\underline{s}$ at that position, cf. Figure 11.4

$$\underline{z}(p_i) = x + \underline{s}(p_i) \tag{11.8}$$

Stating that the attribute has a constant unknown mean over the entire spatial domain implies that matrix $A$ in (11.7) reduces to $A = l$, with $l$ a vector of all ones, i.e. $l = (1, \dots, 1)^T$, of length $m$.

Secondly, it assumes that the covariance is the same over the entire spatial domain under consideration (typically a decaying function of distance, as the example in Figure 11.7 on the right).

The above assumptions, in combination with requirements on unbiasedness and linearity as above, see Appendix B.7, lead to the following Ordinary Kriging system:

$$\begin{pmatrix} Q_{yy} & l \\ l^T & 0 \end{pmatrix} \begin{pmatrix} w \\ \nu \end{pmatrix} = \begin{pmatrix} Q_{yz_0} \\ 1 \end{pmatrix} \tag{11.9}$$

The last row of this matrix-vector system, $l^T w = 1$, ensures that the sum of all weights equals one (and represents the unbiasedness condition). Matrix $Q_{yy}$ is the variance matrix of $\underline{y}$ in (11.7). Vector $Q_{yz_0}$ contains the covariances $\sigma_{s(p_i)s(p_0)}$ between the signal at observation position $p_i$ and the signal at interpolation position $p_0$, and this for $i = 1, \dots, m$; the covariances depend, according to the covariance function, on the distances $d_{i0}$.

Eq. (11.9) is a square system with $m+1$ unknowns, namely $w = (w_1, \dots, w_m)^T$ and Lagrange multiplier $\lambda$, with $\nu = \frac{\lambda}{2}$, cf. Appendix B.7, and can be solved by inverting the $(m+1)$x$(m+1)$ matrix, resulting in

$$\begin{pmatrix} w \\ \nu \end{pmatrix} = \begin{pmatrix} Q_{yy} & l \\ l^T & 0 \end{pmatrix}^{-1} \begin{pmatrix} Q_{yz_0} \\ 1 \end{pmatrix} \tag{11.10}$$

Finally, once vector $w$ has been obtained, the interpolated value is computed with (11.2). Ordinary Kriging is based on Best Linear Unbiased Prediction (BLUP), similar to BLUE in Section 8.3.4. With the weights in $w$ resulting from the above equation, it can be shown, cf. Appendix B.7, that the interpolator (11.2) can be rewritten into

$$\hat{\underline{z}}_0 = \hat{\underline{x}} + Q_{z_0 y} Q_{yy}^{-1} (\underline{y} - l\hat{\underline{x}}) \tag{11.11}$$

with estimator $\hat{\underline{x}}$ as

$$\hat{\underline{x}} = (l^T Q_{yy}^{-1} l)^{-1} l^T Q_{yy}^{-1} \underline{y} \tag{11.12}$$

similar to (8.6) with $A = l$, and $Q_{yy}$ the variance matrix of $\underline{y}$ in (11.7). The estimator $\hat{\underline{x}}$ is also a linear function of $\underline{y}$. Matrix $Q_{z_0 y}$ is a row-vector, containing the covariances $\sigma_{s(p_0)s(p_i)}$.

In Figure 11.6 we show an example of Ordinary Kriging. On the left $m = 5$ observations of terrain height are shown in a two-dimensional spatial domain. The graph on the right shows

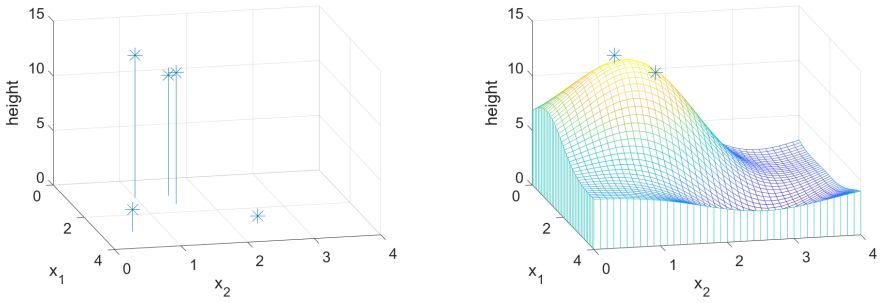



Figure 11.6: Example of Ordinary Kriging. On the left, five observations of attribute height, indicated by asterisks, in a two-dimensional spatial domain $(x_1, x_2)$. The observed heights are 0, 2, 13, 11, 12, at positions $(2\frac{1}{2}, 2\frac{1}{2})$, $(3, \frac{1}{2})$, $(1, 1)$, $(1, 1\frac{1}{2})$, $(1\frac{1}{2}, 1\frac{1}{2})$. On the right the interpolated terrain model - two of the observations are above the interpolated surface, and three are below (and thereby not visible). The used covariance function was $\sigma_{s(p_i)s(p_j)} = ae^{-bd_{ij}^2}$, with $a = 1$ and $b = \frac{1}{2}$, and $d_{ij}$ the Euclidean distance between positions $p_i$ and $p_j$. The measurement noise $\underline{e}$ was set to have a variance of $\sigma_e^2 = \frac{1}{10}$.

the interpolation result, computed for a 0.1×0.1 grid of positions. Using (11.12), the trend $x$ (mean terrain height) is estimated as $\hat{x} = 4.7$. This is quite remarkable, as the mean of the five observations is 7.6. This is caused by the fact that variance matrix $Q_{yy}$ of $\underline{y}$ in (11.7) now contains measurement noise *and* signal noise. The last three observations are closely together (clustered in the spatial domain) and hence, according to a covariance function, like in Figure 11.7 on the right, highly correlated (dependent). In computing the mean height for the full spatial domain, they therefore 'count for' one observation (their mean equals 12, and counts effectively as one observation, rather than three).

With universal Kriging one can use a general polynomial trend model (for instance a sloped line or plane), rather than just a constant, as with Ordinary Kriging.

In the next sub-section we cover a variant of Kriging, with an additional restriction.

### 11.3.3. Simple Kriging

Simple Kriging, as the name suggests, implies a further simplification of Ordinary Kriging. Instead of an unknown trend, it is assumed that the mean value of the attribute is *known*. Parameter $x$ in (11.8) is known.

Simple Kriging will be covered by means of an example. The temperature in a room is kept constant, at a pre-set, known (scalar) value $x$. If this were all available information, predicting the temperature would simply be $\hat{z}_0 = x$ (for any time $t_0$); it equals the known mean, obviously. In practice however, small variations of temperature occur over time. The random signal $\underline{s}$ now represents the deviations from the nominal temperature $x$. This signal has a constant, zero mean, as we had with (11.8) as well.

The temperature in the room is *observed* at regular time instants $t_1, \ldots, t_m$, and the observations are denoted as $y_1, \ldots, y_m$. Compared to the assumed variation in the temperature, the measurement uncertainty is negligible in this example (though measurement noise can also be accomodated with Simple Kriging, similar as with Ordinary Kriging).

Now we would like to interpolate the temperature to some time instant $t_0$ in between the observation time instants, based on all available information, i.e. the known mean $x$, and the observations in vector $y = (y_1, \ldots, y_m)^T$. The result reads

$$\underline{\hat{z}}_0 = x + w^T(\underline{y} - lx) \tag{11.13}$$

where $l$ is again a vector with all ones, $l = (1, \ldots, 1)^T$, and $w = Q_{yy}^{-1}Q_{yz_0}$, similar to (11.10),

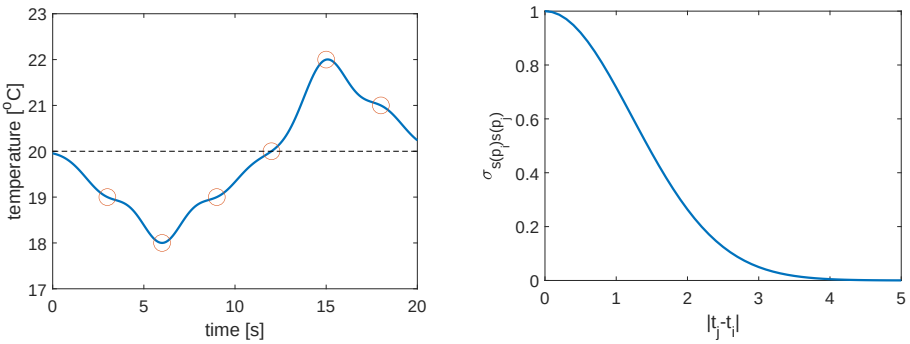



Figure 11.7: Example of Simple Kriging. On the left, six observations (red circles) of attribute temperature (at time instants 3, 6, 9, 12, 15 and 18), and the interpolation result (blue line). On the right the covariance function of the signal $\sigma_{s(t_i)s(t_j)} = ae^{-b|t_j - t_i|^2}$ with $a = 1$ and $b = \frac{1}{3}$; the amount of covariance between the signal at two positions or instants depends on their distance, shown along the horizontal axis, i.e. the covariance depends on how far the positions or instants are apart.

though omitting the last row (this result is given here without proof/derivation).

This equation can — as a side note — be shown to deliver the interpolation value, as a function of both the known mean and the observations, through rewriting it as

$$\underline{\hat{z}}_0 = (1 - w^T l)x + w^T \underline{y} = (1 - \sum_{i=1}^{m} w_i)x + \sum_{i=1}^{m} w_i y_i$$

Substituting the expression for weight-vector $w$ in (11.13) yields

$$\underline{\hat{z}}_0 = x + Q_{z_0 y} Q_{yy}^{-1} (\underline{y} - lx) \tag{11.14}$$

This result is based on Best Linear Prediction (BLP). The result actually looks very similar to (11.11), but in (11.14) the known mean $x$ is used, whereas estimator $\underline{\hat{x}}$ is used in (11.11).

Figure 11.7 shows an example of Simple Kriging, specifically the temperature as a function of time $t$. The known mean equals $x = 20$ and is shown by the dashed line. The actual observations are shown by red circles, $y = (19, 18, 19, 20, 22, 21)^T$. The blue line presents the interpolation result. One can see that in between observations, the interpolation has the tendency of going back to the known mean (dashed line) — the interpolation is a combination of the known mean and the observations cf. (11.13).

The graph on the right shows the covariance function of random signal $\underline{s}$. The amount of covariance between the signal at two positions $p_i$ and $p_j$ depends on the mutual distance $d_{ij}$ between them, or, in this example, the difference in time between instants $t_i$ and $t_j$: $d_{ij} = |t_j - t_i|$. Typically, an exponentially decaying function is used, as $\sigma_{s(p_i)s(p_j)} = ae^{-bd_{ij}^2}$. Parameter $b$ governs the width of the covariance function; a larger value for $b$ yields a function which is more narrow, and a smaller value yields a wider covariance function. The width of the covariance function measures the *smoothness* of the signal; a wide function corresponds to a smooth signal, a narrow function allows for rapid variations. Parameter $a$ represents the signal variance, i.e. $\sigma_{s(p_i)s(p_i)} = a$, and equals $\sigma_z^2$; it is a measure of the *magnitude* of the signal variations with respect to the trend.

The interpolation result in Figure 11.7 on the left, is computed using Eq. (11.14). The

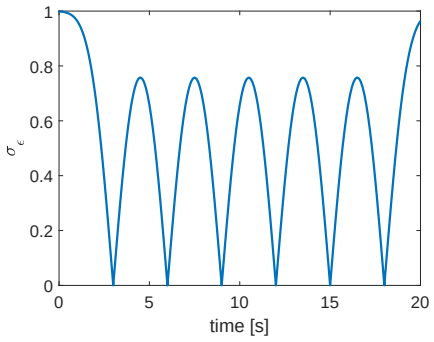



Figure 11.8: Example of Simple Kriging: interpolation error standard deviation $\sigma_{\hat{\epsilon}}$ as a function of the interpolation time instant.

6x6-variance matrix $Q_{yy}$ reads

$$Q_{yy} = \begin{pmatrix} 1 & 0.05 & & & & \\ 0.05 & 1 & 0.05 & & & \\ & 0.05 & 1 & 0.05 & & \\ & & 0.05 & 1 & 0.05 & \\ & & & 0.05 & 1 & 0.05 \\ & & & & 0.05 & 1 \end{pmatrix}$$

For the interpolation $\hat{z}_0$ at time instant $t_0 = 4$, matrix $Q_{z_0 y}$ reads

$$Q_{z_0 y} = \begin{pmatrix} 0.72 & 0.26 & 0 & 0 & 0 & 0 \end{pmatrix}$$

which has only two non-zero values. In line with the covariance function in Figure 11.7 on the right and Eq. (11.14), interpolation at $t_0 = 4$ depends on the known mean, and on the two neighboring observations, at $t_i = 3$ we have $|t_i - t_0| = 1$, and at $t_i = 6$ we have $|t_i - t_0| = 2$. The interpolated value at $t_0 = 4$ becomes $\hat{z}_0 = 18.85$, well in between the two neighboring observations of 19 at $t = 3$, and 18 at $t = 6$.

Figure 11.8 shows the interpolation error standard deviation $\sigma_{\hat{\epsilon}}$ from (B.5), as a function of the interpolation position, in this case, interpolation time instant. At $t_0 = 4$ the standard deviation reads $\sigma_{\hat{\epsilon}(t_0=4)} = 0.66$. Note that the interpolated signal passes exactly through the observations (red circles), cf. Figure 11.7. At the observation positions (time instants), the error standard deviation is zero cf. Figure 11.8, as at these positions there is no uncertainty on the interpolated height/depth (as actually no interpolation is needed at these positions). When there is measurement noise present, the interpolated signal will *not* pass exactly through the observations. In between observation instants, the interpolation error standard deviation will increase. The interpolation error standard deviation as shown in Figure 11.8 initially increases with increasing distance to the observation positions, but levels off at a maximum value at the so-called range distance $\sigma_{\hat{\epsilon}} = \sqrt{a}$ (the level when there would be no observation), the distance beyond which there is no correlation anymore between observables, according to the covariance function used. For more information see e.g. [30].

### 11.3.4. Parametric trend interpolation

In the absence of random signal $\underline{s}$ in (11.7), we have $\underline{y} = Ax + \underline{e}$ and we are basically back at the parameter estimation model of Chapter 8. We suppose that there is no spatial variation of the attribute with respect to the trend. In that sense a one-dimensional temporal interpolation was already presented in Section 10.2. The trend was modeled by a straight line, described by

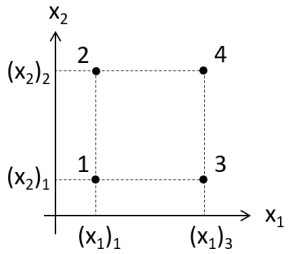



Figure 11.9: Regular measurement lay-out for bi-linear spatial interpolation, with measurements of height at the four corners ($p_1$, $p_2$, $p_3$, $p_4$).

two parameters, namely $x(t_0)$ and $\dot{x}$ ($n$=2); a polynomial of degree 1. With estimates $\hat{x}(t_0)$ and $\hat{\dot{x}}$ one can estimate the position at any time $t_j$ as

$$\hat{x}(t_j) = \begin{pmatrix} 1 & (t_j - t_0) \end{pmatrix} \begin{pmatrix} \hat{x}(t_0) \\ \hat{\dot{x}} \end{pmatrix}$$

At the end of that section even a quadratic model was put forward with $x(t_0)$, $\dot{x}$ and $\ddot{x}$ as unknown parameters ($n$=3). In practice typically a polynomial of low degree, $(n-1)$, is used, with $n \ll m$.

As an example we briefly consider *bi-linear* spatial interpolation. The attribute height $z$ depends on the position coordinates $p = (x_1, x_2)$ through four coefficients $c_{00}$, $c_{10}$, $c_{01}$ and $c_{11}$ as follows

$$z = c_{00} + x_1 c_{10} + x_2 c_{01} + x_1 x_2 c_{11}$$

So, for the interpolated height $z_0$ at $p_0$ we have

$$z_0 = c_{00} + (x_1)_0 c_{10} + (x_2)_0 c_{01} + (x_1)_0 (x_2)_0 c_{11} \tag{11.15}$$

and the four coefficients follow from the heights $y_1$, $y_2$, $y_3$, $y_4$ observed at four positions $p_1$, $p_2$, $p_3$, $p_4$. Following the approach of Chapter 8, model $\underline{y} = Ax + \underline{e}$, or $E(\underline{y}) = Ax$ with $E(\underline{e}) = 0$ becomes

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \underline{y}_3 \\ \underline{y}_4 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & (x_1)_1 & (x_2)_1 & (x_1)_1(x_2)_1 \\ 1 & (x_1)_1 & (x_2)_2 & (x_1)_1(x_2)_2 \\ 1 & (x_1)_3 & (x_2)_1 & (x_1)_3(x_2)_1 \\ 1 & (x_1)_3 & (x_2)_2 & (x_1)_3(x_2)_2 \end{pmatrix}}_{A} \begin{pmatrix} c_{00} \\ c_{10} \\ c_{01} \\ c_{11} \end{pmatrix}$$

where we conveniently took a *regular* measurement lay-out as shown in Figure 11.9, such that e.g. $(x_1)_2 = (x_1)_1$. In this case, the model has $m = n = 4$ observations and unknown parameters, and hence the design matrix $A$ can be just inverted to result into the estimates for the unknown coefficients

$$\begin{pmatrix} \hat{c}_{00} \\ \hat{c}_{10} \\ \hat{c}_{01} \\ \hat{c}_{11} \end{pmatrix} = \frac{1}{((x_1)_3 - (x_1)_1)((x_2)_2 - (x_2)_1)} \begin{pmatrix} (x_1)_3(x_2)_2 & -(x_1)_3(x_2)_1 & -(x_1)_1(x_2)_2 & (x_1)_1(x_2)_1 \\ -(x_2)_2 & (x_2)_1 & (x_2)_2 & -(x_2)_1 \\ -(x_1)_3 & (x_1)_3 & (x_1)_1 & -(x_1)_1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

and with these coefficients interpolated height $z_0$ can be obtained through (11.15). Note that the above model is *not* about fitting plane to the four observed heights $y_1$, $y_2$, $y_3$, $y_4$; for doing that, without redundancy, one would need only three heights.

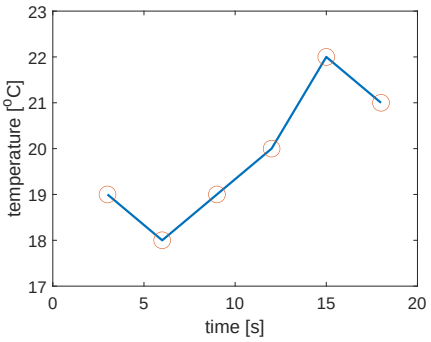



Figure 11.10: Piecewise linear interpolation of observations in the example of Figure 11.7, with Question 1.

## 11.4. Exercises and worked examples

This section presents a derivation-exercise and a simple problem with a worked answer on interpolation.

**Question 1** Piecewise linear interpolation of an attribute, in a one-dimensional spatial, or temporal domain implies just connecting two succesive observations by a straight line, see Figure 11.10. Start from Eq. (11.3) for inverse distance interpolation, and restrict it to $m = 2$, for instance interpolating $z$ for position $p_0$ with $p_0 \in [p_1, p_2]$, using only $y_1$ and $y_2$ in the interpolation. Show that in this case one obtains

$$\hat{z}_0 = \frac{p_2 - p_0}{p_2 - p_1} y_1 + \frac{p_0 - p_1}{p_2 - p_1} y_2 = y_1 + \frac{p_0 - p_1}{p_2 - p_1}(y_2 - y_1)$$

**Answer 1** Restricting $m = 2$ and using solely $y_1$ and $y_2$ in (11.3) yields

$$\hat{z}_0 = \frac{1}{\frac{1}{d_{10}} + \frac{1}{d_{20}}}\left(\frac{1}{d_{10}} y_1 + \frac{1}{d_{20}} y_2\right)$$

with $d_{10}$ the distance between $p_1$ and $p_0$, and $d_{20}$ the distance between $p_2$ and $p_0$. This can be rewritten as

$$\hat{z}_0 = \frac{d_{10} d_{20}}{d_{10} + d_{20}}\left(\frac{1}{d_{10}} y_1 + \frac{1}{d_{20}} y_2\right) = \frac{d_{20}}{d_{10} + d_{20}} y_1 + \frac{d_{10}}{d_{10} + d_{20}} y_2$$

which shows the given equation, when the horizontal coordinate $p$ is introduced (e.g. $d_{20} = p_2 - p_0$).

**Question 2 [*]** Two observations of terrain height are given: $y_1 = 3$ at position $p_1 = 1$, and $y_2 = 4$ at position $p_2 = 3$, see Figure 11.11 on the left. The measurement noise of the observables has a variance of $\sigma^2_{y_1} = \sigma^2_{y_2} = \frac{1}{2}$, and is uncorrelated across the observables. In this spatial domain from $p = 0$ to $p = 5$, the terrain can be assumed flat, though the mean terrain-height is unknown. On the right in Figure 11.11 the covariance function is given, it describes the 'smoothness' of the terrain. It is mathematically given as $\sigma_{s(p_i)s(p_j)} = -\frac{1}{4} d_{ij} + 1$ for $0 \leq d_{ij} \leq 4$, and zero otherwise, with $d_{ij} = |p_j - p_i|$. With all this information, one can perform interpolation using Ordinary Kriging. Compute the interpolated terrain-height for $p = 2$. Also do this for $p = \frac{3}{2}$.

**Answer 2 [*]** The terrain-height exactly halfway the two observation positions can, logically, be expected to be the average of the two observations. Kriging will also provide this

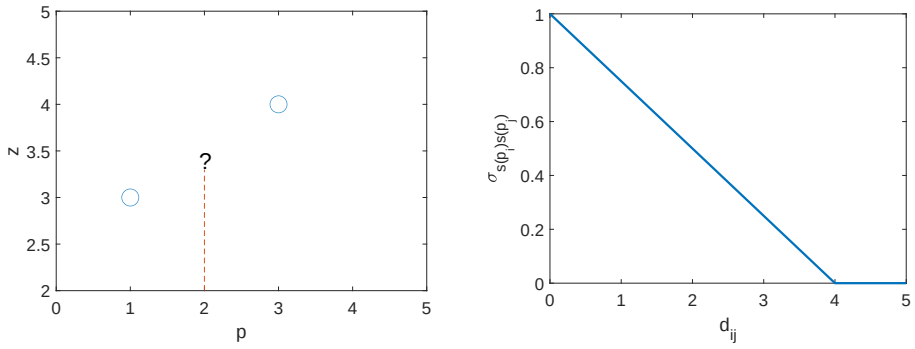



Figure 11.11: On the left, problem statement for Question 2. Two observations of height are given (shown by the circles), $y_1 = 3$ and $y_2 = 4$, at positions $p_1 = 1$ and $p_2 = 3$ respectively, and one is asked to interpolate the height for position $p_0 = 2$. On the right, covariance function $\sigma_{s(p_i)s(p_j)}$ for Question 2.

answer. The variance matrix for the signal at the two observation positions $p_1$ and $p_2$, using the covariance function in Figure 11.11 on the right, reads

$$Q_{ss} = \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$$

Together with the measurement noise $\sigma_{e_1}^2 = \sigma_{e_2}^2 = \frac{1}{2}$, the variance matrix of vector $\underline{y}$, as outlined with (11.7), becomes

$$Q_{yy} = \begin{pmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{3}{2} \end{pmatrix}$$

and its inverse reads

$$Q_{yy}^{-1} = \begin{pmatrix} \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

With (11.12) and $l = (1,1)^T$, the estimated mean terrain height becomes $\hat{x} = \frac{7}{2}$. Then with (11.11), and matrix, or actually vector $Q_{z_0 y}$ as $Q_{z_0 y} = \begin{pmatrix} \frac{3}{4} & \frac{3}{4} \end{pmatrix}$, based on the distance from $p_0$ to $p_1$, and the distance from $p_0$ to $p_2$, and the covariance function of Figure 11.11, we obtain, as expected

$$\hat{z}_0 = \frac{7}{2} + \begin{pmatrix} \frac{3}{4} & \frac{3}{4} \end{pmatrix} \begin{pmatrix} \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} \end{pmatrix} \left[ \begin{pmatrix} 3 \\ 4 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix} \frac{7}{2} \right] = \frac{7}{2}$$

Repeating the exercise for $p_0 = \frac{3}{2}$ yields $\hat{z}_0 = \frac{27}{8}$. The full interpolation result is shown in Figure 11.12. Apparently, interpolation with Ordinary Kriging, does *not* yield a straight line passing through the two observation points. The observations $y_1$ and $y_2$ are subject to measurement error. And also mind that in Ordinary Kriging, one assumes that the attribute under consideration has a constant mean across the entire spatial domain. That is why the result is an in between the straight line and a flat horizontal line, and does not pass through the two observations. As a final exercise, verify yourself that for $p_0 = 1$ you obtain $\hat{z}_0 = 3$, and $\hat{z}_0 = 4$ for $p_0 = 3$, when the measurement noise can be neglected, i.e. when $\sigma_{e_1}^2 = \sigma_{e_2}^2 = 0$. In this case the interpolated result is exactly matching the observations $y_1$ and $y_2$.

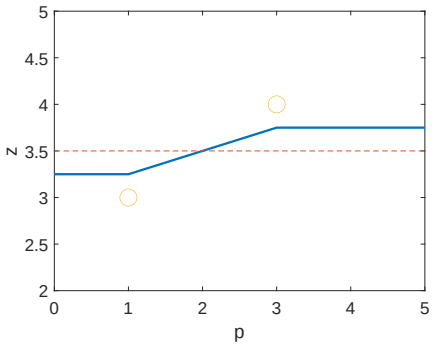



Figure 11.12: Interpolation result using Ordinary Kriging for Question 2, $\hat{x}$ is given by the dashed line, and the interpolation $\hat{z}_0$ by the solid line in blue; the two circles give the observations.

# III

# GPS positioning

# 12

# Introduction

The Global Positioning System (GPS), also known as the NAVigation Satellite Time And Ranging (NAVSTAR) system, is one the most successful satellite systems to date. The first GPS satellite was launched back in February 1978. GPS is a *one-way* radio ranging system which provides real-time knowledge of one's Position and Velocity, and a very accurate Time reference as well (all together referred to as PVT).

GPS provides Positioning, Navigation and Timing (PNT) functionality, which is very valuable not only for the US military, for which it was first developed, but also to a myriad of commercial activities, as well as the general public at large.

The following chapters provide an introduction to GPS positioning. There is much more more of information available on this subject, and the reader is therefore referred to, for instance, the textbooks [31] and [32].

## GPS: system architecture

The GPS system consists of three segments.

1. The space segment, consisting of 24 or more satellites, with accurate atomic clocks on board, continuously transmitting ranging signals to Earth.

2. The control segment, consisting of a number of ground stations, which monitors the satellites, computes their orbits and clock offsets, and uploads this information to the

Figure 12.1: GPS block IIF satellite, built by Boeing. These GPS satellites, 12 in total, have been launched between 2010 and 2016. They have a design lifetime of 12 years. The full GPS constellation nominally consists of 24 satellites. Image courtesy of Boeing [33].

satellites, which in turn encode this information on the ranging signal (the so-called navigation data).

3. The user segment, simply consisting of many GPS receivers, which each track four or more GPS satellites, and compute their own position.

## Market developments

The success of GPS is strongly linked to the ever-decreasing costs of GPS receivers, which primarily consist of electronic hardware. While high-end receivers still cost in the order of $ 1-10k, mass-market receivers, such as those used in smartphones, cost no more than a few dollars each.

Currently there are about as many GPS devices on Earth as people, nearly 7 billion, the vast majority of which are in smartphones. The global Global Navigation Satellite System (GNSS) downstream market revenues, from both devices and services, are currently (2020) around 150 billion Euro, according to the market report by the European GNSS Agency (GSA), now the European Union Agency for the Space Programme (EUSPA) [34].

## Delft perspective on GPS

GPS has found a wide range of applications and has become a standard utility in today's society. An important contribution to the history of GPS, which goes back to the sixties of last century only, was made in Delft. Most, if not all GNSS high-precision positioning equipment in use today on land, in the air, in space and at sea, has a particular 'made in Delft' algorithm inside. This LAMBDA method is the key to high-precision, centimeter-level positioning, and achieving this *very fast*, typically instantaneously, thereby enabling Real-Time Kinematic (RTK) GPS positioning (Section 15.1.1). In the early days of GPS, users were forced to use long observation times (minutes or even hours) in order to get to centimeter accuracy.

A breakthrough came in 1993, when Delft University of Technology professor Peter Teunissen invented the LAMBDA method, by means of which instantaneous carrier phase cycle integer ambiguity resolution became possible.

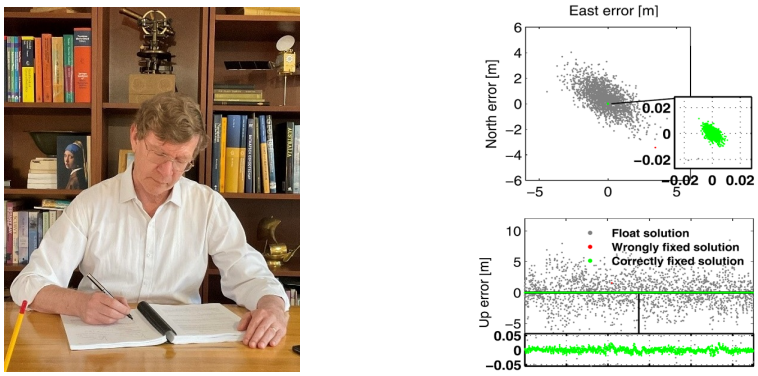



Figure 12.2: Delft University of Technology professor in mathematical geodesy Peter J.G. Teunissen, inventing the Least-squares AMBiguity Decorrelation Adjustment (LAMBDA) method in 1993, at left. At right, East-North scatterplots and Up time-series of instantaneous 3D GNSS position solutions, before (in gray) and after (in green) application of LAMBDA. Note the two-order precision improvement between the solutions in gray and green. Photo at left courtesy of H. Verhoef [35]. Image at right courtesy of R. Odolinski [36].

High-precision positioning with GPS is possible through carrier phase measurements on the signals received from the satellites. The carrier phase provides a millimeter-precise measurement of the satellite-receiver distance, but modulo the signal-carrier wavelength, with

the quotient referred to as the cycle ambiguity, cf. Figure 13.2. Hence, to exploit the high precision of the carrier phase for ranging, one first needs to determine its ambiguity, being the unknown *integer number* of wavelengths that fit into the satellite-receiver distance. As this needs to be done for every satellite, on every signal frequency and for a satellite-receiver geometry that changes only slowly over time (due to the high-altitude satellite orbits), correct resolution of all ambiguities together boils down to a high-dimensional integer estimation problem with a search over a typically elongated hyper-ellipsoidal search space, a situation which was only improved by employing long observation times in the early days. Today the LAMBDA method, by means of a smart linear transformation of the ambiguities, can solve this integer least-squares search problem very efficiently and fast, thereby enabling instantaneous high-precision GPS positioning. More about carrier phase integer ambiguity resolution can be found in Chapter 23 of [31].

## Overview of this part

Chapter 13 presents the basic concepts of the measurement of travel-time of a radio signal from a GPS satellite to a receiver. With these measurements of range as input, Chapter 14 describes the default mode of GPS positioning, referred to as stand-alone or single-point positioning. The next chapter introduces the concept of relative positioning, by means of which high-accuracy, centimeter-level positioning is made possible. Chapter 16 presents, after a brief overview of the four major Global Navigation Satellite Systems (GNSS), an overview of the wide range of applications of GPS/GNSS in today's society.

# 13

# Ranging

## 13.1. Radio signal

The GPS satellites transmit signals in the so-called L-band (i.e. 1 to 2 GHz range) of the electromagnetic radio frequency spectrum. GPS uses Code Division Multiple Access (CDMA) to allow different satellites to send signals at exactly the same center frequency without interfering with each other. The signal consists of a carrier wave on which each satellite modulates its own unique Pseudo Random Noise (PRN) spreading code, see Figure 13.1, and, at a low rate, the satellite orbit and clock information. The signals arrive at the receiver with an unknown *delay* due to travelling all the way from satellite to receiver, and due to the relative velocity of the GPS satellites with respect to a GPS user on or near the Earth's surface, with an unknown *Doppler frequency shift*.

## 13.2. Measurement of range

GPS offers two types of range measurements: pseudorange measurements and carrier phase measurements.

### 13.2.1. Pseudorange measurement

A GPS receiver typically consists of tens to hundreds of so-called channels, and will allocate each of these to a specific GPS (GNSS) satellite. When a GPS receiver first starts up, it will begin to search for a particular GPS satellite on each of its channels, by scanning (trying) for the corresponding spreading codes at different Doppler offsets and time delays. This is done by overlaying the received signal with a local copy or replica of the same code and then (time) shifting it until correlation shows a maximum (best fit, or match). The time shift then directly yields the travel-time measurement.

Once the receiver has locked onto the *spreading code*, it can start regularly taking pseudorange code and Doppler frequency measurements, which are basically the shift in time (delay) and the shift in frequency that are required to maintain the tracking lock (onto the received satellite signal).

Through the pseudorange, the receiver measures the *travel-time* of the radio signal from satellite 's' to receiver 'r':

$$\tau_r^s = t_r - t^s \tag{13.1}$$

where $t^s$ is the time the signal was transmitted by the satellite, and $t_r$ the time the signal was received at the receiver, later noting that these clocks may, to some extent, deviate from

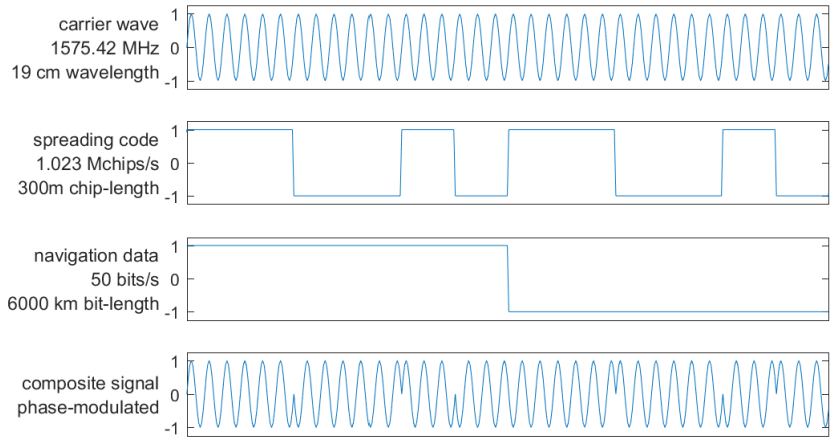



Figure 13.1: The GPS L1 CA-signal is composed of a carrier wave (a sinusoid with a frequency of 1575.42 MHz; not to scale in the above diagram), a spreading code (a sequence of '0' and '1' bits/chips, here represented by values '-1' and '+1', and unique for each satellite), and a low rate navigation data message. Both the spreading code and navigation message are phase-modulated on the carrier wave, through a technique called Binary Phase Shift Keying (BPSK); basically multiplying the carrier by the '-1' and '+1' values of the spreading code and navigation data, and the resulting modulated signal is shown at bottom. For the so-called CA-code on the GPS L1-frequency signal, the spreading codes are all publicly available, and GPS receivers have them built in. CA refers to Coarse Acquisition, but can also be understood as Civilian Access.

the true time. The measured travel-time is converted into the pseudorange, expressed in unit meter, through

$$p_r^s = c\tau_r^s$$

by multiplying by the speed of light $c$ in vacuum ($c \approx 3 \cdot 10^8$ m/s).

The pseudorange represents the travel-time of the signal, and thereby ideally the distance from satellite to receiver. In practice it is affected by the satellite clock offset (known to the receiver through the navigation message), the receiver clock offset, which is unknown, and a number of additional delays, which we cover in the sequel (Figure 14.6), and all multiplied by the speed of light. The clock error is addressed in Section 21.2. In particular the oscillator in the receiver, driving the clock, will not behave perfectly, and hence the receiver clock may run ahead of time, or lag behind. The time shown by the receiver clock is denoted by $t_r(t)$, and it is a function of true time $t$. It equals true time $t$, plus a so-called clock offset $\delta t_r(t)$, hence

$$t_r(t) = t + \delta t_r(t) \tag{13.2}$$

When the receiver measures the travel-time, to eventually produce the pseudorange measurement, it 'reads' the moment of signal arrival at its own clock, and hence this measurement is off by an amount of $\delta t_r(t)$. The travel-time can be conceived as being obtained by 'reading' the receiver clock at signal reception, and 'reading' the satellite clock at signal transmission, hence the *measured* travel-time reads

$$\tau_r^s(t) = t_r(t) - t^s(t - \tau(t))$$

Mind that the (true) travel-time $\tau(t)$ is a function of time, as the receiver may move, and the satellite for sure moves. Substituting here the expression for $t_r(t)$, assuming that the satellite

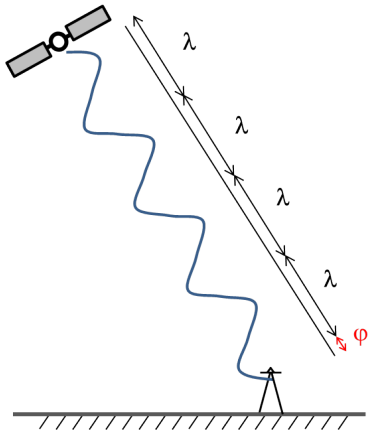



Figure 13.2: Carrier phase measurement: only the *fractional* phase difference can be measured, shown in red in units of length [m] (with $\Phi \in [0, 2\pi\rangle$ when expressed in radians; $\varphi = \lambda\frac{\Phi}{2\pi}$), and the total distance from the satellite to the receiver equals multiple wavelengths $\lambda$ plus the fractional phase difference. The carrier wave is sent continuously, and the receiver cannot distinguish one cycle from another. The unknown integer number of wavelengths, $N$, at the start of signal tracking, is referred to as ambiguity. In this example $N = 4$.

clock is perfectly on time, hence $t^s(t-\tau(t)) = t-\tau(t)$ (the satellites carry atomic clocks), and multiplying by the speed of light now gives:

$$p_r^s(t) = c\tau_r^s(t) = \underbrace{c\tau(t)}_{l_r^s(t)} + \underbrace{c\delta t_r(t)}_{b_r(t)} \tag{13.3}$$

where, in the absence of for instance atmospheric delays, $l_r^s$ denotes the geometric distance between satellite and receiver. This equation shows that the *pseudorange* is a measure for the geometric distance $l_r^s$, apart from the receiver clock offset $b_r$, and hence the term *pseudo*-range.

### 13.2.2. Carrier phase measurement

Additionally, a GPS receiver may measure the *fractional* phase difference between the received carrier wave from the satellite and a locally generated copy (replica). And, it can keep track of the number of cycles of the carrier wave since the start of tracking, together known as the *carrier phase* (CP) measurement. This measurement includes the accumulated number of 'zero-crossings' since lock-on of the signal (for instance, when the fractional phase jumps from $1.99\pi$ to $0.02\pi$, the full period is accounted for and the resulting carrier phase measurement, output by the receiver, is $2.02\pi$).

The carrier wave measurement is a very precise measure of the distance between the satellite and the receiver, but the initial number of carrier wave cycles is unknown, and needs to be estimated before the carrier phase measurements can be effectively used, see Figure 13.2. The much better precision of the carrier phase measurement with respect to the pseudorange code measurement can be understood from Figure 13.1, since the carrier period is much smaller than the code chip duration (for the L1 CA-code signal, 1540 periods of the carrier fit in one chip of the Pseudo Random Noise (PRN) spreading code).

### 13.2.3. Concluding remarks

Linking to the exposition on measuring distances in Chapter 20, the pseudorange measurement corresponds to 'pulse-based'-ranging, and the carrier phase measurement obviously to 'phase-based'-ranging, see Section 20.1 on the principles of ranging, though one should note

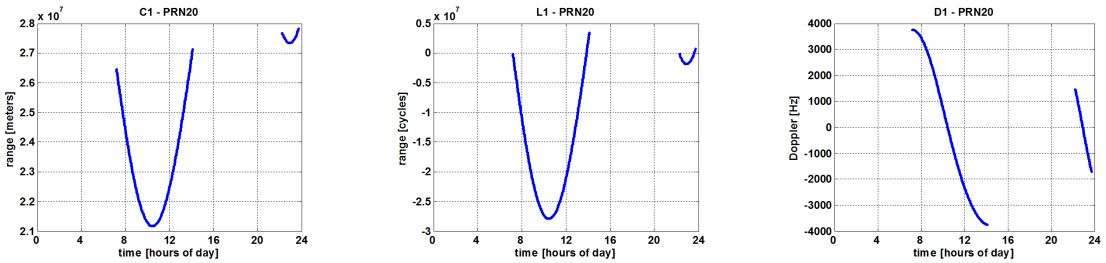



Figure 13.3: Example of time series of (at left) C1 pseudorange measurements, in meter, (in the middle) L1 carrier phase measurements, in cycles, and (at right) D1 Doppler frequency measurements, in Hertz, of a stationary, permanent receiver in Delft, cf. Figure 16.1, tracking GPS satellite PRN20.

that GPS is about *one-way* ranging (rather than two-way ranging, as in Chapter 20).

The receiver can also measure the received *signal-strength*, through the so-called carrier-to-noise-density ratio C/N0, which gives an indication of the quality of the measurement (larger signal-strength yields more precise measurement).

And some receivers output also the measurement of the *Doppler frequency* of the carrier wave, which is a measure for the (relative) velocity of the receiver with respect to the satellite (along the line of sight), see also Section 20.2. The Doppler frequency, multiplied by the wavelength, presents the range-rate $\dot{l}_r^s$, that is, the change in range $l_r^s(t)$ per unit time.

The measurements can be stored, e.g. for the purpose of later analysis and processing, in receiver manufacturer proprietary format or in a generally accepted exchange format, namely RINEX, see Appendix F.

The pseudorange measurement precision is typically at the one or few meter level for low-cost, mass-market equipment, and can get down to the few decimeter level for professional high-end equipment.

The carrier phase measurement precision ranges from the few centimeter to the millimeter level. The carrier phase is an ambiguous measurement of distance, but it is more precise than the pseudorange, typically by two orders of magnitude.

Figure 13.3 shows measurements, collected by a stationary receiver in Delft, on signals received from GPS satellite PRN20, as a function of time. A pass-over of a GPS satellite typically takes several, up to 7 hours. With a nearly circular orbit of the GPS satellite around the Earth, the distance from satellite to receiver is shortest when the satellite is directly overhead. By default actually the negative of the Doppler frequency is output by the GPS receiver (as shown in the graph at right, the measured Doppler frequency is positive (in the interval from about 7-10 hours), while the distance at the same time, as shown in the graph at left, is decreasing).

## 13.3. Multi-frequency ranging

One of the major error sources in GPS is due to the ionosphere, see also Figure 14.6 and Table 14.1. The ionosphere is a ionized part of the Earth's upper atmosphere. There ultraviolet (UV) solar radiation separates electrons from neutral gas atoms and molecules. The free electrons in the ionosphere delay the radio signals, and thus affect the range measurements, with delays in terms of distance ranging from a few meter to hundreds of meters.

The largest delays occur round the geomagnetic equator around local noon, and during solar maxima. The ionospheric delay may be highly variable, as a function of both time and space.

One way of dealing with the ionospheric delay is to track signals from the same satellite on two or more frequencies. The ionosphere delay scales, to a very good approximation,

with the inverse of the square of the radio frequency of the signal, and this relation can be used to create the so-called ionosphere-free range measurements (a linear combination of measurements at two different frequencies, from which the ionospheric delay has been removed). For this reason the GPS satellites were originally designed to transmit ranging signals on both the L1 (1575.42 MHz) and L2 (1227.60 MHz) frequency.

# 14

# Positioning

GPS positioning is based on the concept of multi-lateration (not triangulation). By measuring distances to a number of GPS satellites, as shown in Figure 14.1, and using the known satellite positions, a GPS receiver can compute its own position. To estimate the three position coordinates of the receiver $x_r$, $y_r$, $z_r$, and the receiver clock offset $b_r$, a GPS receiver needs to track at least 4 satellites.

## 14.1. Geometric interpretation

Knowing ones distance to an object (satellite) at a known position, translates into being on a circle (in 2D) or a sphere (in 3D) around this object (with the satellite in the center). As we have seen with (13.3), the GPS pseudorange measurement relates to the geometric range (distance) from satellite to receiver, but, also to an offset caused by the receiver clock! This means, the pseudorange gives us the distance from satellite to receiver, but it may or will be too small, or too large by a certain amount, namely the receiver clock offset $b$. The good news is that the receiver clock offset is the *same* for all pseudoranges measured by the receiver at a specific time. If the receiver clock is ahead of GPS-time, all pseudoranges will be measured too long, and by the same amount. This leads us to the approach of solving for three position coordinates and the receiver clock error at the same time, and hence, requiring pseudorange measurements to at least four satellites (rather than three).

To see the effect of the receiver clock error on the positioning problem at work, we consider a simple two-dimensional positioning example (in which we assume that there is no effect of

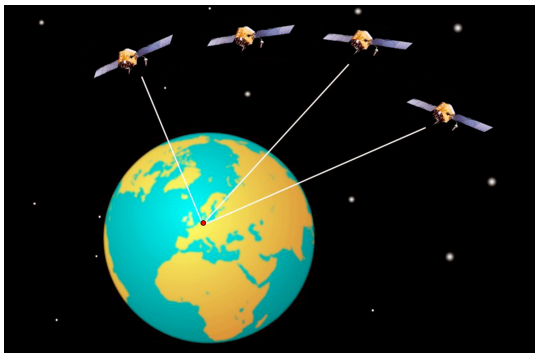


Figure 14.1: GPS positioning — in three dimensions — is based on measuring pseudoranges to at least four satellites, of which the positions are known. Visualization by Axel Smits [37].

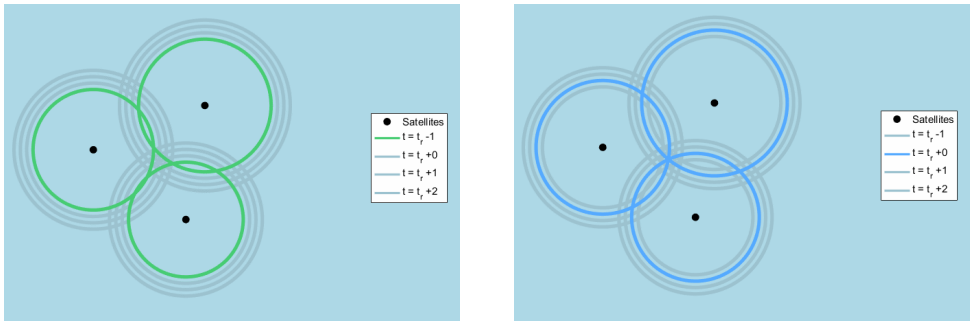



Figure 14.2: Two-dimensional positioning example with three satellites (at known positions, represented by the black dots). The measured pseudoranges are visualized by circles, in green at left, and in blue at right.

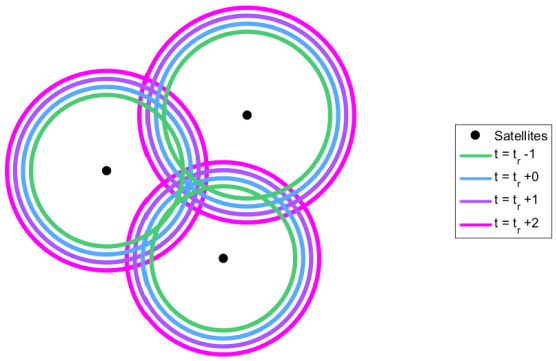



Figure 14.3: The process of determining the receiver clock offset: the measured pseudoranges have to be reduced or enlarged, but all with exactly the same amount, in order to meet at one physical position. The amount to make that happen is the receiver clock offset. The different colors represent different values for the receiver clock offset.

noise present in the pseudorange measurements). So, in two dimensions, we would need to solve for two receiver position coordinates and one receiver clock error, hence in total three unknown parameters, so we need at least three pseudorange measurements.

In Figure 14.2 at left, the measured pseudoranges are shown in green, and apparently these three green circles do not all meet in one point. The pseudoranges are 'too short', the reason obviously being the receiver clock lagging behind. When the radii of the green circles are enlarged, all by exactly the same amount, to yield the blue circles, as shown at right, we arrive at an intersection of all three circles in one point. We have solved for the two position coordinates, and the receiver clock offset as well[1]. This clearly demonstrates that positioning and timing are intimately related!

## 14.2. Pseudorange observation equation

With expanding the one-way geometric range $l_r^s$ between satellite 's' and receiver 'r' as

$$l_r^s = \sqrt{(x^s - x_r)^2 + (y^s - y_r)^2 + (z^s - z_r)^2}$$

[1]When, in this example, the receiver clock would behave perfectly and be exactly aligned with GPS-time, we could solve the two-dimensional positioning problem by measuring just two pseudoranges, which then directly give us two proper distances, though two circles may intersect at two points actually. With GPS this would be no problem, as the satellites are at 20.000 km distance, and the other intersection point will generally be on the other side of the Earth, or even way beyond.

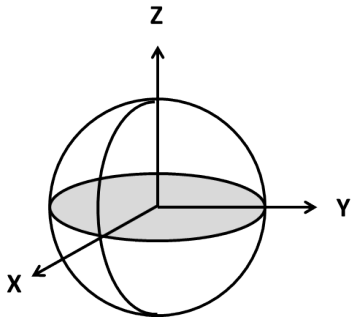



Figure 14.4: Three-dimensional Cartesian Earth-Centered Earth-Fixed (ECEF) coordinate system for GPS positioning.

using a three-dimensional Cartesian coordinate system as shown in Figure 14.4, pseudorange observation Eq. (13.3) turns into

$$\underline{p}_r^s = \sqrt{(x^s - x_r)^2 + (y^s - y_r)^2 + (z^s - z_r)^2} + b_r + \underline{e}_r^s \tag{14.1}$$

where we omitted the argument of time $t$. The satellite position coordinates at time of signal transmission are $x^s$, $y^s$ and $z^s$, and the receiver position coordinates at time of signal reception are $x_r$, $y_r$ and $z_r$. The satellite position, as well as the satellite clock offset, is available to the user through the navigation data message, cf. Figure 13.1. Parameter $b_r$ equals the receiver clock offset $\delta t_r$ multiplied by the speed of light $c$, cf. (13.3). If the receiver clock is ahead of GPS system time, $b_r$ is positive, and the measured pseudoranges are 'too long'. And finally note that we included the (unavoidable) random measurement error $\underline{e}_r^s$ on the right hand side of Eq. (14.1).

## 14.3. Positioning: parameter estimation

In practice, as one typically observes more satellites than the minimum of four, GPS positioning does not actually involve drawing circles or spheres, but employs the principle of least squares estimation. First the observation model is defined, which links the observations to the unknown parameters.

Since the GPS observation model is non-linear, this involves a linearisation with respect to the unknown parameters, around an approximate position, see Section 8.4. The linearized model of observation equations reads

$$\underbrace{\begin{pmatrix} \Delta\underline{p}_r^1 \\ \Delta\underline{p}_r^2 \\ \vdots \\ \Delta\underline{p}_r^m \end{pmatrix}}_{\Delta\underline{y}} = \underbrace{\begin{pmatrix} -u_{r,x}^1 & -u_{r,y}^1 & -u_{r,z}^1 & 1 \\ -u_{r,x}^2 & -u_{r,y}^2 & -u_{r,z}^2 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ -u_{r,x}^m & -u_{r,y}^m & -u_{r,z}^m & 1 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} \Delta x_r \\ \Delta y_r \\ \Delta z_r \\ b_r \end{pmatrix}}_{\Delta x} + \underbrace{\begin{pmatrix} \underline{e}_r^1 \\ \underline{e}_r^2 \\ \vdots \\ \underline{e}_r^m \end{pmatrix}}_{\underline{e}} \tag{14.2}$$

where we assume to have $m$ satellites in view. Section 9.3 presents the linearization of a distance observation equation in two dimensions, and extension into three dimensions is straightforward. The coefficients in the above design-matrix for the coordinate parameters are actually the elements of the unit-direction vector $u_r^s$ from the receiver 'r', pointing to the satellite 's', cf. (9.5). The above model carries $m$ observations and 4 unknown parameters, and hence the redundancy equals $m - 4$.

Next, a least-squares algorithm is used to solve this linearized model, presented in matrix-vector form. When an $m \times m$ variance matrix of the pseudorange observables is involved, a Best Linear Unbiased Estimation solution can be obtained, which minimizes the uncertainty of the solution (see Chapter 8). Then one can also obtain the variance matrix of the parameter estimators, through (8.7), and analyse the precision of the position coordinates.

Most users of GPS are interested in position coordinates $x_r$, $y_r$, and $z_r$. Through knowledge of the receiver clock offset $b_r = c\delta t_r$ one has in fact access to GPS system time, which is an atomic time scale, and thereby also to UTC (Coordinated Universal Time).

Similar to the position coordinates estimation based on pseudorange measurements, the three-dimensional velocity vector of the receiver can be estimated from the measured Doppler shift measurements, cf. Section 13.2.3.

A well-known and widely used format for storing and exchanging GPS (GNSS) Position, Velocity and Time (PVT) solutions is NMEA, see Appendix E.

## 14.4. Reference systems

Relying, by default, on the given satellite positions in the navigation message of the GPS signal, GPS positioning yields Cartesian coordinates $(x, y, z)$ in WGS84, the World Geodetic System 1984, which is an Earth-Fixed, Earth-Centered (ECEF) coordinate system, as presented in Figure 14.4. These Cartesian coordinates can be converted into geographic, or ellipsoidal coordinates latitude $\varphi$, longitude $\lambda$, and ellipsoidal height $h$, see Chapter 29.

In differential mode (introduced in the next chapter), the position coordinates for the user receiver are in the same reference system as the position coordinates of the base, or reference station, generally provided in a local or regional reference system (e.g. ETRS89 in Europe, a realization of the European Terrestrial Reference System).

More information about reference systems and transformations from one to another can be found in Chapters 31 and 34.

## 14.5. GPS accuracy and error sources

The quality of the GPS position solution is largely dependent on the number of available satellites and their geometry with respect to the user. If enough satellites are visible on all sides of the receiver, at high and low elevation angles, a good position accuracy can be expected. The only weakness in the geometry is the fact that there are no satellites visible beneath the receiver, as one cannot track and observe satellites below the local horizon. As a result vertical position accuracy is generally poorer than the horizontal accuracy by about a factor of 1.5.

In many practical situations one or more satellite signals are *blocked* by surrounding buildings or other obstacles which is called shadowing. In this case GPS performance might be significantly degraded. Furthermore, in built-up areas, GPS receivers often experience signal reflections, i.e. signals arrive at the receiver after bouncing off an object. Since the reflected signal path is always longer than the direct path, this causes a corresponding error in the range measurement. It is also possible that both the direct and reflected signals arrive at the receiver, which is referred to as *multipath*. In this case, the receiver must deal with the superposition of these signals, generally resulting in a biased range measurement, see Figure 14.5.

Carefully selecting the location for a survey can help to keep the impact of multipath at a minimum, as well as the use of a good antenna.

The accuracy of *standalone positioning* with GPS, also referred to as single-point positioning, or absolute positioning, according to model (14.2), in the order of 5-15 meters under reasonable satellite visibility, is limited by the accuracy of the range measurements (time can

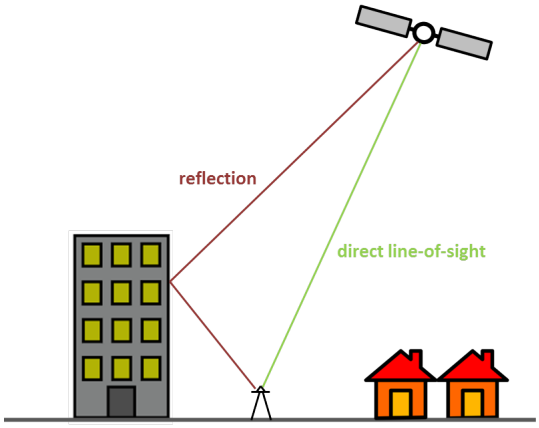



Figure 14.5: Multipath: the direct line of sight signal from the satellite is received, though as well as a signal which has been reflected by the building. The reception of also a reflected signal, which has made a detour, will generally cause a bias in the measurement.

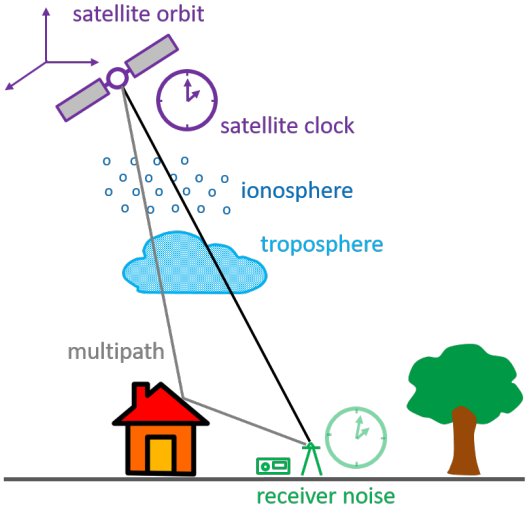



Figure 14.6: GPS error sources. The receiver clock offset (shown in faded green) is accounted for in the observation equation (14.1), and hence not to be considered as an error source.

| error source | 95%-value |
|---|---|
| satellite orbit | 2 m |
| satellite clock | 2-5 m |
| ionosphere | 15-90 m |
| troposphere | 20 m |
| multipath | 1-10 m |
| receiver noise | 1-3 m |
| total range | 5-10 m |

Table 14.1: GPS error budget for standalone positioning, see also Figure 14.6. The errors are given in the range domain, using the satellite (broadcast) navigation data message, and after Klobuchar ionospheric model correction (which in practice yields a 50% reduction of the ionospheric delay error), as well as tropospheric delay correction based on an a-priori (blind) model (which yields about 90% of reduction of the tropospheric delay error). The larger values for ionospheric and tropospheric delay may occur for slant ranges to satellites at low elevation.

Figure 14.7: Survey-marker at TU Delft campus with accurate ground-truth coordinates: $X$ = 3923768.0147 m, $Y$ = 300255.7048 m, $Z$ = 5002640.2228 m (ITRF2014 at epoch 2021.50).

be determined correspondingly with a tens of nanoseconds accuracy). The GPS pseudorange measurements contain errors due to inaccurate satellite orbit and clock information, delays along the path of the radio signal, including atmospheric delays (ionosphere and troposphere), local effects including multipath, and measurement noise, see Figure 14.6 and Table 14.1.

Finally it is mentioned that a GPS receiver, using electromagnetic signals received from the satellites, determines the position of the antenna phase center (typically a point inside or slightly above the antenna) as this is where the radio signals actually arrive. Handling the so-called antenna Phase Center Offset (PCO) with respect to the bottom of the antenna, usually a cm to dm-effect, is important in high-precision positioning discussed in the next chapter.

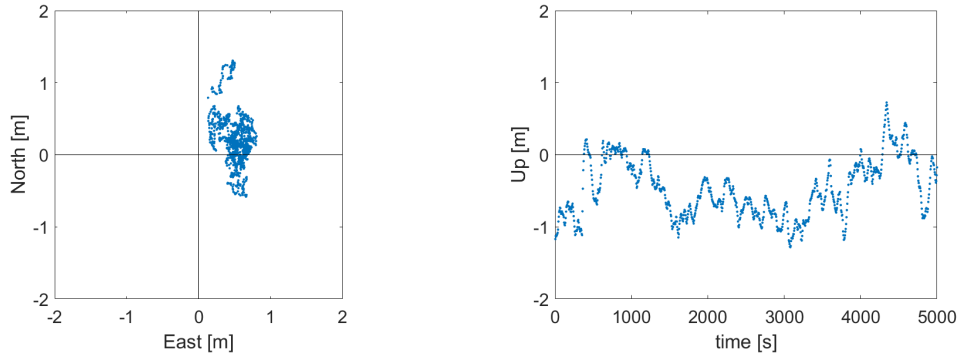



Figure 14.8: Example of GNSS standalone positioning for a duration of 5000 seconds at a 5 second interval, on September 2nd, 2021, with measurements to about 25 GNSS satellites. At left: scatter of horizontal position error, at right: time series of vertical position error.

## 14.6. Standalone positioning: example

With the equipment of Figure 15.7 a short experiment was carried out, lasting 5000 seconds. The antenna was installed on a survey-marker on the TU Delft campus, of which accurate position coordinates were already available, see Figure 14.7. The receiver ran in so-called

| | East | North | Up |
|---|---|---|---|
| mean [m] | 0.51 | 0.23 | -0.47 |
| std [m] | 0.15 | 0.35 | 0.41 |
| rms [m] | 0.53 | 0.42 | 0.62 |

Table 14.2: Empirical mean, standard deviation (std) and root mean square (rms) of position error, based on $N$=1000 GNSS standalone position solutions.

standalone positioning or single-point positioning mode, and every 5th position solution was saved, hence the results shown in Figure 14.8 are obtained at a 5 second interval.

The graph at left shows the horizontal position scatter, North-coordinate versus East-coordinate, and the graph at right shows the vertical position (Up) as a function of time. With the given coordinates of the marker, we actually present the position *error* in Figure 14.8, i.e. the difference of the measured position coordinate and the known ground-truth position coordinate. Hence, the origin of this graph refers to the 'true' position. The position errors are expressed in a local topocentric coordinate system, in terms of local East, North and Up, see Section 29.4.

Table 14.2 presents the resulting empirical mean, standard deviation (std) and the root mean square (rms), which is the square root of the MSE, see Chapter 6, of the position error in East, North and Up, showing a better than 1 meter accuracy of GNSS standalone positioning using over 25 GNSS satellites.

# 15

# GPS positioning modes

Several techniques have been developed to improve on the GPS Standard Positioning Service (SPS) accuracy (standalone positioning, as discussed in the previous chapter). Firstly, GPS satellites broadcast a second, more precise, code on the same carrier wave, to provide the Precise Positioning Service (PPS). However, this code is encrypted and can only be used to full extent by the US military.

Fortunately, even more accurate positioning modes are available, all relying on a kind of *augmentation*. This means that, next to the measurements collected by the user receiver, in addition measurements are used of a nearby permanent GPS receiver, and/or that one relies in addition on data products derived from a network of permanent tracking stations. Such a network could provide precise estimates of the satellite positions for instance (more precise than what we by default encounter in the navigation message on the GPS signal).

## 15.1. Relative positioning, or DGPS

Differential GPS (DGPS) uses a data link to a nearby base or reference station, i.e. another GPS receiver at an accurately known position, and the *relative position* between the two is obtained. Measurement data from this base station are used, to reduce the effects of the atmospheric delays, satellite clock offsets and orbit errors. This can be achieved by differencing the observations from both receivers to the same satellites, which eliminates these (common) errors, which affect both receivers almost identically if the distance between them is small enough, typically in the order of 5 to 10 km, considering that the satellites are at 20.000 km distance, see Figure 15.1.

From the differenced observations, the so-called baseline (vector) between the two receivers can be computed through least-squares estimation. The position of the rover is then obtained by adding the baseline vector to the accurately known coordinates of the reference station. Generally the term 'DGPS' is used for relative positioning, though using only pseudorange measurements.

### 15.1.1. Real-Time Kinematic (RTK)

To obtain the highest possible accuracy from GPS, it is no longer sufficient to use only the pseudorange code measurements, but rather the *carrier phase* measurements, introduced in Chapter 13, are required. As mentioned before, the measurement of fractional phase difference does pose the problem of the unknown initial number of carrier wave cycles, also called the carrier wave *ambiguity*, which need to be estimated together with the other unknown parameters.

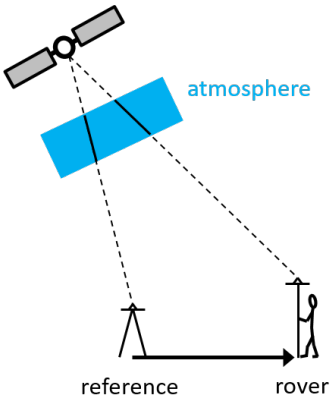



Figure 15.1: Relative GPS positioning combines measurements from a roving receiver with measurements from a reference (or base) station. The position of the rover is actually computed *relative* to the position of the base station. A number of errors, including atmospheric errors, is almost identical for two receivers in close proximity to each other. Hence, these errors cancel in relative positioning.

An ambiguity consists of a fractional part at the satellite (equal for both receivers, and already removed by the differencing between the base station and rover), a fractional part at the receiver (equal for all tracked satellites), and an integer number of whole cycles. This fact of unknown parameters being integers (rather than reals) is exploited in a technique called *Real-Time Kinematic (RTK)* positioning, or Carrier-Phase (CP) based baseline processing (if performed in post processing), by selecting a reference satellite and forming a second difference between the measurement to a reference satellite and those to all other satellites, to eliminate the fractional part at the side of the receiver. In this special case the double-differenced carrier phase ambiguities can be resolved to their integer number very efficiently through integer least-squares estimation. After only a few minutes or within tens of seconds already, centimeter-level position accuracy can be reached.

The requirement for a nearby reference receiver is a disadvantage of RTK, considering effort and/or cost. With RTK the coverage area of a reference receiver or station typically has a radius of ten, or tens of kilometers. In many regions and countries, networks of reference stations, or Continuously Operating Reference Stations (CORS) have been deployed in order to cover the entire area, and in this scenario sometimes the term *network-RTK* is used, see Figure 15.2, where reference stations generally have a 30-40 km interdistance. An example of an application of RTK positioning in road construction is shown in Figure 15.3.

Many high-end GPS receivers have RTK functionality built-in, but it can also be performed with professional software, or even with open source software such as the RTKLIB program package [40].

Today the measurements of the base station are communicated, in real-time, to the rover receiver over an Internet-connection, using NTRIP. Networked Transport of RTCM[1] via Internet Protocol (NTRIP) is an application protocol that supports the streaming of (differential) GNSS data over the Internet, based on Hyper Text Transfer Protocol (HTTP). NTRIP has been developed by the German Federal Agency for Cartography and Geodesy (BKG) [41]. With the measurements of the base station becoming available in real-time at the rover receiver, centimeter accurate position solutions are obtained right at the spot.

[1]Radio Technical Commission for Maritime Services - Special Committee 104 on Differential GNSS


Figure 15.2: Example of network of permanent GPS tracking stations, of a commercial network RTK service provider in the Netherlands, Belgium and Luxemburg. Image obtained with permission from 06-GPS [38].


Figure 15.3: Excavator in the process of constructing a motorway embankment. RTK-GPS provides accurate real-time position information to guide this machine (note the two GPS-antennas on the back of the engine). Image courtesy of Heijmans [39].

### 15.1.2. RTK — carrier phase observation equation [*]

The pseudorange observation equation was presented in (14.1) for the purpose of standalone positioning. The errors discussed in Section 14.5 were basically all ignored.

The carrier phase measurement, Section 13.2.2, is much more precise than the pseudorange — the contribution to the error budget in Table 14.1 by carrier phase multipath and receiver noise would only be at the millimeter to a few centimeter level. The other error sources, like atmospheric delays and satellite related errors are taken into account now, and put together in a delay parameter $d_r^s$. The carrier phase observation equation, for the phase $\varphi_r^s = \lambda \frac{\Phi_r^s}{2\pi}$ expressed in meters, reads

$$\underline{\varphi}_r^s = \underbrace{\sqrt{(x^s - x_r)^2 + (y^s - y_r)^2 + (z^s - z_r)^2}}_{l_r^s} + b_r + d_r^s + \lambda N_r^s + \underline{e}_r^s$$

Parameter $N_r^s$ denotes the carrier phase cycle ambiguity, see Figure 13.2.

### 15.1.3. RTK — carrier phase positioning: parameter estimation [*]

We use *relative* positioning and develop the model of observation equations for a short baseline (i.e. two receivers close together, up to 10-20 km distance). The two receivers 1 and 2 being close together implies that the delays will be very similar $d_1^s \approx d_2^s$ (keeping in mind that the satellite is some 20.000 km away), and in the sequel we assume them to be really equal: $d_1^s = d_2^s$ (and residual errors are assumed to go into the $\underline{e}$-error terms). With the position coordinates of the reference or base station $(x_1, y_1, z_1)$ being known, and taking the difference of measurements across the two receivers, $\varphi_{1,2}^s = \varphi_2^s - \varphi_1^s$, we obtain

$$\begin{pmatrix} \Delta\underline{\varphi}_{1,2}^1 \\ \Delta\underline{\varphi}_{1,2}^2 \\ \vdots \\ \Delta\underline{\varphi}_{1,2}^m \end{pmatrix} = \begin{pmatrix} -u_{2,x}^1 & -u_{2,y}^1 & -u_{2,z}^1 & 1 & \lambda & & \\ -u_{2,x}^2 & -u_{2,y}^2 & -u_{2,z}^2 & 1 & & \lambda & \\ \vdots & \vdots & \vdots & \vdots & & & \ddots \\ -u_{2,x}^m & -u_{2,y}^m & -u_{2,z}^m & 1 & & & \lambda \end{pmatrix} \begin{pmatrix} \Delta x_2 \\ \Delta y_2 \\ \Delta z_2 \\ b_{1,2} \\ N_{1,2}^1 \\ N_{1,2}^2 \\ \vdots \\ N_{1,2}^m \end{pmatrix} + \begin{pmatrix} \underline{e}_{1,2}^1 \\ \underline{e}_{1,2}^2 \\ \vdots \\ \underline{e}_{1,2}^m \end{pmatrix} \qquad (15.1)$$

with $b_{1,2} = b_2 - b_1$, $N_{1,2}^s = N_2^s - N_1^s$ and $\underline{e}_{1,2}^s = \underline{e}_2^s - \underline{e}_1^s$. Note that when we would leave the ambiguities $N$ out, the above model in structure very much resembles model (14.2) for standalone positioning. The goal of RTK-positioning is to estimate the position coordinates of the rover receiver $x_2$, $y_2$, and $z_2$, and this is done while keeping the reference station fixed to the given position coordinates.

In the above model the receiver clock offset parameter $b_{1,2}$, as it is appearing equally in all equations, can be removed by taking differences between measurements, e.g. $\varphi_{1,2}^{1,2} = \varphi_{1,2}^2 - \varphi_{1,2}^1$. The resulting model of taking differences all with respect to the first measurement $\varphi_{1,2}^1$, reads

$$\begin{pmatrix} \Delta\underline{\varphi}_{1,2}^{1,2} \\ \vdots \\ \Delta\underline{\varphi}_{1,2}^{1,m} \end{pmatrix} = \begin{pmatrix} -(u_{2,x}^2 - u_{2,x}^1) & -(u_{2,y}^2 - u_{2,y}^1) & -(u_{2,z}^2 - u_{2,z}^1) & \lambda & & \\ \vdots & \vdots & \vdots & & \ddots & \\ -(u_{2,x}^m - u_{2,x}^1) & -(u_{2,y}^m - u_{2,y}^1) & -(u_{2,z}^m - u_{2,z}^1) & & & \lambda \end{pmatrix} \begin{pmatrix} \Delta x_2 \\ \Delta y_2 \\ \Delta z_2 \\ N_{1,2}^{1,2} \\ \vdots \\ N_{1,2}^{1,m} \end{pmatrix} + \begin{pmatrix} \underline{e}_{1,2}^{1,2} \\ \vdots \\ \underline{e}_{1,2}^{1,m} \end{pmatrix} \qquad (15.2)$$

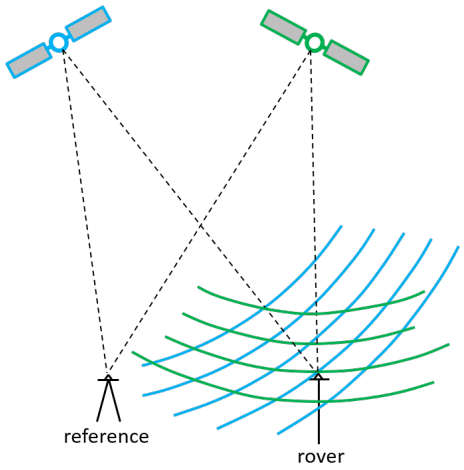



Figure 15.4: Geometric interpretation of relative positioning with carrier phase measurements, which are inherently *ambiguous*. The blue circle arcs, as possible solution for the rover receiver position result from the carrier phase measurement to the blue satellite, and the green circle arcs to those to the green satellite. The arcs are spaced by one wavelength $\lambda$ of the carrier wave.

With carrier phase measurements to $m$ satellites, we have $(m-1)$ of these so-called double difference measurements. The receiver clock offset parameter has been cancelled.

These two optional sections provide a very brief introduction to carrier phase based positioning. For a more in-depth coverage, the reader is referred to [31]. Least-squares estimation of *integer* parameters, such as the ambiguities $N$, is simple when there is only one. Ordinary least-squares estimation yields a real-valued estimate for this parameter, and rounding it to the nearest integer yields the integer least-squares estimate for the ambiguity. With more ambiguity parameters present in the problem at the same time, as in the above model, this becomes a seriously complex problem (for which an adequate solution is provided by the LAMBDA-method [31]).

Figure 15.4 provides a simple geometric interpretation of relative positioning with carrier phase measurements which are inherently ambiguous, as only the fractional phase can be measured. The rover receiver has to lie on one of the blue circle arcs, and at the same time on one of the green circle arcs. The different arcs represent different integer values for the ambiguity. The rover receiver is at one of the intersections, but as long as the ambiguities are not known, it is not known at which one. For this geometric interpretation it is assumed that there is no effect of noise present in the measurements, and the receiver clocks are assumed to behave perfectly ($b_1 = b_2 = 0$).

### 15.1.4. RTK — carrier phase positioning: example

With the equipment of Figure 15.7 a short experiment was carried out, lasting 1000 seconds. Using measurements from a permanent GNSS reference station (only 2 km away, cf. Figure 16.1), received in real-time using NTRIP, the receiver provided so-called RTK-fixed solutions (in ETRF2000). For every epoch, i.e. once every second, a new position solution was computed, and the results are shown in Figure 15.5. The graph at left shows the horizontal position scatter, North-coordinate versus East-coordinate, and the graph at right shows the vertical position (Up) as a function of time. These measurements were taken at a survey-marker of which accurate position coordinates were already available cf. Figure 14.7, so, in the graph of Figure 15.5 we actually present the position *error*, i.e. the difference of the measured position coordinate and the known ground-truth position coordinate. Hence, the origin of this

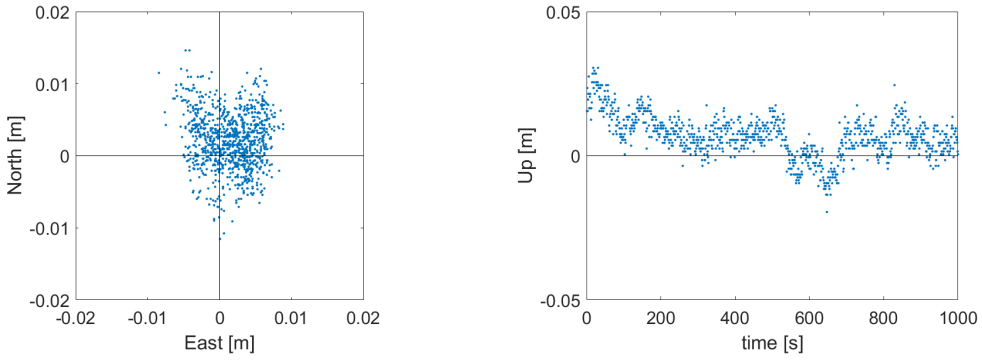



Figure 15.5: Example of Carrier Phase (CP) Real-Time Kinematic (RTK) positioning for a duration of 1000 seconds, on August 27th, 2021, with measurements of about 25 GNSS satellites, and successfully fixing the carrier phase ambiguities (RTK-fixed solution). At left: scatter of horizontal position error, at right: time series of vertical position error.

| | East | North | Up |
|---|---|---|---|
| mean [m] | 0.0016 | 0.0021 | 0.0068 |
| std [m] | 0.0033 | 0.0039 | 0.0072 |
| rms [m] | 0.0037 | 0.0044 | 0.0099 |

Table 15.1: Empirical mean, standard deviation (std) and root mean square (rms) of position error, based on $N$=1000 Carrier Phase (CP) Real-Time Kinematic (RTK) position solutions (with ambiguities fixed).

graph refers to the 'true' position. The position errors are expressed in a local topocentric coordinate system, in terms of local East, North and Up, see Section 29.4.

Table 15.1 presents the resulting empirical mean, standard deviation (std) and the root mean square (rms), which is the square root of the MSE, see Chapter 6, of the position error in East, North and Up, confirming centimeter-accuracy of RTK-GPS positioning. This is an improvement by a factor of 100 compared to the standalone positioning results in Figure 14.8.

### 15.1.5. RTK — carrier phase positioning: Digital Terrain Model (DTM)

Another short experiment was carried out to result in a centimeter-accurate 3D Digital Terrain Model (DTM) of an embankment on the TU Delft campus, see Figure 15.6. The RTK survey of this bank of earth took only 15 minutes (walking with the GNSS-receiver in a grid-like pattern over this bank, and recording measurements every 1 second).

The RTK-fixed position solutions have been interpolated, and the resulting DTM is shown at right in Figure 15.6. With the DTM one can easily evaluate numerically the amount of earthwork needed to create or remove this bank, in this case 442 $\mathrm{m}^3$.

### 15.1.6. Precise Point Positioning (PPP)

In those instances where a nearby reference receiver (or network) is not available or cost-prohibitive, Precise Point Positioning (PPP) is an attractive alternative. PPP only relies on a *global*, very sparse network of reference receivers (e.g. some 40 receivers worldwide, and the nearest reference station can be 1000 km away, or even further), which track the GPS satellites and compute corrections to the errors in the satellite orbits and clocks. Conventional PPP uses dual-frequency data to eliminate the ionosphere delay, while a low-cost variant uses single frequency data with a (predicted) ionosphere model. The fractional carrier phase

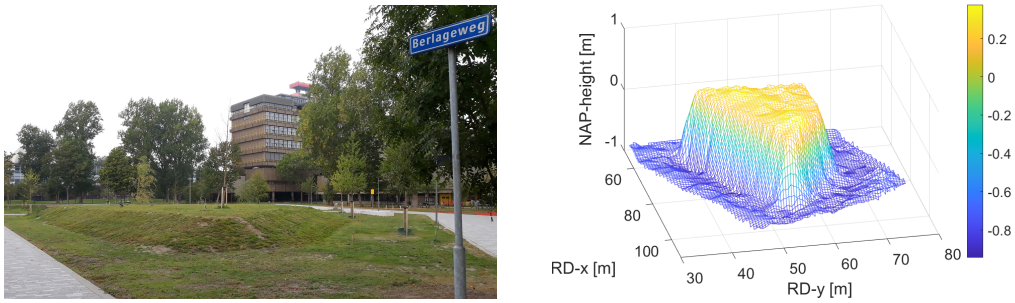



Figure 15.6: Example of a centimeter-accurate 3D Digital Terrain Model (DTM) resulting from Carrier Phase (CP) Real-Time Kinematic (RTK) positioning. The DTM is presented in the national RD-NAP reference system (see Chapter 35), actually with x-85600 m, and y-445900 m. The DTM is viewed from the South-East, like the photo on the left.

ambiguities cannot be eliminated, which means that integer least-squares estimation is not possible. Ambiguities can still be estimated as constant values though, since an ambiguity does not change as long as the receiver keeps tracking the satellite, a fact used in the PPP data processing.

However, because the ambiguities cannot be fixed to integer values, PPP suffers from a longer convergence period than RTK (think of tens of minutes). After a convergence period in which the accuracy of the estimated ambiguities improves gradually, the PPP solution starts relying more and more on the phase measurements. The eventual position accuracy for dual-frequency PPP can reach centimeter, or even millimeter level, while single frequency PPP can reach an accuracy of a few decimeter.

## 15.2. Current developments

Much research effort is spent to try and combine the best aspects of PPP and RTK, i.e. using a sparse (global) reference network and ambiguity resolution to enable precise positioning. Wide Area RTK and PPP-RTK are based on the principles of RTK, but try to stretch the interstation distances to several hundreds of kilometers, while PPP-AR starts from the global PPP network, and tries to solve the problem of ambiguity resolution (AR). The ultimate goal is to achieve high precision positioning across a (very) large area.

Another development is to bring high-accuracy positioning techniques, e.g. RTK and PPP, to low-cost devices. An example is shown in Figure 15.7. The smartphone retrieves GPS differential corrections (or measurements) of a nearby reference station through an Internet-connection using NTRIP, and forwards these to the GPS receiver, which is connected via USB to the smartphone. The firmware on the GPS receiver chip combines the corrections with the measurements of the rover receiver, and delivers a centimeter accurate RTK-position solution, which it relays back to the app on the smartphone. This allows for centimeter accurate navigation, in real-time, with your smartphone.

The antenna of the rover receiver, at right in Figure 15.1, is typically mounted on a lightweight range-pole, for convenience of the survey-job. The position of the antenna on top of the range-pole is being measured (with GPS), and the obtained coordinates are converted into those of the object or marker point occupied by the bottom-tip of the range-pole, using the fact that the range-pole is being held vertically straight-up, and knowing its size.

Recently range-poles with built-in tilt compensation have become available. The tilt angle $\zeta$ is being measured, for instance by means of an inertial measurement unit, and the horizontal displacement or offset is simply found as $l \sin \zeta$, see Figure 15.8.

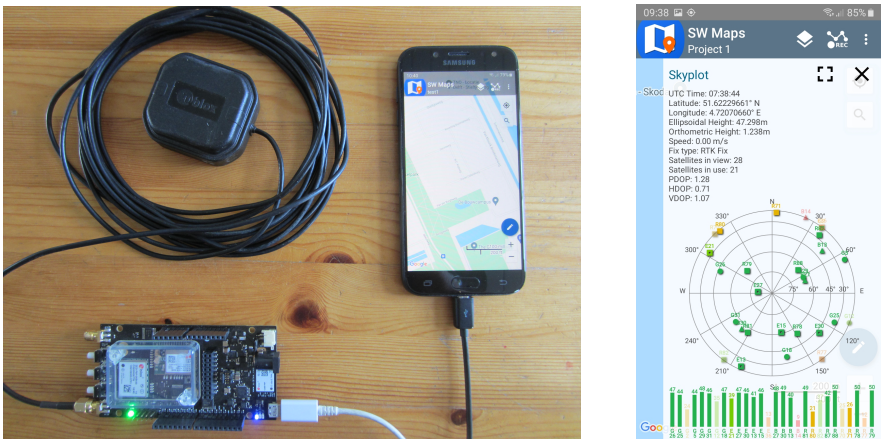



Figure 15.7: At left: dual-frequency, multi-constellation GNSS receiver with receiver board at bottom, small patch-antenna (black) on top, and smartphone with Android positioning app at right; a total equipment cost of below 500 Euro (u-blox ZED-F9P), yet capable of providing cm-accurate RTK-GNSS positioning. At right: screenshot of SW Maps app by Softwel (P) Ltd.

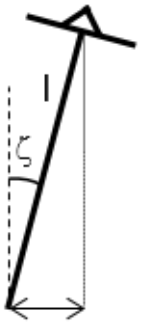



Figure 15.8: Principle of GPS range pole with tilt compensation. Tilt angle $\zeta$ is measured to provide, with known size $l$, the horizontal displacement $l \sin \zeta$.

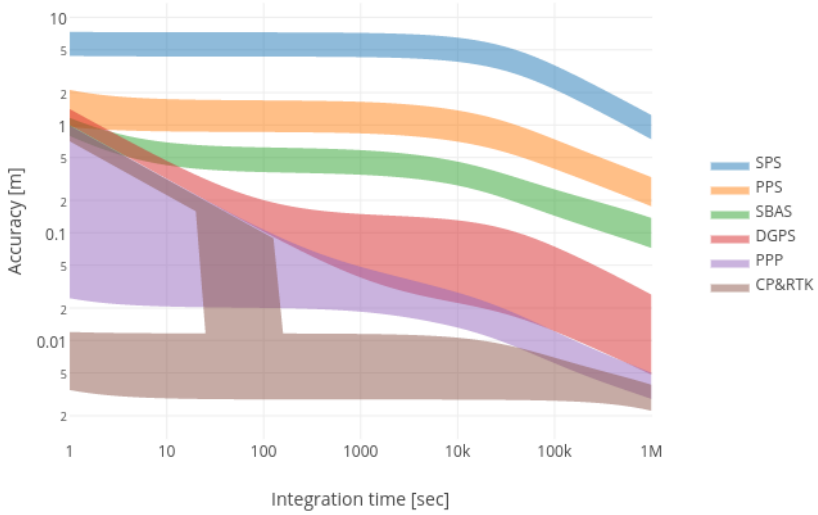



Figure 15.9: Accuracy of various GPS positioning modes for a static receiver. The integration time is the total measurement duration, along the horizontal axis, and the position coordinates accuracy is along the vertical axis. Note the logarithmic scales. CP&RTK stands for Carrier Phase and Real-Time Kinematic positioning.

Satellite Based Augmentation Systems (SBAS), e.g. the European EGNOS system, designed to enable GPS-based aircraft precision approaches, rely on the same principles as PPP. However, given the primary application, the focus is on integrity rather than accuracy (integrity refers to the trust that can be placed in the resulting position solution, the solution is largely fault-tolerant). Carrier phase measurements are only here used to 'smooth' the pseudorange solution. SBAS is a pseudorange code Differential GPS approach for large geographical areas (wide areas). An additional advantage of using SBAS is that the corrections are transmitted on the same radio frequency as GPS signals, so no additional data link is necessary.

## 15.3. Processing strategies, dynamic model and observation period

As already hinted at, the GPS position accuracy improves when the measurement time duration increases. One important factor here is the *dynamic model* of the receiver motion, or how the measurement epochs can be 'linked' to each other.

If the receiver is stationary, the improvement will be most notable, as we can basically estimate a single position from many measurements (a *static* solution). The position accuracy of a static receiver is shown as a function of the measurement duration in Figure 15.9 for each of the previously covered GPS processing strategies.

For a moving receiver the accuracy can also improve over time, if we can exploit the fact that some of the other parameters are constants, e.g. the ambiguities, or, if the movement can be constrained or predicted to some extent based on the current position (e.g. a car driving along a straight line, at constant velocity). This can be implemented with a Kalman filter, or a recursive least-squares data processing algorithm.

In a *kinematic* solution, position coordinates are computed for each measurement instant (for instance every 1 second), to accomodate the fact that the receiver is/was actually moving during the survey. As a result one then obtains a *list* of position estimates, e.g. one every second, instead of one overall position solution as with a static survey. The list describes the *track* or *trajectory* of the moving receiver.

Two related issues are:

1. The difference between *real-time* processing, and *post-processing* (for instance after whole survey has been completed), where post-processed results are generally more accurate, but obviously not available right on the spot, and hence not suitable for certain applications.

2. The measurement rate of the receiver: GPS receivers often take pseudorange code, carrier phase, Doppler shift and signal-to-noise (SNR) measurements once every second, hence at 1 Hz, but depending on the application, 10-20 Hz is also common practice today, and technically up to a 100 Hz measurement rate is possible. To reduce the computational burden, data storage, and power requirement, lower measurement rates (e.g. once per 30 seconds) are common in applications where objects move only very slowly, like in geoscience on measuring tectonic plate motion. The impact of the measurement rate on the position accuracy is marginal (a higher data rate can slightly improve precision), because the measurement errors are generally correlated in time. This means that measurements taken in quick succession are not independent, and thereby do not offer, precision-wise, a lot of new/additional information.

# 16

# GNSS and applications

In this chapter we present a concise overview of Global Navigation Satellite Systems (GNSS), addressing GPS, Glonass, Galileo and BeiDou. Then we briefly touch upon the wide range of applications of GNSS.

## 16.1. Global Navigation Satellite Systems (GNSS)

The Global Positioning System (GPS), developed by the US military and operated by the US Air Force (USAF), is the first Global Navigation Satellite System of its kind. In order not to be dependent on a US military system and/or to get their share of the GNSS market, other countries have developed their own Global Navigation Satellite Systems (GNSS). The result is that today a lot of GNSS satellites can be seen at the same time, anywhere on Earth, anytime. Figure 16.1 shows as an example a so-called skyplot for Delft, with up to 40 GNSS-satellites in view.

Recently we have seen a significant increase in the available Global Navigation Satellite Systems, satellites, radio-frequencies and signals. These developments are briefly reviewed in this section.

### 16.1.1. GPS

GPS is in the process of modernization. This is achieved by following up older satellites by new satellites with expanded and improved capabilities. The civil L2C signal, for improved dual frequency (civilian) performance, becomes available on more and more satellites. Even more importantly, new GPS satellites also transmit an additional (wideband) signal on the L5-frequency (of 1176.45 MHz) primarily designed for safety-of-life applications (higher chiprate, hence shorter chiplength, and more precise pseudorange measurements).

### 16.1.2. Glonass

The Russian GLObal NAvigation Satellite System (GLONASS), has been fully replenished and at present has 24 active satellites. Planned modernizations of GLONASS include an additional signal transmitted on the L5-frequency, and a switch from Frequency Division Multiple Access (FDMA) to CDMA, which will increase interoperability with other GNSSes.

### 16.1.3. Galileo

Galileo, the European GNSS, is still under development, currently with 22 satellites. The full Galileo constellation for Full Operational Capability will consist of 30 satellites. The Galileo system transmits navigation signals on four different carrier frequencies: L1/E1, L5/E5a, E5b

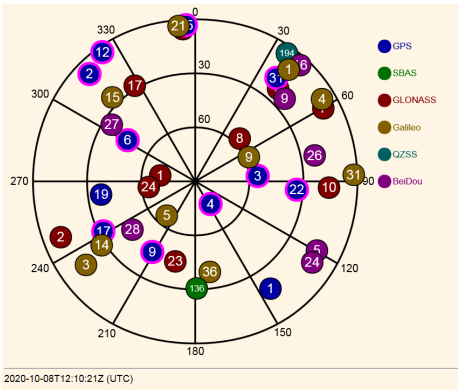



Figure 16.1: Skyplot with GNSS satellites for October 8th, 2020, at 12:10 UTC, in Delft. The skyplot shows the positions of the satellites of the various constellations, like GPS, GLONASS, Galileo and BeiDou, in the sky. The outer circle represents the local horizon in Delft, 360 degrees around (0 is North, 90 East, etc). The smaller circles refer to 30 degrees of elevation, above the horizon, and 60 degrees of elevation. The middle of the skyplot corresponds to the so-called local zenith, which is directly overhead. The skyplot was obtained from the Trimble NetR9 GNSS receiver at the TU Delft observatory, of which the antenna set-up is shown at right.

and E6, two of which (E5a and E5b) can also be tracked together as one extra wideband (AltBOC) signal with unprecedented pseudorange accuracy.

### 16.1.4. BeiDou

The Chinese BeiDou Navigation Satellite System (BDS), sometimes still known as Compass, was designed to provide independent regional navigation in the first stage and global coverage later. The BeiDou (phase III) constellation deployment has been fully completed in 2020, with 30 satellites in orbit, providing global coverage.

### 16.1.5. Concluding remarks

The realized and expected upgrades of and additions to the available GNSS signals can have a range of improvements on many GNSS applications. Some of the more important ones are: the higher pseudorange accuracy of the new signals, the availability of many more satellites at once (more satellites available to combat urban environments, see Figure 16.1), and both the availability of more radio-frequencies and satellites.

Multi-GNSS positioning also brings new challenges, as so-called InterSystem Biases (ISB) are introduced in the model. The system time as maintained by GPS may (will) not be the same as the system time as maintained for Galileo, for instance. Hence one has to account for the fact that these systems may have an offset in time with respect to each other. To use multiple systems simultaneously in an optimal manner, these biases must be studied, and if possible corrected or eliminated.

## 16.2. Applications

There are many different applications of GNSS positioning each with its own requirements and, related to that, a preferred processing strategy.

- smartphones, car navigation, and personal navigation usually have the lowest requirements, and the GPS (GNSS) standard positioning service suffices, cf. Figure 16.2.
- lane specific navigation advice for road users requires sub-meter position accuracy, which can be fulfilled with single frequency PPP.

Figure 16.2: Car navigation, route guidance and fleet management in traffic and transport are popular applications of GNSS positioning, where standard positioning service suffices. In future, assisted and automated driving will call for improved accuracy.

Figure 16.3: Both the on- and offshore part are regularly surveyed, to monitor the development of the Zandmotor (The Sand Engine), at the Dutch North-Sea coast, near Ter Heijde. This 'building with nature' project started in 2011, and at right an aerial photo of the Zandmotor is shown, looking in Southern direction. High-precision RTK-GPS is used for positioning the quad on shore, and the jet-ski in the water (note the GPS-antenna at the back of the jet-ski, in the inset). The measurements by the quad result in a Digital Terrain Model (DTM), and echo sounder depth measurements by the jet-ski result in a seafloor-map. Photo at left by Matthieu de Schipper [42]. Photo at right by Pmblom - own work, May 2016, taken from Wikimedia Commons [9] under CC BY-SA 4.0 license.

- surveying for creating maps and construction works, requires cm to mm position accuracy and will use RTK if available, or PPP otherwise, cf. Figures 16.3 and 15.6.
- deformation monitoring, due to Earthquakes, volcanic activity, mining or extraction of petroleum or natural gas, as well as any number of scientific applications require the highest possible accuracy and use carrier phase based positioning.
- aircraft precision approach and landing requires high integrity positioning, and can use SBAS to obtain this.
- machine guidance, as shown in Figure 15.3 and in particular self-driving vehicles require high accuracy and integrity; this can be achieved by using RTK-GNSS though this is still subject of research, and likely fusion with additional sensors is in order.

There is also a number of GNSS applications, in which the position solution is not the (primary) goal. Accurate time which is obtained through determining also the receiver clock offset $b_r$, is used in timing applications. The standard positioning service allows for timing at the 10-100 ns level, and this is used for instance in telecommunication, cf. Figure 16.4, electrical power grids, and financial networks.

Figure 16.4: A GPS receiver is commonly used to synchronize base stations for telecommunication. Requirements on time-synchronization for this application lie in the order of a $\mu$s. The photo shows a base station with a height of 37 m, providing the full range of mobile services from 2G (GSM) to 5G (NR).

Nuisance parameters such as the atmospheric delays can also be used as observational input e.g. to determine, together with using models, the state of the Earth's ionosphere, or derive troposphere delays, for instance for Numerical Weather Prediction (NWP).

GNSS radio-signals can also be used outside of their intended purpose, e.g. to determine sea-level height by measuring reflected GNSS signals from orbit.

## 16.3. Resources

This part provides an introduction to positioning with GPS/GNSS. For a lot more of technical and mathematical modeling information on GPS and GNSS positioning, navigation and timing, the reader is referred to [31] and [32]. These textbooks also cover a wide range of applications.

The first source of information on GPS, as well as the point of contact is the Navigation Center of the US Coast Guard [43]. Official U.S. government information about GPS is available through GPS.gov [44].

The first source of information on Galileo and point of contact is the European Union Agency for the Space Programme (EUSPA) [45].

The IGS is the International GNSS Service, a voluntary federation of universities and research institutions, operating permanent GNSS stations worldwide, and providing GNSS data and products for high(est)-precision applications [46].

## 16.4. Exercises and worked examples

This section presents a couple of questions and problems with (worked) answers on GPS-positioning.

**Question 1** What are the largest remaining error sources in short-baseline DGPS, explain your answer.

**Answer 1** The atmosphere delays as well as the satellite orbit and clock errors are eliminated in DGPS, cf. Figure 15.1, which leaves multipath and (pseudorange) measurement noise as the largest error sources, cf. Figure 14.6 and Table 14.1.

**Question 2** If a certain application requires decimeter positioning accuracy, which GPS

positioning modes can be considered? And for how long a time would we need to collect measurements?

**Answer 2** Real-Time Kinematic (RTK) provides decimeter or even centimeter accuracy as soon as the ambiguities can be fixed, which generally is (well) within 100 seconds of measurements, and even faster in post-processing. PPP can also provide decimeter accuracy after several minutes. DGPS can reach decimeter accuracy as well, but may considerable time to allow for averaging (with static positioning only), for instance one hour. SBAS and standalone GPS often do not reach decimeter accuracy even after one or several days (averaging with static positioning), especially in the vertical component. An overview of the attainable accuracies can be found in Figure 15.9.

**Question 3** The principle of GPS satellite positioning and navigation consists of determining the range from satellite to receiver through measurement of the signal travel-time. The atomic clock in the satellite is perfectly on time. When the receiver clock is ahead of time by 0.1 $\mu$s, by how much is the measured range to the satellite too long or too short?

**Answer 3** From Eq. (13.2) we can see that clock error $\delta t_r$ is positive, as the receiver clock is ahead of time, hence $t_r > t$. Next, with Eq. (13.3), and $c\delta t_r = b_r = 30$ m, we find that the pseudorange $p_r^s$ is too long by 30 m (compared to the actual distance $l_r^s$).

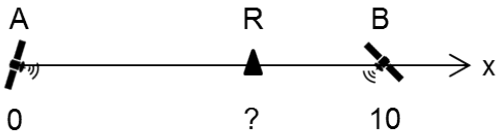



Figure 16.5: One-dimensional GPS positioning (Question 4).

**Question 4** The GPS positioning problem has been simplified to a single dimension. There are two satellites A and B, and the user receiver is at R, see Figure 16.5. The positions of the satellites are known, A is at $x_A = 0$, and B is at $x_B = 10$. The position of the user $x_R$ is unknown. Two pseudoranges have been measured: $p_{AR} = 9$ and $p_{BR} = 7$. Determine the position (coordinate) of the user at R.

**Answer 4** Looking at Figure 16.5 we identify two geometric ranges, namely $l_{AR} = x_R - x_A$ and $l_{BR} = x_B - x_R$ (mind to define these distances to be positive). Then, with Eq. (13.3), we formulate two observation equations:

$$\begin{aligned} p_{AR} &= l_{AR} + b_R \\ p_{BR} &= l_{BR} + b_R \end{aligned}$$

which gives

$$\begin{aligned} p_{AR} &= x_R - x_A + b_R \\ p_{BR} &= x_B - x_R + b_R \end{aligned}$$

and with the given satellite positions, we obtain

$$\begin{aligned} p_{AR} + x_A &= x_R + b_R \\ p_{BR} - x_B &= -x_R + b_R \end{aligned}$$

We have two equations with two unknown parameters, namely $x_R$ and $b_R$, which we can solve, giving $x_R = 6$ and $b_R = 3$. The user position coordinate equals $x_R = 6$, and we are typically not interested in the receiver clock offset. It is easily verified that correcting the measured

pseudoranges for the receiver clock offset yield the actual distances from the two satellites to the receiver: $p_{AR} - b_R = 9 - 3 = 6$ and $p_{BR} - b_R = 7 - 3 = 4$.

**Question 5** The GPS relative positioning problem has been simplified to a single dimension. There is one satellite 'sat' (or just 's') and it is visible at the local horizon. The receivers '1' and '2', and the satellite are all on a straight line (along the x-coordinate axis), see Figure 16.6. The radio-signals from the satellite to the two receivers pass through the Earth's atmosphere (layer 'atm') and get thereby delayed; the delay, expressed in units of range, is denoted by $d^s$. This delay is unknown (but *equal* for the signals to both receivers). The satellite position is known, $x^s = -20$, and the position of receiver 1 as well $x_1 = 5$. Compute the position of receiver 2, $x_2$, based on the pseudorange measurements $p_1^s = 32$ and $p_2^s = 37$. In this case, you can again assume that all clocks run perfectly on time – there are no clock offsets involved.

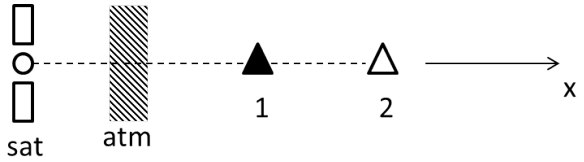



Figure 16.6: Relative positioning in one dimension (Question 5).

**Answer 5** The pseudorange observation equation (13.3) needs to be adapted. There is no clock offset involved at all, so parameter $b_r$ cancels, but now, we face an unknown atmospheric delay $d^s$. Hence

$$\begin{aligned} p_1^s &= l_1^s + d^s \\ p_2^s &= l_2^s + d^s \end{aligned}$$

Looking at Figure 16.6 we identify two geometric ranges, namely $l_1^s = x_1 - x^s$ and $l_2^s = x_2 - x^s$ (mind to define these distances to be positive). Then the two observation equations become:

$$\begin{aligned} p_1^s &= x_1 - x^s + d^s \\ p_2^s &= x_2 - x^s + d^s \end{aligned}$$

where there are two unknown parameters, namely $x_2$ and $d^s$. With the given measurements and coordinates, this is easily solved, to yield $x_2 = 10$ and $d^s = 7$. Alternatively one could take the difference of the two pseudorange measurements $p_2^s - p_1^s = x_2 - x_1$, which gives an identical result for $x_2$, and one is generally not interested in parameter $d^s$.

# IV

# Remote sensing

# 17

# Introduction

The goal of surveying is to gather information about the Earth, the Earth's surface and its topography. In order to gather information about phenomena and processes on Earth, we take *measurements* by surveying and remote sensing. From these measurements we extract the information needed to model the Earth's surface and processes taking place on Earth. Mostly we focus on *geometric* information, hence on '*where* things are'.

The measurements are taken in the real world and therefore start from physical principles. The purpose of this part is to convey the fundamentals of measurements for surveying, including remote sensing. We present the concepts and principles of measurements, primarily from a physics perspective.

## Dutch historical perspective

Important discoveries and contributions in the fields of surveying and remote sensing were made in the past by two Dutch scientists, at a close distance from Delft: Willebrord Snellius from Leiden, and Christiaan Huygens from The Hague.


Figure 17.1: At left: portrait of Willebrord Snellius (1580-1626). Image taken from Wikimedia Commons [9]. Public Domain. At right: memorial plaque at Douzastraat 2a in Leiden, where Snellius lived. In the field of land surveying Snellius carried out triangulation, and he invented the method of positioning by resection (in Dutch: achterwaartse insnijding).

Willebrord Snellius, born Willebrord Snel van Royen (1580-1626), Figure 17.1, was an astronomer and mathematician, and known in English-speaking countries as Snell [47]. His name has been assigned to the law of refraction of light, though this law of refraction was actually discovered much earlier. Snell's law describes the relationship between the angles of


Figure 17.2: At left, Christiaan Huygens (1629-1695), painting by Caspar Netscher (1639-1684), 1671, Haags Historisch Museum. Image taken from Wikimedia Commons [9]. Public Domain. At right, mansion of the Huygens family at Hofwijck, currently housing the Huygens museum in Voorburg.

incidence and refraction, when light is (or waves are) passing through a boundary between two different isotropic media, such as water, glass and air. The law follows from Fermat's principle of least travel time. In Appendix G we review Snell's law.

Christiaan Huygens (1629-1695) was a mathematician and physicist, Figure 17.2. Huygens performed early telescopic studies of the rings of planet Saturn and discovered its moon Titan. He wrote a first treatise on probability theory ('Van reeckening in spelen van geluck'). His invention of the pendulum clock was a breakthrough in timekeeping; time is the very basis of many measurements in surveying and remote sensing today. In 1673, Huygens published the 'Horologium Oscillatorium sive de motu pendulorum', his major work on pendulums and horology. Huygens derived the formula for the period of an ideal mathematical pendulum. His wave theory of light was eventually published in 1690 as 'Traité de la lumière' [48]. We return to the subject of timekeeping — though with today's means — in Section 21.2, on oscillators.

## Overview of this part

This part serves as an introduction into the subject of remote sensing. Technical implementation details and system specifications are not dealt with, nor are operational procedures in practice.

First, the basic principles of the measurement of angle and distance are presented in Chapter 18. Next, a wide range of measurement techniques for surveying and mapping is presented, such as aerial stereophotogrammetry, ranging by means of laser, radar and sonar, and subsequently imaging techniques, such as laser scanning, interferometric synthetic aperture radar and multi-beam echo sounding.

While these chapters focus on *geometric* information, the last chapter of this part, Chapter 25, dedicated to optical remote sensing, covers *radiometric* information. Here the focus is on '*what* kind of things are there' and 'how much is there'.

# 18

# Measurements of geometry

Geometric information can be gathered through measurement of *angle* and *distance*. In this chapter we discuss the basic principles and occurence of both these fundamental measurements in surveying and remote sensing. The last section, titled interferometry, covers the measurement of a *change* in distance.

## 18.1. Measurement of angle

In this section we cover the theodolite, and also optical imaging for photogrammetry. They are both based on the principle of measuring angles.

### 18.1.1. Theodolite

Measurements of azimuth and direction can be made (physically) by reading a 'solid state' scale in the desired direction. The instrument is basically a ring with markers which divide the full 360 degrees (or 400 gon) in equal portions (pies), see Figure 18.1 at left. A (straight) linear scale is made cyclic by 'wrapping' it around a circle (or cylinder), and the circumference of the circle is thereby divided into equidistant parts.

The instrument is centered at the user/observer, and — relying on optical light — he or she has to point the telescope of the theodolite into the desired direction (to the target object), cf. Figure 4.2, and the scale of the device can be read-off mechanically/visually, similar to reading a compass on a vessel in measuring the azimuth (bearing) to for instance a lighthouse.

An electronic theodolite today measures angles by means of electro-optical scanning of digital codes (similar to bar-codes) etched on a glass cylinder or disc within the instrument, see Figure 18.1 at right. More precisely, the instrument can measure a direction to an object of interest, as any direction over the full 360-degrees range has a unique digital code in the instrument. The resolution for high-end equipment is in the order of a few micro-radians, which is about 0.0001 degree or gon.

The theodolite and total station are covered in detail in Chapter 4.

### 18.1.2. Optical imaging: photo camera

With a camera, a three-dimensional situation is mapped, or projected, onto a two-dimensional image (plane), using optical signals (light). Typically, available visible Sun-light is reflected by the topography on the Earth's surface and propagates through the lens and is subsequently captured on a photo film (or sensor array). The *lens* is the central part of the camera. Optical sensing is a *passive* system. The sensor is just sensing (recording amounts of reflected solar radiation); the sensor does not broadcast any signal itself. The sensor can be at another loca-


Figure 18.1: At left: magnetic compass for hiking or sailing with a 360-degrees-scale divided in 2-degrees portions. At right: 360-degrees disc divided in three rings, with each ring in total half of it being black and half white; in this example directions are encoded using (only) 3 bits; when a full turn of 360 degrees is divided by $2^3$ different codes, this yields a 45 degrees resolution for the measurement of direction.

tion than the object or the topography to be surveyed; optical imaging can be done *remotely* (optical remote sensing).

The most basic model of geometric projection for this imaging process holds for an infinitesimally small lens, and is also referred to as the pinhole camera model (and serves as a good approximation for the case with a finite sized lens, capturing distant objects (in principle at infinity)). The projection is shown in Figure 18.2.

A point $P$ in 3D reality is projected onto the 2D image point $p$, in the image plane, through the camera center point $C$, which is the (center of the) lens; point $C$ is the projection center. The camera is an image sensor. In an actual camera, a mirror image is formed *behind* the camera center (the so-called negative image), on photo-film (in the old days), or the image is captured by an array of pixels (Charge Coupled Device (CCD), or Complementary Metal-Oxide Semiconductor (CMOS)).

The corresponding positive image can be thought to stand in *front* of the camera (lens), as shown in Figure 18.2. In three dimensions, the points $C$ (camera center), $P$ (object in the terrain / topography), and $p$ (the image point, either in the negative, or positive image) all lie on one straight line in space; these three points are *collinear*.

The distance between the camera center $C$ and the image plane is the focal distance $f$. Assuming that the camera center $C$ is the origin of the coordinate system, as shown in Figure 18.2, and the image plane is $Z = f$ (the focal distance $f$ away from the lens), point $P$ with coordinates $(X_P, Y_P, Z_P)$ is mapped onto image point $p$ with coordinates $x_p = X_p = fX_P/Z_P$, $y_p = Y_p = fY_P/Z_P$, and $Z_p = f$, see Figure 18.3 (carefully mind the difference in indices between the small letter 'p' and the capital letter 'P'). Coordinates $x$ and $y$ are image coordinates (there is no $z$ coordinate); $X$, $Y$, and $Z$ represent terrain coordinates (sometimes also referred to a world coordinates).

From these simple equations it is clear that the ratio of the focal distance $f$ and the distance of the camera center to the terrain $Z_P$, determines the *scale* of the photo (for the middle of the photo), for instance, the larger $Z_P$ (e.g. the higher the aircraft is flying), the smaller objects get depicted in the photo. A small focal distance gives a wide field of view (wide angle), but objects get depicted at small size in the photo.

The line through the camera center $C$, perpendicular to the image plane is the *principal axis* of the camera (the $Z$-axis). The angle, at the camera center $C$, of the line $PC$ with the

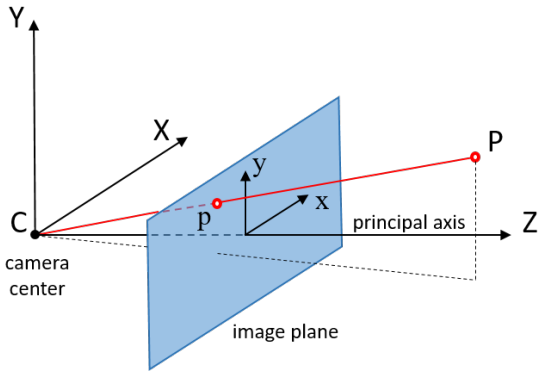



Figure 18.2: The camera projects a three-dimensional situation onto a two-dimensional image (central projection). Through the camera center (lens) $C$, the terrain point $P$ is mapped, or projected, onto the image point $p$ (in principle no matter the distance PC). For convenience we consider here the virtual image, i.e. the positive photo in front of the lens, rather than the actual (negative) image plane behind the lens.

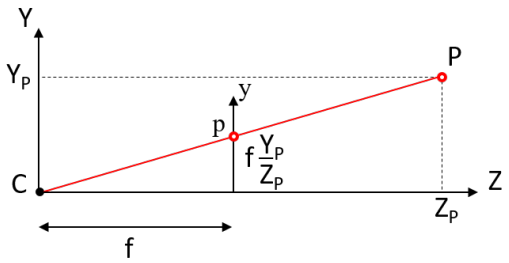



Figure 18.3: Cross-section of the imaging geometry. Real world point $P$ $(X_P, Y_P, Z_P)$ is mapped onto image point $p$ $(x_p, y_p)$. The ratio of the focal distance $f$ and terrain coordinate $Z$ determines the *scale* of the mapping. Point $P$ is assumed to lie in the $Y$,$Z$-plane.

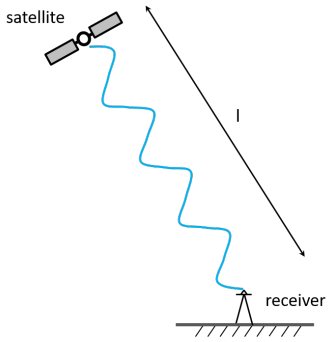



Figure 18.4: Positioning with the Global Positioning System (GPS) is an example of one-way ranging. Both satellite and receiver are equiped with a clock, and the receiver measures the delay of signal arrival, i.e. the difference between time of arrival at the receiver and time of transmission by the satellite. With perfect clocks, and in the absence of error sources, distance $l$ can be directly measured.

principal axis, determines where — at what distance from the origin in the image — the point is shown in the image. Hence, the *measurement* of *direction*, or *angle*, is actually the principle underlying optical imaging. All objects located in the same 'looking'-direction, no matter their distance to the lens, get depicted on the very same spot of the film, or the image (array).

The geometry of (optical) imaging is dealt with in the next chapter, Chapter 19 on photogrammetry, covering the central projection, the geometry of a single image, and also the geometry of an image-pair (stereo-photo). Optical remote sensing is further covered in Chapter 25.

## 18.2. Measurement of distance

In this section we introduce the principle of sensing from a geometric perspective, that is, observing distance by measuring signal travel time. Next, we cover lidar, radar, and sonar, and eventually touch upon imaging using these signals.

### 18.2.1. Sensing: using signals

Classically, a distance is measured by physically spanning a tape or chain along the distance to be measured, from point A to point B, or, for measuring water depth, sounding originally meant using a pre-measured heavy rope or cable, lowered over the ship's side until it touched the seafloor. Today, distances are usually obtained by measurements of *travel-time*, using an electromagnetic or acoustic signal (where electromagnetic signals also include optical signals). A signal is sent by a transmitter to a target (receiver), and basically the time needed to complete the travel from transmitter to receiver is measured, also referred to as the time of flight. Multiplication with the propagation speed of the signal yields the measurement of distance. Measuring distances is then basically a matter of *timing*. Chapter 20 covers the measurement of distance in further detail, distinguishing between measuring the travel-time using a pulse signal, and measuring distance by means of phase comparison (based on a continuous wave signal).

Distance measurements come in two flavours: one-way and two-way ranging. For one-way ranging *two* devices are needed: a transmitter and a receiver, and they need to have their clocks synchronized, in order to properly measure the signal travel-time. Satellite navigation is likely the most popular, and well-known example of one-way ranging, see Figure 18.4. GPS (Global Positioning System) satellites broadcast radio-signals, and a user on Earth (either on foot, in a car, on a vessel, etc.) receives the signal, and determines the travel-time, which can

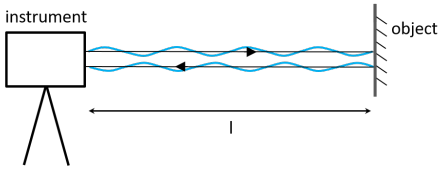



Figure 18.5: Distance measurement by two-way ranging. The signal is sent by the transmitter at left, reflected back by the object at right, and received again by the instrument at left. The signal round-trip travel-time is measured, which is directly proportional to twice the distance, $2l$.


Figure 18.6: Both a laser distometer (at left) and a total station (at right) use two-way ranging, generally by means of a laser. The laser distometer (shown here) uses red laser light with a 635 nm wavelength for ranging with millimeter precision, on distances up to 30 meters. The distance is measured optically, using a pulsed laser as an emitter, and a photo detector as a receiver (integrated in the same device; the two-way traveltime of the laser pulsse is measured). Pulses, with a duration ranging from a few to several tens of nanoseconds, are repeatedly transmitted, with a repetition rate in the order of kHz. The laser light of the distometer reflects directly on the object of interest. With a total station often a reflector is used, as shown at right.

be converted into a measurement of distance, see Part III. GPS is a multi-user system: each satellite transmits a signal down to Earth and the receivers only 'listen'. In principle, there can be an unlimited number of users.

With two-way ranging, the signal round-trip is timed, see Figure 18.5. The signal, after being reflected by the target, returns to the transmitter. The measurement of distance follows from division of the signal round-trip time by two, and multiplication by the travel speed. Typically, transmitter and receiver are integrated into a single device, also referred to as a transceiver. Two examples are shown in Figure 18.6.

Two-way ranging allows for *remote* measurements, this means, provided that the target object can reflect sufficient signal back to the receiver, the target object does not need to be accessed. The distance to the object can be measured from another location, e.g. from an aircraft flying over the terrain.

### 18.2.2. Lidar

The measurement of distance by a total station typically relies on a microwave or infrared signal, or laser, transmitted by the device, and reflected by a prism reflector (a glass corner cube prism), or the object under survey, cf. Figure 18.6. This functionality is referred to as Electronic Distance Measurement (EDM) equipment. The distance is generally measured as the travel-time of a laser pulse, or through phase comparison, for the latter also see Section 4.2.

A laser distometer is an obvious example of two-way ranging. It uses a laser signal (Light Amplification by Stimulated Emission of Radiation - LASER). A laser device emits light coher-

ently (a single frequency, i.e. a single color of light), and this allows the light to be focussed into a *narrow beam*. An example of a long-range laser (airborne, or spaceborne) system is laser altimetry (and the laser terrain profiler). The signal — typically a short pulse — is reflected by the sea- or Earth-surface. The signal travel-time is a measure for the distance.

A laser scanner is a device, set-up like a total station, which can take (two-way) distance measurements in many different directions, shortly after each other, by employing one or two moveable mirrors to steer the laser beam. For every object point in the vicinity of the laser scanner, distance and direction are measured together. The whole scene around the laser scanner can be surveyed in one go, and yields a 3D 'image' of the environment (a point cloud).

Laser ranging and scanning are applications of 'lidar', which stands for: light detection and ranging, similar to the acronym 'radar', covered next. A laser signal is sent, and the purpose is to detect (and measure) the echo from an object. Laser altimetry and scanning are further discussed in Chapter 22.

### 18.2.3. Radar

The acronym *radar* stands for RAdio Detection And Ranging. Radar, with originally a military background, was developed to detect the presence of objects and determine their range (position). A (microwave) radio signal (pulse), transmitted in a particular direction, is reflected by an object (also referred to as a back-scatterer), and the (two-way) travel-time of the signal is measured, which yields a full measurement of position (in 2D or 3D), actually similar to the laser scanner. With a rotating antenna, the full horizon can be sweeped and consequently 'mapped', as done in Air Traffic Control (then implying a measurement of direction as well). Additionally, the amplitude of the reflected signal (intensity of the echo) may contain information about the type of object, after its reflection properties. Radar is an *active* system: a signal needs to be transmitted, in order to 'illuminate' the object to be sensed.

Radio signals can be used, in a set-up very similar to laser altimetry. Also radar altimetry measures the distance between an aircraft or satellite and the surface of the sea or the Earth which reflects the signal.

As radar is based on radio signals, radar may not be able to 'look through' a metal mesh, like a fence, whereas laser will be able to do, and, similarly for a window, radio waves may be reflected, whereas laser (light) may propagate through. On the other hand, laser ranging and scanning can be seriously hindered by rain and fog, whereas radar may 'look through' these, virtually unlimited. Radar may, to some extent, depending on the medium, even propagate through objects. Basic propagation phenomena, like reflection and refraction, are reviewed in Appendix G.2.

Radio ranging is further discussed in Chapters 20 and 23.

### 18.2.4. Sonar

Next to optical and radio-signals, also acoustic signals can be used similarly: the acronym *sonar* stands for SOund NAvigation and Ranging. Sonar is applicable in air and water, and sub-surface as well, but not in space (vacuum), due to the absence of particles needed for the propagation of the pressure variations carrying the sound.

Chapter 24 is dedicated to acoustic sounding.

### 18.2.5. Imaging

So far, we discussed the measurement of distance between two discrete points. Though one should realize that using even a very narrow laser beam from an aircraft flying at 1 kilometer height, yields a *footprint* on the Earth's surface of a few decimeter already, and

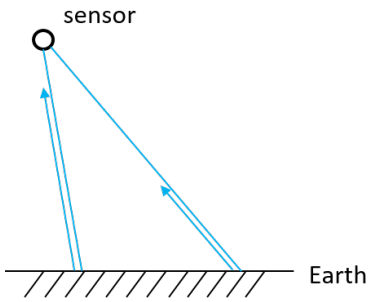



Figure 18.7: Imaging by side-looking radar. From an aircraft or satellite a radio pulse is transmitted sidewards down to the Earth's surface. The spatial separation in the terrain is preserved in the image by difference in time of signal return (indicated by the arrows); topography or objects close by yield early reflections, objects far off late ones. In this image both arrows represent an equal travel time; the reflection on the ground close by is nearly back at the sensor, whereas the reflection on the ground far off just started its return trip.

from a satellite this could be an area with a diameter of tens of meters, hence, not really a discrete point anymore (we return to this observation in Section 24.2). Radio signals, as for instance used with radar, are typically broadcast in a much wider beam (in principle a narrow beam is possibly, but at the complication of a huge antenna). Hence, with radio signals (and acoustic signals as well), one typically illuminates a big part of the scene at once (with a single signal transmission).

Instead of detecting objects in specific directions (possibly by a moving or rotating antenna), a fixed antenna can 'look' sideways and illuminate a whole area in one go, see Figure 18.7. Only one short pulse is transmitted (at a time), and reflections are received during a certain time interval: the *full response* is recorded, with all reflections/echoes. An object nearly underneath the sensor gives a short travel-time, and an object far away (sidewards) results in a longer travel-time. Objects at the same distance to the sensor get depicted on the very same spot in the 'image', no matter their angle ('looking' direction) within the signal beam (an example is given in Section 20.4). Radar imaging is further discussed in Chapters 20 and 23.

From an aircraft or satellite, images are obtained stripwise of ground areas located adjacent to the flight line, in forward direction. Together with side-looking radar this yields, like photography, a two-dimensional image of the Earth's surface and its topography.

### 18.2.6. Array of receivers: angle

Finally, we mention that instead of a single receiver, an *array* of receivers can be used (though, unlike with optical imaging, there is no lens involved). By using the time of arrival as measured by one of the receivers, the distance can be determined, as usual. But, by measuring the travel-time *difference* between multiple receivers, and knowing the lay-out of the receivers, one can also determine the *angle* of arrival of the signal, see Figure 18.8, for a simple two-dimensional representation; the transmitter Tx is assumed to be at large distance here.

## 18.3. Interferometry

Different distances, or *changes* in distance can be observed with an electromagnetic signal at a certain frequency (typically light), and this is the concept of the (Michelson) interferometer.

A light source, in principle coherent, for instance a laser, provides a carrier *wave* signal, that is split in two directions. One light beam goes directly to an optical sensor. The other beam goes to a (movable) reflector, at right, and is then returned to the sensor at bottom,

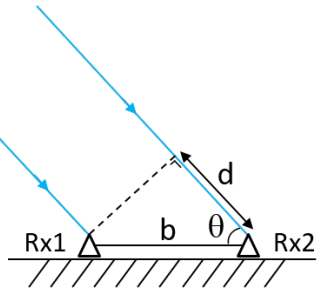



Figure 18.8: The angle of the incoming signal from signal transmitter Tx can be determined by measuring the travel-time *difference* between two receivers Rx1 and Rx2. The travel-time difference of the signal yields distance $d$, and when distance $b$ between the two receivers is known, the angle of arrival $\Theta$ follows from $\cos\Theta = \frac{d}{b}$.

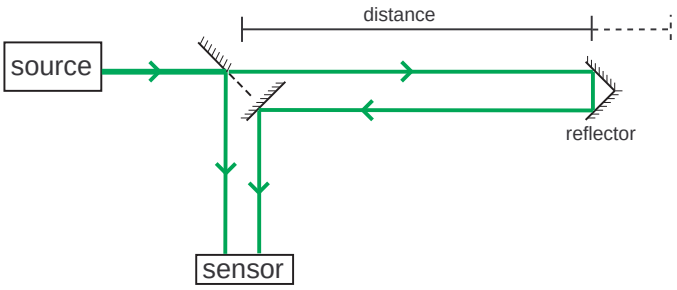



Figure 18.9: Diagram of a laser interferometer. Part of the light (in green) is reflected by the first mirror directly into the sensor; the other part takes a detour to the reflector at right.

see Figure 18.9. The sensor, a photo detector, makes a phase comparison between the direct signal and the signal that made the detour to the (distant) reflector.

The phase difference between the two waves is the result of the difference in path length (in this case the detour to the reflector and back). A path length difference of one wavelength (apart from an integer number of wavelengths) produces a phase difference of $360°$, which is equivalent to no phase difference at all (*constructive* interference, the resultant amplitude is the sum of the amplitudes of the two individual waves). A path length difference of one half wavelength produces a $180°$ phase difference (*destructive* interference; resultant amplitude is the difference of the two individual amplitudes, which is, when these are equal, just zero). Observing a maximum amplitude implies that the measured (two way) distance is longer than the direct signal's path, by a multiple of the wavelength.

A two-way distance can be measured, as indicated in Figure 18.9, but applicability in practice lies usually in measuring (very precisely) a *difference* in distance, by moving the reflector between two distinct positions (indicated by the dashed line). In the classical optics application, a pattern of fringes of light is observed when the reflector is moved (and the sensor counts the light/dark occurrences during the movement). The fringes are caused by constructive and destructive interference of the two waves arriving at the sensor. Such a pattern, similarly produced by the classical two-slit experiment is shown in Figure 18.10.

Changes in distance can be measured very precisely (in the order of 1 $\mu$m) by using directly the wavelength of (optical) light. Laser interferometers have a limited working range and are used in laboratories or with special construction works. The Väisälä interference comparator, with white light, was, and is used to calibrate invar wires for instance for the precise determination of the lengths of so called standard baselines in terrestrial geodetic

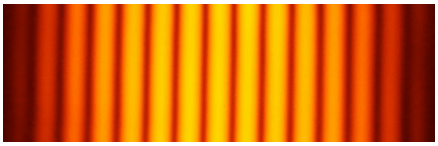


Figure 18.10: Interference pattern from classical two-slit experiment in physics, taken from image by Pieter Kuiper - own work, Public Domain, Wikimedia Commons [9].

networks.

Rather than using light, the principle of interferometry is also applicable with radar (with centimeter - decimeter wavelengths). This allows to measure small changes on the Earth's surface, in the order of millimeters, by means of radar remote sensing, through repeated pass-overs of a satellite. Section 23.2 covers radar interferometry with application in measuring deformations.

This principle of interferometry is also applied in measuring changes of length in fiber optic cables, for instance in constructions expected to be susceptible to deformation. The speed of light in optical fiber is about $2 \cdot 10^8$ m/s, that is roughly 2/3 of the speed of light in vacuum, see Chapter 21.

## 18.4. Exercises and worked examples

Below follow two exercises on measuring distances by means of laser light.

**Question 1** How short does a laser light pulse need to be, in order to be able to observe two objects fully separated, which are located within the same footprint, directly underneath the sensor, and which have a height-difference of 15 m.

**Answer 1** The word 'separately' implies that the reflection received from the first object shall not overlap with the reception of the reflection from the second object. A 15 m height difference translates into a 30 m travel-time difference (two-way travel-time). The laser pulse is travelling at the speed of light (in vacuum), about $3 \cdot 10^8$ m/s, and hence, 30 m, divided by the speed of light yields $10^{-7}$ s, or 0.1 $\mu$s. A pulse of 1 $\mu$s time duration is — through the speed of light (in vacuum) — equivalent to a distance of 300 m in range.

**Question 2** An interferometer is operating at a wavelength of 500 nm. Changes in distance can be observed with a precision of $\sigma$ = 1 $\mu$m. How precise can the fringe be observed? That is, the precision of 1 $\mu$m is equivalent to what part of a fringe (cycle), or how many fringes (cycles)?

**Answer 2** The difference in distance has a precision of 1 $\mu$m, or 1000 nm. Therefore the actual change in distance (observed by measuring/counting the fringes) is 2000 nm (as the light signal makes a two-way trip between instrument and mirror). One full fringe (light-dark cycle) corresponds to a change in the optical path length of one wavelength, hence 500 nm. The measurement precision in terms of fringes is 2000/500=4; the fringe-counting is precise to a standard deviation of 4 fringes.

# 19

# Photogrammetry

The discipline of extracting *metric* information from *photographs* is called photogrammetry. Photogrammetry is the art, science and technology of obtaining information about physical objects and the environment, through the process of recording, measuring and interpreting photographic images, today, mostly digital images. The traditional, and largest, application of photogrammetry is to extract topographic information (e.g. maps and terrain models) from aerial images. Figure 19.1 shows an example of an aerial photograph. Actually a so-called ortho-photo is shown here, which is a geometrically corrected photo (using a ground surface model) as to view the terrain everywhere from directly overhead. In photography a sensing array is used to capture visible light, covering wavelengths of 400 nm (blue) to 700 nm (red).

Figure 19.1: Example of an aerial photograph of a part of the TU Delft campus, taken April 26th 2021, with a Vexcel Ultracam Eagle Mark 3 camera (450 Megapixels), with a focal-distance of $f$=210 mm, and a flying height of slightly over 4 km. Shown is an orthophoto with a ground pixel resolution of 8 cm. Aerial photographs are jointly acquired in the Netherlands, country-wide, by a cooperation of Dutch governmental bodies, led by het Waterschapshuis and het Kadaster, for instance for mapping purposes through stereo-photogrammetry (for example the Basisregistratie Grootschalige Topografie (BGT)). Photo obtained through Beeldmateriaal Nederland [49] under a CC BY 4.0 license. From 2021 on, these aerial photos are available as open data.

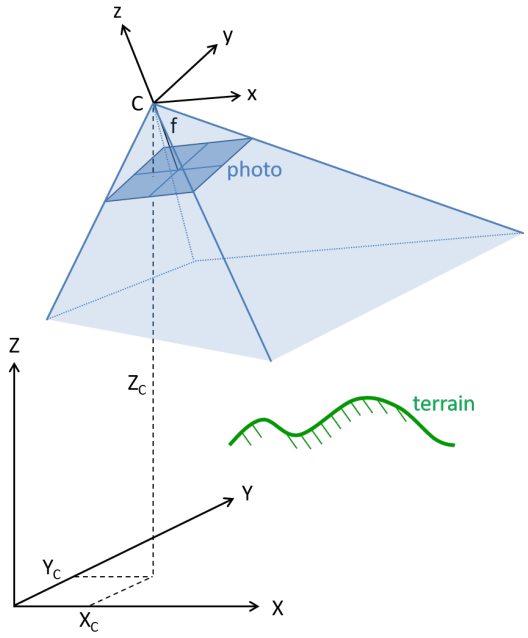



Figure 19.2: Optical imaging: Cartesian camera coordinates $x$, $y$, $z$, and Cartesian ground or terrain coordinates $X$, $Y$, $Z$. The photo-image plane is parallel to the $x$-$y$ plane, and the (positive) image plane is at $z = -f$ (focal distance). The blue shaded volume shows the field of view of the camera.

In this chapter, image geometry is related to actual terrain geometry, through the central projection. Then we consider the fact that (tall) objects and relief (height differences, like dikes and hills) look like laying backward in the image, a phenomenon called relief displacement. Next, we touch upon stereo-photogrammetry, that is, the reconstruction of three-dimensional terrain geometry from a *pair* of photos. A first step with the underlying mathematics of photogrammetry is taken only in the optional Section 19.4.

## 19.1. Central projection

In Chapter 18 we assumed, for simplicity, that the camera center (point C) was the origin, not only of the camera system, but also of the terrain coordinate system, cf. Figure 18.2. In practice the camera center *position* - at the time of capturing the image - is not known, in the terrestrial coordinate system, think for instance of a camera in an aircraft flying over the terrain. And also the *orientation* of the camera (or, attitude) with respect to the terrain is not known. In Chapter 18 the orientation of the camera, again for simplicity, was taken such that the image plane is aligned with the $X$-$Y$-plane of the ground coordinate system.

In practice a relation needs to be established between the terrain coordinates ($X$, $Y$, $Z$), and the image (or camera) coordinates ($x$, $y$, $z$), see Figure 19.2. This is referred to as the *exterior* orientation of the photo. The image coordinate system is fixed to the camera, and the (central) projection center (center of the lens) is the origin. In Figure 19.2 the $z$-axis lies along the principal axis of the camera.

One can *measure* position and orientation of the camera (in the ground coordinate system) during flight, and, establish in the terrain so-called ground control points, which can be well identified in the image, and have known coordinates in the desired terrestrial coordinate system, so that one can *reconstruct* the position and orientation of the camera (in the ground coordinate system). The process of linking image coordinates to position coordinates of ob-

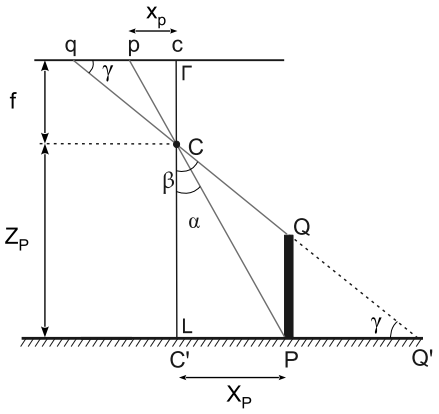



Figure 19.3: Single image: point P is the bottom of the object in the terrain, and mapped to point p in the photo. The top of the object, point Q, is mapped onto point q in the image. Looking from the center of the photo, one can see that the object, for instance a tall tower, leans back in the image (radially outwards from the image center). In the image it looks like the tower has been pushed over, and is now lying backwards on the ground. The tower, which stands vertically upright in the terrain, shows by a certain length in the image and this is referred to as relief displacement. By measuring the displacement, distance $x_q - x_p$ in the photo, one can determine the height $Z_P - Z_Q$ of the object P-Q in, or above the terrain.

jects in the terrain is often called *georeferencing*. It is about 'retrieving where the image was taken'.

The camera coordinate system can be linked to a coordinate system adopted on the ground, in use for surveying and mapping, generally through a three dimensional similarity transformation, that is, using one scale parameter, three translation parameters, and three rotation parameters (see Chapter 28). This is further detailed in the optional Section 19.4.

By using a calibrated camera, one assures to have a known focal distance $f$, the image plane perpendicular to the principal axis of the camera, and the middle of the image (photo) on the principal axis. This concerns the *interior* orientation of the photo. We assume also that we measure positions of objects depicted in the photo directly in the photo/camera coordinate systems $(x, y, z)$. The interior orientation of the photo is the step to correct for lens distortion and distortions in the image array (if any).

Then one will be able to distil, from the image, useful geometric information about the topography.

In the following two sections we take, for pedagogical purposes, again a simple approach, using basically the same coordinate system for both terrain and camera, and only in the optional Section 19.4 we cover the exterior orientation of the camera.

## 19.2. Relief displacement

In this section we consider the retrieval of metric information from a *single image*. We do so again by means of a simple example, in which the camera position and orientation are known. By taking a measurement in the photo-image, specifically the length of an object as it shows in the image, we determine the actual height of the object in the terrain.

The set-up of the example is shown in Figure 19.3. For convenience we show the (negative) image in the camera, rather than the positive image (photo). We use a coordinate system similar to the one in Chapter 18, Figure 18.2. And we consider the vertical $X$-$Z$ cross-section of the situation ($Y = 0$). The origin is the position of the camera center C. The positive $Z$-axis is pointing downwards. The image plane is, for simplicity, nicely aligned with the terrain.

Using the central projection for point P, we consider triangle CC'P, and its image equivalent

Ccp, both with angle $\alpha$. And we consider point Q, with triangle CC'Q', and its image equivalent Ccq, both with angle $\beta$. We have

$$\frac{X_P}{Z_P} = \frac{x_p}{f} \text{ and } \frac{X_{Q'}}{Z_P} = \frac{x_q}{f}$$

which yields

$$X_{Q'} - X_P = \frac{Z_P}{f}(x_q - x_p)$$

where $\frac{f}{Z_P}$ is the photo scale.

Next, we consider triangle QPQ' and triangle Ccq (which has the same shape), and with $\tan\gamma$ we obtain

$$\frac{Z_P - Z_Q}{X_{Q'} - X_P} = \frac{f}{x_q}$$

and combining the two results leads to

$$Z_P - Z_Q = \frac{Z_P}{x_q}(x_q - x_p) \tag{19.1}$$

which says that, knowing the flying height $Z_P$ and measuring the positions of the bottom $x_p$ and top $x_q$ of the tower in the image, one can determine the height of the object in the terrain $Z_P - Z_Q$ (i.e. the difference in vertical coordinates of points P and Q).

In the next section we cover the subject of three-dimensional object reconstruction using *two* images. With (19.1) it seems that we can reconstruct three-dimensional information from just a single image. Mind however that with Figure 19.3, we made some simplifications; we assumed that the camera was pointing straight down, onto a flat, level surface, with tower PQ standing perfectly upright.

With aerial photogrammetry one generally takes photos straight down, as well as possible. In computer vision one uses the *perspective* view, as a result of the central projection, to extract geometric information, exploiting the fact that parallel lines in reality will meet at infinity, in the so-called vanishing point, in the photo. This subject is however beyond the scope of this chapter.

## 19.3. Differential parallax

By taking at least *two images* of an object (or a piece of the Earth's surface), and reconstructing the position and orientation of the camera at times of photo capture, a three-dimensional model of the object can be obtained.

The underlying geometric principle basically is an intersection with *angles*, as shown in Figure 19.4 at left, as a two-dimensional vertical cross-section of an aircraft taking aerial photographs of the terrain. When *angles* $\alpha_1$ and $\alpha_2$ are obtained (reconstructed) from the photo-images (for simplicity the camera is assumed to look straight down here), and the camera positions are given, the position of point P in the terrain can be computed. Similarly the intersection with two measurements of azimuth is covered in Section 9.7.2.

The two photo fragments in Figure 19.4 illustrate relief displacement and differential parallax. These illustrations are details of two of the aerial photos shown in Figure 19.9.

A simple example of extracting 3D geometric information from a *pair* of partly overlapping photos is shown in Figure 19.5 (again we show the vertical X-Z cross-section of the situation).

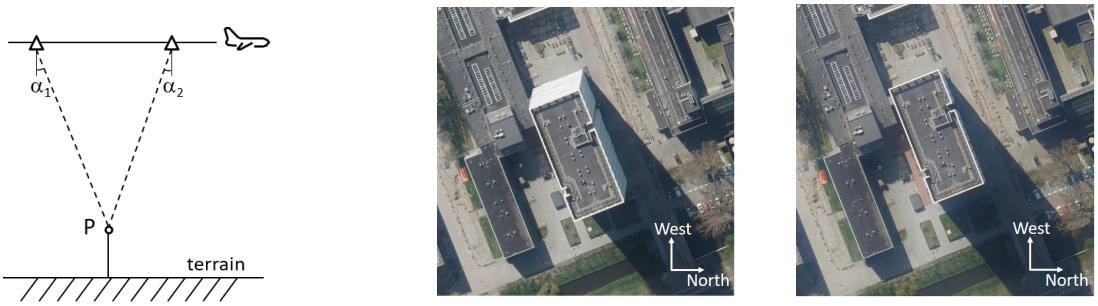



Figure 19.4: At left the principle of measuring angles $\alpha_1$ and $\alpha_2$, in order to determine the position of point P (forward intersection). *Relief displacement* is clearly visible in the photo in the middle, as one can see the West-facade of the white building; the building is 'laying back' in South-East direction as the camera position, when capturing this photo, is 450 m to the North-West of this building. Relief displacement takes place radially outward from the image center. The photo at right shows virtually no relief displacement of this building, as the camera was nearly straight overhead when taking this photo. The different perspective of the same object in these two photos illustrates the *differential parallax*. Photos obtained through Beeldmateriaal Nederland [49] under a CC BY 4.0 license.

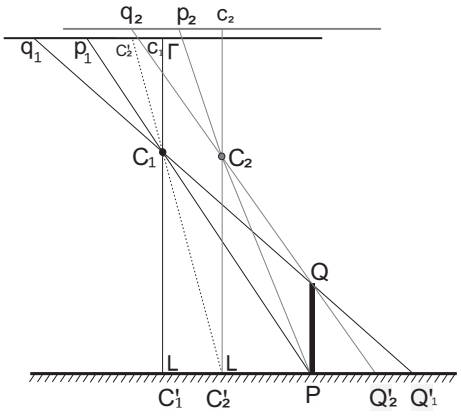



Figure 19.5: Two images: the same object PQ is pictured in two adjacent images, $p_1q_1$ in image 1, and $p_2q_2$ in image 2. Due to a different camera position with respect to the object, the displacement of the tower PQ in the two images is different. By measuring the differential parallax, and knowing the flying height $Z_P$, and the change in position of the camera between the two images $X_{C_2} - X_{C_1}$, one can determine the height of the object $Z_P - Z_Q$.

Objects show up at different locations in the images, caused by a change in camera position; this is *parallax*.

For the photo at left (with projection centre $\mathrm{C}_1$) we can use the relief displacement Eq. (19.1)

$$Z_P - Z_Q = \frac{Z_P}{x_{q_1}}(x_{q_1} - x_{p_1}) = \frac{Z_P(x_{q_1} - x_{p_1})}{x_{p_1} + (x_{q_1} - x_{p_1})}$$

Now, we note that the distance from $\mathrm{C}'_2$ to P, plus the distance from $\mathrm{C}'_1$ to $\mathrm{C}'_2$ together, equals the distance from $\mathrm{C}'_1$ to P in the terrain. And this relation also holds in the photo (just scaling these distances by $\frac{f}{Z_P}$), hence $x_{p_2} + x_{c'_2} = x_{p_1}$, where $x_{p_2}$ is measured in the right photo, and $x_{p_1}$ and $x_{c'_2}$ in the left image, so that the above equation becomes

$$Z_P - Z_Q = \frac{Z_P(x_{q_1} - x_{p_1})}{x_{p_2} + x_{c'_2} + (x_{q_1} - x_{p_1})}$$

And also for the photo at right (with projection centre $\mathrm{C}_2$) we can use the relief displacement Eq. (19.1).

$$Z_P - Z_Q = \frac{Z_P}{x_{q_2}}(x_{q_2} - x_{p_2}) = \frac{Z_P(x_{q_2} - x_{p_2})}{x_{p_2} + (x_{q_2} - x_{p_2})}$$

Setting the two equations (one for the photo left, and one for the photo right) equal

$$\frac{(x_{q_1} - x_{p_1})}{x_{p_2} + x_{c'_2} + (x_{q_1} - x_{p_1})} = \frac{(x_{q_2} - x_{p_2})}{x_{p_2} + (x_{q_2} - x_{p_2})}$$

and solving for $x_{p_2}$ yields

$$x_{p_2} = \frac{x_{c'_2}(x_{q_2} - x_{p_2})}{(x_{q_1} - x_{p_1}) - (x_{q_2} - x_{p_2})}$$

Substituting this in the equation for the relief displacement in the photo at right, we obtain, after some manipulation,

$$Z_P - Z_Q = \frac{Z_P((x_{q_1} - x_{p_1}) - (x_{q_2} - x_{p_2}))}{x_{c'_2} + ((x_{q_1} - x_{p_1}) - (x_{q_2} - x_{p_2}))} \tag{19.2}$$

The term $(x_{q_1} - x_{p_1}) - (x_{q_2} - x_{p_2})$ which appears in both the numerator and the denominator, is the relief displacement in the left image, minus the relief displacement in the right image. And, it equals $(x_{p_2} - x_{p_1}) - (x_{q_2} - x_{q_1})$, which is the *difference* of the parallax, due to moving the camera, between the two points P and Q. In fact we are using here a stereo-photo, from which three-dimensional geometry can be reconstructed, much similar to the way humans and animals perceive three-dimensional geometry. In this way three-dimensional terrain models can be created.

The distance $x_{c'_2}$, also appearing in (19.2), can be measured in the photo. Point $\mathrm{c}_2$ is the middle of the right photo. And in the left photo, the corresponding point can be found, by identifying the same terrain point, indicated by the dotted line, in this photo — this is point $\mathrm{c}'_2$; and the distance from this point to the middle of the left photo can be measured, being $x_{c'_2}$.

Mind that if one image is taken from the left side of the object, and the other image from the right side, the differential parallax $(x_{q_1} - x_{p_1}) - (x_{q_2} - x_{p_2})$ equals the sum of the *magnitudes* of the displacements, as the two displacements are on opposite parts of the $x$-coordinate axis.

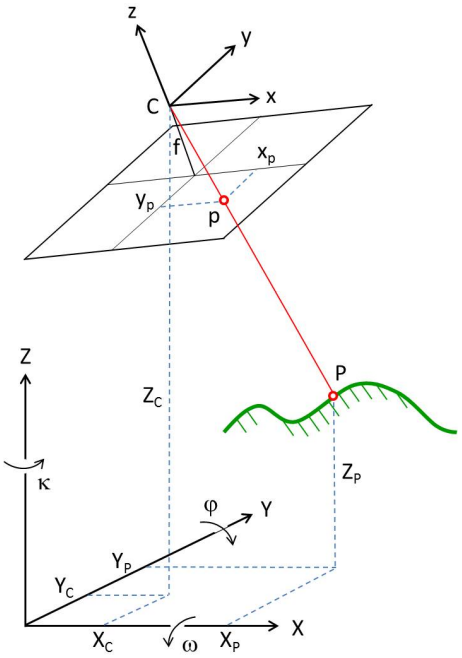



Figure 19.6: The camera coordinate system $(x, y, z)$ and the terrain coordinate system $(X, Y, Z)$ are related through a three-dimensional similarity transformation, involving a translation over vector $(X_C, Y_C, Z_C)$, and three rotations, angles $\omega$, $\varphi$ and $\kappa$. Next a scaling $\lambda$ applies, to depict the terrain scene into the image. Terrain point $P$ with coordinates $(X_P, Y_P, Z_P)$ gets depicted at coordinates $(x_p, y_p)$ in the image, which is at focal distance $f$ from the lens ($z_p = -f$); the positive photo is shown here.

## 19.4. Terrain - camera transformation [*]

In the previous two sections we extracted geometric information about the terrain from a single image, and a pair of images, respectively. We did so, using a simplified example. The general theory for relating geometric information about the terrain to image geometry, is covered in this section; this concerns the so-called exterior orientation of the camera. For an in-depth coverage of the subject, the reader is referred to e.g. [50].

Using the three-dimensional similarity transformation from Section 28.3, one can transform the (source) terrain coordinates $(X_P, Y_P, Z_P)$, cf. Figure 19.6, of point $P$, into the corresponding (target) camera coordinates $(x_p, y_p, z_p)$:

$$\begin{pmatrix} x_p \\ y_p \\ z_p \end{pmatrix} = \lambda \underbrace{R_1(\omega)R_2(\varphi)R_3(\kappa)}_{R} \left( \begin{pmatrix} X_P \\ Y_P \\ Z_P \end{pmatrix} - \begin{pmatrix} X_C \\ Y_C \\ Z_C \end{pmatrix} \right)$$

where we took $\Omega_x = \omega$, $\Omega_y = \varphi$ and $\Omega_z = \kappa$. Mind that $(X_C, Y_C, Z_C)$ refers to the origin of the target coordinate system (camera), expressed in the source coordinate system (terrain) — this is translation vector $t$.

With the rotation angles defined as in Figure 19.6, the 3-by-3 rotation matrix $R$ (cf. (28.9) and (28.11)) becomes

$$R = \begin{pmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{pmatrix}$$

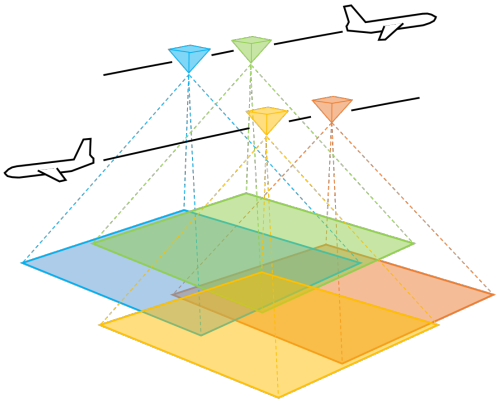


Figure 19.7: Aerial stereo-photogrammetry: adjacent images within one strip typically have a 60% overlap. Between strips this is typically 20%-30%.

$$R = \begin{pmatrix} \cos\kappa\cos\varphi & \sin\kappa\cos\varphi & -\sin\varphi \\ \cos\kappa\sin\varphi\sin\omega - \sin\kappa\cos\omega & \sin\kappa\sin\varphi\sin\omega + \cos\kappa\cos\omega & \cos\varphi\sin\omega \\ \cos\kappa\sin\varphi\cos\omega + \sin\kappa\sin\omega & \sin\kappa\sin\varphi\cos\omega - \cos\kappa\sin\omega & \cos\varphi\cos\omega \end{pmatrix}$$

Through the above equation, and disregarding the scale factor $\lambda$, we relate the coordinates of the terrain point $P$ in the terrain coordinate system, to the coordinates in the camera coordinate system.

The last step is to *scale* from the point in the terrain $P$ to the corresponding point in the image $p$ (in the camera system). This is achieved by demanding that the resulting $z_p$ coordinate, equals the focal distance: $z_p = -f$, cf. Figure 19.6. This — implicitly — determines the value for scale factor $\lambda$.

However, as in the image, we are interested only in two-dimensional coordinates, we eliminate the scale factor $\lambda$ by dividing the expressions for $x_p$ and $y_p$ by the one for $z_p$, and eventually bringing $z_p = -f$ to the right hand side. The result reads:

$$x_p = -f\frac{R_{11}(X_P - X_C) + R_{12}(Y_P - Y_C) + R_{13}(Z_P - Z_C)}{R_{31}(X_P - X_C) + R_{32}(Y_P - Y_C) + R_{33}(Z_P - Z_C)}$$

$$y_p = -f\frac{R_{21}(X_P - X_C) + R_{22}(Y_P - Y_C) + R_{23}(Z_P - Z_C)}{R_{31}(X_P - X_C) + R_{32}(Y_P - Y_C) + R_{33}(Z_P - Z_C)}$$

which are the common equations for the central projection in photogrammetry. They relate the terrain coordinates of object P to the coordinates of this object in the image. They are also known as the collinearity equations, as discussed with Figure 18.2.

## 19.5. Aerial stereo-photogrammetry

The common, and most frequently used way of acquiring large scale topographic maps and three-dimensional Digital Surface Models (DSM) is through aerial stereo-photogrammetry. The three-dimensional geometry of objects and topography on the Earth's surface can be reconstructed through *stereo*-photogrammetry as shown in Figure 19.5; (at least) two photos of the same scene as needed. Or, more precisely: any object to be mapped needs to appear in at least two photos.

An important step in processing the images for this purpose is feature matching, i.e. finding the same object (point) in two, or more adjacent images.


Figure 19.8: Example of flight trajectory to map the TU Delft area through stereo-photogrammetry, using the principle of Figure 19.7. A 7 km by 7 km area is shown. The area is flown strip-wise, and the parallel tracks are about 1.2 km apart (turns of the aircraft are not shown realistically). With a flying-height of over 4 km, the footprint of one photo on ground is typically about 1.3 km x 2 km. The background map is taken from OpenStreetMap (OSM), ©ÒpenStreetMap contributors [51], cf. Appendix J.

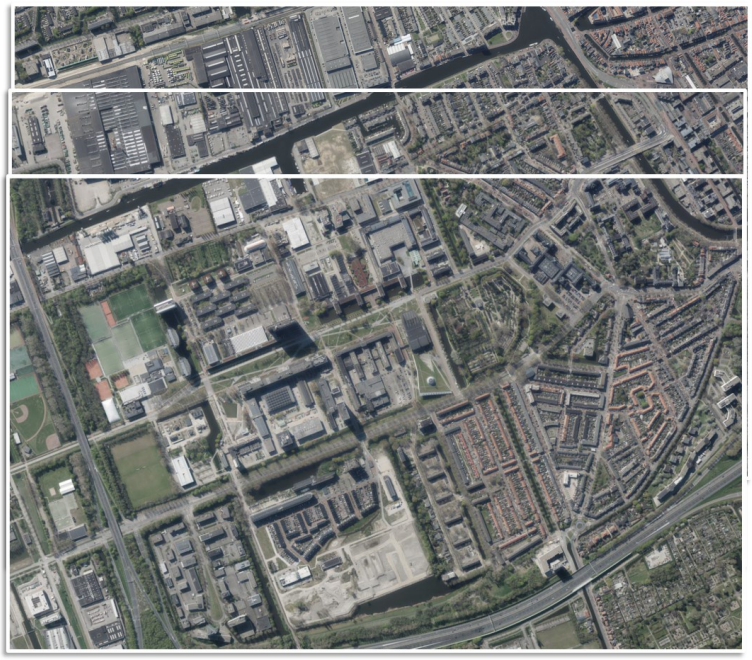


Figure 19.9: Example of three overlapping photos in one strip, for the purpose of mapping the TU Delft campus area through stereo-photogrammetry. The photos have been taken with a Vexcel Ultracam Eagle Mark 3 camera (450 Megapixels), with a focal-distance of $f$=210 mm, at a flying height of slightly over 4 km. Photos obtained through Beeldmateriaal Nederland [49] under a CC BY 4.0 license.

In order to survey an area through aerial stereo-photogrammetry, images are taken with typically a 60% overlap along the flight direction, as shown in Figure 19.7. In this way there is still 20% overlap between photo $i$ and photo $i+2$.

The area is then flown with parallel flight strips, as shown in Figure 19.8, so that the sideward overlap is about 20%-30%. Such flights are typically carried out in early Spring as to avoid objects to be mapped being occluded by vegetation (no leaves on trees yet).

As an example three overlapping photos are shown in Figure 19.9, of the TU Delft campus, acquired in Spring 2021. The aircraft was flying over Delft forth and back, in West-East direction, as shown in Figure 19.8. The photos, as shown here in one strip, are taken about 250 m apart. Note that the photos show West up. The A13 Rotterdam - The Hague highway can be seen at bottom.

## 19.6. Exercises and worked examples

This section contains two exercises on photogrammetry.

**Question 1** An aircraf equipped with a photo-camera is flying — at a 500 m height — over the Earth's surface. Directly underneath the camera, we would like to see/identify single 30x30 cm tiles in the pavement (assuming ideal imaging conditions). What should be the pixel-size in the CCD or CMOS array in the camera-plane, in order to meet this demand (one pixel covering one tile)? The camera has a focal distance of 5 cm.

**Answer 1** Using the equation for the central projection, for instance in the $Y$-coordinate direction, we have $y_p = \frac{fY_P}{Z_P}$, with $f$=0.05 m, and $Z_P$=500 m. The size of the tile (in reality) is $Y_P$=0.30 m (*spatial resolution*), and hence $y_p = \frac{0.05 \cdot 0.3}{500} = 3 \cdot 10^{-5}$ m, or 30 $\mu$m.

Figure 19.10: Stereo-photogrammetry at home: the same scene is pictured twice, with a smartphone camera (Samsung J5 2018 - SM-J530F). The camera was held (approximately) level, at a 'flying height' of 60 cm.

**Question 2** Figure 19.10 shows two photos, taken in a Do-It-Yourself photogrammetry set-up at home, according to Figure 19.5. The camera was held approximately level, at a 'flying height' of 60 cm. Due to different camera positions the relief displacement of both boxes is clearly different in these two images. The image at left clearly shows much more of the front-sides of the boxes, than the image at right. Determine, by measurements of relief displacement in these two photos, the height of both boxes.

**Answer 2** The solution to this problem lies in Eq. (19.2). The flying height is given $Z_P$ = 60 cm, and the camera-displacement between the two photos, in terms of photo-coordinates, was measured to be 540 pixels ($x_{c_2'}$). The relief displacement has been measured, and for the Bonzo-box they are: $x_{q_1} - x_{p_1}$ = 756 and $x_{q_2} - x_{p_2}$ = 348, in the left and right photo

respectively, and for the Heinz-box they are: $x_{q_1} - x_{p_1}$ = 322 and $x_{q_2} - x_{p_2}$ = 159. Then (19.2) results in 25.8 cm for the Bonzo-box, and in 13.9 cm for the Heinz-box. As a verification, direct measurement with a ruler gave 26.2 cm for the Bonzo-box, and 13.6 cm for the Heinz-box. The above measurements have been done in the original photos, in terms of pixels (using a rudimentary image-viewer such as Paint). Doing these measurements in a scaled version of the photos, for instance measuring relief displacement with a ruler in the photos of Figure 19.10 does not impact the outcome, as a scaling factor in $x_{p_1}$, $x_{q_1}$, $x_{p_2}$, $x_{q_2}$ and $x_{c2'}$ actually cancels in the ratio of (19.2). The physical pixel-size of the camera is about 1 $\mu$m, and the total image is about 4000 x 3000 pixels, meaning the the physical image array of the camera in the smartphone measures about 4 x 3 mm. From this, and the measurement tape depicted in the photo, one can reconstruct that the focal length of the camera is about $f = 3.5$ mm.

# 20

# Sensing by measurement of distance

In this chapter we cover the principles of ranging, first from point-to-point, and secondly for imaging. This second section in particular, provides the basis for remote sensing by ranging. The last section provides a brief introduction to the two major aspects of the design of a (spaceborne) remote sensing mission.

## 20.1. Principles of ranging

The purpose of ranging is to measure distances. In this context, ranging is performed using either an acoustic signal, or an electromagnetic signal (including optical such as laser light). Figure 18.5 showed the concept of two-way ranging. The signal travels forth from transmitter to reflector, and then back to the transmitter/receiver, and a measurement of the two-way travel-time $\tau$ is obtained (the total time delay, also referred to as the time of flight). The distance $l$ between the radar transmitter and the reflecting surface/object is obtained through

$$l = \frac{c\tau}{2} \tag{20.1}$$

where $c$ is the speed of light in vacuum (299792458 m/s), and we assume that the signal is traveling actually with this speed of light in vacuum. We refer to Appendix G for a discussion on signal propagation, for instance in the Earth's atmosphere.

In Chapter 18, it was pointed out that we distinguish between *one-way* and *two-way* ranging. In this chapter we start by outlining the two basic principles of ranging, and in the context of radar remote sensing, the discussion is held here in terms of *two-way ranging* (but, very similarly, applies to one-way ranging as well).

As stated in Chapter 18 as well, one can distinguish between two main principles of measuring travel-time: pulse-based ranging, and Continuous Wave (CW) ranging (phase based). In the sequel we will cover both.

### 20.1.1. Pulse based ranging

Figure 20.1 shows the principle of measuring a two-way travel-time by using a pulse signal. For a pulse-based system, the signal travel-time is directly observed through (20.1).

The range *resolution* is directly related to this:

$$\Delta l = \frac{c\Delta\tau}{2} \tag{20.2}$$

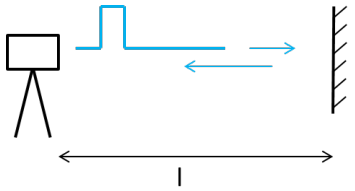



Figure 20.1: Pulse-based ranging: the travel-time of the pulse is measured. The pulse travels from the transmitter to the reflector, and back.

where $\Delta\tau$ is the sampling interval of the clock within the instrument. The sampling interval is typically tuned to the duration of the pulse, and the latter determines the bandwidth of the ranging signal occupied in the frequency domain (a narrow pulse, in time, corresponds to a wide band, in frequency, so it occupies a large part of the spectrum). The resolution determines whether two targets or objects, which are close together, can still be 'seen' separately by the radar or ranging instrument.

The *maximum range* of the radar is dependent on the pulse rate (or conversely, the time duration between two pulses). To avoid mis-interpretation (and without any prior knowledge), it is typically required that no new pulse is transmitted, before the reflection of the earlier pulse is received. Hence,

$$l_{\max} = \frac{c\tau_{\max}}{2} \tag{20.3}$$

where $\tau_{\max}$ is the pulse repetition time interval (or, the time duration between two consecutive pulses). In practise, the pulse repetition time may not be the only, or most important, limiting factor. For example, also the power of the instrument, antenna-pattern, atmospheric conditions, target reflectivity, and receiver/detector sensivitity, will determine the maximum range of operation.

### 20.1.2. Phase based Continuous Wave (CW) ranging

In case of ranging using a continuous wave, the distance measurement is based on the phase difference $\Phi$ (in radians, with $\Phi \in [0, 2\pi\rangle$) between the transmitted and received signal (phase comparison, similar to Section 4.2). Alternatively, the phase difference can be expressed in cycles as $\frac{\Phi}{2\pi}$ (then being dimensionless). The travel-time of the signal $\tau$ is obtained by

$$\tau = \frac{\Phi}{2\pi}T + kT \tag{20.4}$$

Here, $T$ is the period of the wave in seconds, and $k$ is the number of full wavelengths (also referred to as ambiguity, $k \in \mathbb{N}$). The phase difference between transmitted and received signal is a measure for the travel-time, and hence distance, as the received signal is delayed, because it made a trip to the reflector and back.

Using (20.1), and assuming $k = 0$ (or knowing the cycle ambiguity in (20.4)), the measured range $l$ (expressed in one-way) becomes

$$l = \frac{c}{f}\frac{\Phi}{4\pi} \tag{20.5}$$

where $f = 1/T$ is the frequency of the ranging signal carrier wave ([Hz]), and $c = \lambda f$ (in vacuum).

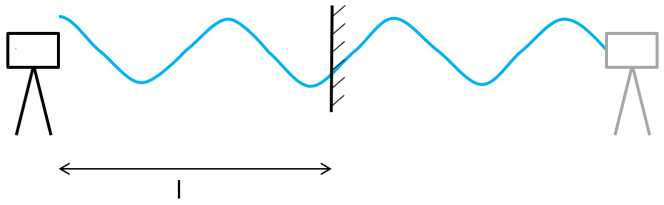



Figure 20.2: Phase-based ranging: the sine carrier wave travels from the transmitter to the reflector, and back, and the way back has, for convenience, been unfolded to the right in this diagram. With $c = \lambda f$ Eq. (20.8) can be turned into $l = \lambda\frac{\Phi}{4\pi} + \lambda\frac{k}{2}$. The fractional phase difference $\Phi = \frac{\pi}{2} - 0 = \frac{\pi}{2}$ (in radian, with $\Phi \in [0, 2\pi\rangle$ as it is the *fractional* phase difference, or $\frac{\frac{\pi}{2}}{2\pi} = \frac{1}{4}$ cycle), and $k = 3$ in this example. Thereby $l = 1\frac{5}{8}\lambda$.

Analogous to the pulse-based ranging, the range *resolution* is

$$\Delta l = \frac{c}{f}\frac{\Delta\Phi}{4\pi} \tag{20.6}$$

Compared to the range resolution of a pulse-based system, the CW-frequency $f$ determines the ranging sensitivity. With increasing $f$, a better range resolution (smaller $\Delta l$) is obtained (as a larger frequency, translates into a shorter wavelength). Generally the phase difference can be determined up to (the order of) 10 milli-radians. Hence, in principle a high frequency is desirable.

However, the *maximum* (unambiguous) *range* for $\Phi_{\max} = 2\pi$ [rad] (a full cycle) is

$$l_{\max} = \frac{c}{f}\frac{\Phi_{\max}}{4\pi} = \frac{\lambda\Phi_{\max}}{4\pi} = \frac{\lambda}{2} \tag{20.7}$$

This relation shows that the maximum range is directly related to the wavelength $\lambda$ of the signal. Hence, a long wavelength (low frequency) is desirable to maximize the range. Beyond this range, one has to deal with an ambiguity. When the cycle ambiguity $k$ (an integer number) in (20.4) is not known, we get, instead of (20.5):

$$l = \frac{c}{f}\frac{\Phi}{4\pi} + \frac{c}{f}\frac{k}{2} \tag{20.8}$$

and the resulting distance, expressed as a one-way distance, has an ambiguity $k\frac{\lambda}{2}$.

Figure 20.2 shows the principle of measuring a two-way travel-time through measuring the phase of a continuous carrier wave.

To overcome the contradicting requirements regarding range resolution and maximum range, often a multi-frequency system is used. The highest frequency signal determines the resolution, whereas the lowest frequency gives the maximum range, see also Section 4.2, which describes the operating principle of an EDM in a total station.

## 20.2. Range-rate: Doppler [*]

The frequency of a signal, measured or observed by a receiver, may differ from the frequency with which the signal was sent by the transmitter, when transmitter and receiver are moving relative to each other, see Figure 20.3. This is the Doppler effect. The most well-known example, in terms of sound waves, is the change in pitch of a police car siren as the car first approaches you and then recedes.

The Doppler effect is used to measure the *rate of change* of a distance, or *range-rate* for short, denoted by $\dot{l}$, where the dot denotes the time derivative $\frac{dl}{dt}$. Practically speaking the

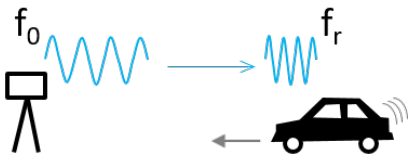



Figure 20.3: The Doppler effect: the receiver is moving toward the transmitter and the observed frequency $f_r$ is larger than the transmitted frequency $f_0$. The Doppler shift equals $f_0 - f_r$.

rate of change of distance $\dot{l}$ is the relative radial speed between transmitter and receiver, i.e. the relative speed along the line connecting the two (or more formally, the relative velocity vector projected onto this line).

Integrating the range-rate $\dot{l}$ (over time) yields the *change in* distance in the time interval $[t_1, t_2]$: $l(t_2) - l(t_1) = \int_{t_1}^{t_2} \dot{l}(\nu) d\nu$.

For convenience we consider a stationary transmitter and a moving receiver (for one-way ranging), as in Figure 20.3. The transmitter is broadcasting an electromagnetic wave with frequency $f_0$, and the receiver measures the frequency to be $f_r$. Then

$$\frac{f_0 - f_r}{f_0} \approx \frac{\dot{l}}{c} \tag{20.9}$$

The measured Doppler frequency (shift) equals $f_0 - f_r$ and when the receiver is moving toward the transmitter $f_0 < f_r$ and correspondingly $\dot{l} < 0$, the distance between transmitter and receiver gets shorter. The above expression is often found as $f_r \approx f_0(1 - \frac{\dot{l}}{c})$. Note that these (approximate) expressions are valid for $\dot{l} \ll c$, see e.g. [52], for electromagnetic waves traveling at $c$ in vacuum.

For two-way ranging the Doppler shift is twice as large, compared to one-way ranging in the same scenario.

## 20.3. Imaging

Instead of detecting objects (targets) in specific directions, possibly by a moving or rotating antenna, as typically used with air traffic control radar, a (fixed) antenna can 'look' sideways and illuminate a whole area in one go. This is the basis of imaging, for instance with radar remote sensing.

Only one (short) pulse is transmitted (at a time), and reflections are received during a certain time span (the *full response* is recorded, with multiple reflections/echoes). An object nearly underneath the sensor gives an early response (as the signal needs to travel only a short distance), and an object far away (sidewards) from the sensor gives a late response (as the signal has to travel a long way). Figure 18.7 already suggested a sideway looking radar.

Figure 20.5 demonstrates the imaging principle. The satellite is flying at altitude $H$ above the Earth's surface. The radar sensor is side-looking; the looking angle of the radar is $\Theta$. Distance $R$ is the slant range, and $R \sin \Theta$ is the corresponding ground range.

The signal beam is indicated in light gray. The Earth's surface, topography and objects within the light gray beam return the signal, and, over a suited time span, the incoming response, with multiple pulse echoes or reflections, is recorded, i.e. periodically sampled, at time instants which are $\Delta\tau$ seconds apart. The sampling time interval is related to the sampling frequency $f_s$, simply through $f_s = \frac{1}{\Delta\tau}$. The samples each represent one pixel, and they are stacked next to each other, and form one row (or line) in the image. The mapping, or projection, from the samples in the time domain, to pixels covering the ground or Earth's

Figure 20.4: Example of radar amplitude image. Amplitude per pixel is shown in gray-scale from black to white. The image shows the area of Delft, about 7 km x 7 km, with the old city center in the middle, and (approximately) North up, acquired, on Nov 1, 2015, by the DLR TerraSAR-X satellite [53] at 500 km altitude. The data are available with a pixel size on ground of about 3 m x 3 m. Water bodies, like the Schie-canal, typically show up in black.

surface, is generally done using a *reference surface*, i.e. assuming a locally flat Earth's surface underneath the radar sensor, as in Figure 20.5, or more sophisticated, a curved Earth's surface (sphere, or ellipsoid).

As the satellite moves on, the next pulse records the next row of the image. In this way a two-dimensional image (or array) is created, with rows and columns of pixels. For ordinary imaging, it is easiest to think of the intensity or *amplitude* of the incoming signal being recorded for each pixel. As outlined in the introduction of Chapter 25, the intensity of the received signal determines the pixel value, to say, whether the pixel gets black, white, or a particular shade of gray.

Figure 20.4 shows an example of a radar image. Per pixel, the amplitude of the reflection is shown: white represents a strong reflection into the direction of the radar sensor, and black only a little or no reflection (such as water bodies).

In Chapter 23 we consider, next to the amplitude, also the *phase* of the radar signal, in order to obtain geometric information (the phase is a measure of distance, cf. (20.5)).

The formation of one pixel is shown in Figure 20.5 through the segment in dark-gray. The pixel size in (slant) range direction, expressed in [m], is

$$P_r = \frac{c}{2f_s} \tag{20.10}$$

similar to the range resolution in (20.2). Figure 20.5 shows that the sampling frequency $f_s$ and the looking angle $\Theta$ determine the pixel size on the ground, expressed in [m].

$$P_g = \frac{P_r}{\sin\Theta} = \frac{c}{2}\frac{1}{f_s \sin\Theta} \tag{20.11}$$

From an aircraft or satellite, with a side-looking instrument, images are obtained rowwise of ground areas located perpendicular to the flight line. Together with the forward motion of the aircraft or satellite, this eventually yields, like photography, a two-dimensional image of the Earth's surface and its topography, see Figure 20.6. The strip length, i.e. the length of the

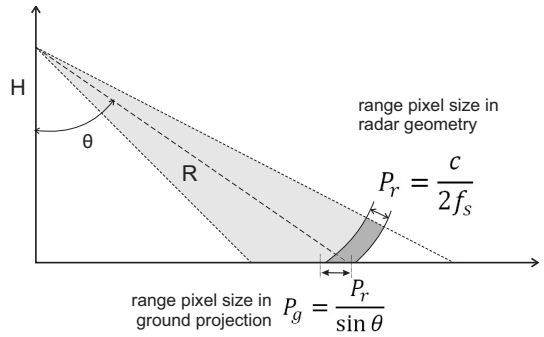



Figure 20.5: Forming image pixels from the received response of the transmitted radar pulse.

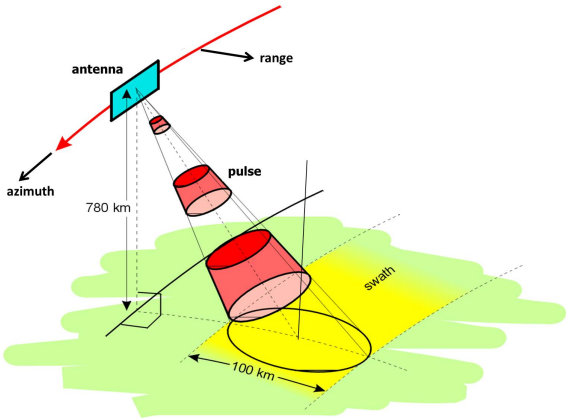



Figure 20.6: Radar remote sensing: with a side-looking radar, images are obtained rowwise, of ground areas located perpendicular away from the flight line. As the satellite moves forward along its track, a two-dimensional image of (a swath of) the Earth's surface and its topography is obtained. The pulse-repetition frequency (PRF) determines the width of the swath (the response area).

rows (in range or across-track direction), is called the swath. The swath width, indicated in yellow in Figure 20.6, is determined by the Pulse Repetition Frequency (PRF), which equals $\frac{1}{\tau_{\max}}$, see with (20.3) and the radar beamwidth is typically set to match the PRF. The longer the interval $\tau_{\max}$ used for recording the echoes is, the wider the swath will be.

The radar azimuth spatial resolution (in the direction of the flight line, i.e. along track), is inversely proportional to the length of the antenna, and hence, can be increased by either using a longer antenna to narrow down the beamwidth. Another approach is to exploit the forward motion of the sensor, by *synthesizing* the effect of a very long antenna, a technique known as SAR (Synthetic Aperture Radar). Features on the Earth's surface are captured in a sequence of images, as the satellite moves forward, see Figure 20.7. Overlapping radar pulse returns from different azimuth positions are combined. By the latter technique the resolution gets essentially independent of the (flying) height of the instrument, and the antenna remains physically unchanged.

## 20.4. Comparison on imaging geometry: photo vs. radar

Although both photography and radar employ electromagnetic waves and yield images, they do rely on *different* principles. From a geometric point of view, photogrammetry, as discussed in Chapter 19, is based on measuring *angles*, and radar (and lidar and sonar as well) on

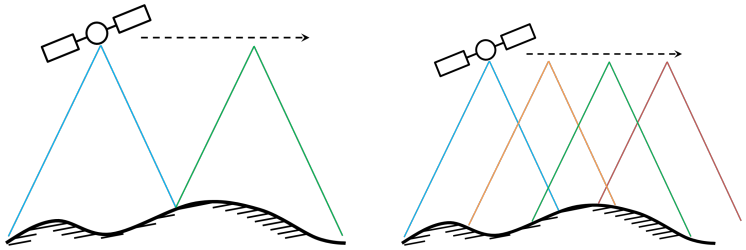


Figure 20.7: Cross-section in azimuthal direction (i.e. along flight track direction) of radar satellite, sensing the Earth's surface. The radio signal beamwidth is inversely proportional to the length of the antenna. In principle (at left), one could opt to capture adjacent rows of the image *non-overlapping*, i.e. take the next row (next pulse) only once the satellite has sufficiently moved forward. With a synthetic aperture antenna (at right) one emulates the effect of using a (very) long antenna, by capturing adjacent rows *largely overlapping*. In the processing of the data effectively a narrow beam is formed. This yields a (much) higher image resolution in azimuth direction.

ranging, hence on measuring *distances*.

This section presents a simple example showing the difference, see Figure 20.8. An artificial cliff appears as a ramp feature on the Earth's surface (only one direction is considered here, a vertical cross-section has been made). The sensor (antenna or lens), carried by an aircraft or satellite, is exactly above the left vertical marker, at a height that equals the distance between the two markers on the horizontal axis (the plate with the photo film has been positioned parallel to the Earth's surface).

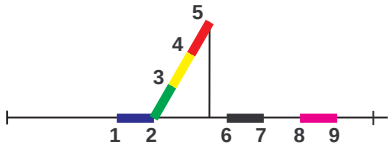



Figure 20.8: Example of simple topography on the Earth's surface to be captured by both photo and radar imaging.

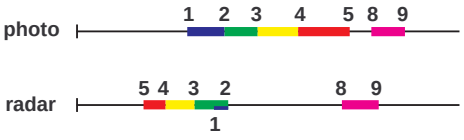



Figure 20.9: Images of the example above, Figure 20.8, the photo image as the line on top, the radar image at bottom.

The graph in Figure 20.9 presents the corresponding optical photo and radar image (same cross-section as above). Along the horizontal axis the distance is given as measured by the radar. For the photo (positive), the angle is converted into distance by the focal distance of the camera (which was taken just equal to the aircraft height for convenience in this example), and consequently is the distance of the image point to the center of the photo. The marker presents again the sensor (or lens, perpendicularly projected onto the photo film).

As can be clearly seen from Figure 20.9, the measurement principles of distance and angle yield two *different* images. The photo depicts features in the terrain as it encounters them in logical order from left to right, from 1 through 9.

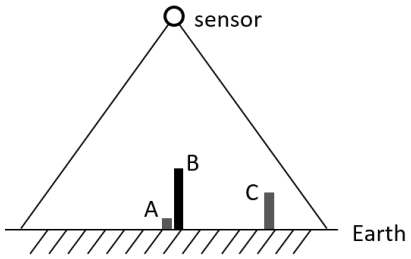



Figure 20.10: Scenario giving clearly distinct images in terms of photo (based on direction) and radar (based on distance).

In a side-looking radar image, the direction of relief displacement is *reversed* compared with an (ordinary) photo. The radar pulse for ranging reaches the top of a vertical feature earlier than the base, and consequently the return signal from the top is received *before* that of the base. This causes a vertical feature to *lay over* or lean *forward* (instead of back). In the radar example point 5 (top) 'comes before' point 2 (base).

Both measurement techniques are based on electromagnetic waves and their propagation — basically geometric optics — which means that objects in the *shadow* of other objects can not be 'seen', as in the example the points 6 and 7. This holds for both the photo and the radar image.

By means of the scenario shown in Figure 20.10, the difference between a photo-image and a radar-image is distinctly marked. When the sensor is an optical camera, the tops of the buildings A and B will be shown directly next to each other in the photo - they are in the same viewing direction. In the radar-image however, they will be clearly apart, as their distances to the sensor are clearly different: object B is much taller than object A, and consequently the top of B is much closer to the sensor than the top of A.

In the radar-image, one will actually see buildings A and C directly next to each other, as the tops of these two buildings have nearly the same distance to the sensor. In the photo-image, buildings A and C are clearly apart, as they are in different viewing directions.

## 20.5. Spaceborne platform: mission design

A radar sensor can be mounted on a spaceborne or airborne platform, or can be operated from the ground. Here, we focus on spaceborne platforms and we provide a short introduction to satellite orbits.

Kepler's laws of planetary motion (Johannes Kepler (1571-1630)) also apply to the motion of a satellite around the Earth. The three laws state that:

1. the orbit is an ellipse with the Earth at one of the two foci (see Section 29.2 for details on the ellipse and its eccentricity; when the eccentricity of the ellipse goes to zero, the orbit turns into a circle, with the Earth in its center)

2. a line segment joining the satellite and the Earth, sweeps out equal areas during equal time intervals (in highly elliptical orbits, the satellite moves very fast when it is close to the Earth (when it is in, or near its perigee), and it moves very slowly in the other part of the orbit (when it is close to its apogee))

3. the square of the orbital period $T$ is proportional to the cube of the semi-major axis $a$ of its orbit

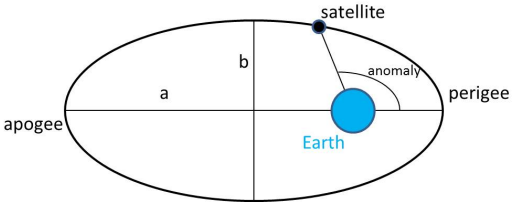



Figure 20.11: The satellite orbit around the Earth is an ellipse.

The third law is conveniently cast in the equation

$$n^2 a^3 = GM \tag{20.12}$$

where $n$ is the mean-motion [rad/s] of the satellite (basically the average rate of the anomaly angle, at the Earth, between the satellite position and the perigee); the mean-motion simply equals $n = \frac{2\pi}{T}$, with $T$ the orbital period [s] (it takes $T$ seconds to complete a full turn of $2\pi$ [rad]). The semi-major axis of the orbital ellipse is denoted by $a$ ([m]), see Figure 20.11, and $GM$ is the Earth's gravitational constant ($GM = 3.986005 \cdot 10^{14}$ $\mathrm{m}^3/\mathrm{s}^2$), see Chapter 31.

These laws also follow from Newtonian mechanics, as a solution to the so-called two-body problem (the satellite and the Earth, where the mass of the Earth is by far much bigger than the mass of the satellite, and the satellite is falling around the Earth).

Ideally the satellite orbital plane (the plane in which the ellipse resides) is fixed in inertial space. The two main parameters of this orbital plane are the inclination, which is the angle of the plane, with the Earth's equator, and, the longitude of the ascending node, which is the angle between the vernal equinox and the longitude, along the equator, of the point where the satellite — on its way to the Northern hemisphere — crosses the equator. The vernal equinox refers to the direction to the Sun basically at March 21st; an equinox occurs when the plane of Earth's equator passes (contains) the center of the Sun (around March 21st, and around September 21st).

An inclination of 90 degrees yields a so-called *polar* orbit (the satellite will each time pass over both poles). And an inclination of 0 degrees, together with an orbital period of 23h56m, yields a so-called *geo-stationary* orbit; the satellite keeps — as seen from the Earth — a fixed position in the sky. In a polar orbit, a (single) satellite can, while orbiting the Earth, as the Earth rotates underneath the satellite, in principle observe all places on Earth — it can get directly overhead at any place on Earth. A satellite in a polar orbit can, in a couple of full orbits around the Earth, observe the entire Earth's surface.

One can see that the larger the orbital radius is (larger $a$), the larger the orbital period $T$ will be, and hence, the longer it takes before the satellite completes one orbit around the Earth. The orbital period is an important parameter in the satellite's mission design, as it drives the orbit repeat time; together with the Earth's rotation (one full turn, in inertial space, in 23h56m) it determines when the same part of the Earth is imaged again by the satellite, i.e. the time between two successive image captures of the same area (and thereby the *temporal resolution* of an image sequence).

In general the orbital plane does not keep its orientation in inertial space, and in particular it may rotate about the Earth's rotation axis (a phenomenon known as precession, due to oblateness of the Earth). In case the orbital plane makes one full turn in one year, the satellite is in a so-called Sun-synchronous orbit. During the course of the Earth's journey around the Sun in one year, the orientation of the orbital plane with respect to the Sun, stays the same. This is very convenient as the incidence angle of Sun-light onto the remotely sensed Earth's surface is always the same. The illumination of the scene is always the same.

## 20.6. Concluding remarks

The principles covered in this chapter apply to radar remote sensing using electromagnetic radio signals and to sounding using acoustic signals. Radar remote sensing is further covered in Chapter 23, and acoustic remote sensing is further covered in Chapter 24.

The principles of ranging in Section 20.1 also apply to laser altimetry and scanning using optical signals. Laser scanning is further covered in Chapter 22.

## 20.7. Exercises and worked examples

This section presents several exercises and worked answers about satellite remote sensing.

**Question 1** With a radar altimeter on a satellite, the distance from the satellite to the Earth's surface is measured. Suppose that a distance accuracy of 1.5 cm is wanted, what is then the corresponding requirement on the timing accuracy, to observe the signal travel-time?

**Answer 1** A radar altimeter works through ranging, i.e. by transmitting a signal, which is reflected by the Earth's surface (or another object), and measuring the total travel-time — upon reception on board of the satellite — of the reflected signal. The relation of the two-way travel-time $\tau$ and the geometric range or distance $l$ is given by Eq. (20.1). The word 'accuracy' can here be interpreted, in the assumed absence of biases and systematic offsets, as precision, quantified by the standard deviation. Hence, the question can be reformulated as: compute $\sigma_\tau$, given a requirement of $\sigma_l$ = 1.5 cm. The relation between $\tau$ and $l$ is a linear one: $l = \frac{c}{2}\tau$, and with the propagation laws of Chapter 7, we simply have $\sigma_l = \frac{c}{2}\sigma_\tau$. Inserting $\sigma_l$ = 1.5 cm, leads to $\sigma_\tau$ = 0.1 ns (1 nanosecond). This requirement on the travel-time accuracy is independent from the flying altitude of the satellite.

**Question 2** A satellite is flying at an altitude of 20.200 km above the Earth's surface. Compute the orbital period. The radius of the Earth can be taken as $R$ = 6378 km.

**Answer 2** The semi-major axis of the satellite's orbital ellipse is $a = 20200 + 6378 = 26578$ km. In this question we actually consider an even simpler case, namely of an orbital ellipse with zero eccentricity, hence, just a circle. With $GM = 3.986005 \cdot 10^{14}\ \mathrm{m^3/s^2}$, and the third law of Kepler $n^2a^3 = GM$, the mean motion becomes $n = 1.457 \cdot 10^{-4}\ \mathrm{s^{-1}}$. With $n = \frac{2\pi}{T}$, we obtain $T$=43122 s. The satellite considered is actually a GPS satellite, which orbits the Earth twice per 23h56m.

**Question 3** Given is the (vertical) cross section of the terrain and the radar imaging satellite in Figure 20.12. The satellite is side-looking, and flying straight in (or out of) the paper (the azimuth direction is perpendicular to the paper). Construct the radar image of the shown terrain.

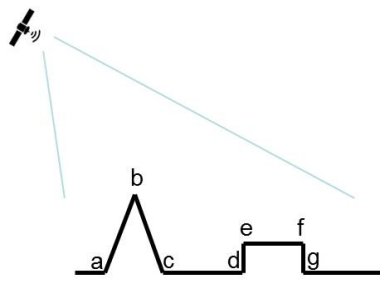



Figure 20.12: Vertical cross section of terrain to be imaged through radar remote sensing (Question 3).

**Answer 3** Radar remote sensing relies on ranging, hence, on measuring *distances*. One

way to construct the radar image is to measure distances from the satellite to all points of interest in the terrain. And then lay-out all these distances along a straight line — this yields the row of interest in the image. One should realize that points c and g will not be visible in the image, as they are in the shadow — the radar cannot see them. The alternative is to note that the satellite is generally at a large distance from the terrain (typically flying at several hundreds of kilometers altitude), and thereby radar signal wavefronts, shown in gray, are — locally seen — (nearly) parallel, see Figure 20.13. Mind that this drawing is not to scale, see also the dimensions given in Figure 20.6. The row of interest in the image is then constructed by projecting the points of interest, along these wavefronts, onto the line connecting the satellite and the terrain. The points in the image are indicated with a prime, hence b', a', e', d' and f'. The first ramp (or hill) with point b as its top, *leans forward* in the image — this means that the top (b') is shown in the image, before the base (a').

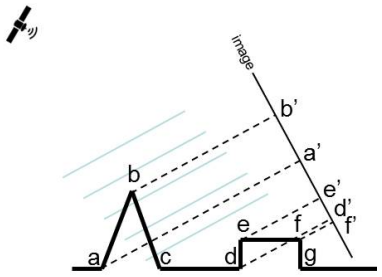



Figure 20.13: Process of imaging terrain with radar remote sensing (Answer 3).

# 21

# Signals and hardware

In surveying we use electromagnetic (EM) signals, and audio signals (sound). Electromagnetic waves are produced when free electric charges accelerate (for instance in an antenna), or when electrons bound to atoms and molecules make transitions to lower energy states. Electromagnetic signals propagate well in space (vacuum) and in air, over large distances, but only very limited in liquid (e.g. water). Sound is generated by motions of particles in a medium (this causes minute variations in pressure and density, for instance in air). These compressions and dilations propagate from transmitter to receiver, as a longitudinal wave. The presence and propagation of sound relies on the presence of particles, as air-molecules. Acoustic signals can not be used in space, but they can be used well in liquid (e.g. water). For the physical background on these subject, see e.g. [52], which actually serves as a reference textbook for the entire field of physics. A short exposition on signal propagation is given in Appendix G.

The second part of this chapter provides a basic overview of the working principles of two crucial hardware components as present in most of today's survey equipment. The oscillator and the antenna — delivering electromagnetic waves — are discussed.

## 21.1. Spectrum

For analysis purposes signals are decomposed in terms of *periodic* signals. That is, signals which vary sinusoidally with time $t$, and repeat after $T$ seconds in time. The strict definition of a signal $s$ being periodic, reads $s(t) = s(t + kT)\ \forall\ t \in \mathbb{R}$ with $k \in \mathbb{Z}$. $T$ is the period in [s], and $f = \frac{1}{T}$ is the frequency in [1/s], which is [Hz], sometimes referred to as 'cycles per second'. Mind that in Chapter 19, symbol $f$ was also used, to denote the focal distance of a camera.

The periodic behaviour of a parameter or quantity can be interpreted as a vector, which's endpoint traces a (unit) circle. In radians, a full turn corresponds to $2\pi$, and hence $\omega = 2\pi f$ yields the angular velocity of the rotating vector, or angular frequency in [rad/s].

### 21.1.1. Electromagnetic spectrum

By the speed of light (or any EM-signal) in [m/s], the frequency $f$ and the wavelength $\lambda$ in [m] are related according to $c = \lambda f$. As a medium we consider here vacuum, and symbol $c$ is reserved specifically for the speed of light in vacuum. In other media the speed of light, and hence the wavelength $\lambda$, may be different (though the speed in air may still be very close to the one in vacuum). When an electromagnetic field is, as a wave, propagating through space, the period in time translates into a period in space, and this is the wavelength.

| designation | frequency range [Hz] | wavelength range [m] | remark |
|---|---|---|---|
| Extremely Low Frequency (ELF) | $3 \cdot 10^0 - 3 \cdot 10^1$ | $10^8 - 10^7$ | |
| Super Low Frequency (SLF) | $3 \cdot 10^1 - 3 \cdot 10^2$ | $10^7 - 10^6$ | |
| Ultra Low Frequency (ULF) | $3 \cdot 10^2 - 3 \cdot 10^3$ | $10^6 - 10^5$ | |
| Very Low Frequency (VLF) | $3 \cdot 10^3 - 3 \cdot 10^4$ | $10^5 - 10^4$ | |
| Low Frequency (LF) | $3 \cdot 10^4 - 3 \cdot 10^5$ | $10^4 - 10^3$ | long waves |
| Medium Frequency (MF) | $3 \cdot 10^5 - 3 \cdot 10^6$ | $10^3 - 10^2$ | medium waves |
| High Frequency (HF) | $3 \cdot 10^6 - 3 \cdot 10^7$ | $10^2 - 10^1$ | short waves |
| Very High Frequency (VHF) | $3 \cdot 10^7 - 3 \cdot 10^8$ | $10^1 - 10^0$ | |
| Ultra High Frequency (UHF) | $3 \cdot 10^8 - 3 \cdot 10^9$ | $10^0 - 10^{-1}$ | micro waves |
| Super High Frequency (SHF) | $3 \cdot 10^9 - 3 \cdot 10^{10}$ | $10^{-1} - 10^{-2}$ | micro waves |
| Extremely High Frequency (EHF) | $3 \cdot 10^{10} - 3 \cdot 10^{11}$ | $10^{-2} - 10^{-3}$ | |
| | | | |
| thermal, or far-infrared | $10^{12}$ | $10^{-4}$ | |
| infrared | $10^{13}$ | $10^{-5}$ | |
| near-infrared | $10^{14}$ | $10^{-6}$ | |
| visible light | $4 \cdot 10^{14} - 7 \cdot 10^{14}$ | $7 \cdot 10^{-7} - 4 \cdot 10^{-7}$ | |
| ultraviolet (UV) | $10^{16}$ | $10^{-7}$ | |
| X-ray | $10^{18}$ | $10^{-9}$ | |
| gamma-radiation | $> 10^{19}$ | $< 10^{-11}$ | |

Table 21.1: The electromagnetic spectrum divided after frequency. Wavelength $\lambda$ and frequency $f$ are related by the speed of light in vacuum $c$, according to $c = \lambda f$, with $c = 299792458$ [m/s], roughly $3 \cdot 10^8$ [m/s]. The speed of light in vacuum is a universal physical constant. For thermal-infrared, infrared, near-infrared, ultraviolet, X-ray and gamma-radiation only approximate mid-values are given (for frequency and wavelength).

After the wavelength, or conversely the frequency, the whole range of the electro-magnetic spectrum can be sub-divided, see Table 21.1.

Note that the designations, or bands, for the radio frequencies, from ELF through EHF, each cover one order of magnitude; for example, VHF ranges from 30 to 300 MHz in frequency ($3 \cdot 10^7$ - $3 \cdot 10^8$), corresponding to 10 to 1 meter in wavelength ($10^1$ - $10^0$).

Electromagnetic waves from ELF to EHF are usually referred to as *radio*-waves.

Apart from AM-radio in the medium range and short waves in the high frequency range, the VHF category contains most radio and television signal transmissions (up to the UHF-band). FM-radio and DAB digital radio are in the VHF-band (88–108 MHz, 174–240 MHz resp.), and TV channels use the VHF-band (54–88 MHz) and (174–216 MHz) and the UHF-band (470–862 MHz), the latter is also used by DVB-T digital television.

GPS satellite navigation operates in the UHF-band ((L2-frequency) 1.2276 GHz and (L1-frequency) 1.57542 GHz), and so do microwave ovens (2.45 GHz), and Wireless Local Area Networks WLAN/WiFi (2.4 GHz). Also mobile communication takes place in this part of the electromagnetic spectrum, initially around 900 MHz (GSM) and today also at 1.8, 1.9 and 2.1 GHz (UMTS, 4G), and also 3.5 GHz (5G), with bands up to several tens of MHz wide.

Radar, with millimeter through decimeter wavelengths in the 1–100 GHz band, is used for remote sensing. Specifically the so-called C-band is often used (with wavelengths in the order of 4-7 cm) and the X-band (with wavelengths around 3 cm). Radar supplies its own source of energy to illuminate objects of interest. Electromagnetic radar waves travel two ways, from sensor to object, and back. It is an active measurement system. Radar can also be used to detect precipitation in the Earth's atmosphere, as for instance rain showers. Reflectivity of the signal depends on the type of precipitation (rain, snow, hail, etc.), precipitation rate and on employed frequency. Weather radar usually operates in the 4–10 GHz range.

| designation | frequency range [Hz] | wavelength range [m] | remark |
|---|---|---|---|
| audible | $2 \cdot 10^1 - 2 \cdot 10^4$ | $2 \cdot 10^1 - 2 \cdot 10^{-2}$ | |
| ultrasound | $2 \cdot 10^4 - 2 \cdot 10^9$ | $2 \cdot 10^{-2} - 2 \cdot 10^{-7}$ | |

Table 21.2: The sound spectrum divided after frequency. Wavelength $\lambda$ and frequency $f$ are related by the speed of sound in *air* $v$, according to $v = \lambda f$, with $v \approx 340$ [m/s].

| designation | frequency range [Hz] | wavelength range [m] | remark |
|---|---|---|---|
| audible | $2 \cdot 10^1 - 2 \cdot 10^4$ | $7 \cdot 10^1 - 7 \cdot 10^{-2}$ | |
| ultrasound | $2 \cdot 10^4 - 2 \cdot 10^9$ | $7 \cdot 10^{-2} - 7 \cdot 10^{-7}$ | |

Table 21.3: The sound spectrum divided after frequency. Wavelength $\lambda$ and frequency $f$ are related by the speed of sound in *water* $v$, according to $v = \lambda f$, with $v \approx 1480$ [m/s].

Electromagnetic waves for which the human eye is sensitive, lie in the range from 700 to 400 nm (wavelength), from red, via green, to blue; longer waves are in the infrared range and shorter waves are ultraviolet radiation. By the virtue of reflected Sun light we can observe features at daylight. Laser generally uses wavelengths between 0.5 and 1 $\mu$m. Electro-optical EDMs typically operate on near-infrared (0.8–0.9 $\mu$m) and on red laser ($\approx 0.6$ $\mu$m).

Far infrared, with wavelengths of about $100$ $\mu$m may alternatively be referred to as thermal infrared. All objects or media at temperatures above absolute zero (0 Kelvin, -273.15 °C) continuously emit electromagnetic radiation, associated with heat. The object's surface temperature is a key parameter to the amount of radiated energy. Passive thermal scanners rely on this type of electromagnetic radiation, see Chapter 25.

### 21.1.2. Audio spectrum

As electromagnetic waves virtually do not penetrate in water, acoustic signals are generally used in hydrography instead. Techniques and systems based on acoustic waves are referred to as *sonar*. The acronym stands for SOund NAvigation and Ranging. Sound (like human speech and music) usually refers to oscillations of air. Man's ear is sensitive to frequencies in the range of 20 Hz to 20 kHz and human voice typically ranges from 300 to 3600 Hz. A bat uses an audio signal in the 100—200 kHz range to navigate and locate a prey. The frequencies used for sonar in water typically range from a few kHz to several hundreds of kHz.

The propagation speed of an acoustic wave in air, as in the Earth's atmosphere, is about $3.4 \cdot 10^2$ [m/s], but considerably larger in liquid and solid media. In water it lies in the order of $1.5 \cdot 10^3$ [m/s], and depends on temperature, pressure and salinity. Tables 21.2 and 21.3 present the spectrum of audio-signals in air, and in water.

The basic instrument in hydrography is the echo sounder that measures depth. A transducer attached to the bottom of a vessel measures the time duration between departure of the (transmitted) signal and arrival of the signal reflected by the seafloor, similar to Figure 18.5, but then in a vertical sense. At present, side scan sonars and multiple sonar (multibeam) systems are used for seafloor mapping. Acoustic waves are then not used only for depth measurements, but merely for positioning under water in general. Finally we state that the word 'sounding' is generally used for all types of depth measurements, including those that do not use sound.

Figure 21.1: Prototype of pendulum clock developed and used by Christiaan Huygens in the tower of the Old Church in Scheveningen (Oude Kerk in de Keizerstraat). This original pendulum clock is now housed in the Huygens museum in Voorburg, see Figure 17.2.

## 21.2. Oscillator

An oscillator delivers a periodically repeated signal, phenomenon or event, think for instance of a pendulum clock, invented by Christiaan Huygens in 1656, with the pendulum 'endlessly' swinging forth and back, see Figure 21.1. The oscillator, the heart of most measurement equipment, is responsible for basically generating a constant *frequency*, or conversely a sequence of constant time durations (constant period). The oscillator drives both timekeeping in the instrument, as well as signal generation and processing.

A clock is basically a *counter*; it counts the number of events or cycles produced by the oscillator, and dividing the count by the (nominal) frequency yields the lapsed time. For example, counting 50 cycles from an oscillator running at 10 Hz (10 cycles per second), yields at time duration of 50/10 = 5 seconds.

In this section we outline the principle of deriving time from an oscillator signal, and we show how frequency instability of the oscillator leads to timing errors. As a signal we consider just a basic monotone carrier signal. The carrier signal supplied by the oscillator is a (periodic) sinusoidal wave $s(t) = \cos\phi(t)$. The sine, or cosine, carries the phase $\phi(t) = \omega t = 2\pi f t$ as the argument in radians, or $\phi(t) = ft$ in cycles. The angular frequency is denoted by $\omega$, and frequency $f$ is in Hertz, and time $t$ in seconds. Up to now it was tacitly assumed that the oscillator behaved perfectly and that the phase (in cycles) followed as

$$\phi(t) = ft$$

with frequency $f$ constant and $\phi(t = 0) = 0$.

The frequency realized by the oscillator may vary with time however, $f(v)$, and more

precisely, the (difference in) phase in cycles (over a time span from $t_o$ to $t$) reads

$$\phi(t) - \phi(t_o) = \int_{t_o}^{t} f(\nu)d\nu \tag{21.1}$$

where the time argument of the integral $\nu$ runs from $t_o$ to $t$, and at time $t$ the cosine (or sine) with argument would read

$$\cos 2\pi(\int_{t_o}^{t} f(\nu)d\nu + \phi(t_o))$$

Any deviation in the phase of the oscillator transfers into the *phase* of the *signal* carrier. The frequency in Hertz (or number of cycles per second) is just the time derivative of the phase.

In practice the oscillator can not maintain the specified/nominal frequency exactly. The actual frequency $f(\nu)$ can be split into $f(\nu) = f_o + \delta f(\nu)$, the nominal (and constant) frequency $f_o$, and the frequency deviation $\delta f(\nu)$. In the sequel we first consider how frequency deviation causes a time error, and then we return to timing stability.

### 21.2.1. Clock error

In most measurement equipment, frequency and time are due to the very same oscillator. The oscillator generates the signal, and also drives the clock. We will introduce therefore the clock error (or offset) of the equipment or device.

Any deviation in frequency translates directly into a timing error. The change in phase (21.1) over the time span $[t_o, t]$ as realized by the oscillator of the device, divided by the nominal frequency $f_o$, yields the change in time $\bar{t}$ as kept and *indicated* (or displayed) by the device.

$$\begin{aligned}
\bar{t}(t) - \bar{t}(t_o) = \frac{\phi(t) - \phi(t_o)}{f_o} &= \frac{1}{f_o}\int_{t_o}^{t} f(\nu)d\nu \\
&= \frac{1}{f_o}\int_{t_o}^{t} (f_o + \delta f(\nu))d\nu \\
&= (t - t_o) + \frac{1}{f_o}\int_{t_o}^{t} \delta f(\nu)d\nu \\
&= (t - t_o) + \delta t(t) - \delta t(t_o)
\end{aligned}$$

Disregarding the (absolute) origin of time (mankind can actually measure only *time duration*), the instantaneous relation reads

$$t = \bar{t}(t) - \delta t(t)$$

The ideal proper or true time $t$ is obtained by reading the clock of the device $\bar{t}(t)$, and correcting it for the clock error/offset $\delta t(t)$, see Figure 21.2.

As in Figure 18.4, clocks are used to measure signal travel-time. If the clock of the receiver runs ahead, or lags behind, the observed signal travel-time will be too long, or too short, respectively. When, in measuring the signal travel-time from transmitter i to receiver j, $t_{ij} = t_j - t_i$, the clock of receiver j is involved, the observation shall be corrected for the clock error $\delta t_j(t)$ multiplied by $c$, or equivalently, when unknown, the clock error $\delta t_j(t)$, multiplied by $c$, appears on the right hand side in the observation equation (we assume here that transmitter i is perfectly synchronized, and has zero clock error).

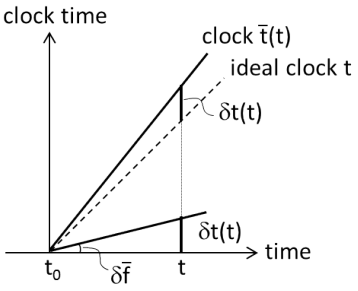



Figure 21.2: The clock offset $\delta t$ is the offset between the time $\bar{t}$ shown by the device and the true time $t$. In this example the clock drift is caused by a (constant) frequency offset $\delta\bar{f}$. Clock $\bar{t}$ is running too fast.

### 21.2.2. Timing stability [*]

Fluctuations $\delta f(\nu)$ in the frequency of an oscillator about its nominal frequency $f_o$, referred to as frequency accuracy (or, actually frequency inaccuracy), can result from perturbations such as thermal noise in electric components, instrumental ageing, and environmental variations (temperature, pressure, vibrations). As just shown, these fluctuations cause the time, displayed by the device, to deviate from the true time.

Departures $\delta f(\nu)$ from the nominal frequency $f_o$ are typically expressed as a relative measure, through

$$\delta\bar{f}(\nu) = \frac{\delta f(\nu)}{f_o}$$

which is the *relative* frequency departure, or fractional frequency error. It is a dimensionless quantity.

From $\frac{1}{f_o}\int_{t_o}^{t} \delta f(\nu)d\nu = \delta t(t) - \delta t(t_o)$, setting $t_o$ to $-\infty$, and taking the derivative

$$\delta\bar{f}(t) = \frac{\delta f(t)}{f_o} = \frac{d(\delta t(t))}{dt}$$

one can deduce basically that the relative frequency error equals the rate of change of the clock error.

Ideally $\delta\bar{f}(t) = 0$ and the oscillator runs exactly at the right, nominal frequency. The clock time error will be constant (not change). The clock may run late or early, already from the start, but it runs at the right pace, and its offset stays the same.

If the relative frequency error is constant, $\delta\bar{f}(t) = k$, with $k \neq 0$, the oscillator constantly runs too fast (or too slow), and the clock offset will change (linearly) as time goes by, see Figure 21.2. The frequency is not correct $f(t) = (1 + k)f_o \neq f_o$, but this oscillator still is perfectly *stable* (running at a constant frequency).

In practice frequency instabilities are more of a concern, i.e. the variation in $\delta\bar{f}(t)$. In the sequel we want to quantify (measure) the frequency (in)stability.

Comparing the readings of a certain clock with a calibrated/proven standard, we could compute the Mean Squared Error (MSE) of the observed time differences, as a measure of accuracy. In the timekeeping community instead a slightly different approach is followed, by using the Allan-variance [54] as a measure of time-*stability*.

Based on a series of discrete relative frequency error measurements $\delta\bar{f}_i$ at times $t_i$ with $i = 1, \ldots, N$, equidistantly in time, $\Delta t = t_{i+1} - t_i$, (these measurements may originate from a series of time error measurements $\delta t_i$ at times $t_i$ with $i = 1, \ldots, N + 1$, specifically *differences*

| oscillator | stability |
|---|---|
| pendulum | $10^{-5}$–$10^{-6}$ |
| quartz | $10^{-6}$–$10^{-10}$ |
| atomic - rubidium | $10^{-12}$ |
| atomic - cesium | $10^{-13}$ |
| hydrogen maser | $10^{-13}$–$10^{-15}$ |

Table 21.4: Oscillator types and their long term (in)stability, over a full day. The time (in)stability is expressed as the Allan deviation, which is the square root of the Allan variance.

in time error $\delta\bar{f}_i = \frac{\delta t_{i+1} - \delta t_i}{\Delta t}$) the Allan variance, which is the mean of a series of two-sample variances, reads

$$\hat{\sigma}^2_{\delta\bar{f}} = \frac{1}{2}\frac{\sum_{i=1}^{N-1}(\delta\bar{f}_{i+1} - \delta\bar{f}_i)^2}{N-1}$$

The Allan-variance takes into account the *time duration* over which the *stability* is kept, in this case $\Delta t$, sometimes referred to as the averaging time. For oscillators in practice short term stability may clearly differ from long term stability. The Allan variance is a dimensionless quantity. The square root of the Allan variance is the Allan deviation, and it can be thought to have units 'seconds per second'.

### 21.2.3. Timekeeping

Table 21.4 presents several common types of oscillators and their timing (in)stability. Given is the stability over one day of 24 hours. Stability characteristics of an oscillator may differ significantly on short, mid and long term. The classical mechanical pendulum looses 1 to 0.1 second a day. For comparison it is mentioned that the daily variation of the Earth's rotation rate is at the level of a few milliseconds per day. The Earth's timing accuracy lies consequently on the order of $10^{-8}$ s/s. For a long time, (civil) time has been based on (astronomial) celestial motions, and in particular the Earth's rotation.

For two-way ranging systems, actually only short-term timing stability is critical. As transmitter and receiver are just a single physical device, and sharing the same clock/oscillator, timing stability is — in principle — needed over just the two-way travel-time duration. Though note that when using EM-signals, just one nanosecond timing error, already translates in a 0.30 m error, by the speed of light.

At present an oscillator is based on either a vibrating crystal (piezo-electric effect; quartz oscillator) or (energy) state transitions of (usually cesium or rubidium) atoms, (atomic clock). A cheap quartz oscillator (often found in GPS receivers) looses 1 $\mu$s every second, which is close to 0.1 second a day. The stability of an ordinary quartz crystal oscillator (XO) is improved to $10^{-8}$ and $10^{-10}$ by respectively the temperature compensated crystal oscillator (TCXO) and the oven-controlled crystal oscillator (OCXO). Atomic clocks possess superior stabilities. They are based on rubidium, hydrogen (maser) and (at present primarily) cesium. Using strontium a stability of $10^{-18}$ has been demonstrated with an experimental atomic clock. Confined ion clocks may be used in future and optical frequency standards on the long run.

Only since a few decades, (civil) time is kept by atomic frequency standards; the Coordinated Universal Time (UTC) is related to the International Atomic Time (TAI). The national time standard in the Netherlands is maintained by National Metrology Institute VSL in Delft, see Figure 21.3. For further reading on the art of timing, we refer to [54].

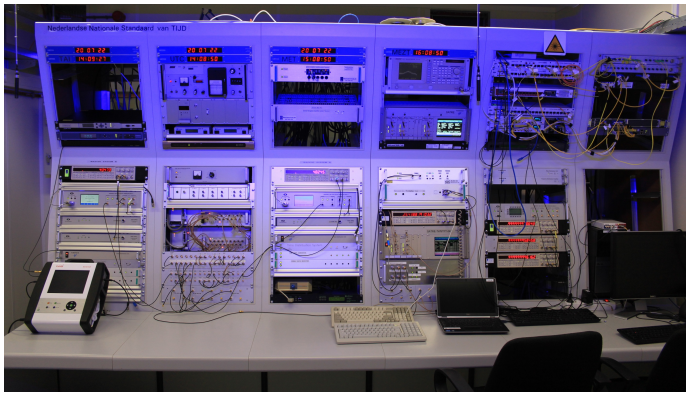


Figure 21.3: Control room at National Metrology Institute VSL in Delft for maintaining the official time standard in the Netherlands. This time is driven by four atomic clocks at VSL, which are continuously being compared against atomic clocks around the world (more than 400 of these clocks at about 90 time laboratories). This comparison relies on satellite signals, as for instance navigation satellite signals (such as those of GPS and Galileo). The Coordinated Universal Time (UTC) is eventually realized as a weighted mean of an ensemble of atomic clocks around the world, including the ones at VSL. Image courtesy of National Metrology Institute VSL in Delft [55].

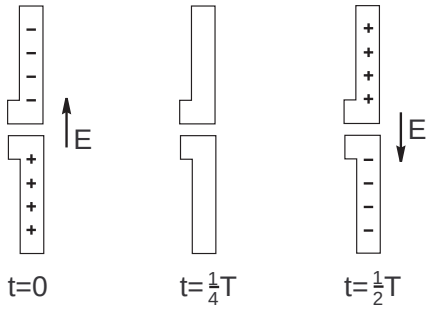



Figure 21.4: When charged particles (e.g. electrons) are accelerated through a conductor, a (varying) electric field is induced. An electric dipole antenna is fed, at the two short ends on the left, with an *alternating* current, causing the charged particles to oscillate in the two antenna elements. The electric field, shown by vector $E$, is changing accordingly and the (typically sinusoidal) wave propagates away from the antenna to the right, see also Figure G.4. The magnetic field is not shown. The polarization is linear.

## 21.3. Antenna

In this section we briefly describe how electromagnetic waves are generated, and in particular the class of radio-waves. The acoustic transducer is not covered here.

The antenna converts transmitter's time-varying electric currents into electromagnetic waves, representing an analog signal, that can propagate through space, and the Earth's atmosphere, to the receiver, at which the reverse process takes place. Figure 21.4 shows one of the elementary antenna types, namely the dipole antenna. It consists of two pieces of conducting material (wires or rods). In Figure 21.4 the two pieces are aligned vertically, in the plane of the paper.

For several applications it would be ideal would the antenna radiate in any direction, and with the same amount, this would be an isotropic antenna, but this is physically not possible. The bottom line is that the electric and magnetic field strengths (and hence the power flow as well) depend on spatial direction. The relative distribution of radiated power as a function of the spatial direction is given by the antenna's *radiation pattern* and an example is shown in Figure 21.5. It presents the radiation pattern of the very elementary dipole antenna. In

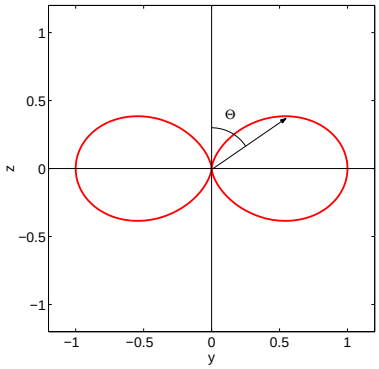



Figure 21.5: The radiation pattern of a dipole antenna. The radiated power, indicated by the length of the arrow, is proportional to $\sin^2\Theta$. Along the antenna axis it is zero and at maximum in the plane perpendicular to the antenna axis, as shown here in the direction of the $y$-axis.

this case the radiated power depends only on zenith angle $\Theta$, not on the azimuth angle. The pattern is rotation symmetric about the $z$-axis. The antenna is positioned in Figure 21.5 just as it is shown in Figure 21.4.

The spheres in Figure 21.5 show that the antenna transmits the strongest signal along the horizontal axis, and that signal-strength decreases with increasing angle with the horizontal axis. No signal power is transmitted in the direction aligned with the antenna elements (zenith angle equal to zero).

A *dipole antenna* consists of two rods of conducting material as shown in Figure 21.4. In order to radiate electromagnetic power efficiently, the minimum size of the antenna must be comparable to the wavelength. Each of the rods in Figure 21.4 is one-quarter of a wavelength long. They are placed end to end with a small spacing at the center. The antenna has consequently a size of half a wavelength.

A dipole antenna can receive the electromagnetic field shown in Figure 21.4, when it is orientated parallel to the direction of the electric field, shown by vector $E$, so that an alternating current is induced in the receiving antenna.

Electromagnetic signal transmission is covered in further detail in Appendix G.1.

## 21.4. Exercises and worked examples

This section presents one exercise on the impact of a clock error on measuring distances with GPS.

**Question 1** GPS satellite navigation is based on observing distances from satellites to a receiver by measuring the travel-times of the radio-signals, and multiplying them by the speed of light. If the simple oscillator in the GPS receiver, for instance embedded in a smartphone, has a stability of only $10^{-6}$, how quickly does the bias in the observed pseudorange increase (or decrease)? To say, what is the rate of change in the pseudorange, induced by the oscillator in the receiver?

**Answer 1** The answer is simply evaluated as $10^{-6}$ s/s, multiplied by the speed of light $c$ $3 \cdot 10^8$ m/s. A stability of $10^{-6}$ means that the clock looses (or gains) 1 $\mu$s every second, and 1 $\mu$s is equivalent to 300 m in terms of range.

# 22

# Lidar

Laser ranging relies on electromagnetic signals with wavelengths in the range roughly from 500 nm (visible light) to 1500 nm (short-wavelength infrared), see Section 21.1.1 on the spectrum. Most often red and green colored laser light is used. Lidar stands for 'light detection and ranging'. A laser signal is sent, and the purpose is to detect the echo from an object. Both principles of ranging are applied with laser ranging: pulse based ranging as discussed in Section 20.1.1, and phase based ranging as discussed in Section 20.1.2. In this chapter we cover lidar remote sensing.

As lasers operate in the *optical* domain, a clear line of sight is needed from instrument to target. Dust, smoke, fog and clouds hamper the propagation of the laser signal. In particular green lasers can penetrate into water to some extent, up to several tens of meters in clear water, and this is useful for bathymetry of shallow waters, for instance in coastal areas.

## 22.1. Laser ranging

Laser ranging basically amounts to determining the *distance*, by measuring the travel-time of the laser signal, from instrument to target and back, hence, the two-way travel-time, see Figure 18.5. The principle of a laser ranger is similar to that of an EDM on a total station (Section 4.2), though typically with laser ranging no special reflector is needed — the object under survey itself reflects the laser light. The amount of reflected laser light depends on certain properties of the object, for instance surface material and color. The instrument can be mounted in an aircraft, helicopter or satellite, and point down to the Earth's surface. While the air- or spacecraft moves forward, laser ranging is performed, and a profile of the terrain is obtained. This is referred to as laser altimetry, see Figure 22.1.

Laser beams can be very narrow, with little divergence only. Over distances of tens of

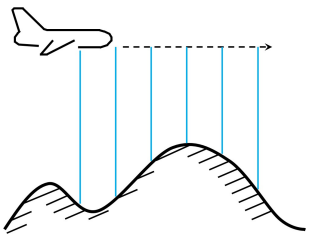


Figure 22.1: Laser altimetry: with a downward pointing laser instrument on an aircraft a profile of the terrain is obtained, once position and attitude of the aircraft are determined.

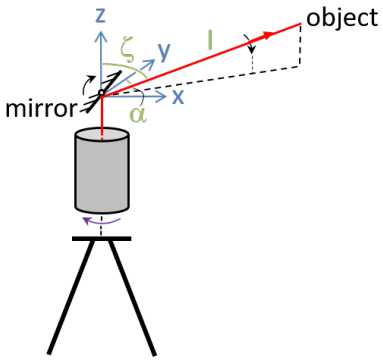



Figure 22.2: Principle of a laser scanner: with measurement of vertical angle $\zeta$, horizontal angle $\alpha$ (with respect to the y-axis), and distance $l$ to an object, 3D-position coordinates, in the local laser scanner coordinate system $(x, y, z)$ are obtained as: $x = l \sin\alpha \sin\zeta$, $y = l \cos\alpha \sin\zeta$, and $z = l \cos\zeta$, cf. Figure 29.6. This is the principle of *spherical coordinates*.

meters the footprint diameter (beam diameter) is only a few millimeter, and for example, for a spaceborne laser (the Global Ecosystem Dynamics Investigation (GEDI) instrument carried by the International Space Station (ISS)) at 400 km orbital altitude, the footprint diameter on the Earth's surface is 30 meter. With these figures we assume that the reflecting surface is perpendicular to the inciding laser beam.

In addition to the distance, also the amplitude of the response is recorded, which is a measure for the reflectivity of the target surface. Snow for instance reflects very well, and trees and sand reflect moderately. Generally white objects are better laser reflectors than black objects. In this respect it should also be noted that sending a laser signal from overhead onto a forest (e.g. with airborne laser altimetry) generally yields multiple echoes: part of the signal's energy is reflected by the top of the canopy (tree crowns), another part by lower branches and bushes, and finally the last part by the ground surface.

## 22.2. Laser scanning

In this section we introduce the principle of laser scanning. By means of a simple example we analyze the precision of geometric information obtained with laser scanning, and we briefly the interpretation and georeferencing of the obtained point cloud.

### 22.2.1. Principle

A laser scanning instrument can measure, much similar to a total station covered in Chapter 4, the horizontal and vertical angle, as well as the distance to an object. The principle of a laser scanner is shown in Figure 22.2. The instrument is set-up local level, such that the z-axis is pointing to the zenith (and the axis is aligned with the direction of the local gravitational acceleration). The optical unit of the instrument, also referred to as the head, rotates about the vertical z-axis, such that horizontal angle $\alpha$ varies, and objects around can be measured. The mirror, shown in Figure 22.2, deflects the laser signal (in red) into a direction with a vertical angle of $\zeta$ with respect to the local zenith. The mirror is spinning such that angle $\zeta$ varies, and objects around can be measured. The laser signal returned by an object (not shown here) basically travels the same path back into the instrument to a photo detector.

A terrestrial laser scanner can scan 360 degrees around (horizontally, angle $\alpha$), and 135 degrees (or more) (vertically, angle $\zeta$); field of view. This allows one to build a full, three-dimensional picture of the environment — the whole scene is surveyed in just a couple of

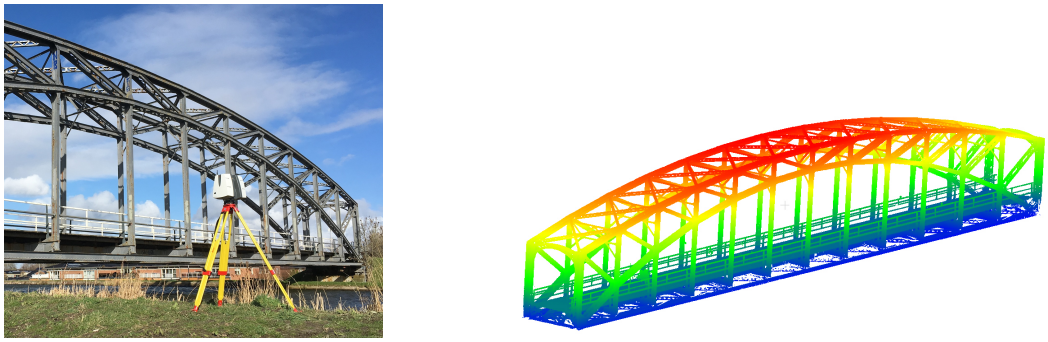


Figure 22.3: At left, Leica P40 terrestrial laser scanner in front of De Trambrug, an iron-arch bridge in Schipluiden, near Delft, crossing the Vlaardingse Vaart. The measurement principle of a terrestrial laser scanner is similar to that of a total station, and consists of measuring vertical angle, horizontal angle, and distance to objects. Measurements with a laser scanner are acquired at an astonishingly high rate! This laser scanner can measure up to 1 million points per second, with a maximum range of 270 m. At right, combined point cloud resulting from 14 scans with the laser scanner occupying positions on both sies of the bridge, and also on the bridge deck. The full 3D-model shown here contains about 42 million points. Photo at left [56] and point cloud at right [57] by Linh Truong-Hong, 2019.

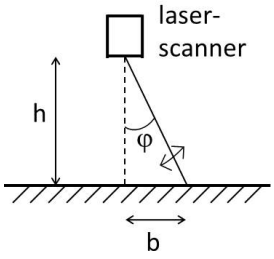



Figure 22.4: Set-up of a terrestrial laser scanner. Error analysis for a special case (vertical cross-section), with a fixed (known) height $h$, scanning objects on a flat ground surface. An error in the measured angle $\varphi$ then translates into an error in distance $b$.

minutes. Such a scanner is referred to as a panoramic scanner, see the instrument shown in Figure 22.3 at left. An example of such a full scan, resulting in a so-called *point cloud*, is shown in Figure 22.3 at right. Laser scanning generally yields large amounts of data.

### 22.2.2. Example: analysis of scanning precision

As an example we consider a special case in a two-dimensional situation, shown in Figure 22.4. The scanner is set-up at height $h$, above a flat ground surface; height $h$ is assumed to be perfectly *known* here, and the reference direction is straight down (direction $\varphi = 0$).

The positions of objects on the ground surface, that is, distance $b$, can now be determined by just measuring angle $\varphi$ (for simplicity we do not use the measured slant distance in this example). Distance $b$ is obtained through

$$b = h \tan \varphi$$

A high-end laser scanner has an angular precision of about $\sigma_\varphi = 0.003^\circ$ ($\approx 5 \cdot 10^{-5}$ rad). The goal of this example is to propagate the standard deviation of angle $\varphi$, into the standard deviation of $b$. As the above relation is non-linear in $\varphi$, we need to take recourse to the

| $\varphi$ [°] | $\sigma_b$ [mm] |
|---|---|
| 0 | 0.1 |
| 15 | 0.1 |
| 30 | 0.1 |
| 45 | 0.2 |
| 60 | 0.4 |
| 75 | 1.6 |

Table 22.1: Standard deviation of coordinate $b$ for various values of scan angle $\varphi$, for the set-up of Figure 22.4, with $\sigma_\varphi = 0.003°$ and $h = 2$ m.

variance propagation law for the non-linear case (7.12) in Section 7.4. With

$$\frac{\partial(h \tan\varphi)}{\partial\varphi} = h\frac{1}{\cos^2\varphi}$$

we obtain

$$\sigma_b^2 \approx h\frac{1}{\cos^2\varphi}\ \sigma_\varphi^2\ h\frac{1}{\cos^2\varphi}$$

and hence

$$\sigma_b = h\frac{1}{\cos^2\varphi}\sigma_\varphi$$

with angular standard deviation $\sigma_\varphi$ in radians. Table 22.1 shows the resulting standard deviation coordinate $b$ for various values of angle $\varphi$. This shows that the uncertainty in the direction (angle $\varphi$) alone, has an impact on the horizontal coordinate which gets much larger with larger angle $\varphi$, i.e. further (sidewards) away from the scanner. Directly underneath the scanner, $\varphi = 0°$, we have $\sigma_b = h\sigma_\varphi$, and the standard deviation of $\varphi$ scales just linearly by height $h$ into the standard deviation of coordinate $b$.

### 22.2.3. 3D laser scanning

In a laser scanner, the laser light is directed in virtually any possible direction, physically, by a rotating mirror, cf. Figure 22.2. The horizontal and vertical directions are measured very accurately, typically with a precision of 0.003°. The laser beam width is typically larger by a factor of 2. Generally the laser scanner is programmed to scan (a large number of) discrete directions, with a particular (user selectable) angular increment between them. The smallest possible angular increment, or resolution, is generally in the order of 0.01°.

The angular increment, together with distance, determines the *point density* on the surface to be scanned (assuming here that the object surface is perpendicular to the laser beam). With $\alpha$ the angular increment, and instrument height $h$ (cf. Figure 22.4), the distance $D$ between two points is given by Eqs. (24.1) and (24.2), see also Figure 24.3. The density scales, in one dimension, with $\frac{1}{D}$, and in two dimensions with $\frac{1}{D^2}$.

A professional terrestrial laser scanner can reach a distance accuracy of a few millimeters, over distances up to several tens of meters. The maximum distance can be hundreds of meters, to even several kilometers.

For a detailed full survey of for instance a structure, as in Figure 22.3, a laser scanner is set-up (static) at several (discrete) positions around the building.

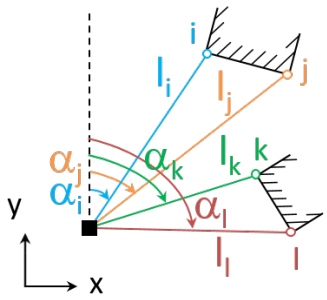



Figure 22.5: Topview of a point cloud obtained by a laser scanner (as shown in Figure 22.2, set-up at the black square). Distances $l$ and horizontal angles $\alpha$ are observed to target (object) points i, j, k and l.

### 22.2.4. Point cloud

As suggested by Figure 22.5, showing a horizontal cross-section, knowing the position and orientation of the laser-scanner instrument, and measuring horizontal direction, as well as distance, allows one to determine the position of the target point, here in two dimensions (polar coordinates), and in practice, when also measuring the vertical direction, in three dimensions (spherical coordinates). The result of scanning the environment with a single set-up of the instrument, is a *point cloud* in the local coordinate system of the laser-scanner. It is just a (large) collection of individual points, which result from signal reflections on surfaces of objects around the scanner. The points initially bear no relation between them (in principle we do not know, even for points close together, whether they belong to the same physical object or structural element, or not).

In processing laser scan data, two steps are typically taken. First, point clouds are aligned with an existing 3D-model, or with other point clouds, see Figure 22.3 at right, where 14 point clouds were aligned and connected together. In the example of Figure 22.6 two point clouds have been observed (from two different positions), one is shown in blue, the other in red. To connect and align point clouds and to mark reference and control points sometimes retro-reflective targets are used, similar to the one in Figure 4.18 at right. Then, point clouds can be, quite directly, visualized (rendered), and human interpretation of the result may be quite appealing. Though an automated mathematical reconstruction of the surfaces and objects is generally challenging. A segmentation is done in order to group points using some similarity criterion. Points in a point cloud can be converted into a three-dimensional surface by means of Delaunay triangulation (see Chapter 11), resulting in a triangular mesh-model. Alternatively, assuming that multiple points physically lie on the same surface, one may want to fit geometric primitives, like planes, spheres and eventually cubic volumes to multiple neighboring points. In some applications one detects and fits cylinders or cones to the observed data, or even other types of volumes. By separating terrain points from non-terrain points a Digital Elevation Model (DEM) of the terrain, or a seafloor model can be created (cf. Chapter 24).

Good progress has been made in automatically fitting structural elements like beams and walls to the millions of distance and direction measurements, linking points to straight lines, circle arcs and planes, enabling automated gathering and generation of structural models. Laser scanning can be used to create a 3D-model of an existing building or structure, for which for instance no model or design is available (anymore).

Repeated laser scan surveys can be used to *monitor* structures, for instance to detect deformation or damage, as cracks and loss of material surface and volume.

Figure 22.6: Example of laser scanning the Symbio-bridge. Two point clouds, one shown in red, the other in blue, were merged, resulting in the visualization at right. This bridge — designed by TU Delft student Rafail Gkaidatzis — was opened in 2016, in the Delft Technopolis Science Park area. Point cloud image at right [58] by Roderik Lindenbergh, 2018.

### 22.2.5. Georeferencing

As shown with Figure 22.2, measurement of vertical angle $\zeta$, horizontal angle $\alpha$, and distance $l$ to an object, 3D Cartesian coordinates of that object become available, in the local coordinate system of the laser-scanner instrument. The position of the laser-scanner may be unknown, and the $y$-axis will generally have an arbitrary unknown orientation (and for instance not be aligned with 'map North' of the national coordinate system, see also Figure 9.2). Furthermore one may wish to combine point clouds obtained with a laser scanner at different positions, and possibly with different orientations.

This calls for georeferencing the point cloud, much similar to dealing with the exterior orientation of a photo camera in Chapter 19, and Section 19.4 in particular, linking the instrument coordinate system to the terrain coordinate system. The three-dimensional similarity transformation to do this, is covered in Section 28.3. Generally, the scale factor in this transformation will be $\lambda = 1$ or $\lambda \approx 1$.

## 22.3. Application: AHN

In Figure 22.1 we considered a laser ranger, with a fixed looking direction, mounted on an aircraft. In practice also a laser scanner can be mounted on an aircraft or helicopter, though scanning is then typically restricted to one direction, sidewards, as the motion of the platform will cover the other direction. This is referred to as *line-scanning* - the terrain is scanned, line-by-line, as the aircraft moves forward. Similarly such a laser scanner can be mounted on a road vehicle for the application of mobile mapping built environment.

With the principle of laser altimetry shown in Figure 22.1 a DEM can be created, which can be considered as a 'blanket draped over the terrain'. The laser signal reflects on the top surface of objects, and thereby underpasses under buildings are not present in these models. They are also referred to a 2.5D models, rather than 3D models; each (horizontal) position has one and only one height.

In a *Digital Terrain Model* (DTM) commonly the ground-level surface is represented (maaiveld) — the bare ground surface without any objects. A *Digital Surface Model* (DSM) includes also natural and built features and objects of the environment. A Digital Elevation Model (DEM) is used as a generic term for both DTM and DSM. You would use a DTM for flooding analysis and drainage modelling, and a DSM for landscape and city modeling.

The Actueel Hoogtebestand Nederland (AHN) has been collected with airborne laser scanning. The initiative was taken in 1996 and the AHN-2 has been flown from 2007 to 2012, for all of the country, with a point density of 6 to 10 points per square meter. The flying height of the aircraft or helicopter is typically around 400 meter. The elevations in the AHN pertain to the terrain ground surface ('maaiveld'), the so-called filtered version (DTM). In the unfiltered

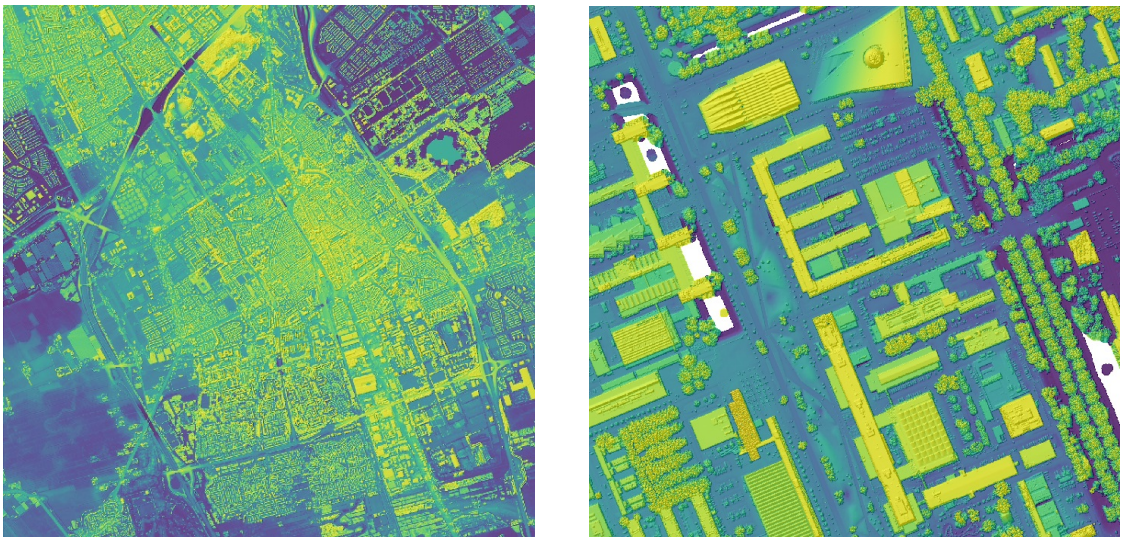


Figure 22.7: Example of laser scan product (AHN-3). The height per pixel is shown in color-scale from blue through green and yellow to red, with heights ranging in this example from -7 m to 97 m (NAP). The image at left shows the area of Delft, about 7 km x 7 km, with the old city center in the middle, and North up (5 m x 5 m pixel). The image at right shows a more detailed scene of the TU Delft campus, with the Aula and library on top (0.5 m x 0.5 m pixel). The images present the Actueel Hoogtebestand Nederland (AHN) - version 3, at a half meter resolution, unfiltered (with vegetation and buildings present) - Digital Surface Model (DSM), shaded relief. AHN by Rijkswaterstaat; data retrieved from PDOK [59] under CC0 license.

version also vegetation is present (DSM).

The AHN-2 is available as a Digital Elevation Model (DEM) with a 0.5 m grid-size. In total this amounts to over 135.000.000.000 points. The elevations have a precision of about 5 cm (standard deviation), with at most a 5 cm of systematic offset. Since 2014 the use of the AHN is for free.

Data collection for the AHN-3 started in 2015, and was released in 2019, also at a 0.5 m grid-size. More information about the AHN can be found at [60]. The AHN-4 will have an even higher point density (up to 20-25 points per square meter) and will be completed in 2022 for all of the Netherlands. An example of AHN-3 of the Delft-area is shown in Figure 22.7.

# 23

# Radar

In this chapter we consider radar remote sensing, and in particular we focus on radar interferometry. This is a rapidly developed new technology for Earth observation — it started to emerge in the nineties of last century, see e.g. [61]. The technique is very much suited to detect and precisely measure *deformations* of the Earth's surface and topography, including the built-environment and infrastructure. Surface changes between subsequent images can be measured at millimeter scale, thanks to the centimeter wavelength carrier radio waves.

As discussed in Chapter 20, radar and Synthetic Aperture Radar (SAR) can be used to create a two-dimensional image. The radar is side-looking and illuminating a strip of the Earth's surface, and as the instrument is flying forward, it creates, strip by strip, a two-dimensional array of pixels. In this chapter we focus on extracting *geometric* information from these radar images.

## 23.1. Fractional phase of radar signal

In addition to what is stated in Section 20.3 about recording the intensity/amplitude, also the *phase* of the carrier wave of the incoming signal for each pixel is recorded, with the phase being a (precise) measure for the geometric range. The use of phase (or phase difference) for the measurement of distance was outlined in Section 20.1.2. The travel-time of the signal from transmitter to reflector, and back to the receiver, determines the phase difference of the received radar signal carrier and the carrier on board the transmitter/receiver. As the used radar wavelength is generally in the order of a few centimeter, cf. Chapter 21, the distance can be measured, in principle, very precisely (in the order of millimeters). Though, generally the number of full wavelengths occuring in the difference, the so-called ambiguity, $k$ in (20.4), is not known. Only the *fractional* phase difference is measured. Any actual/physical phase difference (between incoming and outgoing signal) is cut to fit in the interval $[-\pi, +\pi\rangle$ radians; the actual phase difference got *wrapped* into this interval.

Figure 20.4 showed the amplitude for each pixel, in a gray-scale image. Figure 23.1 shows the corresponding phase. The image seems to show just random phase values. Different pixels in the terrain will have different heights, thereby different distances to the radar sensor, and hence lead to different phase values. As the ambiguity per pixel is unknown, and only the fractional phase is shown, the image looks randomized. No useful interpretation can be given to this image.

As will be clear from Figure 20.5, a pixel does not represent a single discrete point in the terrain, but instead a certain *area*. Typically, for spaceborne radar, the pixel size on ground is in the order of a few meters to tens of meters. Consequently the amplitude and/or phase

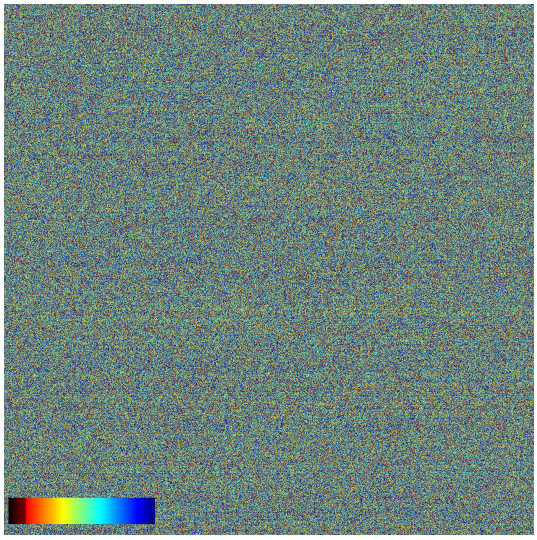


Figure 23.1: Example of radar phase image. Fractional phase per pixel is shown in a color-scale from red to blue, representing $-\pi$ and $+\pi$ radians respectively. The image shows the area of Delft, acquired, on Nov 1, 2015, by the DLR TerraSAR-X satellite [53], cf. Figure 20.4. The frequency of the radar signal is 9.65 GHz (X-band), and the corresponding carrier wavelength is 3.1 cm.

for a pixel result from measurement on the summation of all reflections from that area. For a pixel, two extreme cases of reflection can be distinguished: point scattering and distributed scattering. With point scattering, a strongly reflecting object (resulting in a large amplitude signal) is dominating the measurement; for instance there is a facade or roof of a building within the pixel area on the ground, at appropriate angle, to reflect a lot of the incident radar signal. Such an object is referred to as a *point scatterer*. With distributed scattering a large number of small scattering objects form the total response together; for instance a dike with a top layer constructed of cobble stones, see Figure 23.2.

## 23.2. Radar interferometry

A radar, as it is based on the measurement of *distance*, may not be able to distinguish between objects at different locations with different heights: the signal response from a low building near the satellite ground track and the response from a tall building further away from the ground track may arrive at the very same time, and end up in the very same pixel, see also the discussion with Figure 20.10.

In order to resolve objects at different elevations, another radar image is needed, which is

- acquired at the same time from a different orbital track (e.g. by a tandem radar satellite mission - single pass interferometry), or,
- acquired at a separate time with the same satellite (hence, at another pass-over of the satellite - multi-pass interferometry)

Then the principle of *interferometry* is applied, cf. Figure 18.9. An interferometric image is created, in which the *phase* measurements of the *two images*, with slightly different imaging geometries, are combined (effectively differenced). The solution of using *two* images, rather than one, is the same as applied to photogrammetry. With a single optical photo, objects in the same *direction* (but possibly at different elevations) can not be distinguished, and therefore a second optical image is used (taken not too far away from the first one), to allow for three-

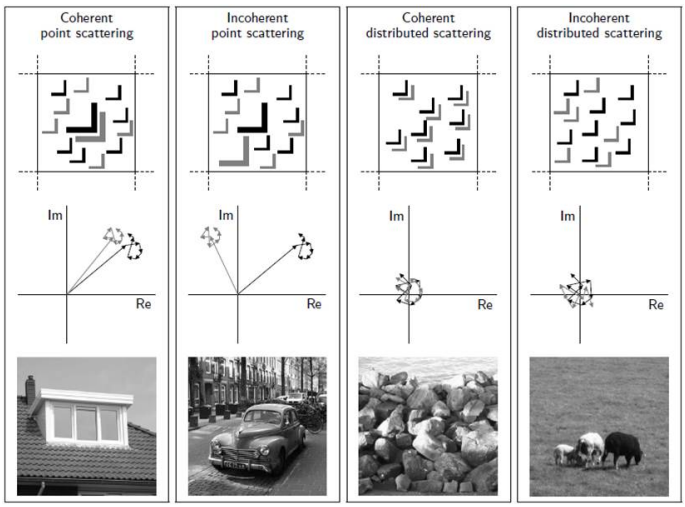



Figure 23.2: Classification of reflecting objects in radar imaging [62]. A point scatterer is a strongly reflecting object, as the roof of a building. Distributed scattering occurs for instance on a dike with cobble stones. Coherent scatterers are permanent — they are constantly present in sequences of radar image; incoherent scattering is caused by objects which may move in between two image acquisitions.

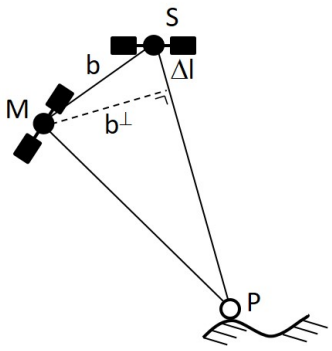



Figure 23.3: Geometry of radar image acquisition for interferometry. The positions M and S indicate the two radar sensor positions (forming the so-called baseline $b$), and point P is the - to be imaged - object in the terrain. In reality, this plane, with points M, S and P, is positioned in the three dimensional space.

dimensional reconstruction of the imaged Earth's surface and topography (see Section 19.3 — stereo-photogrammetry/vision).

The imaged scattering objects preferably should not change between the two radar image acquisitions — the scattering should be *coherent*. The images should be sufficiently similar. Therefore — to form an interferogram — the two images should have only a slightly different geometry, as the radar backscatter on objects is generally much dependent on the viewing direction. Vegetation is a typical example of yielding incoherent scattering, as it will change over time (grow). The objective of radar interferometric analysis is the retrieval of (geometric) information from pixels showing sufficient coherent scattering.

Figure 23.3 shows the geometry of two radar image captures, to create a radar interferogram. Assuming a small angle at the object P, the difference in range from M to P, and S to P is $\Delta l$. This difference is represented by the interferometric phase difference, similar to Eq. (20.5). However, as noted in the previous section, only the *fractional* phase difference (between the two images) can be measured, and there is an ambiguity present, as indicated in (20.4) and (20.8); there is an unknown number $k$ of full wavelengths involved.

Figure 23.4 shows an example of a radar interferogram, constructed from two radar (phase) images, of which an example was shown in Figure 23.1; the interferogram shows actually the phase difference (the two contributing phase measurements are still wrapped, and the phase difference is kept within the interval from $-\pi$ to $+\pi$).

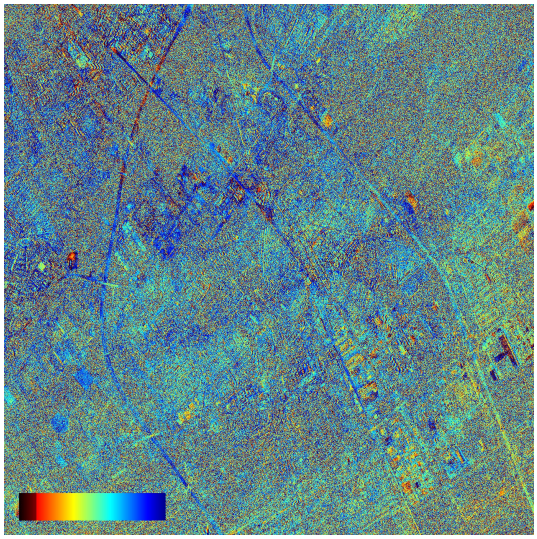


Figure 23.4: Example of radar interferogram. Fractional phase per pixel is shown in a color-scale from red to blue, representing $-\pi$ and $+\pi$ respectively. The image shows the area of Delft, about 7 km x 7 km, with the old city center in the middle, and (approximately) North up, acquired by the DLR TerraSAR-X satellite [53], cf. Figure 23.1. The interferogram follows as the *difference* in phase between two image acquisitions, namely on Nov 1 and Nov 12, 2015.

Neighboring pixels in the terrain may have different heights, and this leads to the conclusion that, in principle, every pixel will have its own unknown phase ambiguity; the phase ambiguities, even of neighboring pixels, may be unrelated. Phase unwrapping is needed to obtain a continuous interferometric phase map. Though some topography can be recognized already in Figure 23.4, still not too much information can be retrieved from just a single interferogram alone.

### 23.2.1. Digital Elevation Model (DEM) [*]

As a side step, we mention that from a radar interferogram, constructed from two radar (phase) images, in principle a Digital Elevation Model (DEM) can be reconstructed. With two different satellite positions, as shown in Figure 23.3, essentially a 'stereoscopic' effect is realized (based on a resection with *distances*). The phase difference from one pixel to the next, may be different in the two images, and this allows one to reconstruct a height difference.

The longer the baseline between the two image acquisitions (cf. Figure 23.3), the larger this 'stereoscopic' effect is, i.e. the better height differences in topography can be observed. The height difference accuracy is currently in the order of meters to tens of meters. The counter-effect of a longer baseline is that phase unwrapping gets more difficult.

## 23.3. Measuring deformations

In differential interferometry the effect of topography is removed by modeling the phase using a reference surface or an existing Digital Elevation Model (DEM) and the orbit parameters of the satellite. Hereby the 'stereoscopic' effect is removed from the interferogram. What remains are *phase differences* due to *changes* in the positions of objects or the Earth's surface level

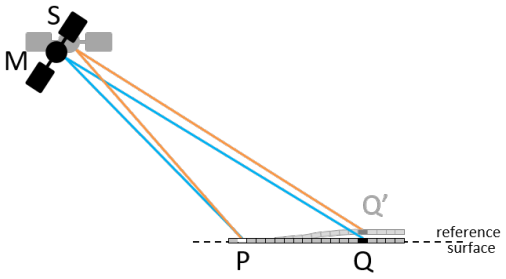



Figure 23.5: Principle of radar interferometry for deformation monitoring. The radar takes an image of the surface with points P and Q in the original state, indicated in blue, and another image, indicated in orange, after a slight land-uplift has taken place in the area around point Q (point Q moved to point Q').

(deformation). In case of small changes they are (well) within a wavelength (of the used radar signal). Otherwise assumptions about the spatio-temporal smoothness of the deformation are needed to resolve the phase ambiguities. A sudden jump, of one full wavelength (or multiples), in one pixel may go unnoticed.

In Figure 23.5 the principle of measuring deformation of the Earth's surface is shown. In this simple example we omit the fact that the phase measurement is ambiguous; for the moment the distance readily follows as (20.5). Based on the first image acquisition, shown in blue, from satellite position M, the received radar signal response is mapped onto a row of pixels using a simple flat plane as a reference Earth's surface here (cf. Figure 20.5). Similarly, once point Q has moved to Q', while point P is a stable reference point, a second image is acquired, shown in orange, from satellite position S. In this example satellite positions M and S nearly coincide. Though usually the so-called baseline $b$ between M and S is short, cf. Figure 23.3, in practice the interferometric phase is corrected for the difference in viewing geometry by using a reference surface. After that, the deformation of point Q to Q' (for its component in the direction of the satellite) can be retrieved, based on the phase difference between pixel P and Q observed at M, and the phase difference between P and Q' at S. As radar relies on the measurement of distance, object or surface position changes in the direction perpendicular to the look-direction (line of sight) of the radar, can*not* be detected (that is in the direction of $b^{\perp}$). With differential interferometry deformations are obtained in a *relative* sense, i.e. with respect to another point within the image (Q with respect to P).

Radar interferometry can offer full spatial coverage of the imaged area, rather than so-called point measurements as for instance taken with GPS. And, as a remote sensing technique, radar interferometry can do this remotely. Radar signals penetrate through clouds, and hence the technique offers all-weather operation, day and night.

A couple of aspects needs attention though. The first one is spatial decorrelation, and refers to the fact that from different viewing angles, the same scene may look different. The viewing directions of two images used to form an interferogram should be not too different, in order to maintain sufficient coherence of the interferometric phase. In Figure 23.3 the distance between points M and S should be not too large. The second one is temporal decorrelation, and refers to the fact that the scene itself may change over time, and hence coherence of the interferogram may get lost — the two images got too different. For instance in a vegetated area, vegetation may grow, and the scene no longer looks the same. Current radar satellite missions have a so-called repeat time in the order of several days to several weeks. The third aspect is about atmospheric delays. Radar signals travel from a satellite, through the Earth's atmosphere, to the Earth's surface (and back), and the signal may, thereby, get extra delayed,

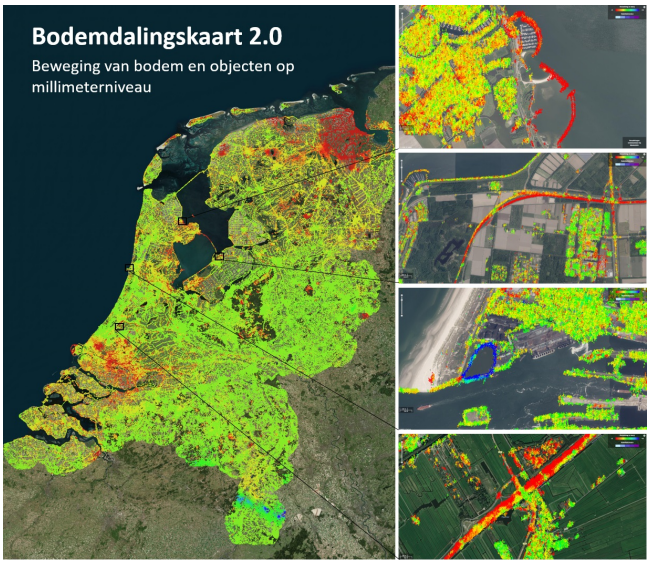



Figure 23.6: The Dutch Ground Motion Service (in Dutch: Bodemdalingskaart Nederland): the map shows the rate-of-change in millimeters per year, of land surface level, with millimetric precision, covering all of the Netherlands, and is updated a few times per year. Image courtesy of Nederlands Centrum voor Geodesie en Geo-informatica (NCG) [64].

resulting in an additional phase offset. The atmospheric delay may change over the area, and hence the corresponding phase delay may vary across the image, and, it will also change over time, just like the weather. Since the time interval between two image acquisitions is generally one week or more, the atmospheric phase delay is typically uncorrelated here in time (this is likely also the effect we see in Figure 23.4).

Persistent Scatterer Interferometry (PSI) is a multi-epoch interferometric radar analysis technique to overcome — to some extent — spatial and temporal decorrelation. PSI relies on consistently *coherent* scatterers, which are typically buildings (in urban areas) or infrastructure of concrete and steel (like roads and rail), with high-amplitude reflections. They are pixels which can be easily identified in the image, and represent discrete objects, referred to as *point*-wise coherent scatterers. They represent relatively stable ground targets, and these permanent scatterers pose minimal spatial and temporal decorrelation through a whole stack or *time series of images*. PSI can use all acquired images, hence a full time series over a period of months or even years, to estimate the kinematic (deformation) behaviour for consistently coherent points. It allows for monitoring deformation behaviour of for instance buildings, embankments, dikes, dams and railroads with sub-centimeter accuracy.

### 23.3.1. Dutch Ground Motion Service

As an application of differential interferometry, or differential Interferometric Synthetic Aperture Radar (InSAR), next to GPS and absolute gravity measurements, the 'actuele bodemdalingskaart Nederland' (The Dutch Ground Motion Service - surface and object motion map of the Netherlands) was launched in 2018, see [63].

Pumping water out of polders, extracting natural resources and construction of underground infrastructure, like a tunnel, may all cause subsidence of the land surface, and filling old mines with water, water suppletion in closed underground mining systems, may cause actually a rise of the land surface. This is shown in Figure 23.6.

# 24

# Sonar

As mentioned in Section 21.1.2, electromagnetic waves do not really propagate well in water, and therefore acoustic signals are used instead for remote sensing in water. In this chapter an introduction is given to acoustic sounding. An acoustic signal is transmitted, it travels forth to a reflector (e.g. the seafloor), gets reflected, and travels back to the receiver, just as shown in Figure 18.5. The distance is measured by means of two-way ranging. This is still the main idea here with active sonar.

In this chapter we consider single beam and multi-beam echosounders, and address the concept of footprint. We touch on seafloor surveying of the North Sea. Then optionally, we go in more detail on the shape of the pulse, once it got reflected by the seafloor and received by the sonar, that is, we consider the shape of the received echo. Finally we briefly address seafloor classification.

Some preliminaries on the audio frequencies and the propagation of acoustic waves in a medium, like water, are covered in Section 21.1.2, and in Section G.3 of the Appendix respectively.

## 24.1. Introduction

In order to create hydrographic maps and Digital Elevation Models (DEM) of the seafloor, for instance for the purpose of underwater construction works (e.g. offshore platforms, pipe-laying and dredging), bathymetric surveys are performed. Bathymetry is concerned with the underwater depth of lake-, river-, sea- and oceanfloors. The name '*bathymetry*' has its origin in Greek, and literally means 'measuring depths'. Its purpose is to acquire information about the underwater topography. The reference level for measuring depth at sea is covered in Section 35.4.

In the old days, bathymetry involved the measurement of ocean depth through depth sounding. A pre-measured heavy rope was lowered over the ship's side until it touched the sea bottom. Today sounding is done using sonar. The title of this chapter could read 'sounding using sound', to refer correctly to the process of measuring water depth using acoustic signals. A type of sonar is the echo sounder (in Dutch: echolood), which is an instrument mounted beneath, or over the side of a boat, transmitting an acoustic signal, a so-called 'ping', downward to the seafloor, and measuring the two-way signal travel-time, to determine the distance (again cf. Figure 18.5, but then in a vertical sense, rather than horizontal). The echo sounder is generally looking straight down (similar to laser altimetry in Figure 22.1), but it can also look sidewards (side-scan), or in multiple directions simultaneously (multi-beam).

Figure 24.1 at left shows two compact and portable echo sounders for use in shallow water,

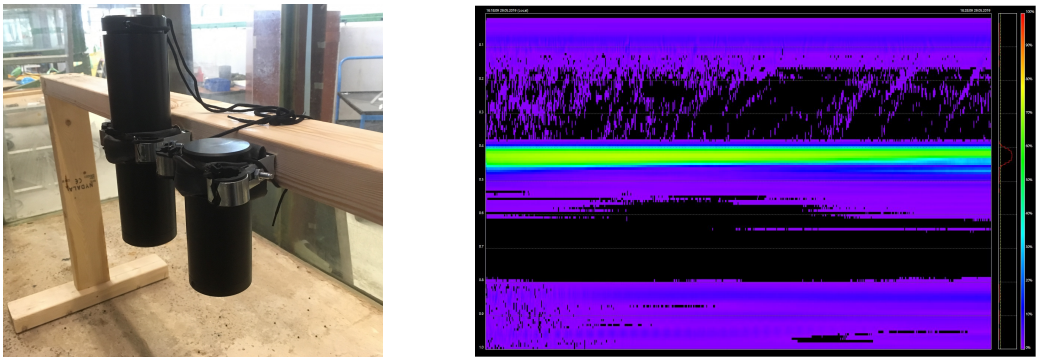


Figure 24.1: At left: Echo logger EA400 and AA400 echo sounder installed in tank for testing. Next, a sandbed is supplied on the floor of the tank, and the tank is filled with water. At right: actual depth measurements by the EA400 instrument in the tank as a function of time (intensity of the received signal is shown according to the color scale at right). The sandbed surface is about 0.40 m below the echo sounder. Photographs by Ruurd Jaarsma [65], July 2019.

e.g. for sediment monitoring. They work at a frequency of 450 Hz, with a beam-width of 5 degrees (single beam), and their operating range is from 0.15 m to 100 m (50 m).

The graph at right, in Figure 24.1 shows the depth measurements as a function of time (time axis horizontal, covering 10 minutes of time). The vertical axis goes down from 0.0 to 1.0 m. The graph shows, in this case, a clear single reflection on the sandbed, at a depth of about 0.4 m. The intensity of the received signal is indicated by color, according to the color-bar at right, from blue, through green, to red.

In practice, the depth is measured from a shipborn sensor, and this means that, in order to deliver useful results, during bathymetric surveys the position (and attitude) of the ship or vessel need to be determined, its draught, and possibly also the actual sea-level as well. These subjects are not covered in this chapter.

## 24.2. Single Beam Echo-Sounder (SBES)

In practice, one can not create an infinitesimally small (narrow) beam with acoustic sounding. First of all the transducer antenna has finite dimensions, and secondly, the radiation pattern of the antenna does not allow for a transmission in one single discrete direction, cf. Figure 21.5. The signal beam has a certain width, indicated by angle $\alpha$ in Figure 24.2. Therefore the distance measurement is no longer a point-to-point measurement. The transmission 'illuminates' a certain area of the seafloor, called the *footprint*. For the radius of the footprint area, it holds that $\frac{D}{2} = h \tan \frac{\alpha}{2}$, where we assume, as shown in Figure 24.2, that the acoustic beam incides the seafloor surface perpendicularly. For small angles $\alpha$, we can also write

$$D \approx h \tan \alpha \tag{24.1}$$

Some values are presented in Table 24.1 for a beam-width of 3 degrees. Obviously, the larger the water depth, the larger the footprint.

Typically audio waves with frequencies in the order of 10 - 100 kHz are used for sounding. A larger frequency bandwidth yields a better range resolution and accuracy, see Section 20.1.1 (pulse-based ranging). For precision bathymetry a transducer operating around 200 kHz is used, as it is suitable for up to 100 meters in depth. Deeper water requires a lower frequency (down to 10-20 kHz). The higher the frequency of the audio waves, the more they suffer from absorption/attenuation (see also Appendix G.1.1 on link budget), and the less far they penetrate in the medium (water). A sub-bottom profilers uses an even lower frequency, and

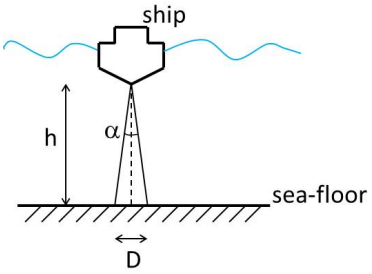



Figure 24.2: Single Beam Echo Sounder (SBES). The beam is typically 2.5 to 3.0 degrees wide, and this leads to a footprint diameter $D$. In this diagram the footprint is a circular area.

| $h$ [m] | $D$ [m] |
|---|---|
| 10 | 0.52 |
| 20 | 1.05 |
| 50 | 2.62 |
| 100 | 5.24 |

Table 24.1: Diameter of footprint $D$ (in [m]) as a function of water depth $h$ (in [m]) for a beam-width of $\alpha = 3^\circ$, according to (24.1).

these waves can penetrate into the seafloor beyond the surface by tens to hundreds of meters. Reflection seismics uses (even) lower frequencies (in the order of 10-1000 Hz) to go even deeper.

## 24.3. Multi-Beam Echo Sounder (MBES)

In practice for large area seafloor mapping, a multi-beam system is used, rather than a single beam. Figure 24.3 shows the principle of a Multi-Beam Echo Sounder (MBES). In this figure only two beams are shown, the one straight down, and on 'side-looking' beam. In practice there are many (tens of) beams, regularly spaced, as to integrally cover a certain swath.

As can be seen from this figure, the footprint diameter for a 'side-looking' beam is larger

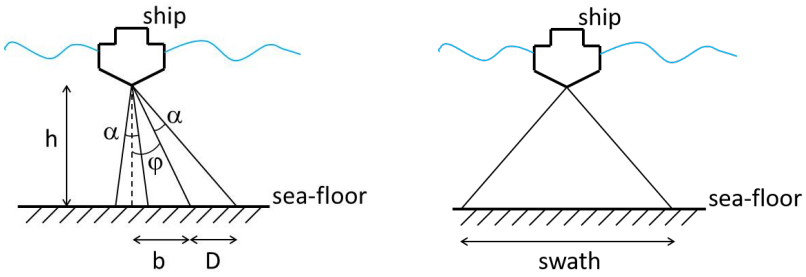



Figure 24.3: Multi-Beam Echo Sounder (MBES). At left, a single 'ping' is transmitted, and by means of a receiver *array*, consisting of multiple antennas, the return signal is split into many different beams. This is done through mathematical signal processing, and relying on a principle similar to the one shown in Figure 18.8 with $d = b \cos\theta$; signals from different directions arrive with different time intervals at Rx1 and Rx2. Beam widths are typically in the order of $\alpha = 0.5^\circ$ to $1.5^\circ$, and the range (angle $\varphi$) can reach up to $75^\circ$ in both sideward directions. At right, the swath covered by the MBES is shown.

| $h$ [m] | $D$ [m] ($\varphi = 0°$) | $D$ [m] ($\varphi = 75°$) |
|---|---|---|
| 10 | 0.17 | 2.79 |
| 20 | 0.35 | 5.57 |
| 50 | 0.87 | 13.94 |
| 100 | 1.75 | 27.87 |

Table 24.2: Diameter of footprint $D$ (in [m]) as a function of water depth $h$ (in [m]) for a beam-width of $\alpha = 1°$, according to (24.2) for look angle $\varphi = 0°$ and $\varphi = 75°$.

than for the beam looking straight down. With $b = h \tan\varphi$ and $\tan(\varphi + \alpha) = \frac{b+D}{h}$ we obtain

$$D = h(\tan(\varphi + \alpha) - \tan\varphi) \tag{24.2}$$

Some values for the diameter are presented in Table 24.2 for a beam width of 1 degree. Note that for a 'side-looking' beam, the area is no longer circular. Note that for $\varphi = 0$, Eq. (24.2) reduces to (24.1).

Using the goniometric identity $\tan(\varphi+\alpha) = \frac{\tan\varphi+\tan\alpha}{1-\tan\varphi\tan\alpha}$, with $\tan\varphi = \frac{b}{h}$ yields, after some manipulations,

$$D = \frac{(h^2 + b^2)\tan\alpha}{h - b\tan\alpha} \approx \frac{(h^2 + b^2)\tan\alpha}{h} = (\frac{b^2}{h} + h)\tan\alpha$$

where the approximation holds for (very) small angles $\alpha$. The diameter $D$ is now expressed as a function of angle $\alpha$.

Surveys with a MBES typically result in point clouds, quite similar to the result of laser scanning in Chapter 22.

## 24.4. Seafloor surveying

Figure 24.3 at right shows that a MBES, when it is regularly measuring, while the vessel sails forward, by its swath-width, will cover a strip of the sea-floor. The area is surveyed much similar as shown in Figure 19.8 with airborne photogrammetry.

Rijkswaterstaat is responsible for periodic surveying and monitoring the North Sea coast, inner waters with large water bodies such as the Waddenzee and IJsselmeer and major rivers and canals, and the approach routes to the Rotterdam and Amsterdam/IJmuiden harbors, see [66].

The yearly survey of the North Sea coast consists of measuring height and depth profiles along transects (in Dutch: raaien) about every 250 m, from just over the first dune into the North Sea, typically up to about 800 m into the water, which corresponds to a depth of about 12 m (-12 m NAP; these water depths are reported in NAP). These profiles are referred to as Jarkus, short for the yearly survey of the coast by means of profiles (in Dutch: jaarlijkse kustprofielen). These measurements are typically done by laser altimetry on shore (Figure 22.1) and by echo sounding offshore, with a precision of better than 1 dm.

Once every few years, a full area survey is carried out of each section of the Dutch North Sea coast down to a depth of about 20 m (in Dutch: vak-lodingen programma). The results are provided as a seafloor map interpolated to a 20 m x 20 m grid.

## 24.5. Received pulse shape [*]

In this analysis we use, as a signal, a simple rectangular pulse with amplitude $A$ and time duration $\Delta\tau$, it is shown in Figure 24.4.

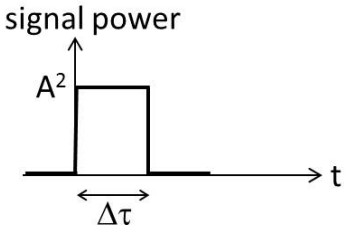



Figure 24.4: Power of transmitted simple pulse signal as a function of time $t$, with amplitude $A$ and duration $\Delta\tau$.

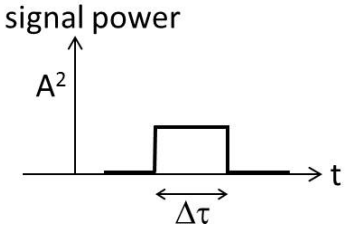



Figure 24.5: Power of received signal, which is delayed and attenuated, but still of the same shape as the transmitted signal.

In ideal circumstances, and considering ranging from point-to-point, the received signal — the echo — is delayed (because of the travel-time). In practice it is also attenuated, as shown in Figure 24.5.

When in addition there is not a single point of reflection, but instead a whole area, the footprint, which reflects the transmitted audio signal, see Figure 24.2, the pulse gets even a different shape.

We assume that the transmit antenna distributes the signal power/intensity uniformly over the full beam width (equally in all directions) The signal first reaches the seafloor in the center of the footprint, at time $t = \frac{h}{v}$, where $v$ is the signal propagation speed. Then, as a function of time, the reflection point moves radially outward; the 'illuminated' area is a circle which gets bigger and bigger. When the radius of the circle is $r$, the slant range is $\sqrt{h^2 + r^2}$, and the pulse just arrives there at time $t = \frac{1}{v}\sqrt{h^2 + r^2}$. For small angles $\alpha$ this can, through $t = \frac{1}{v}\sqrt{h^2(1 + \frac{r^2}{h^2})}$ with $\frac{r}{h}$ small, be approximated (first order Taylor series) by

$$t \approx \frac{h}{v}(1 + \frac{r^2}{2h^2})$$

Hence, once the signal has reached the seafloor the circle radius $r$ increases by the square root $\sqrt{t}$ of time $t$. Thereby, in this approximation, the 'illuminated' area ($= \pi r^2$) increases linearly with time $t$ (and the rate of change is $2\pi vh$). Next, we assume a 100% reflectivity, i.e. all incident acoustic signal energy is reflected back to the receiver on the ship (perfect backscatter). Finally we need to account for the fact that the pulse has to travel back as well to the receiver. For the signal reflected in the center of the footprint this takes (another) $\frac{h}{v}$ seconds, and for the signal reflected at radius $r$ this takes $\frac{1}{v}\sqrt{h^2 + r^2}$ seconds. The received pulse consequently gets shaped as shown in Figure 24.6.

Note that we assumed that the pulse duration $\Delta\tau$ is such that the full footprint area gets

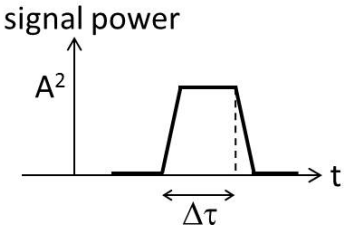



Figure 24.6: Power of received signal, which got deformed due to reflection by the footprint area. Any signal loss in the medium is ignored (no attenuation). The total signal power (area under the curve) is the same as for the transmitted signal shown in Figure 24.4. The slope of the increase is $\pi v h$.

illuminated. If the pulse is short, then the end of the pulse may already arrive in the footprint center, whereas the start of the pulse has no yet left the footprint area (not yet at the boundary) - in that case the shape of the signal in Figure 24.6 is different, but still similar.

The shape in Figure 24.6 already shows that in practice measuring the travel-time (and thereby the depth) by means of leading edge detection on the received signal is less trivial than one would initially assume. The rectangular pulse gets deformed, in these still ideal conditions, and the perfectly right leading edge gets lost.

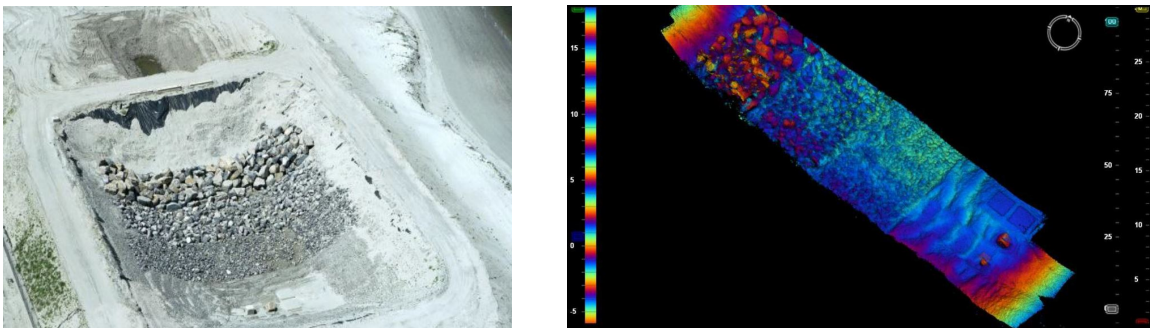


Figure 24.7: At left a calibration dock for acoustic surveys, upon construction and not yet flooded. The rock material of different dimensions can be clearly seen, as well as the concrete blocks and the two 2 m x 2 m calibration plates in the front-center. At right the corresponding image, of the center part, resulting from the acoustic survey, once the dock was flooded. Image courtesy of Eric Peeters, Operations Manager Survey at Van Oord Dredging and Marine Contractors [67]; both images taken from the presentation 'PUMA - Project organisatie Uitbreiding Maasvlakte', 27 September 2012.

## 24.6. Sea-floor classification

Though a bathymetric survey is primarily focussed on measuring depth (i.e. geometric information), one might be interested — through audiometry — to acquire also thematic information. The seafloor can consists of different materials, from sludge, mud and sand, to clay, gravel, cobble stone and rock for instance (according to increasing hardness). Mud has a low impedance contrast with water and consequently yields only little signal reflection, whereas rock has a high impedance contrast with water. Therefore, next to travel-time of the acoustic signal, its amplitude or intensity provides us with information revealing the type of sediment. To obtain even more information from backscatter intensity, the survey can be carried out using *multiple* frequencies, as the (acoustic) impedance depends on the frequency of the audio wave. For example, a higher frequency signal is reflected by the top mud layer, while a lower

frequency signal is reflected only by the underlying hard surface.

The seafloor generally presents relief, also on a local scale, and the surface is not smooth, cf. Figure 24.7. The interaction between acoustic wave and seafloor is also driven by the size of grain and stone/rock. The shape of the echo, as a function of time, as discussed in the previous section may reveal seafloor roughness. With stones and rock for instance, part of the signal may penetrate *in between* the stones and rock, and get reflected by an underlying layer, thereby resulting in a different shape of the total echo.

# 25

# Radiometric sensing

Remote sensing is about collecting information about an object without being in physical contact with the object, that is, collecting information from a distance. So far this part focussed on *geometric* aspects of remote sensing, in order to determine position, shape and/or orientation of objects. For *radiometric* aspects one is concerned with measurement of *radiant energy*. The primary parameter of the received electromagnetic signal is the amplitude (intensity), rather than its time delay or phase offset as in earlier chapters. The subject of this chapter can be termed as 'interpretative' remote sensing.

Optical remote sensing was briefly introduced in Section 18.1.2, and the geometric aspects were covered in more detail in Chapter 19 on photogrammetry. Optical remote sensing makes use of visible, near-infrared and short-wave infrared sensors, cf. Table 21.1, to acquire images of the Earth's surface and its topography by recording solar radiation reflected by the surface and objects on it.

The process of image acquisition, using a sensor array, is shown in Figure 25.1. For every pixel, the brightness, or intensity, in a spectral band of interest, is measured and visualized here in a gray-scale (from zero intensity (0) to maximum intensity (255)). Rather than using a two-dimensional array, satellites often employ a linear or one-dimensional array, perpendicular to the flight direction. Through the forward motion of the satellite, and regularly repeating the acquisition using the single line of sensors, eventually a two-dimensional image is obtained (so-called push-broom image acquisition).

A digital image basically is a two dimensional array of individual picture elements — pixels — arranged in rows and columns. The pixel is the smallest entity or physical unit in the image. Each pixel represents a certain area on the Earth's surface. The pixel has an *intensity* value, and a location address in the two dimensional image (row and column number). In remote sensing the intensity value represents the measured physical quantity, such as the (amount of) solar radiance in a certain wavelength band (e.g. visible light) reflected from the ground, or emitted infrared radiation. In a digital image, the intensity value is stored as a digital number. For example with an 8-bit representation, there are $2^8 = 256$ possible (intensity) values (from 0 to 255). The number of bits determine the *radiometric resolution* of the image.

In this chapter, we cover the interaction of radiation, as used in remote sensing, with an object, be it a brick in road surface, a water droplet in the atmosphere, or a leaf of a tree. Radiometric sensing allows one to make thematic inferences about the object being sensed (as for instance color or water-content of vegetation).

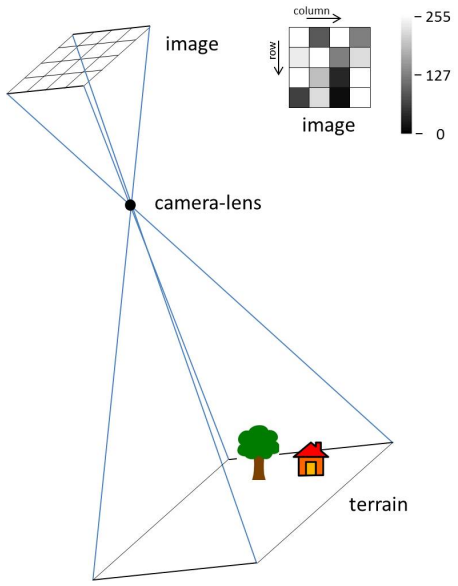



Figure 25.1: Image acquisition using a 4-by-4 array of Charge Coupled Device (CCD) sensors. Each pixel is exposed to incident radiation and builds up an electric charge proportional to the intensity of the incident radiation. The charge of each pixel is subsequently converted into a digital value (represented here by 8 bits), resulting in a digital image, which is actually a *matrix* of intensity values.

## 25.1. Introduction

All objects absorb and reflect light in different ways. The amount of light an object reflects in some wavelengths and absorbs in others, is an indication of its properties and can tell us a lot about that object. As an example, suppose there are two leaves on a table: one is a healthy leaf and another one is a dry leaf (Figure 25.2). You are asked to judge which one is the healthy leaf and which one is the dry one without touching the leaves. Most people would simply come to the conclusion that the green leaf is healthy and the yellow or brown one is the dry leaf, based on its color and texture.

In this example, the photons (light) bouncing back off the leaves to your eyes enable you to determine whether the leaf is healthy, and you do not need to come in physical contact with the leaves to make that judgment. Remote sensing works in a similar way. Remote sensors collect data about objects by detecting the amount of light that is reflected from the objects without having any form of physical contact. Observing and measuring physical (e.g. shape, texture) and chemical attributes of the Earth's surface and the atmosphere help scientists to analyze and understand the changes of the Earth's environment and to develop models and plans for the future.

Sensors onboard remote sensing satellites measure and record the electromagnetic energy in visible light (blue, green, and red light), and also wavelengths of radiant energy that human eyes cannot see (e.g. ultraviolet and infrared 'light'). We classify radiant energy by wavelength, which we typically measure in micrometers (or nanometers). The spectrum of electromagnetic radiation was presented in Table 21.1.

Remote sensors measure the reflected and emitted light from objects at specific wavelengths. 'Spectro-radiometers' are used to detect *specific wavelengths* of light, also called 'bands' or 'channels'. Scientists design 'spectro-radiometers' to be particularly sensitive to the bands that tell them most about the objects of interest, based on their knowledge of how objects interact with light at certain wavelengths. Specific bands reveal a lot of details about vegetation, while other bands are suitable for obtaining information about the ocean surface

Figure 25.2: Healthy leaf vs dry leaf.

or clouds.

Down here on Earth, we can easily identify vegetated areas or distinguish healthy thick vegetation from dying sparse or stressed crops. However, it is not that easy for a satellite observing the Earth from an altitude of about 700 kilometers. Instead, scientists obtain information about vegetation covers from remote sensing images by interpreting the recorded signal (i.e. the *amount* of *energy* received) by remote sensors.

We can distinguish remote sensing in active and passive remote sensing. For *active* remote sensing, the instrument first 'illuminates' the object of interest by transmitting radiation, and next receives back the reflected radiation, as for instance with radar in Section 20.3. With *passive* remote sensing one relies either on reflected radiation of the Sun, or on radiation by the body/object itself (thermal radiation). In this chapter we primarily focus on *passive remote sensing*.

## 25.2. Example: Sentinel-2 imagery

Before we go into more detail on the physics of electromagnetic radiation, this section presents a basic example of radiometric remote sensing using data acquired with the multi-spectral sensor on board the two ESA Sentinel-2 satellites [68]. These two satellites are in a Sun-synchronous orbit at almost 800 km altitude, cf. Section 20.5.

Radiometric measurements are done in specific ranges of wavelength (or conversely, ranges of frequency) of the electromagnetic spectrum, cf. Table 21.1. These spectral windows are referred to as bands. Each of the two Sentinel-2 satellites carries a multi-spectral sensor, which can observe in 13 spectral bands. Figure 25.3 shows the measured intensity in three of these bands, all in the visible light part of the spectrum. These bands each cover typically a bandwidth of 30-60 nm. The intensity (of the reflected Sun-light) is represented (originally) by a 12-bit number, allowing for a measurement range from 0 (nothing received at all) to 4095 (maximum intensity received), in this band, cf. the matrix in Figure 25.1. Actually, with Sentinel-2, the image is captured row-by-row as the satellite is flying forward (covering a 290 km swath, or field-of-view). This is referred to as the push-broom principle. The sensor consists of a linear array of elements or detectors, rather than a full matrix.

Next, a 'true-color' image is created, by combining (merging) the three images acquired in the blue, green and red parts of the visible spectrum. The result is shown in Figure 25.4. Now each pixel carries an intensity value for Red, Green and Blue (RGB).

Similarly, figure 25.5 shows a 'true-color' image of the South-Western part of the Netherlands. This is obtained using three full images, in blue, green and red, as acquired by the Sentinel 2 satellite. One image covers about an area of 110 km by 110 km, at a 10 m pixel

Figure 25.3: Measured intensity in band 2 (492 nm for blue, at left), in band 3 (560 nm for green, in middle), and in band 4 (665 nm for red, at right). Data from ESA Sentinel-2 satellite (Copernicus Sentinel data 2019) obtained through the Sentinel-Hub [1], CC BY-NC 4.0, covering the area of Delft, 7 km x 7 km, with the old city center in the middle, and North up.

Figure 25.4: 'True color' image, created by combining the three images of Figure 25.3. Data from ESA Sentinel-2 satellite (Copernicus Sentinel data 2019) obtained through Sentinel-Hub [1], CC BY-NC 4.0, covering the area of Delft, 7 km x 7 km, with the old city center in the middle, and North up.

Figure 25.5: 'True color' image, created by combining the Red-Green-Blue (RGB) images, of the South-Western part of the Netherlands. Data from ESA Sentinel-2 satellite (Copernicus Sentinel data 2019) obtained through the Sentinel-Hub [1], CC BY-NC 4.0, image taken on August 24th, 2019 (10:56:43 UTC), with Coordinate Reference System (CRS) WGS84 in UTM zone 31N projection.

resolution. The data shown are so-called level 1C data, which is reflectance from the Earth, observed by the satellite in space (see actually (25.5), as a passive sensor relies on reflected Sun-light). Level 1C data refers to the Earth, including the atmosphere. So-called level 2 data is also available, which has been corrected to the bottom of the atmosphere, and is thereby better suited for observation and interpretation of the Earth's surface. In the example shown in this section, a cloud-free day has been selected (cloud coverage of only 0.4%), so that effectively, there is not much difference between the two. L1C data has been re-sampled and geometrically corrected (ortho-rectification).

## 25.3. Physics of electromagnetic radiation

The sun, as the main source of energy for the Earth, radiates electromagnetic energy in a wide range of wavelengths. The electromagnetic energy is mainly reflected and absorbed by objects in the 0.4 - 3 $\mu$m spectral range. At longer wavelengths, all objects at temperatures above zero Kelvin emit *thermal radiation* according to Planck's law.

A black body, defined as an idealized substance that absorbs all electromagnetic radiation falling on it, and emits (an equal amount of) electromagnetic radiation according to its (constant) temperature, follows Planck's blackbody equation:

$$L(\lambda, T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda k T}} - 1} \tag{25.1}$$

where $L(\lambda, T)$ is the energy emitted per unit time (Watt per steradian per cubic meter) at wavelength $\lambda$ (meter), $h$ is the Planck constant ($6.626 \cdot 10^{-34}$ Js), $k$ is the Boltzmann constant ($1.381 \cdot 10^{-23}$ J/K), $c$ is the speed of light ($2.998 \cdot 10^{8}$ meter/second), and $T$ is the blackbody's temperature in degrees Kelvin. The Planck law gives a distribution of the emitted energy as a function of wavelength $\lambda$, at a given temperature $T$ (see Figure 25.6).

Figure 25.6 illustrates that the distribution of emitted energy, at different temperatures (red, green, blue and black curve), peaks at different wavelengths. The relationship between

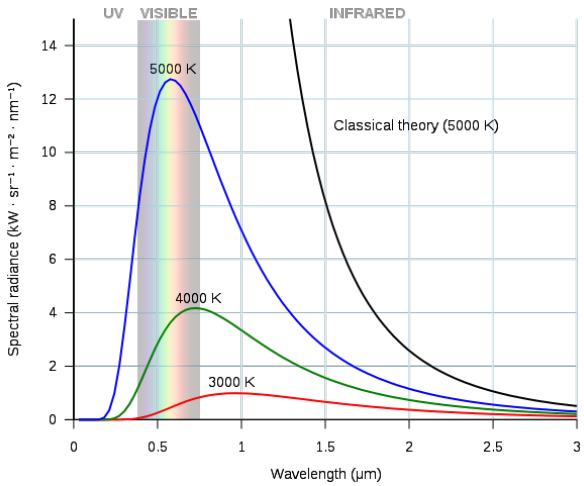



Figure 25.6: Black body radiation distribution of emitted energy at different temperatures, presented as radiance in kW per steradian, per squared meter, per nanometer. The effective temperature of the Sun is about 5800 K. Graph by Darth Kule - own work, taken from Wikimedia Commons [9]. Public Domain.

the temperature ($T$) of a blackbody and the wavelength in [m] of the peak of the radiation distribution is given by Wien's law:

$$\lambda_{max} = \frac{2.898 \cdot 10^{-3}}{T} \tag{25.2}$$

According to Wien's law, hotter objects emit most of their radiation at shorter wavelengths and that is why they appear to be 'bluer'. This means that objects of higher temperature have more energy, and therefore emit photons of higher energy (shorter wavelength). For a constant velocity (the speed of light here), wavelength is inversely proportional to its frequency ($c = \lambda f$). Therefore, the shorter the wavelength, the higher the frequency. The energy content of light varies with frequency (or wavelength). The relationship between energy and frequency is described by the Planck-Einstein relation:

$$E = hf \tag{25.3}$$

where $E$ is the energy of a photon [J], $h$ is the Planck constant, and $f$ is the frequency. The Planck-Einstein equation implies that a shorter wavelength has higher energy. For instance, a photon of blue light (shorter wavelength) has more energy than a photon of red light (longer wavelength).

The total radiation being emitted at all wavelengths by a blackbody is explained by the Stefan-Boltzmann law, which is the area under the Planck's law curve in Figure 25.6. The Stefan-Boltzmann law shows that the total radiant energy emitted (per time interval) by a blackbody is proportional to the fourth power of its temperature:

$$P_r = \sigma T^4 \tag{25.4}$$

where $\sigma$ is the Stefan-Boltzmann constant ($5.6703 \cdot 10^{-8}\ \frac{\text{Watt}}{\text{m}^2\text{K}^4}$), and $T$ is the temperature in Kelvin; $P_r$ is in [W/$\text{m}^2$]. To conclude, Planck's law describes the behavior of blackbody radiation (the curves in Figure 25.6), while Wien's law explains the shift of the peak toward shorter wavelengths with increase in temperature, and the Stefan-Boltzmann law describes the higher level of the curves as temperature increases.

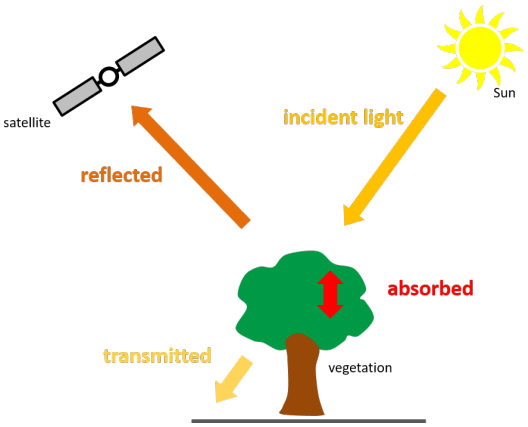



Figure 25.7: Three different interactions between solar light and vegetation: transmission, absorption and reflection.

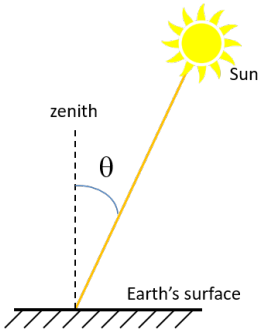



Figure 25.8: Sun-light incident on the Earth's surface; $\theta$ is the zenith angle of the Sun.

## 25.4. Interaction of electromagnetic radiation with objects

When electromagnetic energy arrives at an object or a target, three different types of interaction may occur: transmission, absorption and reflection, see also Appendix G on signal propagation. For an in-depth coverage of these subjects in the context of remote sensing we refer to [69]. From the incident light, there is a portion which is absorbed, another which is transmitted, and a third portion that is reflected (see Figure 25.7).

Reflectance is defined as the percentage, or ratio of reflected to incident light, and varies between zero (no light coming back from the object) and 1 (all of the incident light is reflected to the sensor). Formally, *reflectance* in remote sensing is defined as

$$R(\lambda) = \frac{\pi L(\lambda)}{E(\lambda)\cos\theta} \tag{25.5}$$

with $E(\lambda)$ the solar irradiance (Sun light incident on the Earth's surface) and $L(\lambda)$ the backscattered radiance (Sunlight reflected by the Earth's surface, eventually observed by remote sensing). Both $E(\lambda)$ and $L(\lambda)$ are functions of wavelength $\lambda$. Angle $\theta$ is the zenith angle of the sun, cf. Figure 25.8, and accounts for the orientation of the surface with respect to the incoming sun-light (for instance, when $\theta = 90°$ the surface receives no solar radiation). Reflectance $R(\lambda)$ is a *relative* measure, and defined as in (25.5), it is a dimensionless quantity. Reflections can — in limiting cases — be either specular or diffuse (scattered), cf. Figure G.10 (Appendix G.3).

The reflected photons reaching our eyes are the reason that we can see the objects. Furthermore, the color that we perceive is, in fact, the combination of wavelengths that are

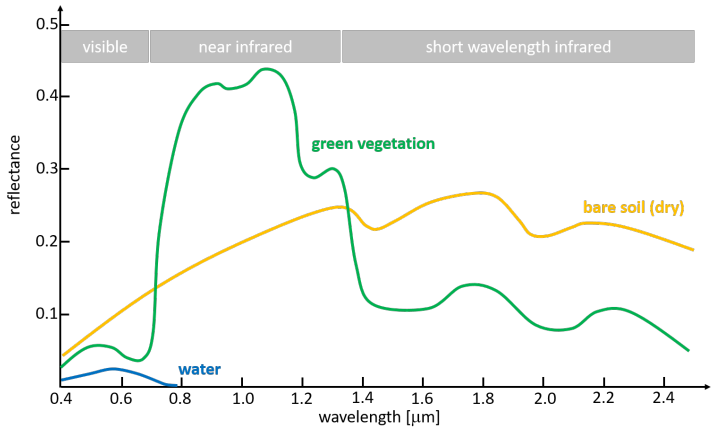



Figure 25.9: Spectral signatures of three different surface types over the 0.4 - 2.5 $\mu$m spectral range of electromagnetic radiation.

not absorbed by the object and reflected back to us. The reflected fraction of the light is the part that we are interested in, for instance while sensing vegetation from space. The light reflected from vegetation carries information about chemical and physical vegetation characteristics, as well as the background soil. The key to decoding this information resides in understanding the interactions of solar light with different elements in vegetation and soil, and their architecture and spatial arrangement.

The reflectance values of different features of the Earth's surface and topography may be plotted in so called 'spectral response curves' or 'spectral signatures' as a function of wavelength of the electromagnetic radiation. Differences in the spectral response of different surface features make it possible to *classify* them. That is because the spectral signatures of similar features have similar shapes. The *spectral signature* is dependent on a number of factors including coverage, composition, shape, texture etc. For instance, healthy green vegetation has a unique spectral reflectance curve (see Figure 25.9) that allows it to be easily distinguished from other surface types, like soil and water, in the visible and near-infrared region of the spectrum.

Figure 25.9 shows a sketch of spectral reflectance curves of three basic surface coverages: green vegetation, dry bare soil and water. As shown in the figure, there are distinct differences between the spectral curves of the surface types which help to differentiate among them. To do so, *spectral resolution* has to be carefully chosen to correspond to such distinctive regions. Spectral resolution is defined as the sensitivity of a remote sensor to respond to a specific range of wavelengths, characterized through the number of spectral bands and their widths. Multispectral sensors often collect data — at the same time — at several broad spectral bands ($\sim$ 100 nm wide; typically between 3 and 10 spectral bands), while hyperspectral sensors provide measurements in hundreds of continuous narrow spectral bands ($\sim$ 10 nm).

Multispectral sensors, mostly onboard the missions listed in Table 25.1, are widely used for different applications in agriculture, forestry resources, mapping, geology, hydrology, coastal resources and environmental monitoring. On the other hand, the high spectral resolution of hyperspectral sensors allows for detection and quantification of surface materials and their chemical compositions. Although hyperspectral sensors were originally developed for mining and geology, they are now used in a wide range of applications from precision agriculture, crop monitoring for water-stress, disease, and insect attack, to military applications. The three major Earth remote sensing satellite missions at present, are listed in Table 25.1.

| | |
|---|---|
| Landsat | Landsat is a joint NASA (National Aeronautics and Space Administration) and USGS (United States Geological Survey) program. So far Landsat is by far the longest running satellite Earth-imagery program. The first satellite was launched in 1972. |
| MODIS | MODIS, which stands for Moderate Resolution Imaging Spectroradiometer, is a NASA program. It consists of payloads on board of two satellites, namely Terra (EOS AM), launched in 1999, and Aqua (EOS PM), launched in 2002. |
| Sentinel | The Copernicus program, as part of ESA's Earth Observation missions, consists of multiple satellite missions in the context of the Sentinel project, with the first launch in 2014. The Sentinel missions include satellite radar and super-spectral imaging for land, ocean and atmospheric monitoring. |

Table 25.1: Current major Earth remote sensing satellite missions.

## 25.5. Interaction of solar radiation with vegetation

The optical properties of a soil-vegetation system depend on a large number of components of vegetation and the background soil. We now give more details on the interaction of light with vegetation characteristics and background soil. The spectral reflectance of vegetation from 0.4 to 2.5 $\mu$m can be subdivided into three regions, visible (0.4 - 0.7 $\mu$m), near infrared NIR (0.7 - 1.3 $\mu$m), and short wavelength, infrared (1.3 - 2.5 $\mu$m), see also Table 21.1. Another spectral region for vegetation characterization, known as 'red edge' (see Figure 25.9), is the abrupt transition (of the green curve) between 0.69 and 0.74 $\mu$m, caused by the transition from strong chlorophyll absorption to strong near-infrared leaf scattering.

The leaf pigments control the radiation regime of vegetation in the visible domain by strong *absorption* of light, resulting in lower reflectance for healthy vegetation in this domain. Among leaf pigments, chlorophyll (a and b) are the main absorbers of radiation in which they can absorb up to 70% - 90% of the incident solar radiation. Chlorophyll gives leaves their green color by absorbing red ($\sim$ 0.67 $\mu$m) and blue ($\sim$ 0.5 $\mu$m) spectral wavelengths as a result of electronic transitions in its molecular structure. Chlorophyll only influences the visible spectrum and is transparent to infrared radiation.

The absorption in the infrared part is restricted to the dry matter compounds such as cellulose, lignin and other structural carbohydrates. Instead, leaf tissue structure explains the main optical properties in the near infrared part. The high *reflectance* of leaves in the infrared is mainly due to the spongy mesophyll cells placed in the back or interior of a leaf. The number of intercellular air spaces in the lower mesophyll layer is the main factor responsible for the scattering. The red edge transition zone is important to monitor plant activity. Besides, the position of the red edge is used for the estimation of the canopy chlorophyll content. Strong absorption by leaf water content characterizes the middle or short-wavelength infrared region with absorption bands centered at 1.44 $\mu$m and 1.95 $\mu$m. The middle infrared region is crucial for vegetation stress identification due to drought.

The development over time or evolution, of vegetation or another phenomenon on the Earth's surface can be monitored and analysed by using a *time series*, or so-called stack of remote sensing images.

Figure 25.10 shows the difference of (amount of) healthy green vegetation between Winter (at left) and Summer (at right), for a rural/agricultural area. These images are based on the above 'red edge' transition. The reflectance is measured in the visible red-part of the spectrum (600-700 nm), as well as in the Near-Infra Red (NIR) part (800-900 nm). The so-called Normalized Difference Vegetation Index (NDVI) equals the difference of reflectance measured in the NIR-part and the reflectance in the red-part, divided by the sum of the two; the NDVI ranges from $-1$ to $+1$. Accordingly, the larger this index, the 'greener' the area.

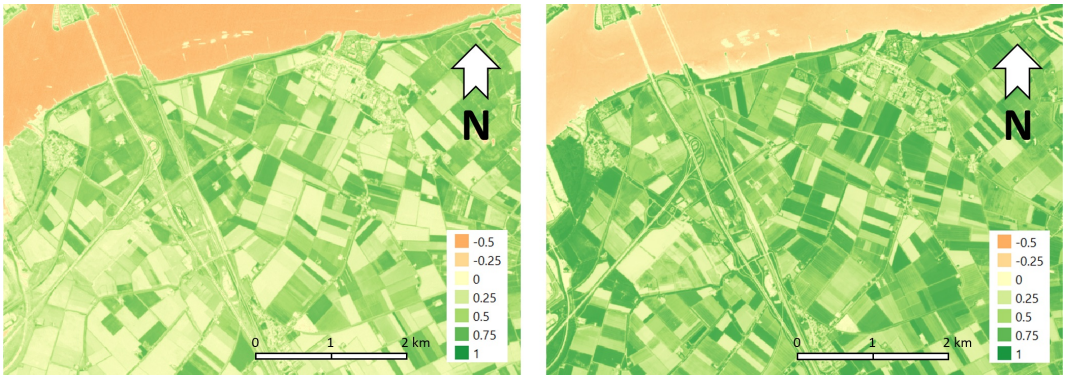



Figure 25.10: Map of Normalized Difference Vegetation Index (NDVI) in Winter 25-JAN-2021 (left) and Summer 14-JUN-2021 (right), for the area around the villages Moerdijk, Lage Zwaluwe and Zevenbergsche Hoek, near the Hollandsch Diep river (on top, in orange). Sentinel 2 (level 1C) data (Copernicus Sentinel data 2021) was retrieved from the Sentinel Hub [1], CC BY-NC 4.0. The NDVI is based on reflectance measured in band 4 (B4) and on reflectance measured in band 8 (B8), according to NDVI = (B8-B4)/(B8+B4).

The area in Figure 25.10 is clearly more 'green' in Summer than in Winter.

Vegetation water content is a measure of root zone water availability. It is an important indicator of water stress for vegetation canopies. This parameter is crucial for drought monitoring and canopy fuel moisture content. The latter, in turn, plays an important role in predicting the occurrence and spread of wildfire. Information about vegetation water content has also widespread utility in agriculture, forestry, and hydrology. It was already mentioned that water absorbs radiant energy throughout the near-infrared (0.7 - 1.3 $\mu$m) and middle or short wavelength infrared (1.3 - 2.5 $\mu$m) spectral regions. A detailed analysis of the vegetation spectral signature provides useful information on the vegetation growth stage, stress, and chemical processes taking place in vegetation.

## 25.6. Example: land cover classification [*]

With reference to Figure 25.9 we demonstrate, by means of a small and simplified example, how automated classification of multi-spectral remote sensing images can be done, in terms of Earth's surface land cover.

The images consist of 5×5 pixels. We use two spectral bands, and Figure 25.11 shows, at left an image based on measured reflectance $R$ in a spectral band around $\lambda_1 = 600$ nm, and similarly in the middle, an image for the band around $\lambda_2 = 800$ nm. In Figure 25.9 one can see that the three different types of land cover considered here, water, dry soil and green vegetation, clearly present a different reflectance in these two bands. In Figure 25.11 the result of classification is shown at right, where blue refers to water, yellow to dry soil, and green to green vegetation.

Classification in this example is based on the k-Nearest Neighbor (kNN) method. This method is an example of *supervised* learning. This implies that a *training dataset* is used to train the classifier. In this example the training dataset consists of the 3×3 pixels in the top left of the image. The training data are labeled, meaning that the actual land-cover type is known for these pixels, for instance upon inspection on-site. Once training has been completed, the classifier will predict the classes of other/further data, in this case, the remaining 16 pixels of the image, leading to the land cover map shown at right, in Figure 25.11.

Classification is done on the basis of *features*. Features are quantities or metrics, used by the classifier, to assign observations to individual classes. In this example we use two features, namely the reflectance at 600 nm wavelength, $R(\lambda_1)$, and, the reflectance at 800 nm

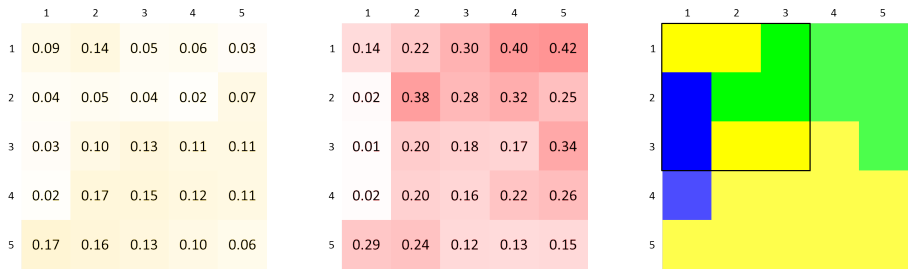



Figure 25.11: At left, and in the middle, remote sensing image based on measured reflectance in spectral band around $\lambda_1 = 600$ nm and $\lambda_2 = 800$ nm respectively. At right the resulting land cover classification, with blue for water, yellow for dry soil, and green for green vegetation.

wavelength, $R(\lambda_2)$. The core of classification is getting to know how features are mapped onto the so-called response, i.e. the variable we would like to predict; in the present example, how the two reflectances $R(\lambda_1)$ and $R(\lambda_2)$ are mapped onto the land cover type (water, dry soil and green vegetation).

Figure 25.12 shows, as a graph, the features of all 25 pixels, with reflectance $R(\lambda_1)$ along the horizontal axis, and reflectance $R(\lambda_2)$ along the vertical axis. Classification then takes place, in this coordinate system, according to distances from the pixel to-be-classified, to the training pixels. In this example we consider the k=3 nearest training pixels, i.e. k=3 nearest neighbors. For each to-be-classified pixel, the distances to all training pixels are computed, then sorted ascendingly, and the smallest k=3 distances are considered. The class which is occuring most frequently among the k=3 nearest neighbors, then sets the class for the pixel under consideration. The k nearest neighbors are used to predict the label of the 'new' observation $(R(\lambda_1), R(\lambda_2))$. Note that the word distance here, refers to feature distance, not to a (geometric) distance between pixels in the image or map.

In a geometric interpretation, kNN-classification is about a circle centered at the coordinates of the observed features, hence for pixel (2,5) $R(\lambda_1) = 0.07$ and $R(\lambda_2) = 0.25$, and growing this circle, until k=3 training pixels are contained, and then determining the majority of the classes of three training pixels in the circle. In this case, see Figure 25.12 at right, there are two green pixels and one yellow pixel, hence by its two features, pixel (2,5) gets classified as green vegetation.

k-Nearest Neighbor (kNN) is a very basic, very common example of supervised classification, especially suited for categorical data, as in this example, with classes like water and green vegetation. In this section, we just demonstrate its working - we did not assess the accuracy of prediction.

There are many more methods for automated classification, also being referred to as pattern recognition. Classification is a true subject on its own, and a comprehensive exposition is beyond the scope of this book.

## 25.7. Soil reflectance [*]

If vegetation is not very dense, solar radiation can penetrate the vegetation canopy and reach the soil surface. In that case, part of the reflected radiation from the soil will contribute to the observed reflectance from the top of the canopy. Therefore, it is important to study soil factors affecting the output reflectance. The optical properties of soil are a complex function of the soil main physical and chemical properties such as soil moisture content, surface roughness, mineral and organic compositions. Soil mineral grains of different sizes and shapes absorb solar radiation causing lower reflectance, mostly in the middle infrared, depending on their chemical

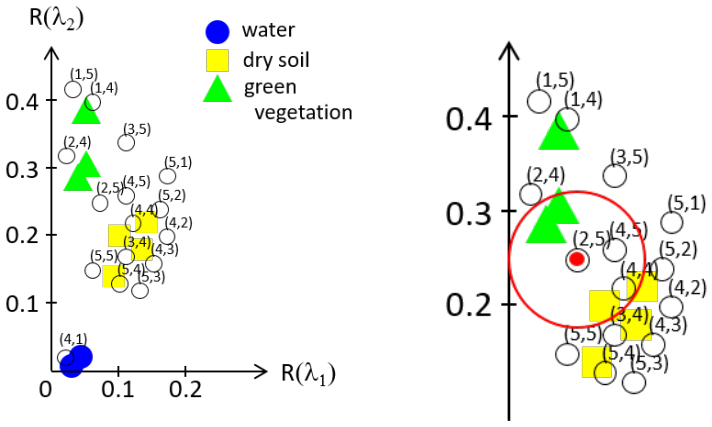



Figure 25.12: k Nearest Neighbor (kNN) classification at work, with k=3, for the 16 pixels in Figure 25.11, using the 3×3 training pixels. The training pixels are shown by means of a symbol: blue dot for water, yellow square for dry soil and green triangle for green vegetation. The pixels which are to be classified are shown by a black circle, together with their position in the image (row, column); for each to-be-classified pixel the k=3 nearest neighbors determine, by majority voting, the class for the pixel.

compositions such as carbonates, sulfates and clay minerals. Organic matter influences the chemical and physical properties of soil and mainly affects soil reflectance indirectly due to the variation in the structure and water retention capacity of soil. The overall soil reflectance decreases with the increase of soil organic matter over the entire spectral range from 0.4 to 2.5 $\mu$m. Soil surface roughness is another important factor due to the shadowing effects of soil aggregates. In general, the rougher the soil, the lower the reflectance is. The soil moisture observable by optical remote sensing is approximately the moisture content within a very thin layer at the soil surface. Soil moisture is a vital factor for many different applications and influences the exchange of energy between soil, vegetation and the atmosphere. It has a significant impact on the soil spectrum over the entire wavelength range. The effect of soil moisture is stronger in the middle infrared part (also referred to as Short Wavelength IR) of the spectrum, cf. Figure 25.9. The higher the soil moisture content, the lower the reflectance.

## 25.8. Exercises and worked examples

Below two questions are presented on the underlying physics of remote sensing.

**Question 1** The sun is nearly a blackbody. The temperature at its surface is about 5600 degrees Celsius (5873 Kelvin). Calculate the total amount of radiation of the sun and compare it to the radiation of the Earth (assume the Earth is a blackbody in this example, however it is not in reality!) at a temperature of 30 degrees Celsius. Find the wavelength of the peak for both curves in the spectral domain from 0.2 to 10 $\mu$m. What are the differences between the two peaks? Discuss the reason.

**Answer 1** Using Eq. (25.4) with $T = 5873$ K, we find $P_r = 6.746 \cdot 10^7$ [W/m$^2$] for the sun, and with $T = 303$ K for the Earth $P_r = 4.779 \cdot 10^2$ [W/m$^2$]. Using equation (25.2) we find the peak at $\lambda_{max} = 0.493$ $\mu$m for the Sun, and $\lambda_{max} = 9.564$ $\mu$m for the Earth. For the solar radiation the peak is in the visible light domain. The radiation of the Earth is infra-red. The atmosphere is — as a side-note — strongly absorbant in these wavelengths, meaning that this energy is kept in the Earth's system (rather than emitted into space), causing, what is known as the 'greenhouse'-effect. The full radiation distribution — resulting in a plot like Figure 25.6 — can be found using Eq. (25.1).

**Question 2** Leaves turn into yellow in Fall because of loss of chlorophyll. We know that the yellow color is combination of red and green. Since leaves are green in Summer, discuss which region of the spectrum (see Figure 25.9) is most affected by chlorophyll.

**Answer 2** We see a healthy leaf as colored green, as most of its reflected radiation is in the green wavelengths. And it absorbs other wavelengths, in particular red light. When the leaf turns yellow, it means that now both green and red light are reflected back of the leaf, the latter caused by the reduced amount of chlorophyll in the leaf. Or, in other words, in Summer the leaf absorbs red and blue light (but not green), and in Fall — in addition — it no longer absorbs red light.

# V

# Reference systems

# 26

# Introduction

With the advent of the Global Positioning System (GPS) technology in smartphones and other navigational devices, almost anyone, anywhere on Earth, at any time, can determine a three-dimensional position accurately 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, individuals and professional users alike can achieve, with relative ease, three-dimensional positions with centimeter accuracy. This feat, was until recently achievable only for a small community of land surveyors and geodesists.

Our increased ability to collect accurate position data, and 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 becoming available to users.

However, sharing position 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 ...?*', '*Why do our heights differ by 2.31 m (at the Dutch-Belgium border)?*' These are just a few frustated outcries one will hear from users (including civil engineering students). The reason is simple: users may have opted for different coordinate reference systems (CRS). Positions, including heights, 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 for good reasons. The solution is straightforward, but not simple: knowing the name and identifier of the reference system is key. If you have position 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 part will provide you with the background and terminology that is commonly used. For the actual transformations you can use (freely) available software.

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 means 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, allowing multiple visualizations and analysis of the data in Geographic Information Systems (GIS). Surveying means 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 located on the surface of the Earth. They are used to depict topographic features, boundaries for ownership, locations of buildings, location of subsurface features (pipelines), and are often used to monitor changes (deformation, subsidence). Points may only exist on

Figure 26.1: The blue-white line on the Delft University of Technology campus visualizes latitude $\varphi = 52.0000000^\circ$ North. The 16-millimeter-wide black line, in the middle of the white band, indicates the 'exact' position of the $52^\circ$ latitude, in the International Terrestrial Reference System (ITRS), on 1 January 2018. Time matters because the $52^\circ$ parallel shifts due to plate tectonics, as we cover in Chapter 34. The width of the black line represents the shift per year. To illustrate the time effect further, six gray lines (not visible in this picture) have been painted parallel to the blue line. These gray lines indicate significant events related to Delft. The first line is 2.79 metres to the North, and marks the foundation of TU Delft in 1842, Delta, 27 August 2018. Photo courtesy of Conny van Uffelen [70].

paper, or in a Computer-Aided Design (CAD) system, when they represent points to be staked out for construction work.

The process of assigning positions to geographical objects is often referred to as *georeferencing*. Georeferencing also applies to maps and (aerial) photographic and remote sensing images, whereby the internal coordinate system of the image (pixels) is related to a geographic coordinate system and this information is typically stored with the image (either inside the image file or as a separate file). This is called georeferenced image data.

Geocoding and reverse geocoding are related to addresses. Geocoding is the process of converting (street) addresses to geographic coordinates (so that they can be displayed on a map). Reverse geocoding does the opposite: it links geographic coordinates to human-readable addresses. When no addresses are available you can use the *Open Location Codes*, which are also known as *plus codes*. The Open Location Code (OLC) is used for identifying an *area* anywhere on the world. They are derived from latitude and longitude coordinates, are similar in length to a phone number (11 characters), and can be used anywhere on the world to indicate an area with a resolution of 3.5 m at the equator. Nearby locations have similar codes; therefore codes can be shortened, and/or combined with the name of a town or municipality. For example, look up the the plus-code 'X9XG+H6 Delft' in Google maps.

To describe the position of points, a mathematical framework is needed. This mathematical framework consists 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 part several of these mathematical frameworks are described, as well as the way they are used in surveying and mapping.

This book was written close, very close, to $52.0000^\circ$ 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 on the campus with a blue-white line, see Figure 26.1. Check out this news item in the university newspaper Delta [71] for a full story of how the line was created and

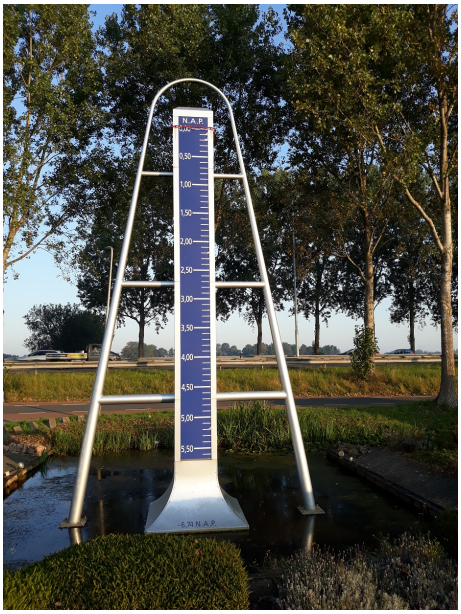



Figure 26.2: Monument for the lowest point in the Netherlands, in the Zuidplaspolder near Nieuwerkerk aan den IJssel, right next to the A20 highway from Rotterdam to Gouda. The Zuidplaspolder, created in 1841, is a former lake. The lake itself was formed as the result of peat extraction. The bottom of the monument, which carries the inscription -6.74 m NAP, corresponds to the lowest point in a nearby meadow. In 2005 the value for the lowest point was adjusted to -6.76 m NAP (2 cm lower).

how it has moved over the campus.

## Overview of this part

In Chapters 27 and 28 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 29. The *map projections*, which result in easy to use 2D Cartesian coordinates, are covered in Chapter 30.

Coordinate conversions and geodetic datum transformations are discussed in Chapter 31. 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.

Height always plays a special role in coordinate reference systems. Height is also closely associated with the flow of water and gravity. For a low lying country like the Netherlands, with its many polders below sea level and lowest point at -6.74 m NAP (Figure 26.2), a precise and reliable height reference system is of vital importance. Height coordinate systems are discussed in Chapter 33, while in Chapter 32 basic background information on the Earth's gravity field is given.

Finally, in Chapter 34 and 35 several important, commonly used reference systems are described. They include the well-known World Geodetic System (WGS84) used by GPS, the International Terrestrial Reference System (ITRS), and the European Terrestrial Reference System (ETRS89) in Chapter 34, 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 at sea, in Chapter 35.

In this part vectors and matrices are systematically typeset in bold. For example a position vector is denoted as $\mathbf{r}$, and the length or norm of this vector is denoted by scalar as $r = \|\mathbf{r}\|$.

# 27

# 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 30.

## 27.1. 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 27.1(a). 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 $P_i$,

$$\mathbf{r}_i = x_i \mathbf{e}_x + y_i \mathbf{e}_y \ , \tag{27.1}$$

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 $\alpha_{12}$ between two points $P_1$ and $P_2$ are defined as,

$$\begin{aligned} d_{12} &= \|\mathbf{r}_2 - \mathbf{r}_1\| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \\ \alpha_{12} &= \arctan \frac{x_2 - x_1}{y_2 - y_1} \end{aligned} \tag{27.2}$$

The azimuth $\alpha$ is given in angular units (degrees, radians, gon) while the distance $d$ is expressed in length units (meters), see also Appendix H. For practical computations the $\arctan$ in Eq. (27.2) should be replaced with the $\mathrm{atan2}\,(x_2 - x_1, y_2 - y_1)$ function in order to obtain

[1] an easy way to remember this is the *right hand rule*: have the thumb of your right hand pointing up (in the direction of the positive z-axis, cf. Chapter 28), the fingers now point from the x-axis to the y-axis

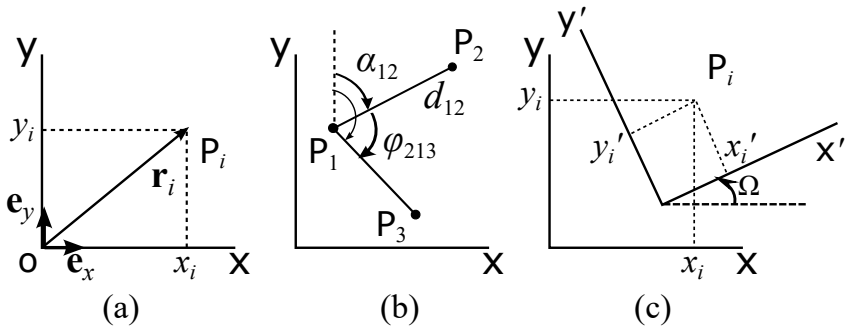



Figure 27.1: 2D Cartesian coordinate system (a), definition of azimuth, angle and distance (b) and a 2D coordinate transformation (c).

the right quadrant for the azimuth $\alpha_{12}$[2]. The angle $\angle P_2P_1P_3$ between points $P_2$, $P_1$ and $P_3$ is,

$$\varphi_{213} = \angle P_2P_1P_3 = \alpha_{13} - \alpha_{12} = \arccos \frac{< \mathbf{r}_2 - \mathbf{r}_1, \mathbf{r}_3 - \mathbf{r}_1 >}{\|\mathbf{r}_2 - \mathbf{r}_1\| \|\mathbf{r}_3 - \mathbf{r}_1\|} \tag{27.3}$$

with $< \mathbf{u}, \mathbf{v} >$ the dot (inner) product of two vectors. See Figure 27.1(b). 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 27.1(b). 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 textbook 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 $\alpha_{oi}$ 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 the origin in the instrument and y-axis in an arbitrary, yet to be determined, direction (representing the zero reading of the instrument, see Figure 4.26).

## 27.2. 2D coordinate transformations

In this section we first discuss shape preserving transformations, followed by affine and polynomial transformations.

[2]the atan2 function is the four quadrant version of the arctangent function with two input values, and output value in the range of all four quadrants (full circle) with $-\pi \leq \text{atan2}\,(dx, dy) \leq \pi$, compared to $-\pi/2 \leq \arctan \frac{dx}{dy} \leq \pi/2$, with $dx = x_2 - x_1$ and $dy = y_2 - y_1$

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

### 27.2.1. 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 27.1(c). 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 $\Omega$, a scale change by a scale factor $s$ and a translation by two origin shift parameters $(t_{x'},t_{y'})$ via a so-called (2D) similarity transformation,

$$\begin{pmatrix} x'_i \\ y'_i \end{pmatrix} = s \begin{pmatrix} \cos\Omega & \sin\Omega \\ -\sin\Omega & \cos\Omega \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} + \begin{pmatrix} t_{x'} \\ t_{y'} \end{pmatrix} . \tag{27.4}$$

The translation parameters $(t_{x'},t_{y'})$ give the origin of the source system in target system coordinates. A different way of writing the same transformation is

$$\begin{pmatrix} x'_i \\ y'_i \end{pmatrix} = s \begin{pmatrix} \cos\Omega & \sin\Omega \\ -\sin\Omega & \cos\Omega \end{pmatrix} \begin{pmatrix} x_i - t_x \\ y_i - t_y \end{pmatrix}. \tag{27.5}$$

with $(t_x,t_y)$ the origin of the target system in source system coordinates .

The transformation preserves angles, for instance in a triangle with three points, i.e. angles are not changed by the transformation. This also means that shapes are preserved. This transformation is known as the *similarity transformation* or 2D Helmert transformation. Distances are not necessarily preserved in similarity transformation, unless the scale factor $s$ equals one, but ratios of distances are preserved.

The transformation involves four parameters: two translations $t_x$ and $t_y$, a rotation $\Omega$, 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 $\Omega$. 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.

### 27.2.2. 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,

$$\begin{pmatrix} x'_i \\ y'_i \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} + \begin{pmatrix} t_{x'} \\ t_{y'} \end{pmatrix} . \tag{27.6}$$

with the $2$-by-$2$ transformation matrix containing $4$ different elements. Affine transformations introduce so-called *shearing* between the coordinate axes. 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. As example, the 2D second order polynomial transformation is

$$\begin{aligned} x'_i &= t_{x'} + ax_i + by_i + ex_i^2 + 2fx_iy_i + gy_i^2 = \\ &= t_{x'} + (a + ex_i + fy_i)x_i + (b + fx_i + gy_i)y_i = t_{x'} + \bar{a}(x,y)x_i + \bar{b}(x,y)y_i \\ y'_i &= t_{y'} + cx_i + dy_i + hx_i^2 + 2kx_iy_i + ly_i^2 = \\ &= t_{y'} + (c + hx_i + ky_i)x + (d + kx_i + ly_i)y = t_{y'} + \bar{c}(x,y)x_i + \bar{d}(x,y)y_i \end{aligned} \tag{27.7}$$

with six coefficients per coordinate. As is shown in Eq. (27.7) polynomial transformations can be written in the form of an affine transformation, conform Eq. (27.6), but with the transformation parameters $\bar{a}$, $\bar{b}$, $\bar{c}$ and $\bar{d}$ as functions of the coordinates themselves. This means that straight lines, and shapes, are not necessarily preserved in polynomial transformations. Polynomial transformations are sometimes used as approximate transformations to match satellite 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. The 2D affine transformation of Eq. (27.6) is actually the same as a first order polynomial transformation.

## 27.3. 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 axes 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 of 1 for the y-coordinate 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 or a laser distometer) and also any measurement error that was made in this measurement is 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.

## 27.4. 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 can be interpreted from a linear algebra perspective.

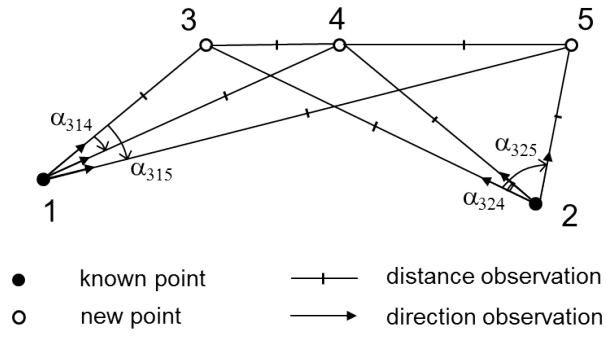



Figure 27.2: Two-dimensional survey network with 4 angle measurements at two known points, points 1 and 2, and 8 distance measurements, to determine the coordinates of points 3, 4 and 5.

### 27.4.1. 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 27.2 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.

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 27.3, a 2D Cartesian coordinate system is uniquely defined by four parameters: the scale $s$, orientation $\Omega$, 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 27.2 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 to be 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).

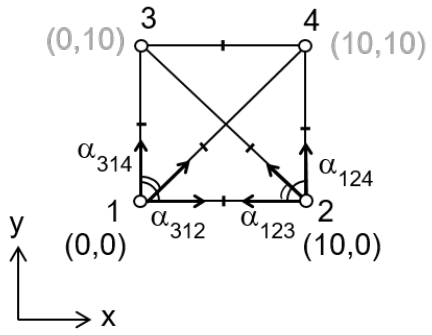



Figure 27.3: Two-dimensional simple survey network with 4 angle measurements and 6 distance measurements.

### 27.4.2. Algebraic analysis

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

The network is shown in Figure 27.3. This artificial network, being just a square, allows for a convenient and simple algebraic analysis. There are four angle measurements $\alpha_{314}$, $\alpha_{312}$, $\alpha_{123}$ and $\alpha_{124}$. 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

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{27.8}$$

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 $\Delta$-symbol). Further details can be found in Chapters 8 and 9. For the artificial geometry in Figure 27.3, this system of equations becomes:

$$\begin{pmatrix} \Delta\alpha_{314} \\ \Delta\alpha_{312} \\ \Delta\alpha_{123} \\ \Delta\alpha_{124} \\ \Delta l_{12} \\ \Delta l_{13} \\ \Delta l_{14} \\ \Delta l_{23} \\ \Delta l_{24} \\ \Delta l_{34} \end{pmatrix} = \begin{pmatrix} \frac{10}{200} & \frac{10}{200} & 0 & 0 & -\frac{10}{100} & 0 & \frac{10}{200} & -\frac{10}{200} \\ \frac{10}{100} & \frac{10}{100} & 0 & -\frac{10}{100} & -\frac{10}{100} & 0 & 0 & 0 \\ 0 & -\frac{10}{100} & -\frac{10}{200} & \frac{10}{200} & \frac{10}{200} & \frac{10}{200} & 0 & 0 \\ 0 & -\frac{10}{100} & -\frac{10}{100} & \frac{10}{100} & 0 & 0 & \frac{10}{100} & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 & 0 & 0 & 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \Delta x_1 \\ \Delta y_1 \\ \Delta x_2 \\ \Delta y_2 \\ \Delta x_3 \\ \Delta y_3 \\ \Delta x_4 \\ \Delta y_4 \end{pmatrix} \tag{27.9}$$

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

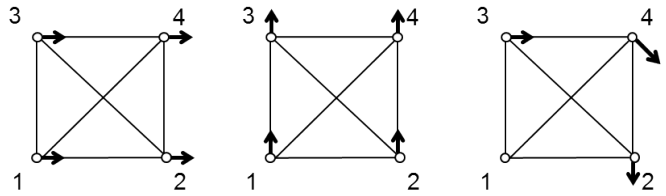



Figure 27.4: Interpretation of the null-space of matrix $\mathbf{A}$, vector $\mathbf{v}_1$ at left, vector $\mathbf{v}_2$ in the middle, and vector $\mathbf{v}_3$ at right.

null-space of matrix $\mathbf{A}$ can be spanned by the following three (linearly independent) vectors:

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \; ; \; \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \; ; \; \mathbf{v}_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 1 \\ -1 \end{pmatrix} \tag{27.10}$$

These three vectors can be stored together, in $8$x$3$ 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 Eq. (27.8), then $\mathbf{x}' = \mathbf{x} + \mathbf{V}\boldsymbol{\beta}$, with $3$x$1$-vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$, is also a solution, namely

$$\mathbf{y} = \mathbf{A}\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{A}\mathbf{V}\boldsymbol{\beta} = \mathbf{A}\mathbf{x} \tag{27.11}$$

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 27.4. 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 $\Omega$ of the network about point 1.

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. $(x_2, y_2) = (., 0)$. 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 $10$x$5$, 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.

## 27.5. Exercises and worked examples

This section presents a worked example on the use of the 2D similarity transformation.

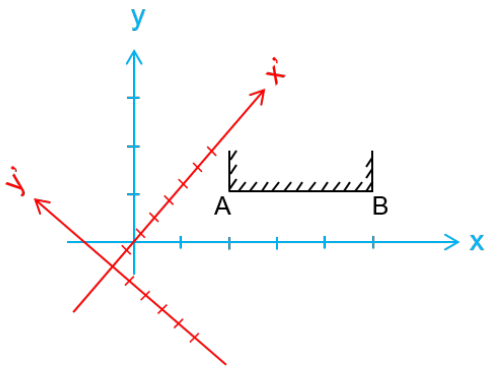



Figure 27.5: Setup for measuring points A and B of the facade of a building (Question 1).

**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 Figure 27.5. 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 $\Omega$, and translations $t_{x'}$, $t_{y'}$.

**Answer 1** A clever approach to solving this problem is using Eq. (27.4) 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. (27.4), 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 $(t_{x'}, t_{y'}) = (\sqrt{2}, 0)$.

# 28

# 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 29. 3D Cartesian coordinate systems can be considered a straightforward extension of 2D Cartesian coordinate systems by adding a third axis.

## 28.1. 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 28.1(a). 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 H). Capital letters $X$, $Y$, $Z$ are used for the coordinates to set them apart from the 2D coordinate system in Chapter 27, 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 29.4. 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, with geocentric and topocentric coordinates, the coordinate system is somehow tied to the Earth, but this is not necessary. In another commonly used variant 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.

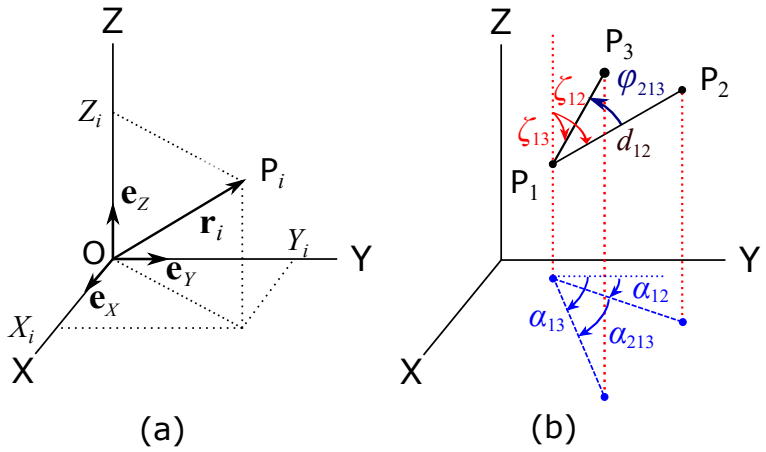



Figure 28.1: 3D Cartesian coordinate system (a) and definition of azimuth $\alpha_{12}$, horizontal angle $\alpha_{213}$, vertical angle $\zeta_{12}$, angle $\varphi_{213}$ and distance $d_{12}$ (b).

## 28.2. 3D Cartesian coordinates

The 3D topocentric Cartesian coordinate system can be considered a straightforward extension of a 2D Cartesian coordinate system. Just imagine in Figure 27.1 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

$$\mathbf{r}_i = X_i\mathbf{e}_X + Y_i\mathbf{e}_Y + Z_i\mathbf{e}_Z \ , \tag{28.1}$$

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 28.1(a). As shown in Figure 28.1(b), the distance $d_{12}$ between two points $P_1$ and $P_2$ is

$$d_{12} = \|\mathbf{r}_2 - \mathbf{r}_1\| = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2 + (Z_2 - Z_1)^2} \ , \tag{28.2}$$

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

$$\varphi_{213} = \angle P_2P_1P_3 = \arccos \frac{< \mathbf{r}_2 - \mathbf{r}_1, \mathbf{r}_3 - \mathbf{r}_1 >}{\|\mathbf{r}_2 - \mathbf{r}_1\| \ \|\mathbf{r}_3 - \mathbf{r}_1\|} \tag{28.3}$$

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 $\alpha_{12}$ and vertical angle $\zeta_{12}$ between points $P_1$ and $P_2$ can be defined as,

$$\begin{aligned} \alpha_{12} &= \arctan \frac{X_2 - X_1}{Y_2 - Y_1} \\ \zeta_{12} &= \arctan \frac{\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}}{Z_2 - Z_1} = \arccos \frac{Z_2 - Z_1}{\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2 + (Z_2 - Z_1)^2}} \end{aligned} \tag{28.4}$$

See also Figure 28.1(b). For practical computations the $\arctan$ in Eq. (28.4) 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 $\alpha_{12}$. 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. (28.4) cannot be used for global geocentric terrestrial coordinate systems. The angle $\varphi_{213}$ of Eq. (28.3) is not the same as the horizontal angle $\alpha_{213}$ in Figure 28.1(b). The horizontal angle is defined as $\alpha_{213} = \alpha_{13} - \alpha_{12}$.

## 28.3. 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.

### 28.3.1. 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).

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \underbrace{\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}}_{\mathbf{R}} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_{x'} \\ t_{y'} \\ t_{z'} \end{pmatrix} \tag{28.5}$$

The 3D affine transformation is specified by a total of 12 parameters, consisting of 9 parameters for matrix $\mathbf{R}$ plus 3 translation parameters $(t_{x'}, t_{y'}, t_{z'})$.

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).

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \lambda \mathbf{R}(\Omega_x, \Omega_y, \Omega_z) \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_{x'} \\ t_{y'} \\ t_{z'} \end{pmatrix} \tag{28.6}$$

The 3D similarity transformation is specified by a total of 7 parameters, three rotation parameters $(\Omega_x, \Omega_y, \Omega_z)$, a scale factor $(\lambda$ )and three translation parameters $(t_{x'}, t_{y'}, t_{z'})$.

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 $\lambda = 1$

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \mathbf{R}(\Omega_x, \Omega_y, \Omega_z) \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_{x'} \\ t_{y'} \\ t_{z'} \end{pmatrix} \tag{28.7}$$

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.

### 28.3.2. 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 scale factor $\lambda$,

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \lambda \cdot \mathbf{R}(\Omega_x, \Omega_y, \Omega_z) \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_{x'} \\ t_{y'} \\ t_{z'} \end{pmatrix} \tag{28.8}$$

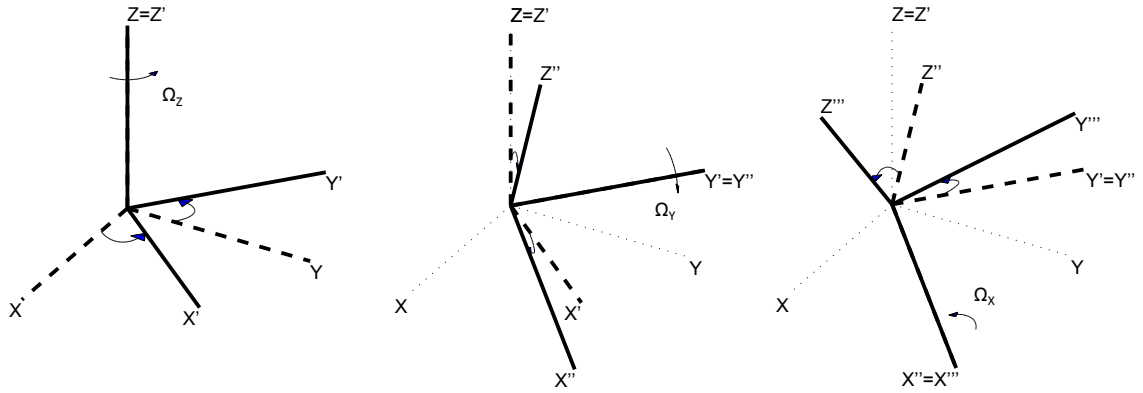



Figure 28.2: Definition of rotation angles for the 7-parameter similarity transformation. The figure on the left shows a rotation $\Omega_z$ about the Z-axis, the middle figure a rotation $\Omega_y$ about the newly obtained Y'-axis, and the figure on the right a rotation $\Omega_x$ about the final X"-axis. 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.

with $(X, Y, Z)$ the coordinates in the source coordinate system and $(X', Y', Z')$ the coordinates in the target coordinate system. The translation vector $(t_{x'}, t_{y'}, t_{z'})$ has to be added to the source coordinates after rotation. The translation vector $(t_{x'}, t_{y'}, t_{z'})$ gives the coordinates of the origin of the source coordinate system with respect to the target coordinate system. The 7-parameter similarity transformation can also be written as,

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \lambda \cdot \mathbf{R}(\Omega_x, \Omega_y, \Omega_z) \cdot \begin{pmatrix} X - t_x \\ Y - t_y \\ Z - t_z \end{pmatrix} \tag{28.9}$$

whereby the translation vector $(t_x, t_y, t_z)$ represents the coordinates of the origin of the target coordinate system with respect to the source coordinate system. The relation between the two translation vectors is,

$$\begin{pmatrix} t_{x'} \\ t_{y'} \\ t_{z'} \end{pmatrix} = -\lambda \cdot \mathbf{R}(\Omega_x, \Omega_y, \Omega_z) \cdot \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} \tag{28.10}$$

The rotation matrix $\mathbf{R}(\Omega_x, \Omega_y, \Omega_z)$ is defined as a sequence of so-called Euler rotations. The order in which the rotations are applied, and over which axis, matters. For instance, a 3-2-1 series of Euler rotations gives the rotation matrix

$$\mathbf{R}_{321}(\Omega_x, \Omega_y, \Omega_z) = \mathbf{R}_1(\Omega_x) \cdot \mathbf{R}_2(\Omega_y) \cdot \mathbf{R}_3(\Omega_z) =$$
$$\begin{pmatrix} \cos\Omega_z \cos\Omega_y & \sin\Omega_z \cos\Omega_y & -\sin\Omega_y \\ \cos\Omega_z \sin\Omega_y \sin\Omega_x - \sin\Omega_z \cos\Omega_x & \sin\Omega_z \sin\Omega_y \sin\Omega_x + \cos\Omega_z \cos\Omega_x & \cos\Omega_y \sin\Omega_x \\ \cos\Omega_z \sin\Omega_y \cos\Omega_x + \sin\Omega_z \sin\Omega_x & \sin\Omega_z \sin\Omega_y \cos\Omega_x - \cos\Omega_z \sin\Omega_x & \cos\Omega_y \cos\Omega_x \end{pmatrix} \tag{28.11}$$

with $\Omega_z$, $\Omega_y$ and $\Omega_x$ 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

$$\mathbf{R}_3(\Omega_z) = \begin{pmatrix} \cos\Omega_z & \sin\Omega_z & 0 \\ -\sin\Omega_z & \cos\Omega_z & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{R}_2(\Omega_y) = \begin{pmatrix} \cos\Omega_y & 0 & -\sin\Omega_y \\ 0 & 1 & 0 \\ \sin\Omega_y & 0 & \cos\Omega_y \end{pmatrix}, \quad \mathbf{R}_1(\Omega_x) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\Omega_x & \sin\Omega_x \\ 0 & -\sin\Omega_x & \cos\Omega_x \end{pmatrix} \tag{28.12}$$

whereby the right-handed rotation is positive, which is, when viewed along the axis towards the origin, a counter-clockwise rotation. See Figure 28.2. This sense of rotation is the same as was used in the 2-dimensional case, see Eq. (27.4). Imagine in Figure 27.1(c) a z-axis pointing out of the paper, then the rotation matrix of Eq. (27.4) is essentially $\mathbf{R}_3(\Omega_z)$ (which does not change the z-axis or z-coordinates) and the rotation angle $\Omega$ of Eq. (27.4) is actually $\Omega_z$ in the 3D-transformation. The rotation $\Omega_z$ around the z-axis is actually a rotation of the x-axis (and also y-axis) by $\Omega_z$. 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.

Changing the order of the Euler rotations in Eq. (28.11) 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. (28.11), gives a 1-2-3 sequence of Euler rotations with rotation matrix,

$$\mathbf{R}_{123}(\Omega'_x, \Omega'_y, \Omega'_z) = \mathbf{R}_3(\Omega'_z) \cdot \mathbf{R}_2(\Omega'_y) \cdot \mathbf{R}_1(\Omega'_x) = \begin{pmatrix} \cos\Omega'_z \cos\Omega'_y & \cos\Omega'_z \sin\Omega'_y \sin\Omega'_x + \sin\Omega'_z \cos\Omega'_x & -\cos\Omega'_z \sin\Omega'_y \cos\Omega'_x + \sin\Omega'_z \sin\Omega'_x \\ -\sin\Omega'_z \cos\Omega'_y & -\sin\Omega'_z \sin\Omega'_y \sin\Omega'_x + \cos\Omega'_z \cos\Omega'_x & \sin\Omega'_z \sin\Omega'_y \cos\Omega'_x + \cos\Omega'_z \sin\Omega'_x \\ \sin\Omega'_y & -\cos\Omega'_y \sin\Omega'_x & \cos\Omega'_y \cos\Omega'_x \end{pmatrix} \tag{28.13}$$

with $\Omega'_x$, $\Omega'_y$ and $\Omega'_z$ 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 $\Omega'_x$, $\Omega'_y$ and $\Omega'_z$ are different from the rotation angles $\Omega_x$, $\Omega_y$ and $\Omega_z$ used in Eq. (28.11).

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. (28.11), 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 $\mathbf{R}_{321}(\Omega_x, \Omega_y, \Omega_z)^T = \mathbf{R}_{123}(-\Omega_x, -\Omega_y, -\Omega_z)$. Compare the terms in Eqs. (28.11) and (28.13), 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.

Therefore Eqs. (28.11) and (28.13) are often used as forward and inverse transform pairs (or vice versa), whereby only the sign of the rotation angles changes. However, you can also use the same formula for the inverse rotation, but then the values of the rotation angles for the forward and inverse transformation differ by more than only the sign.

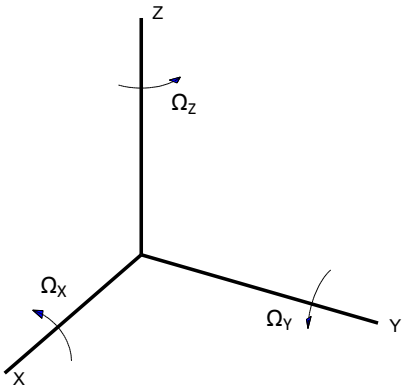



Figure 28.3: Definition of infinitesimal small rotation angles $\Omega_x$, $\Omega_y$ and $\Omega_z$ for the *Helmert* transformation.

### 28.3.3. 7-parameter Helmert (small angle) transformation [*]

In case the rotation angles are very small, with $\cos\Omega \simeq 1$ and $\sin\Omega \simeq \Omega$ (with $\Omega$ in radians), the rotation matrix $\mathbf{R}(\Omega_x, \Omega_y, \Omega_z)$ is

$$\mathbf{R}(\Omega_x, \Omega_y, \Omega_z) \simeq \begin{pmatrix} 1 & \Omega_z & -\Omega_y \\ -\Omega_z & 1 & \Omega_x \\ \Omega_y & -\Omega_x & 1 \end{pmatrix} \tag{28.14}$$

with the rotation angles as defined in Figure 28.3. The 7-parameter similarity transformation of Eq. (28.8) in its simplified form (whereby products $\mu\Omega_i$ can be safely neglected), is

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_{x'} \\ t_{y'} \\ t_{z'} \end{pmatrix} + \begin{pmatrix} \mu & \Omega_z & -\Omega_y \\ -\Omega_z & \mu & \Omega_x \\ \Omega_y & -\Omega_x & \mu \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{28.15}$$

with $\mu = \lambda - 1$ the differential scale factor. When the scale factor $\lambda$ is close to one, which is often the case, the differential scale factor $\mu$ will be a small number and is sometimes expressed in parts-per-million (ppm), with 1 ppm = $10^{-6}$.

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 book 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 axes, 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.

### 28.3.4. 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

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = (1+\mu)\cdot \mathbf{R}(\Omega'_x, \Omega'_y, \Omega'_z)\cdot \begin{pmatrix} X - X_0 \\ Y - Y_0 \\ Z - Z_0 \end{pmatrix} + \begin{pmatrix} t'_x \\ t'_y \\ t'_z \end{pmatrix} + \begin{pmatrix} X_0 \\ Y_0 \\ Z_0 \end{pmatrix} \tag{28.16}$$

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. (28.16) uses the same sign convention as Eq. (28.8). Note that, whereas many publications use for Eq. (28.8) the same sign convention as this book, most publications use for Eq. (28.16) the opposite convention. You are warned!

## 28.4. 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 28.3. This transformation preserves angles and distance ratios, 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 uniquely defined by assigning coordinates for (at least) three points, using a similar approach as we did in Section 27.3. 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 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 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, with 7 parameters, and then minimize in a 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 network. This is called a *free-network*. The underlying mathematical theory on S-system and S-transformation (in Dutch: schrankingsstelsel en schrankingstransformatie) are due to Delft University of Technology professor W. Baarda (1917-2005), see Figure 5.1.

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

## 28.5. Exercises and worked examples

Below follows a simple exercise on setting up a three-dimensional rotation matrix.

**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. (28.12). 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 target system II (e.g. with coordinates $(1, 0, 0)$).

# 29

# 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 the origin at the center of mass of the Earth, the Z-axis is coinciding with the Earth's 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 the scale tied to the SI definition of the meter. Considering that the Earth radius is about 6400 km, then to represent positions on the surface of the Earth in meters 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.

## 29.1. Geocentric coordinates (spherical coordinates)

For example, assuming a sphere with radius $R$ approximating the Earth surface, spherical coordinates $\psi$, $\lambda$ and $r$ (with $r = R + h'$, and $R$ = 6371 km the mean radius of the Earth) can be defined, see Figure 29.1. The relationship between Cartesian and spherical coordinates is given by,

$$\begin{aligned} X &= r\cos\psi\cos\lambda \\ Y &= r\cos\psi\sin\lambda \\ Z &= r\sin\psi \end{aligned} \tag{29.1}$$

The inverse relationship is given by,

$$\begin{aligned} \psi &= \arctan\left(\frac{Z}{\sqrt{X^2+Y^2}}\right) \\ \lambda &= \arctan\left(\frac{Y}{X}\right) \\ r &= \sqrt{X^2+Y^2+Z^2} \end{aligned} \tag{29.2}$$

The spherical coordinates $\psi$ and $\lambda$ 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 $\psi$ the *geocentric latitude* and $\lambda$ the *longitude* of the point. In Eqs. (29.1) and (29.2) 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).

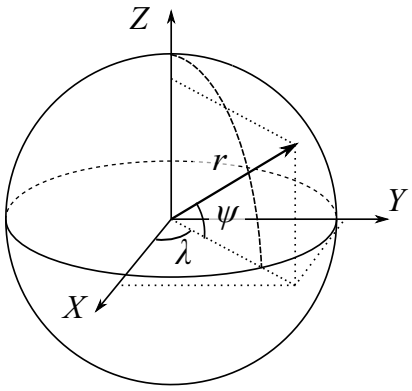



Figure 29.1: Spherical coordinates $\psi$, $\lambda$, $r$ and Cartesian coordinates $X$, $Y$, $Z$.

## 29.2. Geographic coordinates (ellipsoidal coordinates)

As shown by Newton in his Principia, 1687, see e.g. [52], 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

$$f = \frac{a-b}{a}\,,\quad e^2 = 2f - f^2 = \frac{a^2-b^2}{a^2}\,,\quad b = a(1-f) = a\sqrt{1-e^2} \tag{29.3}$$

For the Earth the semi-major axis $a$ is about 6378 km and semi-minor axis $b$ about 6357 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.

### 29.2.1. Relation between geographic and Cartesian coordinates

The position of a point with respect to an ellipsoid is given in terms of *geographic* or *geodetic* latitude $\varphi$, longitude $\lambda$ and height $h$ above the ellipsoid, see Figure 29.2. The relationship between Cartesian and geographic coordinates is given by,

$$\begin{aligned} X &= (\bar{N}+h)\cos\varphi\cos\lambda \\ Y &= (\bar{N}+h)\cos\varphi\sin\lambda \\ Z &= (\bar{N}(1-e^2)+h)\sin\varphi \end{aligned} \tag{29.4}$$

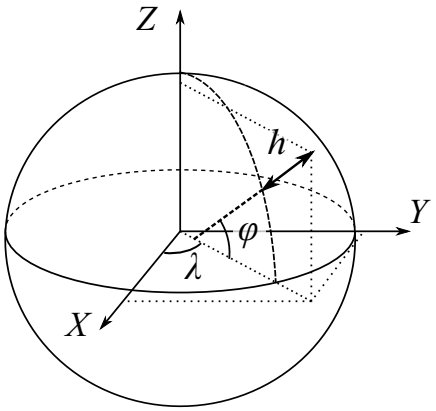



Figure 29.2: Ellipsoidal and Cartesian coordinates. The ellipsoidal latitude $\varphi$ is also known as *geodetic* or *geographic* latitude. The ellipsoidal coordinates $\varphi$ and $\lambda$ are also called geographic coordinates.

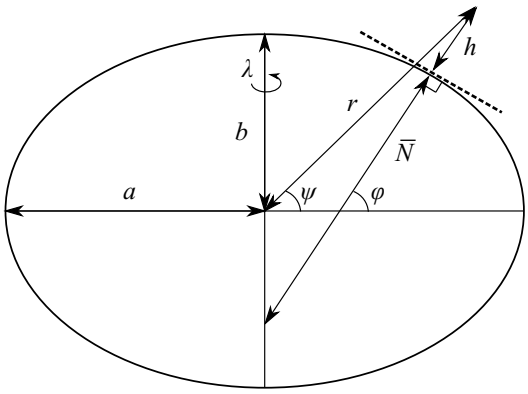



Figure 29.3: Ellipsoidal, geodetic or geographic latitude $\varphi$, geocentric (or spherical) latitude $\psi$, radius of curvature $\bar{N} = \bar{N}(\varphi)$, radius $r$, ellipsoidal height $h$, semi-major axis $a$ and semi-minor axis $b$ of the ellipsoid. The dashed line shows the local tangent plane to the ellipsoid.

The inverse relationship is given by,

$$\begin{aligned} \varphi &= \arctan\left(\frac{Z + e^2 \bar{N} \sin\varphi}{\sqrt{X^2 + Y^2}}\right) \\ \lambda &= \arctan\left(\frac{Y}{X}\right) \\ h &= \frac{\sqrt{X^2 + Y^2}}{\cos\varphi} - \bar{N} \end{aligned} \tag{29.5}$$

$\bar{N}$ in Eqs. (29.4) and (29.5) is the radius of curvature in the prime vertical, as shown in Figure 29.3.

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

$$\bar{N}(\varphi) = \frac{a}{\sqrt{1 - e^2 \sin^2 \varphi}}$$
$$\bar{M}(\varphi) = \frac{a(1 - e^2)}{(1 - e^2 \sin^2 \varphi)^{3/2}} \tag{29.6}$$

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 29.4.

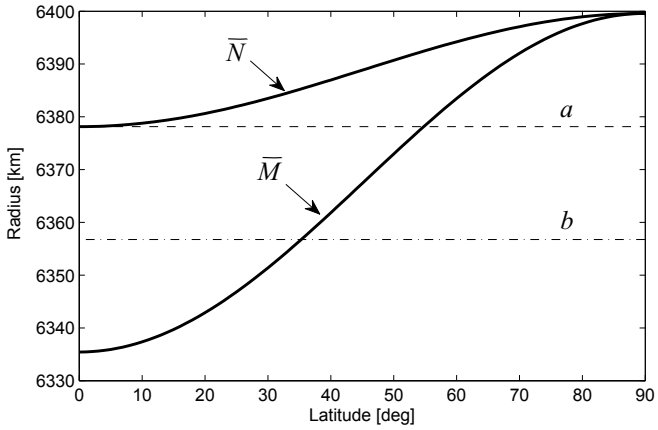



Figure 29.4: Radius of curvature $\bar{N}(\varphi)$ and $\bar{M}(\varphi)$ as function of latitude $\varphi$. The dashed lines represent the semi-major axis $a$ and semi-minor axis $b$.

### 29.2.2. Relation to units of length

As a side-step we consider Northing and Easting as a means to express small differences in latitude and longitude, between two points on the surface, in units of length, rather than angles.

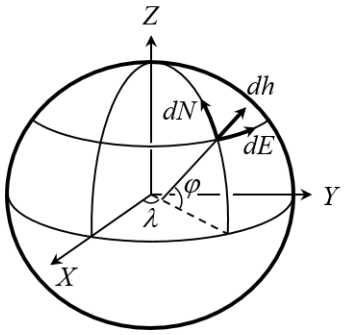



Figure 29.5: Local topocentric curvi-linear coordinate system, with origin at location $(\varphi, \lambda, h)$, with Easting and Northing, $dE$ and $dN$, and ellipsoidal height difference $dh$.

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

$$\begin{aligned} dN &= (\bar{M}(\varphi) + h)\ d\varphi \\ dE &= (\bar{N}(\varphi) + h)\cos\varphi\ d\lambda \end{aligned} \tag{29.7}$$

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. (29.6) and Figure 29.4. Both $dN$ and $dE$ are in units of meters and are often referred to as *Northing* and *Easting*, see Figure 29.5. The relations in Eq. (29.7) 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. (29.7) can be replaced simply by the radius $R$ of the spherical Earth.

### 29.2.3. Computational aspects

The geographic latitude in Eq. (29.5) must be computed by an iteration process as the geographic latitude $\varphi$ appears both in the left and right hand side of the equation. Also note that the radius of curvature $\bar{N}$ in Eq. (29.6), which is a function of the geographic latitude (that still needs to be computed by Eq. (29.5)), can be computed by the same iteration process. The iterative procedure, whereby in the first iteration $N' = \bar{N}\sin\varphi$ is approximated by $Z$, reads

$$\begin{aligned} &N'_0 = Z \\ &\text{for } i = 1, 2, \ldots \\ &\quad \varphi_i = \arctan\left(\frac{Z + e^2 N'_{i-1}}{\sqrt{X^2 + Y^2}}\right) \\ &\quad \bar{N}_i = \frac{a}{\sqrt{1 - e^2 \sin^2 \varphi_i}} \\ &\quad N'_i = \bar{N}_i \sin\varphi_i \end{aligned} \tag{29.8}$$

Usually four iterations are sufficient. For points near the surface of the Earth $\varphi$ can also be computed using a direct method of B.R. Bowring [72],

$$\varphi = \arctan\left(\frac{Z + e'^2 b \sin^3\mu}{\sqrt{X^2 + Y^2} - e^2 a \cos^3\mu}\right) \tag{29.9}$$

with $e'^2$, also called the *second eccentricity*, and $\mu$ given by

$$\begin{aligned} e'^2 &= \frac{e^2}{1 - e^2} = \frac{a^2 - b^2}{b^2} \\ \mu &= \arctan\left(\frac{aZ}{b\sqrt{X^2 + Y^2}}\right) \end{aligned} \tag{29.10}$$

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 $\psi$ and the geodetic (or geographic) latitude $\varphi$ for a point on the surface of the Earth, see Figure 29.3, is

$$\psi(\varphi) = \arctan(\frac{\bar{N}(1-e^2)+h}{\bar{N}+h}\tan\varphi) \simeq \arctan((1-e^2)\tan\varphi) \mid h \ll \bar{N} \tag{29.11}$$

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.

## 29.3. 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 $(\varphi, \lambda)$ is given by the unit direction vector $\bar{\mathbf{n}}$

$$\bar{\mathbf{n}} = \begin{pmatrix} \cos\varphi\cos\lambda \\ \cos\varphi\sin\lambda \\ \sin\varphi \end{pmatrix} \tag{29.12}$$

The ellipsoidal normal does not pass through the centre of the ellipsoid, see Figure 29.3, 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 Chapter 32. The direction of the true vertical $\mathbf{n}$ is given by the *astronomical* latitude $\phi$ and longitude $\Lambda$,

$$\mathbf{n} = -\frac{\mathbf{g}}{g} = \begin{pmatrix} \cos\phi\cos\Lambda \\ \cos\phi\sin\Lambda \\ \sin\phi \end{pmatrix} \tag{29.13}$$

with $g = \|\mathbf{g}\|$. The astronomical latitude and longitude can be determined through (zenith) measurements to the stars. The astronomical latitude $\phi$ is the angle between the equatorial plane and the true vertical at a point on the surface; the ellipsoidal, geodetic or geographic latitude $\varphi$ is the angle between the equatorial plane and the ellipsoidal normal. A similar distinction exists for the astronomical longitude $\Lambda$ and ellipsoidal longitude $\lambda$. 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,

$$\begin{aligned} \xi &= \phi - \varphi \\ \eta &= (\Lambda - \lambda)\cos\varphi \end{aligned} \tag{29.14}$$

Astronomical latitude $\phi$ and longitude $\Lambda$ are obtained from astronomical observations to stars whose positions (declination $\delta$ and right ascension $\alpha$ ) in a celestial reference system are

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

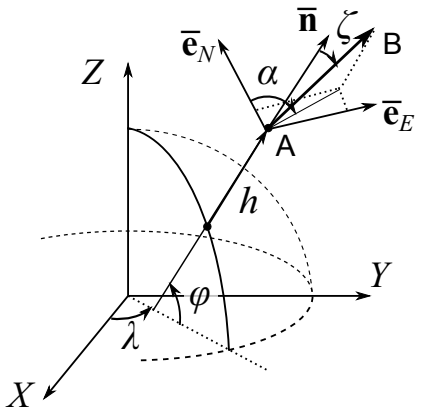



Figure 29.6: Local right-handed Cartesian topocentric system in point A, with ellipsoidal coordinates $(\varphi, \lambda, h)$, with the local azimuth $\alpha$ and zenith angle $\zeta$ for the direction $\vec{AB}$. The vertical (ellipsoidal normal vector) $\bar{\mathbf{n}}$ is the third axis of the local right-handed system , $\bar{\mathbf{e}}_E$ is the first 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 second axis completing the topocentric system. Vectors $\bar{\mathbf{e}}_N$ and $\bar{\mathbf{e}}_E$ together span the local *tangent* plane to the ellipsoid, see Figure 29.3. The bar on top of these three vectors denotes that they are related to the ellipsoid.

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. We briefly elaborate on the deflection of the vertical in Section 32.6.

## 29.4. 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 axes with respect to the local vertical and geographic North. This type of 3D Cartesian coordinate system is called a local *topocentric* coordinate system and its coordinates are called topocentric coordinates. In Figure 29.6 the origin of the local Cartesian topocentric system is the (observation) point A with geographic coordinates $(\varphi, \lambda, h)$. The vectors $\bar{\mathbf{e}}_E$, $\bar{\mathbf{e}}_N$ and $\bar{\mathbf{n}}$ form the three axes of a right-handed local topocentric system centered at A, with the third axis along the normal of the ellipsoid $\bar{\mathbf{n}}$, the first axis $\bar{\mathbf{e}}_E$ orthogonal to the plane of the meridian and positive to the East, and the second 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 Cartesian topocentric system are denoted by $E$ (East), $N$ (North) and $U$ (Up). The system itself is also called a East-North-Up (ENU) coordinate system.

In principle it is possible to work with coordinates $(E, N, U)$ which are defined with respect to the origin $A$ in Figure 29.6 and use the 7-parameter similarity transformation of Eq. (28.8) to convert between $(E, N, U)$ and $(X, Y, Z)$ coordinates. However, when dealing with two points (and neither one the origin $A$), it is more convenient to work with coordinate *differences* $(\Delta E, \Delta N, \Delta U)$ and $(\Delta X, \Delta Y, \Delta Z)$ . The relation between the differential coordinates $(\Delta X, \Delta Y, \Delta Z)$

[2]Astronomical latitude is not to be confused with declination, the coordinate astronomers use 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.

and $(\Delta E, \Delta N, \Delta U)$ is,

$$\begin{pmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{pmatrix} = \begin{pmatrix} -\sin\lambda & -\sin\varphi\cos\lambda & \cos\varphi\cos\lambda \\ \cos\lambda & -\sin\varphi\sin\lambda & \cos\varphi\sin\lambda \\ 0 & \cos\varphi & \sin\varphi \end{pmatrix} \begin{pmatrix} \Delta E \\ \Delta N \\ \Delta U \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{e}}_E & \bar{\mathbf{e}}_N & \bar{\mathbf{n}} \end{pmatrix} \begin{pmatrix} \Delta E \\ \Delta N \\ \Delta U \end{pmatrix} \tag{29.15}$$

with $\bar{\mathbf{e}}_E$, $\bar{\mathbf{e}}_N$ and $\bar{\mathbf{n}}$ the three axes of a right-handed local topocentric system centered at the observation point $(\varphi, \lambda, h)$. With coordinate differences any translation cancels and the transformation matrix of Eq. (29.15) can be found by two rotations as $R_3(-\lambda - \frac{\pi}{2})R_1(\varphi - \frac{\pi}{2})$, cf. Eq. (28.12) and Figure 29.6.

The inverse relation of Eq. (29.15) is,

$$\begin{pmatrix} \Delta E \\ \Delta N \\ \Delta U \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{e}}_E & \bar{\mathbf{e}}_N & \bar{\mathbf{n}} \end{pmatrix}^{-1} \begin{pmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{e}}_E & \bar{\mathbf{e}}_N & \bar{\mathbf{n}} \end{pmatrix}^{T} \begin{pmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{pmatrix} \tag{29.16}$$

where we used the property that the inverse of a rotation matrix is the transpose of the matrix. Here the transformation follows as $R_1(\frac{\pi}{2} - \varphi)R_3(\lambda + \frac{\pi}{2})$.

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 $\alpha$ of $0°$, and a point to the East $+90°$, see Figure 29.6. The azimuth $\alpha$ and zenith angle $\zeta$ are defined by,

$$\begin{pmatrix} \Delta E \\ \Delta N \\ \Delta U \end{pmatrix} = s \begin{pmatrix} \sin\alpha\sin\zeta \\ \cos\alpha\sin\zeta \\ \cos\zeta \end{pmatrix} \tag{29.17}$$

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

$$\begin{aligned} \alpha &= \arctan(\frac{\Delta E}{\Delta N}) = \arctan(\frac{<\mathbf{e}_E, \mathbf{s}>}{<\mathbf{e}_N, \mathbf{s}>}) \\ \zeta &= \arctan(\frac{\sqrt{\Delta E^2 + \Delta N^2}}{\Delta U}) = \arccos(\frac{\Delta U}{s}) = \arccos(\frac{<\bar{\mathbf{n}}, \mathbf{s}>}{s}) \\ s &= \sqrt{\Delta E^2 + \Delta N^2 + \Delta U^2} = \sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2} = \sqrt{<\mathbf{s}, \mathbf{s}>} \end{aligned} \tag{29.18}$$

with $\mathbf{s} = (\Delta X, \Delta Y, \Delta Z)^T$. Compare this to Eq. (28.4), of Chapter 28, 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 $\alpha$ and zenith angle $\zeta$ are computed in point A, the origin of the topocentric system, as shown in Figure 29.6, then $(\Delta E, \Delta N, \Delta U)$ in Eqs. (29.15), (29.16), (29.17) and (29.18) 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. (29.4).

The order of the coordinates in Eqs. (29.15) and (29.16) 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 practice for geographic coordinates of giving the latitude before the longitude.

The coordinates $E$ and $N$ ($\Delta E$ and $\Delta N$) are sometimes also referred to as *Easting* and *Northing*, however, Northing and Easting have been defined before in Eq. (29.7) of Section 29.2.2, and the two are not exactly the same.

In Section 29.2.2, differential ellipsoidal coordinates $d\varphi$ and $d\lambda$), given in radians, were expressed in Northing $dN$ and Easting $dE$ in meters, using the relation of Eq. (29.7)

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

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. (29.6) and Figure 29.4. $d\varphi$ and $d\lambda$ must be given in units of radians, but $dN$ and $dE$ are in units of meters.

The difference with $\Delta N$ and $\Delta E$ is that $dN$ and $dE$ are curvilinear coordinates, whereas $N$ and $E$ are Cartesian coordinates. For small values of $\Delta N$ and $\Delta E$ we have $dN \simeq \Delta N$ and $dE \simeq \Delta E$, but although $dh \simeq \Delta U$, the surface $dh = \text{const}$ represents a curved surface, whereas $\Delta U = \text{const}$ 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 $\alpha$ and zenith angle $\zeta$ 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.

To add to the possible confusion, map-coordinates, resulting from a map projection, are also often called Easting and Northing, see Chapter 30. However, the two definitions of Easting and Northing that were discussed in this section are actually examples of two different map projections.

## 29.5. 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 $\varphi$ should never be confused with the geocentric or spherical latitude $\psi$, or astronomical latitude $\phi$, which are two different types of coordinates. Also you should not confuse geodetic longitude $\lambda$ with astronomical longitude $\Lambda$. For the ellipsoidal height $h$ it is a different story. Because the ellipsoid is off from an equipotential surface by up to 100 meter, 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 33 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 32). This leaves us with a couple of options to represent positions

- use Cartesian coordinates $X$, $Y$ and $Z$ to represent positions in three dimensions

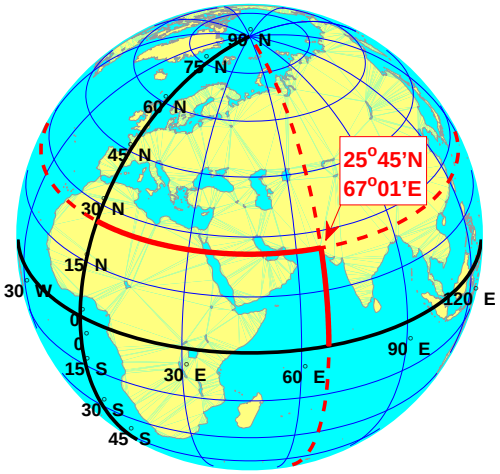



Figure 29.7: Latitude and longitude grid as seen from outer space (orthographic azimuthal projection). The prime meridian (through Greenwich) and equator are in black with latitude and longitude labels. The meridian and parallel through Karachi, Pakistan, $25°45'$N $67°01'$E, are the dotted lines in 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).

- use geographic coordinates and/or height
  - ellipsoidal latitude $\varphi$ and longitude $\lambda$ 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 29.7. 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 $+180°$) 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 $+90°$), 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

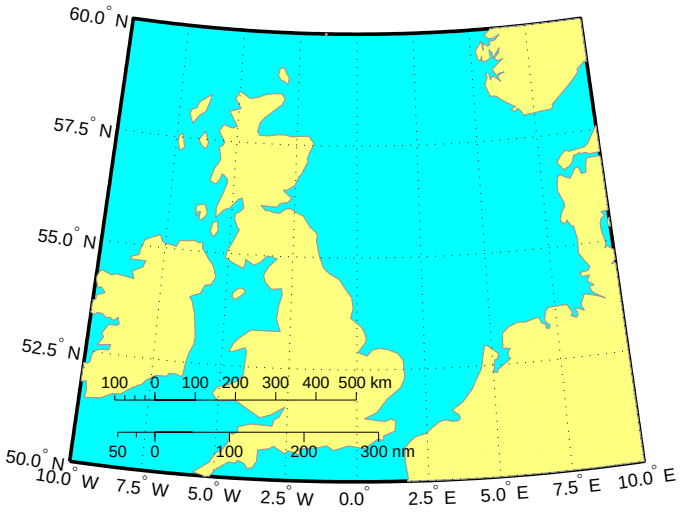



Figure 29.8: The latitude and longitude grid over the North-Sea in an equidistant conic projection of uniform scale. One degree of latitude is about $60$ nm (Nautical miles), or more precisely $111.2$ km at $50°$ and $111.4$ km at $60°$ latitude. However, one degree of longitude is much shorter, it varies 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.

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 (1852 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. (29.7) (on account of the term $\cos\varphi$, which result in the meridians converging at the poles), as is shown in Figure 29.8 for the North-Sea area. From Eq. (29.7) 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 $52°$ 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 compu-

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

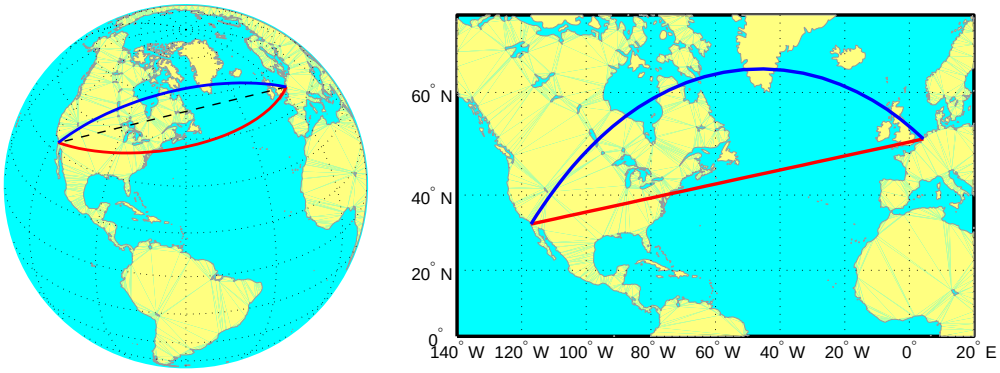



Figure 29.9: Great circle (blue), rhumb line or loxodrome (red) and straight line (black dashed) between Delft, NL, 52°N 4.37°E and San Diego, CA, USA, 32.8°N 117.1°W. The plot on the left uses an orthographic azimuthal projection, with the Earth as seen from outer space, while the plot on the right 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.

tation 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 29.6, map projections are discussed in Chapter 30.

## 29.6. 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 geometry is the shortest path between points on a sphere or ellipsoid, which is called *geodesic* (in Dutch: *geodetische lijn*). On a sphere geodesics are great circles. This is illustrated in Figure 29.9. 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 $\varphi_1$ and longitude $\lambda_1$ of point P1, and the azimuth $\alpha_1$ and distance $s_{12}$ from point P1 to P2, determine the latitude $\varphi_2$ and longitude $\lambda_2$ of point P2, and azimuth $\alpha_2$ in point P2 to P1.

**Inverse (second) geodetic problem** Given the latitude $\varphi_1$ and longitude $\lambda_1$ of point P1, and latitude $\varphi_2$ and longitude $\lambda_2$ of point P2, determine the distance $s_{12}$ between point P1 and P2, azimuth $\alpha_1$ from P1 to P2, and azimuth $\alpha_2$ 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 segment with azimuth $\alpha$ and length $ds$ to differential ellipsoidal coordinates $(d\varphi, d\lambda)$,

$$\begin{aligned} \frac{d\varphi}{ds} &= \frac{\cos\alpha}{\bar{M}(\varphi)} \\ \frac{d\lambda}{ds} &= \frac{\sin\alpha}{\bar{N}(\varphi)\,\cos\varphi} \end{aligned} \tag{29.19}$$

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. (29.6) and Figure 29.4, and with $\bar{N}(\varphi)\cos\varphi$ the radius of the circle of latitude $\varphi$. See also Eq. (29.7) 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

$$\frac{d\alpha}{ds} = \sin\varphi\frac{d\lambda}{ds} = \frac{\tan\varphi}{\bar{N}(\varphi)}\sin\alpha \tag{29.20}$$

Eqs. (29.20) and (29.19) 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 book.

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 $\alpha_1$ and $\alpha_2$, and distance $s_{12}$ along the great circle, are

$$\begin{aligned} \tan\alpha_1 &= \frac{\sin\lambda_{12}}{\cos\phi_1\tan\phi_2 - \sin\phi_1\cos\lambda_{12}} \\ \tan\alpha_2 &= \frac{\sin\lambda_{12}}{-\cos\phi_2\tan\phi_1 + \sin\phi_2\cos\lambda_{12}} \\ \cos\sigma_{12} &= \sin\phi_1\sin\phi_2 + \cos\phi_1\cos\phi_2\cos\lambda_{12} \end{aligned} \tag{29.21}$$

with $\lambda_{12} = \lambda_2 - \lambda_1$[5]. The distance is given by $s_{12} = R\,\sigma_{12}$, where $\sigma_{12}$ is the central angle (in radians) between the two points and $R$ the Earth radius. For practical computations the

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

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 WGS84 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 $\alpha_1$ and distance $s_{12}$ along the great circle, the latitude and longitude of P2 are found by,

$$\begin{aligned}
\tan\phi_2 &= \frac{\sin\phi_1\cos\sigma_{12} + \cos\phi_1\sin\sigma_{12}\cos\alpha_1}{\sqrt{(\cos\phi_1\cos\sigma_{12} - \sin\phi_1\sin\sigma_{12}\cos\alpha_1)^2 + (\sin\sigma_{12}\sin\alpha_1)^2}} \\
\tan\lambda_{12} &= \frac{\sin\sigma_{12}\sin\alpha_1}{\cos\phi_1\cos\sigma_{12} - \sin\phi_1\sin\sigma_{12}\cos\alpha_1} \\
\tan\alpha_2 &= \frac{\sin\alpha_1}{\cos\sigma_{12}\cos\alpha_1 - \tan\phi_1\sin\sigma_{12}}
\end{aligned} \tag{29.22}$$

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

The corresponding formulas for on the ellipsoid can be found in e.g. [73].

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*.

## 29.7. Exercises and worked examples

This section presents several exercises on working with ellipsoidal coordinates.

**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. (29.3), 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 31.1.

**Answer 4** Converting geographic coordinates into Cartesian coordinates is done through Eq. (29.4), with the expression for the radius of curvature in the prime vertical in Eq. (29.6). 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 31.1), the geographic coordinates of this location.

**Answer 5** The formal computation goes through Eqs. (29.5) and (29.6), and requires an iteration, see Eq. (29.8). 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 29.4.

**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. (29.7) 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 29.4 or Eq. (29.6). 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. (29.7) 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 34 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 31.

# 30

# 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. A curved surface is mapped onto a flat plane. This is known as a *map projection*. From then on simple 2D Euclidian geometry can be used.

## 30.1. 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 $(\varphi, \lambda)$ into a set of planar coordinates $(x, y)$ 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 book a *map projection* is defined as the mathematical transformation

$$\begin{aligned} x &= g(\varphi, \lambda, h) \\ y &= f(\varphi, \lambda, h) \end{aligned} \tag{30.1}$$

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

Many different map projections are in use all over the world for different applications and for good reasons. 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 software packages are specifically written for one specialized map projection, and give incorrect results when using coordinates from a different type of projection.

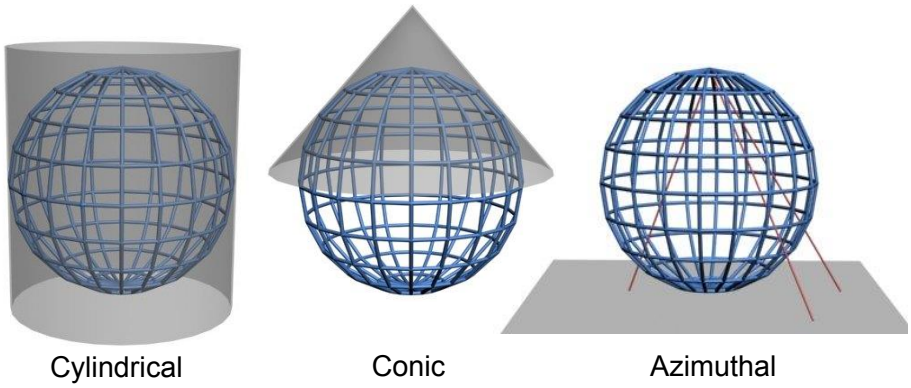



Figure 30.1: Cylindrical, conic and azimuthal map projection types. Image of cylindrical, conic and azimuthal map projection, by Traroth, March 2005, taken from Wikimedia Commons [9], under CC BY-SA 3.0 license.

## 30.2. Map projection types and properties

The properties of a map projection depend mainly on the type and position of the projection surface and the projection origin that is used.

### 30.2.1. Projection surface

Map projections can be grouped into four groups depending on the nature of the projection surface, see Figure 30.1,

**Cylindrical map projections** The projection surface 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 cylindrical map projection is the Mercator projection. Figure 29.9 (right part) of the previous chapter is a Mercator projection.

**Conic map projections** The projection surface 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 29.8 of the previous chapter.

**Azimuthal map projections** The projection surface 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 35, and the orthographic azimuthal projection used in Figure 29.7 of the previous chapter.

**Miscellaneous projections** Mostly used for cartographic purposes.

Any projection can be applied in the *normal*, *transverse* and *oblique* position of the cylinder, cone or tangent plane, as shown in Figure 30.2 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 30.4.2 and Figure 29.9. In the transverse case the axis of projection is in the equatorial plane (orthogonal to the minor axis), for example, in the *Universal Transverse Mercator* (UTM) projection small strips are mapped on a cylinder wrapped around the poles and with a specific meridian as line of contact. In the oblique case the axis of projection does not coincide with the semi-minor axis or equatorial plane.

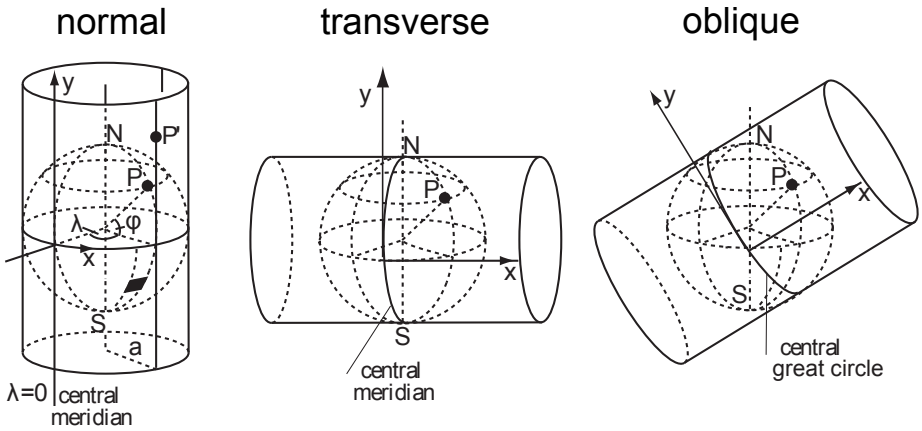



Figure 30.2: Normal, transverse and oblique projection for a cylinder. Image on cylindrical projection aspects by Peter Mercator, own work, November 2009, taken from Wikimedia Commons [9]. Public Domain.

### 30.2.2. Projection origin

Map projections also differ in the *point of perspective* that is used. Figure 30.3 shows three common choices for azimuthal projections. 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 30.1 on the right. On the other hand, for the *orthographic azimuthal projection*, which was used for Figure 29.7 and in Figure 29.9 (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. In Figure 30.3 the mapping plane is *tangent* to the sphere. If the mapping plane is shifted slightly into the sphere, the map projection is called *secant*.

### 30.2.3. Properties

Some distortion in the geometrical elements, distance, angles and area, is inevitable in map projections through transformation Eq. (30.1). In this respect map projections are divided into

**Conformal projections** preserve the angle of intersection of any two curves.

**Equal Area (equivalent) projections** preserve the area or scale.

**Equi Distance (conventional) projections** preserve 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 27.2) in an infinitesimally small region. A conformal mapping

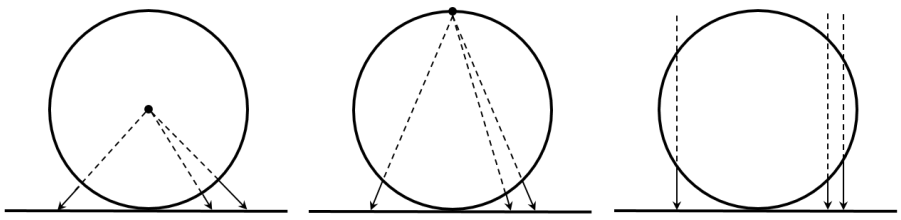


Figure 30.3: Cross-section of azimuthal map projection, the mapping surface being a flat plane, with central point of projection, so-called gnomonic (left), stereographic projection (middle), and orthographic projection (right).

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 (1841) ellipsoid on the sphere, the second step a stereographic projection onto a plane tangential to the ellipsoid with the center at Amersfoort, see also Figure 35.2.

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 31.4)
- latitude of natural origin or standard parallel ($\varphi_0$) for cylindrical and azimuthal projections, or, the latitude of first standard parallel ($\varphi_1$) and second standard parallel ($\varphi_2$) for conic projections
- longitude of natural origin (the central meridian) ($\lambda_0$)
- 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.

## 30.3. 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 30.4.5, 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, $31$N and $32$N, 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 19th and early 20th 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.

## 30.4. 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 30.4. The map projection turns (spherical) coordinates $(\varphi, \lambda)$[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 $(\varphi = 0, \lambda = 0)$. The Earth's surface is approximated by a sphere with radius $R$. The middle part of Figure 30.4 shows a top view of the equatorial plane. If we express $\lambda$ in radians, the distance from $(\varphi = 0, \lambda = 0)$ *along* the equator to an object at longitude $\lambda$ 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 same meridian have a constant $x$ value.

The function $y = y(\varphi)$ to project latitude $\varphi$ onto $y$ values is still open, it can be any one from an unlimited number of functions. In Figure 30.4, on the right, one such function is illustrated: the central cylindrical projection.

### 30.4.1. 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 30.4 shows a meridianal cross section of the Earth at longitude $\lambda$. 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

$$\begin{aligned} x &= \mu R(\lambda - \lambda_0) \\ y &= \mu R \tan \varphi \end{aligned} \tag{30.2}$$

with $\mu$ a scaling factor and $\lambda_0$ the central meridial (e.g. Greenwhich meridian with $\lambda_0 = 0$), with $\lambda$ and $\lambda_0$ expressed in radians.

The true scale on the equator is unity for $\mu = 1$. Everywhere else the linear scale is stretched by a factor of $1/\cos\varphi$ in the $x$-axis direction, and $1/\tan\varphi$ in the $y$-axis direction.

The central cylindrical projection is neither conformal or equal area. Distortion increases so rapidly away from the equator, see Figure 30.5, that the central cylindrical is seldomly used

[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 does not matter for a graphical representation. However, formally, geographic coordinates should first be projected onto spherical coordinates, before the map projection is applied.

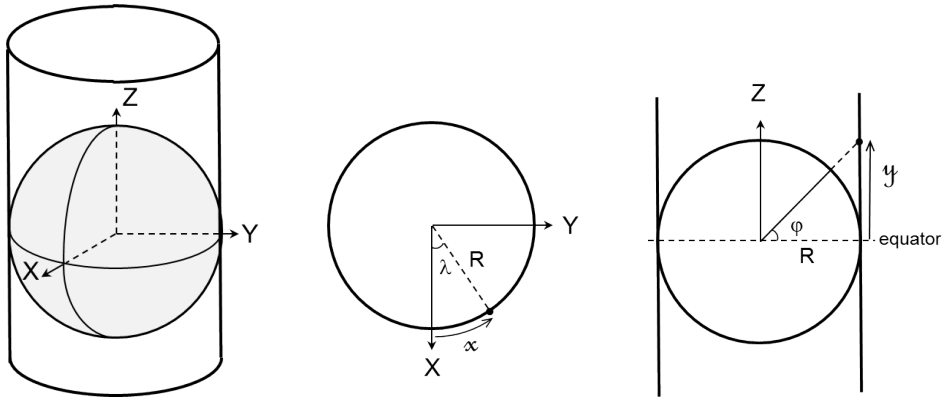



Figure 30.4: For the cylindrical projection, the mapping plane is wrapped around the Earth like a cylinder (left), longitude $\lambda$ turns into map-coordinate $x$ (middle; horizontal cross-section of equator plane), and latitude $\varphi$ turns into map-coordinate $y$ (right; vertical cross-section).

for practical maps. Its vertical, latitudinal, stretching is even greater than that of the Mercator projection, which we discuss next.

### 30.4.2. Mercator projection

The Mercator projection is a cylindrical map projection, presented by the Flemish geographer and cartographer Gerardus Mercator, in 1569, see also Figure 37.9. 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 29.9, 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 $1/\cos\varphi$ 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 $y'(\varphi) = R/\cos\varphi$[2]. Integrating this equation gives

$$y(\varphi) = R\ln\left[\tan(\frac{\pi}{4} + \frac{\varphi}{2})\right] \tag{30.3}$$

This function is illustrated in Figure 30.5. The map projection formulas for a basic normal Mercator projection are thus

$$\begin{aligned} x &= \mu R(\lambda - \lambda_0) \\ y &= \mu R\ln\left[\tan(\tfrac{\pi}{4} + \tfrac{\varphi}{2})\right] \end{aligned} \tag{30.4}$$

with $\mu$ a scaling factor, $\lambda_0$ the central meridian. The angular units are radians. The true scale on the equator is unity for $\mu = 1$. Everywhere else the linear scale is stretched by a factor

[2]the derivative for the central cylindrical projection is: $y'(\varphi) = R/\cos^2\varphi$

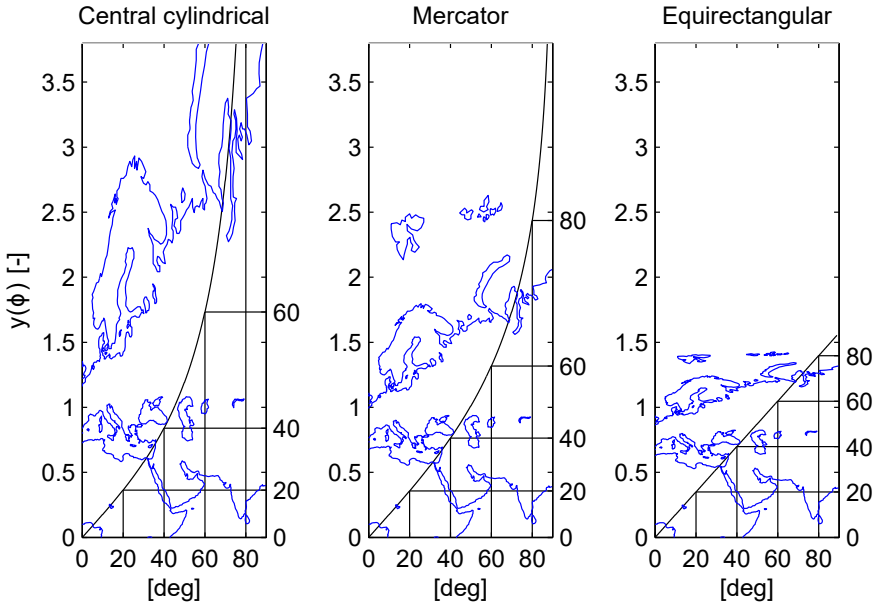



Figure 30.5: Mapping function $y = y(\varphi)$ for the central cylindrical, Mercator and equirectangular (Plate Carrée) projections (showing (part of) Europe, Russia, the Middle East and Africa). The function $y(\varphi)$ with, $R = 1$, is the black line. The x-axis is $\varphi$ in degrees, the y-axis 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.

of $1/\cos\varphi$. This distorts the size of geographical objects far from the equator; objects like Greenland and Antarctica appear to be much larger than they in reality are, see Figure 29.9, and also Figure 30.5. The Mercator projection is conformal, it preserves angles, but it is definitely not an equal area projection. By choosing a value of $\mu$ 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 $(\lambda, \psi)$ instead of geographical coordinates $(\lambda, \varphi)$. 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 $(\lambda, \varphi)$ is the geographic longitude and latitude, the Mercator projection must be modified to remain conformal. The map projection formula in case of the ellipsoidal model are

$$\begin{aligned} x &= \mu R(\lambda - \lambda_0) \\ y &= \mu R \ln\left[\tan(\tfrac{\pi}{4} + \tfrac{\varphi}{2})\left(\tfrac{1-e\sin\varphi}{1+e\sin\varphi}\right)^{\frac{e}{2}}\right] \end{aligned} \tag{30.5}$$

with $e$ the eccentricity of the ellipsoid and all angles expressed in radians.

### 30.4.3. 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

$$\begin{aligned} x &= R(\lambda - \lambda_0) \\ y &= R\varphi \end{aligned} \tag{30.6}$$

with angles expressed in radians. The parallels and meridians are being equidistant in the map and form a square grid[3], as can be seen in Figure 30.5. 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 at $\varphi_1$ North and South of the equator, are

$$\begin{aligned} x &= R(\lambda - \lambda_0)\cos\varphi_1 \\ y &= R(\varphi - \varphi_1) \end{aligned} \tag{30.7}$$

with angles in radians. 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 $\varphi_1$. 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.

### 30.4.4. 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. It uses the same spherical formulas of Eq. (30.4) as the standard Mercator, however, the Web Mercator uses the spherical formulas with the geographical coordinates $(\lambda, \varphi)$ in the WGS84 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 agencies have declared this map projection to be unacceptable for any official use.

### 30.4.5. 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 30.2, 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

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

180 km on each side of, and about parallel to, the central meridian ($\arccos 0.9996 = 1.62°$ 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.

## 30.5. Exercises and worked examples

This section presents two simple exercises on projecting the Earth's surface on a plane.

**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.

**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 30.6. Eventually you may want to apply a scale factor $\mu$, so that $x = \mu X$ and $y = \mu Y$.

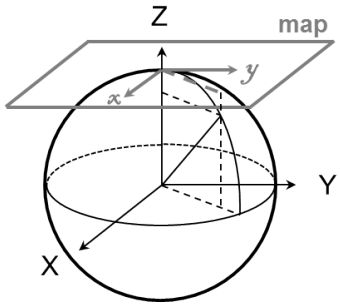



Figure 30.6: Orthographic azimuthal map projection (Answer 1). The point of tangency of the mapping plane is the North Pole.

**Question 2** The left part of Figure 30.7 shows an orthographic azimuthal map projection, where the point of tangency of the mapping plane is set to the middle of The Netherlands, close to Amersfoort, indicated by A. This map projection does not preserve distances. The Netherlands in West-East direction is about 160 km wide, so over the Earth's surface the distance from W to A is 80 km, and identically from A to E. By how much is distance W-E shorter on the map? You can assume that the Earth is a perfect sphere, with radius $R$=6378 km.

**Answer 2** The distance W-E over the Earth's surface equals $2s$ as shown in the right part of Figure 30.7. The distance W-E on the map equals $2d$. Angle $\alpha$, in radians, is easily found as $\alpha = \frac{s}{R}$, see Figure 30.7. We also have $\sin\alpha = \frac{d}{R}$, and we can use the first two terms of the Taylor series of the sine, which provides a good approximation for small angles, as in this case. Hence, with $\sin x \approx x - \frac{x^3}{3!}$, we have $\frac{s}{R} - \frac{s^3}{3!R^3} \approx \frac{d}{R}$, and $s - d \approx \frac{s^3}{6R^2}$. With $s$=80 km and $R$=6378 km, we arrive at $s - d \approx 2.1$ m, hence, in the map the Netherlands have shrunk by 4.2 m.

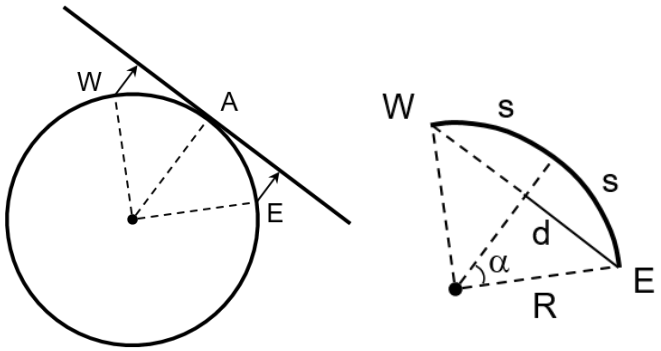



Figure 30.7: Orthographic azimuthal map projection (left) for Question 2, and the method for determining distance $d$ for given distance $s$ over the Earth's surface (right) used in the answer.

# 31

# Datum transformations and coordinate conversions

In practice geospatial projects often involve coordinates of Earth's surface topography and objects 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 called *geodetic datum transformations*. In this chapter we discuss geodetic datum transformations and coordinate conversions.

## 31.1. Geodetic datum

In 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. For instance, take the example of Cartesian and ellipsoidal coordinates first. For 3D Cartesian coordinates we need to define seven parameters: three for the origin, three for the orientation of the axes, and one for scale (see Section 28.4). For ellipsoidal coordinates, i.e. geographic latitude, longitude and ellipsoidal height, we need to define first the shape of the ellipsoid, i.e. the length of the semi-major axis and flattening (or the length of the semi-minor axis or inverse flattening) of the chosen ellipsoid (2 parameters), and secondly the position of the ellipsoid with respect to the Earth, i.e. origin, orientation and scale of the ellipsoid (7 parameters). These definitions constitute the *geodetic datum*.

The whole of the coordinate system, the geodetic datum, the type of coordinates that are used, and their parameters, is what defines a spatial coordinate system, or a *coordinate reference system* (CRS).

When countries developed their national coordinate systems at the end of the 19th and beginning of the 20th century, each country chose 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 with respect to the Earth. Table 31.1 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 differs.

Therefore, each spatial coordinate system, or coordinate reference system, has a geodetic *datum*. The geodetic datum, or *datum* for short, specifies how a coordinate system is linked to the Earth: it consists of parameters that describe how to define the origin of the coordinate

| 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 31.1: 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.

axes, how to orient the axes, and how 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*.

## 31.2. Coordinate operations

It is common that spatial reference systems rely on 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 axes 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 31.1.

*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 31.1, 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 31.1. These affine or polynomial transformations are mostly coarse 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 31.1 is very straightforward: this is accomplished by simply dropping the height coordinate.

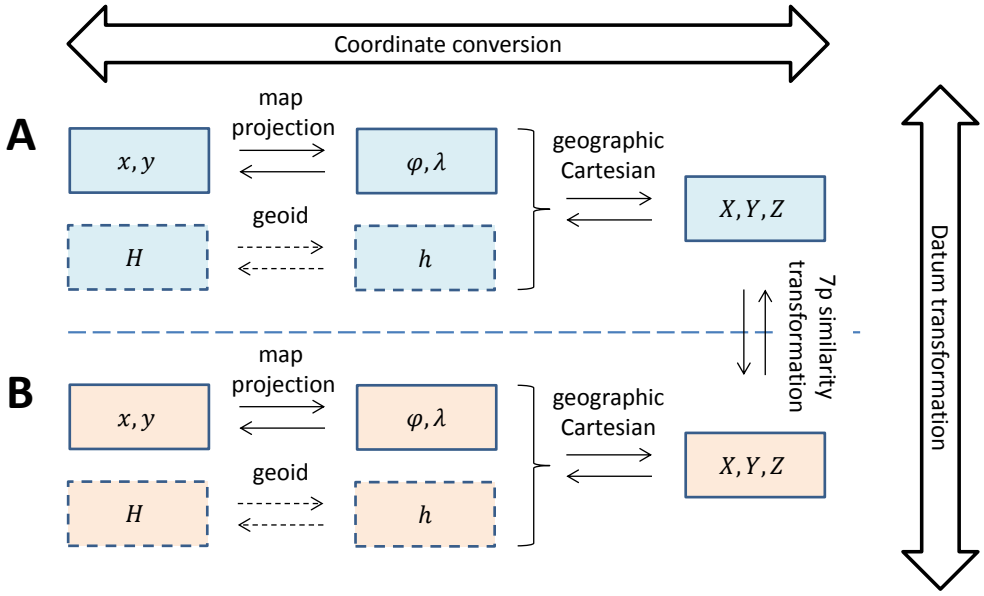



Figure 31.1: Coordinate conversions and datum transformations. Horizontal operations represent coordinate conversions. The vertical operations are datum transformations from system A to B. Not shown in this diagram are polynomial transformations (approximations) directly between map coordinates or geographic coordinates of the two systems.

The reverse, from 2D to 3D, is indeterminate. This is an issue when 2D coordinates (geographic or grid coordinates) have to be transformed into another datum or reference system, as this involves 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 positions 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 is 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 numerical values assigned to the coordinates of 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 previous estimates.

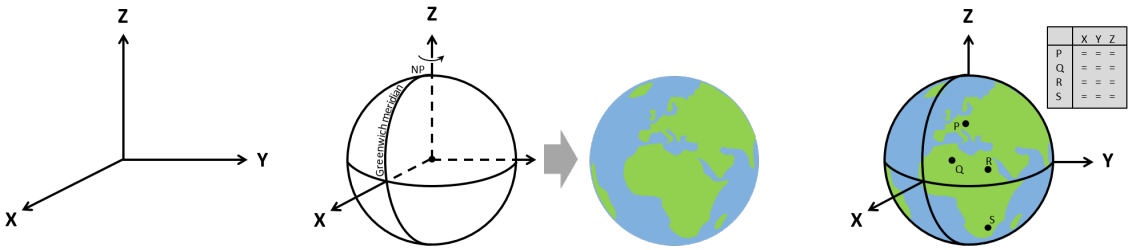



Figure 31.2: At left a coordinate system, which is purely a mathematical concept. In the middle a reference ellipsoid and parameters have been *defined* to model the Earth, and together with the coordinate system, this forms a *reference system*. A reference system is a theoretical concept, which is still to be connected to physical Earth. A reference system is only realized in practice once numerical coordinate values are assigned to physical points and objects on Earth; this realization is called a *reference frame*, shown at right. The origin, orientation and scale of the coordinate system is referred to as *geodetic datum*. The geodetic datum definition is part of a reference system, but its realization is only through a reference frame. A reference system and its realization in a reference frame are together referred to as a *coordinate reference system* (CRS) in practice.

3. The third possibility is that there is no third party that has determined the transformation parameters, and you as a 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 should all come from the same realization, or, reference frame.

An illustration and summary of the most important terminology in this chapter so far, are given in Figure 31.2.

## 31.3. 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 the 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 offshore work and to define the European borders. The satellite era saw the development of a number of worldwide 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 100 m of the geoid worldwide. 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 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, result in unnecessary costs and in worst case a disaster happening.

## 31.4. EPSG dataset and WKT-CRS

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 [74]. For instance, the EPSG code for the Dutch RD coordinate system is EPSG:28992 (Amersfoort / RD New - Netherlands). The EPSG website also provides the equations for the various mappings that have been stored in the EPSG database [75]. The EPSG database, although extremely useful, has no official status, and sometimes contains only approximate parameters.

Another format to describe spatial reference systems is the Well-Known Text representation (WKT or WKT-CRS). This is a text markup language for representing spatial reference systems and transformations between spatial reference systems. The formats were originally defined by the Open Geospatial Consortium (OGC) and is an ISO standard. For example, the WKT below describes a two-dimensional geographic coordinate reference system with a latitude axis first, then a longitude axis. The coordinate system is related to Earth by the WGS84 geodetic datum (example taken from Wikipedia under CC BY-SA license [76]).

```
GEODCRS["WGS 84",
  DATUM["World Geodetic System 1984",
    ELLIPSOID["WGS 84", 6378137, 298.257223563, LENGTHUNIT["metre", 1]]],
  CS[ellipsoidal, 2],
    AXIS["Latitude (lat)", north, ORDER[1]],
    AXIS["Longitude (lon)", east, ORDER[2]],
    ANGLEUNIT["degree", 0.0174532925199433]]
```

The WKT format can also be used to describe the operation methods and parameters to convert or transform coordinates between two different coordinate reference systems (see next section).

## 31.5. Coordinate conversion and transformation software [*]

Software for map projections, coordinate conversions and datum transformations is provided for instance by the open source PROJ package [77] 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. This program is limited to converting within one datum. 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. The `cct` program is a 4D equivalent to the `proj` and `cs2cs` programs to perform coordinate conversion and transformation that include time.

Besides PROJ strings, PROJ can also use Well-Known Text (WKT) and as spatial reference ID's (such as EPSG codes) to describe the coordinate reference system. WKT or spatial reference ID's are preferred over PROJ strings as they can contain more information about a given CRS. If you know the EPSG identifiers these can be used to specify the source and destination CRS in `cs2cs`. E.g. `cs2cs +init=epsg:4326 +to +init=epsg:28992 ...` will transform geodetic coordinates in the WGS84 reference frame to RD coordinates. Though PROJ supports the Dutch RD coordinate system, through the EPSG code EPSG:28992, users should nevertheless be extremely careful. For this particular example you probably do not have coordinates in WGS84, but in one of the ITRF's or in ETRS89 (see e.g. Section 14.4 and Chapter 34), each with its own EPSG code. Another concern is that the EPSG database is actively maintained, and your local version may not be fully up to date, or the grid correction file may be missing. For visualizations on a map or GIS system these details do not matter, but they do matter for when an accuracy better than 1 meter is required.

In Chapter 35 we will come back to the matter of transforming coordinates into the Dutch RD system and heights into the NAP system using the RDNAPTRANS™procedure. The latest version of RDNAPTRANS™, called RDNAPTRANS™2018, is fully implemented in PROJ using a pipeline of `+proj` strings.

## 31.6. Exercises and worked examples

This section presents just a quick exercise on comparing the semi-major axes of two ellipsoids.

**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 31.1). Hence, the difference is 739.845 m. That is a nearly a kilometer difference at the equator!

# 32

# 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 first illustrated by means of the very simple example of the Earth being a perfect sphere, and initially restricting the discussion to gravitation. Eventually we introduce the geoid and the so-called deflection of the vertical.

## 32.1. Introduction

Gravitation, 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, see e.g. [52], 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 H.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) $\ddot{\mathbf{r}}$

$$\mathbf{F} = m\ddot{\mathbf{r}} \tag{32.1}$$

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

## 32.2. Earth's gravitational 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 $\ddot{\mathbf{r}}$. Acceleration has unit [m/s$^2$].

Weight is defined as the force of gravitation on an object. Its magnitude is (also) not a constant value over the Earth or above it. If the force of gravitation 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 vector $\mathbf{g}$ is directed towards the center of Earth, and causes the satellite to maintain a circular orbit. An object located extremely far from the Earth (and any other body) would be truely weightless (but still have the same mass).

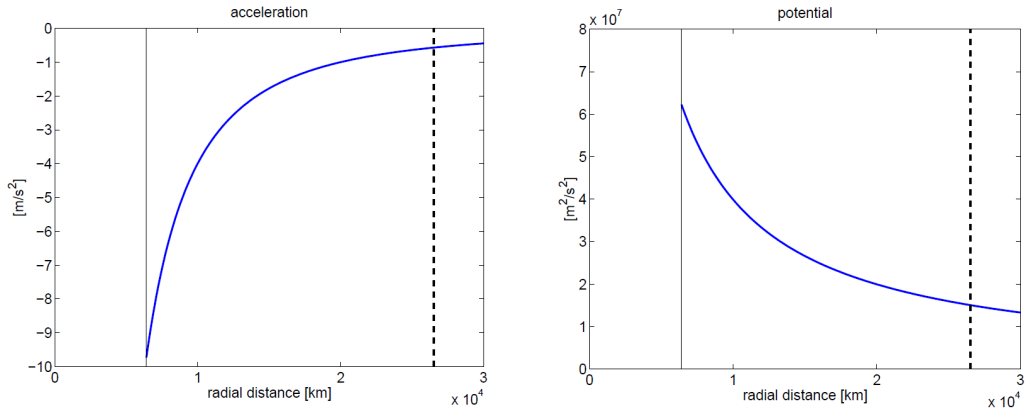



Figure 32.1: The (size of) acceleration $-g$ (left) and the potential $W$ (right) both approach zero as radial distance $r$ goes to infinity. A radial distance of zero corresponds to the Earth's center. The curves start at the Earth's surface (the thin vertical line at $r$=6378 km). The dashed line indicates the radial distance of the orbit of a GPS satellite; there 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 gravitation. 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

$$\mathbf{F} = -\frac{GMm}{r^2}\frac{\mathbf{r}}{r} \tag{32.2}$$

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

$$\mathbf{g} = -\frac{GM}{r^2}\frac{\mathbf{r}}{r} \tag{32.3}$$

The force in Eq. (32.2) has been divided by mass $m$ to obtain the acceleration in Eq. (32.3), which consequently could be interpreted as the force per unit mass. The *magnitude* of the gravitational acceleration $g = \|\mathbf{g}\|$ in Eq. (32.3) decreases with increasing height above the Earth's surface, and reduces to zero at infinite distance, see Figure 32.1 (at left). The acceleration (vector) $\mathbf{g}$ also points toward the Earth's center.

## 32.3. Gravitational 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

$$W = \int_A^B \mathbf{F} \cdot d\mathbf{s} \tag{32.4}$$

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. (32.4), which is actually the acceleration in Eq. (32.3), and interpreting the work in Eq. (32.4) 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. (32.4) 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. (32.2)) and (32.4), the potential of gravitation becomes

$$W = \frac{GM}{r} \tag{32.5}$$

for a pointmass or spherical Earth (we continue with the very simple example of the previous section and in addition assume here that the Earth is non-rotating and at rest in inertial space).

The above 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 32.1 at right. Substituting Eq. (32.5) in Eq. (32.3), gives the relation between gravitational acceleration and gravitational potential

$$\mathbf{g} = -\frac{W}{r}\frac{\mathbf{r}}{r} \tag{32.6}$$

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 32.1 equals the acceleration shown at left.

From a physics perspective $W$ in Eq. (32.5) 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. (32.5)).

So far we used a (very) simple example of an ideal spherical(ly layered) Earth. In the next section we consider in addition that the Earth is rotating, and so are objects on its surface. With an Earth-fixed reference, vector $\mathbf{g}$ will be the result of gravitational and centrifugal acceleration, and is referred to as the acceleration of gravity.

## 32.4. Gravity and geoid

According to Eq. (32.5), 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, for the simple example in the previous section. Surfaces where the gravitational potential $W$ is constant are equipotential surfaces, and the gravitation vector $\mathbf{g}$ is everywhere orthogonal to them (dictating the local level, according to which geodetic instruments are set up).

This concept is now extended to gravity. The surface of *reference* in a vertical sense for physical phenomena on Earth like water flow is the *geoid*, the gravity equipotential surface at mean sea level (MSL).

A geoid is shown in Figure 32.2. 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 in ITRS which, for practical purposes is almost identical to WGS84 [79]. Three versions of EGM are published: EGM84 with degree and

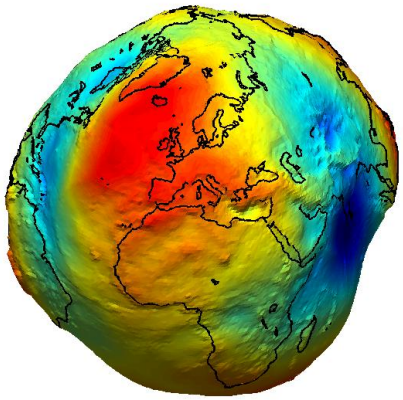


Figure 32.2: The height of the geoid with respect to the best fitting Earth ellipsoid (GRS80). The geoid height $N$ is defined in Figure 33.1. The color-scale ranges from about -100 m (blue) to +70 m (red). This geoid is based on GRACE data. Image taken from ESA, October 2004 [78]. Released publicly.

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 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 35.3), 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 so far as being (perfectly) spherical, just like in Section 29.1. 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 29.2, 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 groundwater level changes.

The gravity acceleration experienced on Earth in practice (and hence observable) consists of, first, gravitational acceleration due to the mass of Earth, as discussed before, but also of Sun and Moon (tidal acceleration) and secondly, centrifugal acceleration due to the Earth's rotation (this effect is largest at the equator, and absent at the poles; objects on the Earth's surface co-rotate with the Earth). 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 (in Dutch: zwaartekracht), with gravity vector $\mathbf{g}$, is commonly defined as the sum of *gravitational* acceleration and *centrifugal* acceleration, where in the first one the part due to the attraction of Sun and Moon is discounted.

## 32.5. Gravimetry [*]

With leveling, increments (distances) are measured along the (local) direction of gravity. Gravity determines the direction of the height system (up and down); the surface perpendicular

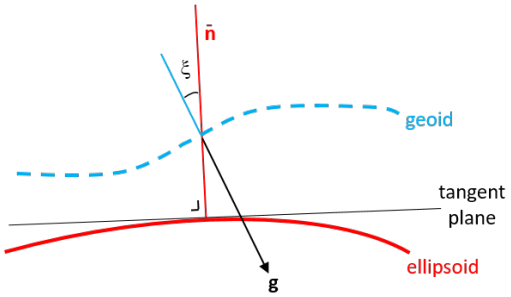



Figure 32.3: The gravity field acceleration vector $\mathbf{g}$, which is orthogonal to the geoid (in blue) deviates slightly from the normal to the ellipsoid $\bar{\mathbf{n}}$ (in red). The deviation $\xi$ is called deflection of the vertical.

to the vector of gravity represents points at equal height (no water flow), cf. Figure 3.6, and may be locally approximated by a tangent plane.

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

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 32.1,

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

Gradiometers measure second order (positional) derivatives of the gravitational 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 the separation made between geometric observables and physical gravity observables is not a strict one. They are unified in the theory of general relativity: the path of a light ray for instance (as used for electro-optical measurements of distance) will bend as it travels through a (strong) gravity field.

## 32.6. Deflection of the vertical

As shown in Figure 32.3 the gravity field acceleration vector $\mathbf{g}$ is orthogonal to the geoid (and to equipotential surfaces in general). The direction of $\mathbf{g}$ may, and will in practice, deviate slightly from the normal to the ellipsoid, denoted by the vector $\bar{\mathbf{n}}$ in Figure 29.6. The angle $\xi$ in Figure 32.3 shows the difference between the two, and is referred to as the *deflection of the vertical* (in Dutch: schietloodafwijking), see also Eq. (29.14). The deflection of the vertical has two components, the deflection of the vertical $\xi$ along the meridian, and the deflection of the vertical $\eta$ along the parallel. Figure 32.3 shows only one component.

In the Netherlands, the deflection of the vertical is very small, in the order of $0.001°$, or $2 \cdot 10^{-5}$ radian at most (with the GRS80 ellipsoid). In mountainous areas the deflection of the vertical can be much larger. A surveying instrument, like a total station, is set-up with its vertical axis aligned with the direction of local gravity. Eventually one may want to

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

Table 32.1: 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.

use coordinates for mapping in a local East-North horizontal plane, as shown in Figure 29.6, hence a plane tangent to the ellipsoid, rather than the local plane perpendicular to the gravity vector. Over small distances the effect will be negligible. In terms of height, the effect, with the earlier given deflection of the vertical in the Netherlands, will be 2 mm over a distance of 100 m.

## 32.7. 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 29.2, 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 32.2 and Table 32.1. Table 32.1 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 groundwater level changes.

## 32.8. Exercises and worked examples

This section presents a few exercises on gravity acceleration and potential.

**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. (32.3), 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 $GM = 3986004.418 \cdot 10^8 \text{m}^3/\text{s}^2$ (Table 31.1). At the equator the distance to the Earth's center equals $a = 6378137.0$ m (Table 31.1, semi-major axis of WGS84 ellipsoid), and at a pole $b = 6356752.314$ m, see Question 3 in Section 29.7. 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 $v = 2\pi a/T = 465.1$m/s. The acceleration becomes $a = 0.034$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 $W = \frac{GM}{r}$ (32.5), 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.

# 33

# Vertical reference systems

Until now the focus has been on the *geometry* of points on the Earth's surface, 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, and in particular add a *physics* perspective on the matter.

## 33.1. 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 29.2, or, to use a different - physics inspired - definition of height, which we will do in the next section.

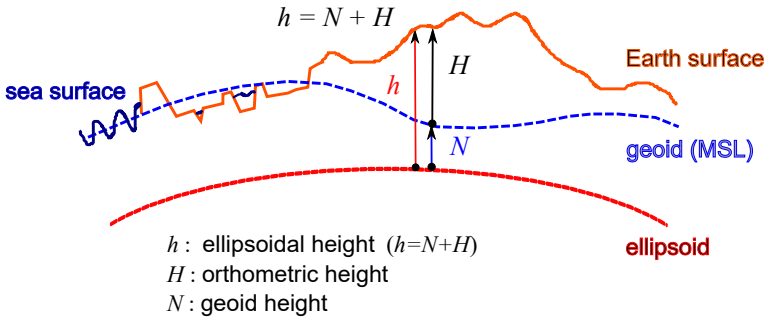



Figure 33.1: Relation between ellipsoidal height $h$, orthometric height $H$ and geoid height $N$ (arrows indicating positive heights).

The ellipsoid is a *geometric* shape. Ellipsoidal heights are a relatively new concept, which can only be measured using space geodetic techniques such as GPS. 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; water flow is a concept from *physics*. 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.

## 33.2. Orthometric and normal heights

In order to deal with water flow, which is a physics concept, it is most appropriate to use the gravity potential $W$ or potential differences $\Delta W$. The reader is referred to [80] for an in-depth discussion. For a quick review of gravity and potential numbers the reader is referred to Chapter 32. Here it suffices to recapitulate from Chapter 32 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. (32.3) and (32.5) follows that the potential difference $\Delta W$ and height difference $\Delta H$ are related,

$$\Delta W = -g\ \Delta H \tag{33.1}$$

with $g$ the gravity acceleration in [$m/s^2$]. The gravity acceleration is a positive number: the minus sign in Eq. (33.1) is because the gravity potential $W$ decreases with increasing height, see Figure 32.1 at right, and therefore $\Delta W$ and $\Delta H$ have opposite sign. Note that the gravity acceleration $g$ is not a constant but depends on the location on Earth and height. Eq. (33.1) relates potential difference and physical height difference, and therefore allows for easier interpretation in practice (that is, height expressed in meters), rather than working with potential or a potential difference.

*Orthometric heights* are defined by the inverse of Eq. (33.1),

$$H_{\text{orthometric}} = -\frac{1}{g}(W - W_0)\ . \tag{33.2}$$

with $W_0$ the potential of the chosen reference equipotential surface (with $H_0 = 0$).

*Normal heights* are based on the *normal* gravity $\gamma$ instead of the (actual) gravity acceleration $g$,

$$H_{\text{normal}} = -\frac{1}{\gamma}(W - W_0) \tag{33.3}$$

with $\gamma$ the *normal gravity* from a normal (model) gravity field that matches gravity 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 (or normal) height $H$ and ellipsoidal height $h$ is given by the following approximation

$$h = N + H \tag{33.4}$$

with $N$ the height of the geoid above the ellipsoid. This is illustrated in Figure 33.1. 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', which is the vertical distance to any reference surface, one often finds the words 'altitude' or 'elevation'. Altitude of for instance an aircraft is the height relative to the geoid (or Mean Sea Level), elevation refers to the height of a point on the Earth's surface relative to the geoid (or Mean Sea Level).

## 33.3. Height measurements

In this section we consider two commonly used techniques for height measurements, leveling and GPS.

### 33.3.1. Spirit leveling

Spirit leveling is one of the most precise techniques to measure height differences. To measure the height difference between two points, as is shown in Chapter 3, vertical rods are set up at each of these points and the height difference is obtained from two rod readings by a leveling instrument positioned between the rods. In fact, a difference in vertical (geometric) distance is observed, cf. Figure 3.5. When a loop (circuit) is measured, starting on a point A, to B, C, ..., finally ending on A, i.e. multiple leveling sections that close again on the starting point, one would expect the mathematical sum of the height differences to be zero. This is *not* the case for large leveling loops, even for perfect observations without measurement errors! The reason is that gravity is not the same at every point on Earth.

To solve this problem the leveled height differences $\Delta H_i$ are converted to potential differences, $\Delta W_i = -g_i \Delta H_i$ conform Eq. (33.1), using gravity acceleration $g$ measured at the surface. Then, for measurements with perfect precision, the sum of the potential differences should be zero. Thus leveling networks can only be adjusted after the observed height differences have been converted into potential differences, otherwise the model is strictly not correct. The output from the network adjustment are potential differences $W_i - W_0$, with $W_0$ the (chosen) potential at a reference point.

To compute the orthometric height Eq. (33.2) is used. This sounds simple, but the problem with Eq. (33.2) is in the value for $g$ that should be used: this is the value of $g$ along the plumb line between the $W_0$ and $W_i$ equipotential surface, meaning it is a value of gravity that is inside the Earth (as usually the geoid is below the surface) and hence density variations in the Earth crust start to play a role, complicating matters very quickly. This is one of the main reasons why many countries have chosen to use normal height instead of orthometric height for their height reference system, as is it much easier to compute the normal gravity $\gamma$.

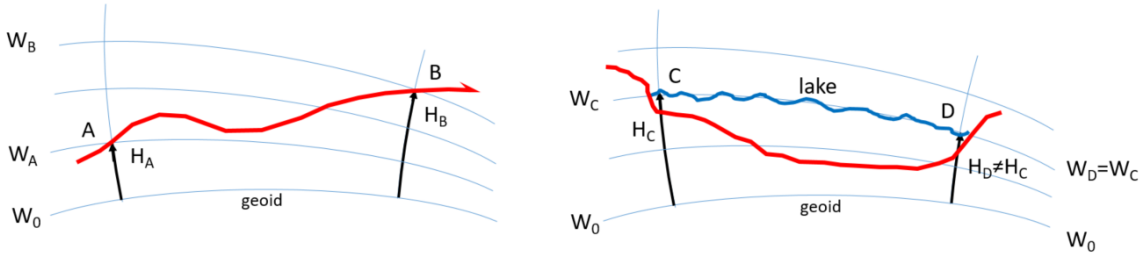



Figure 33.2: The orthometric height is the height difference along the plumb line between an equipotential surface $W$ and $W_0$. As is shown on the right, equipotential surfaces are not parallel. This means that for a large body of water, as shown on the right, the orthometric height of the water level is not constant, $H_C \neq H_D$.

The main benefit of orthometric height is that it is the vertical distance (along the plumb line) between two equipotential surfaces. But is this what we want? Not necessarily. Since gravity is not constant over large areas, the orthometric height of an equipotential surface (other than the reference surface) is not constant, as is shown in Figure 33.2. In other words, equipotential surfaces are not parallel. This means, when working with large bodies of water (e.g. a lake or river system) that orthometric height is not the best choice, or needs a correction to maintain the same height for a level body of water. One possible choice is

to use so-called *dynamic heights*. The dynamic height is simply computed by dividing the geopotential number by a constant $\gamma_{45}$ (the normal gravity at $45°$ latitude).

Another option, especially for small areas and/or countries with little variation in gravity, is to use uncorrected leveled heights, without the conversion to geopotential numbers. This kind of height is used for instance by the Netherlands and Belgium. The differences between uncorrected leveled and orthometric heights are very small for the Netherlands and the only noticable differences occur in the South of Limburg.

### 33.3.2. GPS leveling

Heights measured by GPS (GNSS) are always with respect to the reference ellipsoid. To convert these into orthometric height the inverse of Eq. (33.4) must be used, i.e. the geoid height $N$ needs to be subtracted, and to convert ellipsoidal height to normal height the quasi-geoid height should be subtracted.

Problems with GPS height measurements is that these are not as precise as spirit leveling (see also Chapter 15), which is in particular the case for short observation times, and that a precise (quasi-)geoid is needed to convert ellipsoidal height into orthometric or normal height. However, when a precise (quasi-)geoid is available, and centimeter accuracy for the height is sufficient, GPS leveling is much more cost effective than spirit leveling, especially over larger distances.

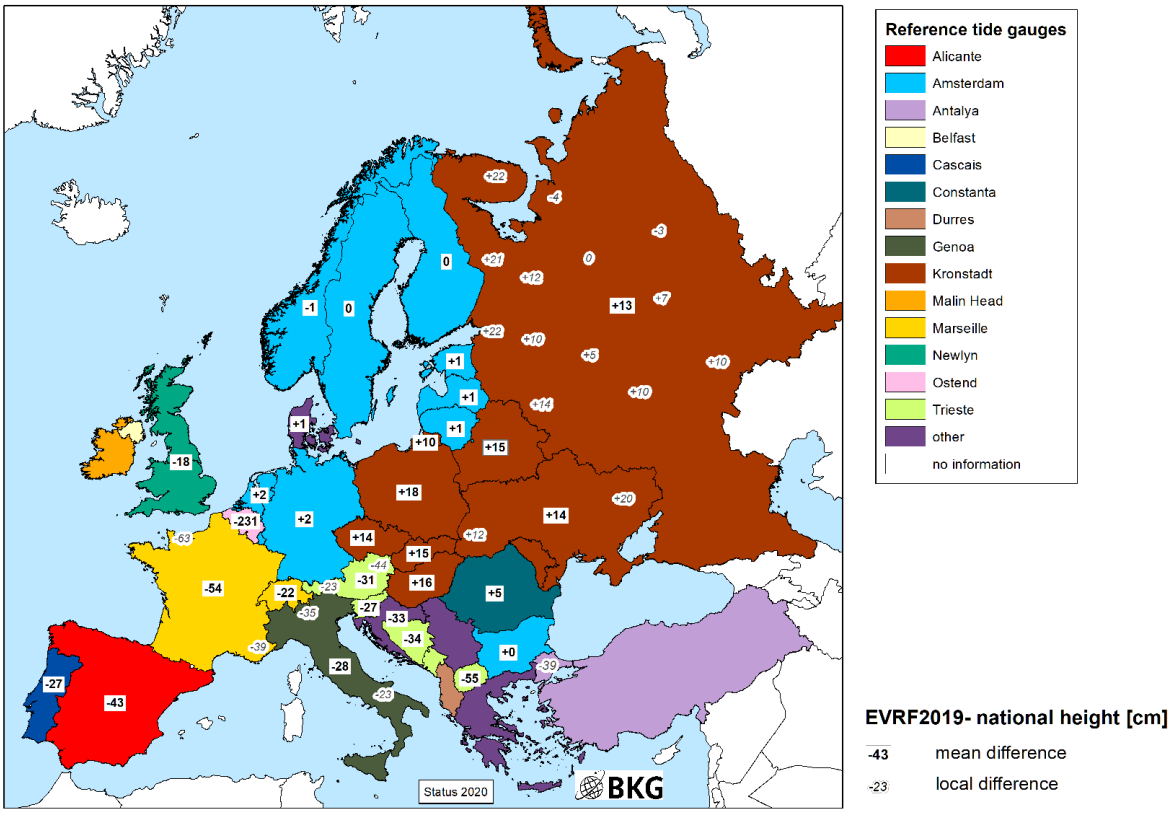



Figure 33.3: Differences between national height datums in Europe and reference tide-gauges in centimeters. Also note that different countries use different kinds of heights. The Netherlands and Belgium use uncorrected leveled heights. Normal heights are in use in for instance France and Germany, and Spain and Italy use orthometric heights. Image courtesy of Federal Agency for Cartography and Geodesy (BKG, Germany) [81], 2020.

## 33.4. 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). For the European Vertical Reference System (EVRS), based on an equipotential surface (constant Earth's gravity field potential), the datum point is Normaal Amsterdams Peil (NAP) [81]. The EVRS serves to harmonize the vertical reference of spatial coordinates in Europe. In Figure 33.3 the reference tide-gauges used in different European countries are shown, together with the differences in the height datum with respect to the European Vertical Reference Frame 2019 (EVRS2019). The differences have been computed from the European re-adjustment of precise levelings. The effects of using different tide-gauges, and the differences between Mean Sea-Level for the North Sea, Baltic Sea, Mediterranean, Atlantic Ocean and Black Sea are clearly visible. Also some countries, for instance Belgium, do not use mean sea-level to define their height datum but use low water spring as a 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 a reference surface[1]. This is *not* an equipotential surface as this reference surface also depends on the tidal variations.

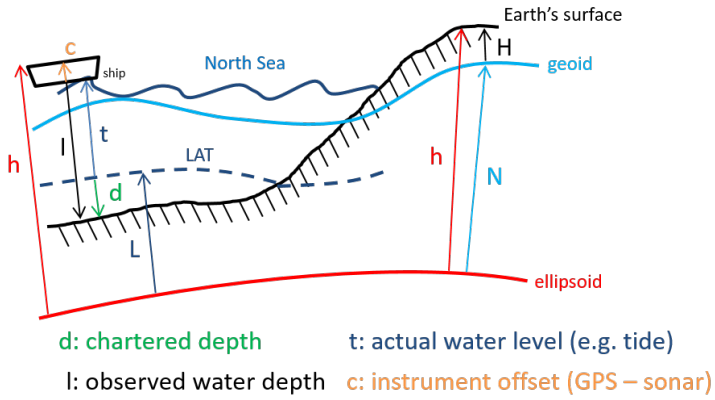



Figure 33.4: LAT, geoid and ellipsoid reference surfaces with the relations between chartered depth $d$, observed water depth $l$, actual water level $t$, ellipsoidal height $h$, orthometric height $H$, height of the LAT reference surface $L$ and geoid height $N$. They are related as $h - c - l - L = -d$, and arrows show positive direction.

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 (and one practically avoids having less depth or clearance than chartered, hence preventing grounding of ships), although in practice lower tides may exceptionally 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, as is shown in Figure 33.4. The depth of water, at a given point and at a given time, is then calculated by adding the charted depth $d$ to the height of the tide $t$, or by subtracting the drying height from the height of the tide, with all heights and depths given with respect to LAT. Height (or depths) with respect to LAT can be converted into heights with respect to the geoid or ellipsoid using gridded correction data.

[1] Lowest Astronomical Tide (LAT) is only one of many reference surfaces used in hydrography; for an overview of existing reference surfaces in hydrography see e.g. [73]

## 33.5. Exercises and worked examples

This section presents a number of simple exercises of working with different heights.

**Question 1** Modeling the Earth as a sphere with radius equal to the semi-major axis of the WGS84 ellipsoid (see Table 31.1), and assuming that all mass is concentrated in the Earth's center, compute the gravitational acceleration at the surface.

**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 32, Eq. (32.2). 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 (leveled) height $H$ is $h = H + N$, with $N$ the geoid height. This relation can also be exploited in a height-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 leveled height difference is about 4.6 m smaller than the ellipsoidal height difference. With the leveled height of Terschelling being $H_T$ = 14.695 m, the leveled height of Eijsden becomes $H_E$ = 57.793 m.

**Question 3** The position of a ship is measured with GPS; the (ellipsoidal) height (of the antenna) is $h$=46 m. The height of the GPS-antenna on the ship with respect to the bottom of the ship (the keel) is $v$=7 m. The chartered depth of the waterway at the ship's location, retrieved from a hydrographic map, is $d$=4 m (given with respect to the LAT reference surface). The height of the LAT reference surface with respect to the ellipsoid is $L$=42 m. Compute the clearance of the ship (i.e., the height of the ship's keel above the bottom of the waterway. Note that positive height is upward, and positive depth is downward, as is indicated in Figure 33.4.

**Answer 3** The keel of the ship is $h - v$ above the ellipsoid, $h - v$=46-7=39 m. The waterway-floor is at $L - d$ above the ellipsoid, $L - d$=42-4=38 m. Hence, the clearance is: 39-38=1 m.

# 34

# International reference systems and frames

In this chapter a number of common international reference systems and frames is discussed. We start with the well known worldwide 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.

## 34.1. 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 (see also Table 14.1 and Figure 14.8). 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 coincide with the ITRS and its realizations ITRFyy, see Section 34.2, at the decimeter level. For these WGS84 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 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'.

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

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

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

## 34.2. 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 (ITRF), which provides coordinates of a set of points located on the Earth's surface [82]. 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 by the International Earth Rotation and Reference Systems Service (IERS) [83], 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.

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 geodetic measurement 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. ITRF2020 will be the next realization, using more recent data, reprocessing of old data, improved models and processing software.

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.

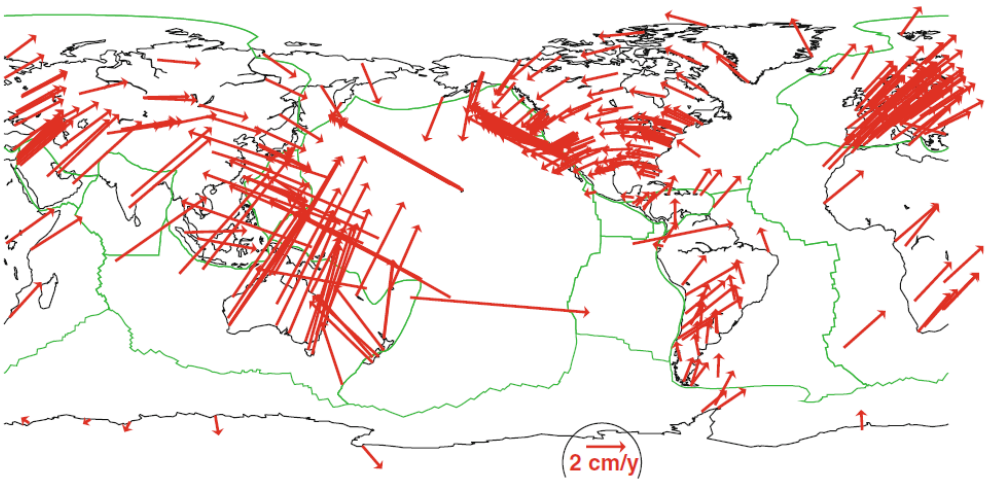



Figure 34.1: ITRF2008 horizontal velocity field at space geodetic observatory locations due to plate tectonics with major plate boundaries shown in green. Continents are shown in black lines, with Europe and Africa on the right. Image taken from [82] under CC BY-NC license, see also [84].

- 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. Secular motion is practically best described as 'straight line' motion. It refers to persistent change over a longer time period (a year or longer). In particular a secular motion does *not* include periodic and/or tidal components. The position $\mathbf{r}(t)$ at a specific epoch $t$ is given by

$$\mathbf{r}(t) = \mathbf{r}(t_0) + \dot{\mathbf{r}} \cdot (t - t_0) \tag{34.1}$$

The ITRF2008 velocities are given in Figure 34.1. 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 [85]. For more information on the realization of ITRF2008 see [82].

In Figure 34.2 a time series of station positions for a GPS receiver in Delft is shown, cf. Figure 16.1 at right. 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 34.2, 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 34.3 shows the same time series, but in the European ETRF2000 reference frame, which is discussed in Section 34.3.

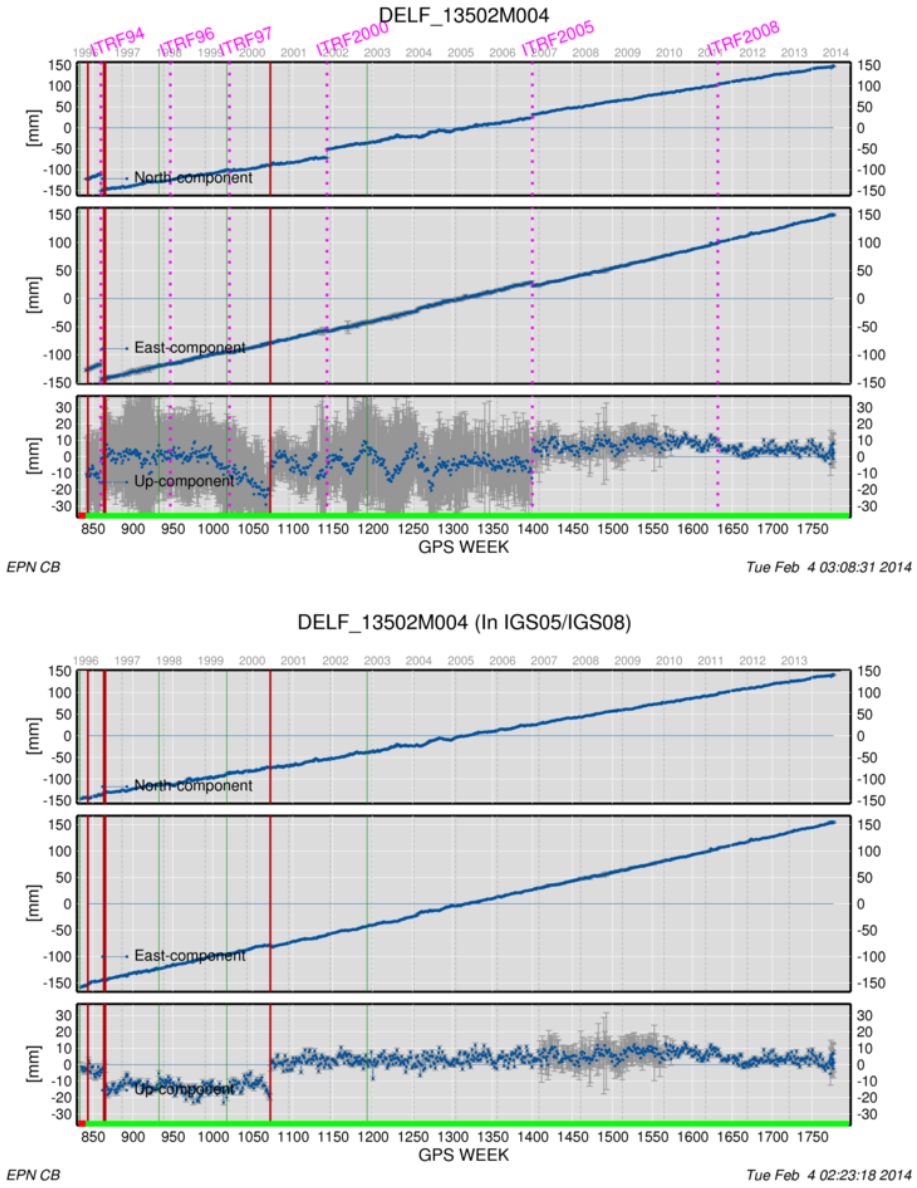



Figure 34.2: Time series of station positions of a permanent GPS receiver in Delft from 1996-2014. The top figure shows the time series in the ITRFyy reference frame that was current at the time the data was collected. The bottom figure shows the data after re-processing in the IGS05/IGS08 reference frame that is based on the most recent ITRF2008 frame. The vertical red lines indicate equipment changes. Image taken from EUREF Permanent GNSS Network (EPN) [86], for private, educational and scientific purpose.

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 essentially based on the Satellite Laser Ranging (SLR) time series of Earth orbiting satellites.

[4]The International Terrestrial Reference Frame (ITRF) is maintained by the International Earth Rotation and Reference Systems Service (IERS) [83]. The IERS is also responsible for the International Celestial Reference Frame

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 transformation parameters and formula are accessible through [83]. The transformation formula is essentially Eq. (28.15), except for a different sign of the transformation parameters.

Web-based Precise Point Positioning (PPP) services (cf. Section 15.1.6), 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 31.1. This is the same ellipsoid as used for instance by WGS84.

## 34.3. 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 34.1 and 34.2.

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

(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 [46]. The main product of IGS are precise orbits for GNSS satellites (including GPS), satellite clock errors and station positions all in the ITRF.

[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 [87] 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 are time-variant in both systems, and can be given at any epoch. 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 are also used for a wide range

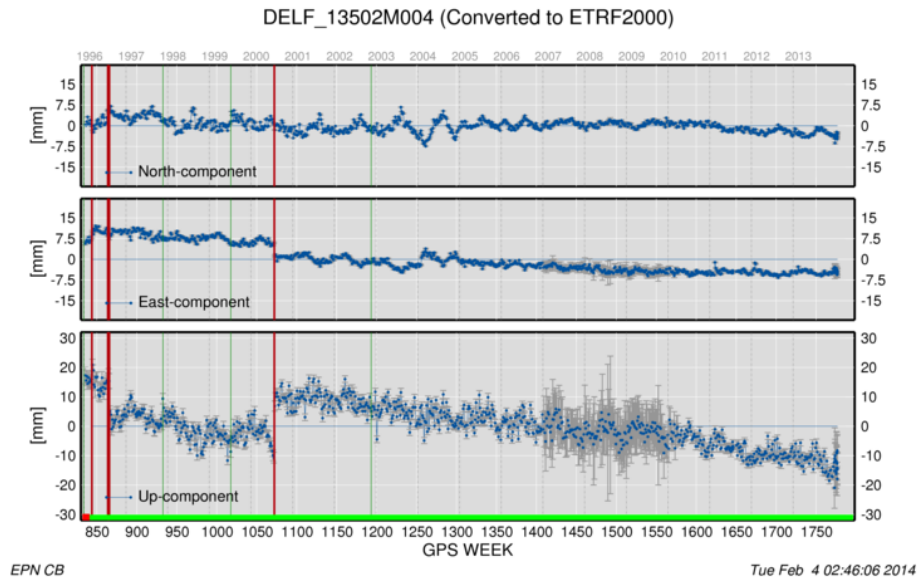



Figure 34.3: Time series of station positions of a GPS receiver in Delft from 1996-2014 in the European reference frame ETRF2000. The horizontal station velocity In ETRS89 is at the few mm level. The vertical red lines indicate equipment changes which cause jumps of a few mm in the time series. Image taken from EUREF Permanent GNSS Network (EPN) [86], for private, educational and scientific purpose.

reference stations with precisely known station positions and velocities in the ETRS89, or through one of many national or commercial GPS networks which realize ETRS89 on a national scale.

Station velocities in ETRS89 are generally very small because ETRS89 is fixed to the stable part of the Eurasian plate. Compared to ITRS, with station velocities in the order of a few centimeter/year, station velocities in ETRS89 are typically smaller than a few mm/year. This is clearly illustrated in Figure 34.3 for the permanent GPS station in Delft that was also used for Figure 34.2. 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

of scientific applications such as the monitoring of ground deformations, sea-level, space weather and numerical weather prediction.

[8]there is no ETRF2008

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 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, see e.g. Figure 16.1 at right. 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. Data for the AGRS.NL and other GNSS receivers is available from the Dutch Permanent GNSS Array (DPGA) website operated by our university [88]. 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, see e.g. Figure 15.2. The Dutch network RTK services are certified by the Dutch Kadaster [89] 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 [87] does recommend to use one of three selected projections: *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 national systems is provided, see Chapter 35. A service to convert coordinates from ITRS to ETRS89, and vice versa, is provided at the EPN website [86]. Specifications for the transformation procedure and reference frame fixing can be found in the EUREF Technical Note 1 [90].

## 34.4. Exercises and worked examples

Position coordinates in the International Terrestrial Reference Frame (ITRF) are time-variant, in the first place due to plate motion. Below a simple exercise is presented to propagate position coordinates given at a specific epoch to another epoch in time.

**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, $Y = 443305.133$, $Z = 5064884.818$.

# 35

# 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 Netherlands, with the remark, that many other countries have undergone similar developments and adopted similar approaches.

## 35.1. 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.

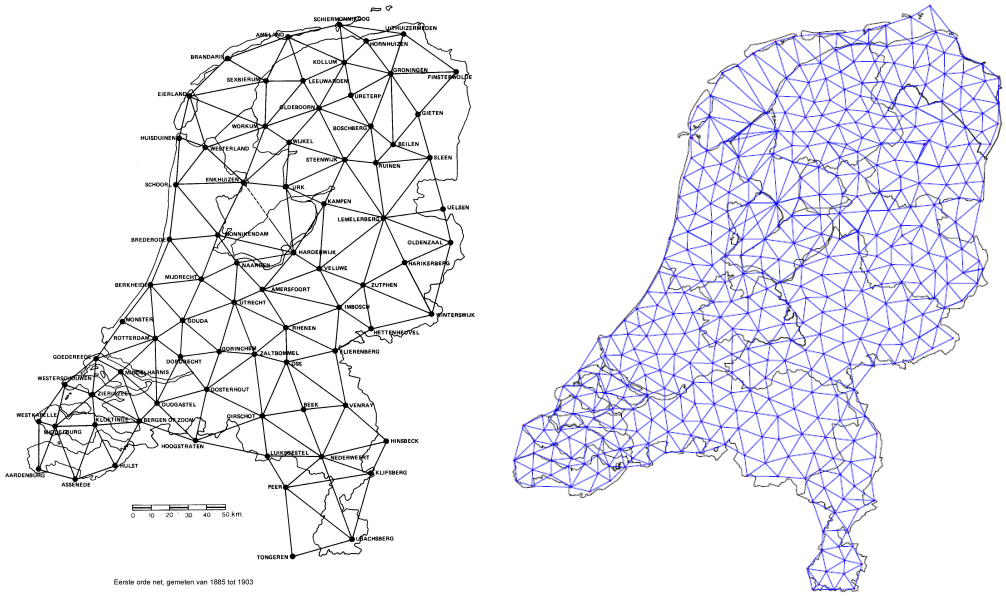



Figure 35.1: First order triangulation network for the Netherlands of 1903 (left) and 'GPS-Kernnet' (GPS base station) network of 1997 (right) [91].

### 35.1.1. RD1918

The first-order triangulation grid was measured in the years between 1885 and 1904 (Figure 35.1). The church tower of Amerfoort was selected as the origin of the network and as reference ellipsoid the Bessel (1841) ellipsoid of 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 in 1918 by Heuvelink [92]. The projection consists of a Gauss-Schreiber conformal projection of Bessel's (1841) ellipsoid onto a sphere, followed by a oblique stereographic projection of the sphere to a tangential plane, as shown in Figure 35.2.

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 (secant instead of tangent, see Figure 30.3), the scale distortions at the edges of the projection domain will be within reasonable limits.

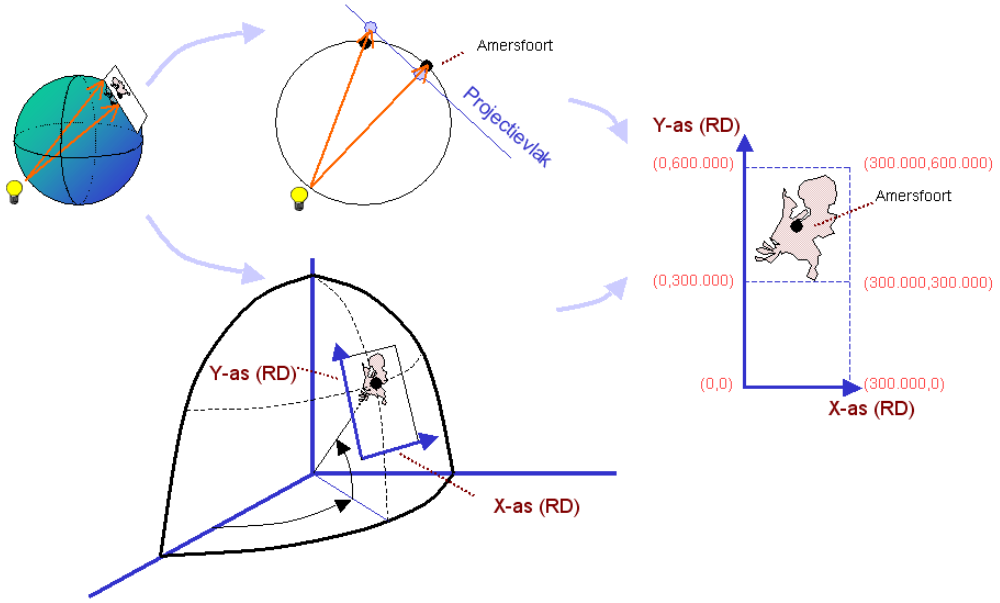



Figure 35.2: RD double projection (Bessel (1841) ellipsoid → Sphere → Plane) and definition of RD coordinates Image by T. Nijeholt, August 2007, taken from Wikimedia Commons [9], under CC BY-SA 3.0 license.

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 with distances of 2.5-4 km between each other. 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-Kernnet' (GPS base network) of 418 points was established to offer GPS users a convenient way to connect to the RD system. See Figure 35.1. Most of traditional triangulation points, many of which are church spires or towers, are not accessible to GPS measurements. The points in the GPS-Kernnet have an unobstructed view of the sky and are easily accessible by car. The GPS-Kernnet 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 leveling 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 35.3 for the province of Friesland. In the era of GPS this may seem as a large number, but in 1918 this was an excellent accomplishment.

Today, about 105 out of the original 418 'Kernnet' points, nowadays called GNSS-Kernnet, are maintained and regularly checked by the Kadaster.

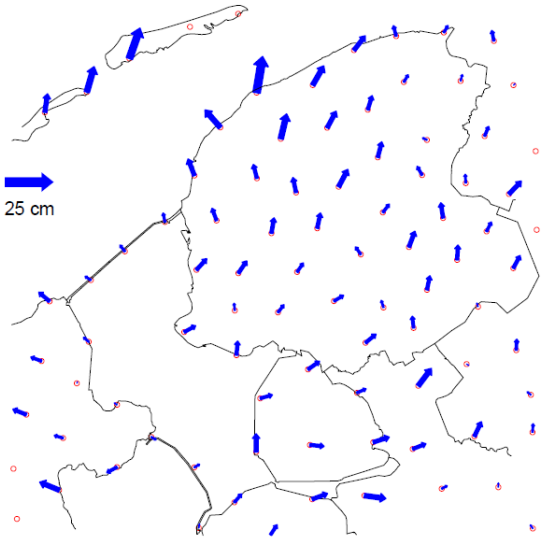



Figure 35.3: Differences between RD and ETRS89 based coordinates for the GPS-Kernnet in the province of Friesland, showing significant regional correlation between the vectors [91].

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 became noticeable. This led to a major revision

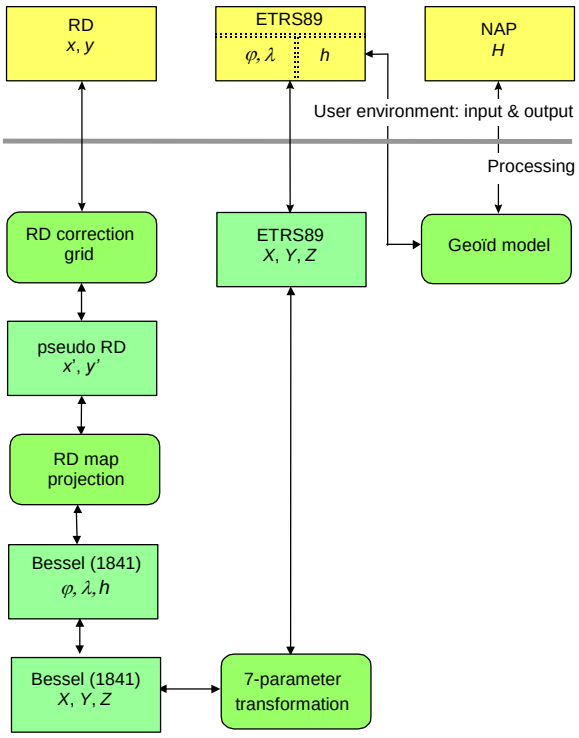



Figure 35.4: RDNAPTRANS transformation procedure until and including RDNAPTRANS™2008. The figure outlines the relationships and transformations between ETRS89, RD and NAP (Figure after [91]). The coordinates below the line, with the exception of Cartesian coordinates in ETRS89, are used only for computational purposes and should never be published or distributed to other users. The procedure used for RDNAPTRANS™2018 is different in a couple of aspects, see Figure 35.5.

in the definition of RD in 2000.

### 35.1.2. RD2000 and RDNAPTRANS

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 35.4 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 above the GRS80 ellipsoid of ETRS89, which will be discussed next in Section 35.2.

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 $h = 0$ m when transforming from RD to ETRS89 and $h = 43$ m when transforming from ETRS89 to RD (which are close to $NAP = 0$) so that one gets the same result after transforming back and forth. In this case 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 2018. The minor revisions were related to changes in the European reference frame, which affected the 7-parameter transformation, the introduction of an improved NLGEO2004 geoid in 2004, and a small height offset in 2008. 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, published in 2005 and respectively 2009. 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.

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.

### 35.1.3. 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 (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. In RDNAPTRANS™2018 the correction grid is applied to the latitude and longitude coordinates instead of the pseudo RD coordinates, the NLGEO2018 quasi–geoid is used instead of NLGEO2004

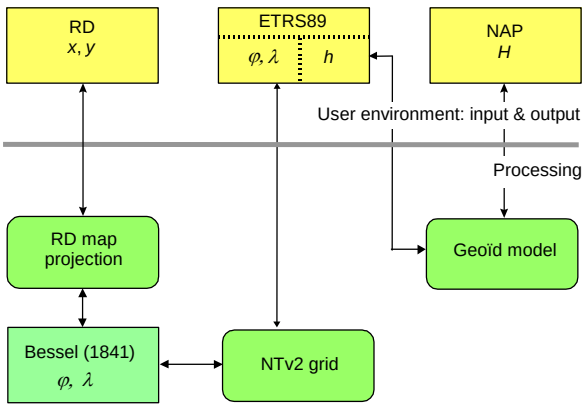



Figure 35.5: NTv2 transformation procedure used by RDNAPTRANS™2018. The figure outlines the relationships and transformations between ETRS89, RD2000 and NAP using the proposed NTv2 procedure, in variant 2 of RDNAPTRANS™2018 where the datum transformation is included in the correction grid. The coordinates below the line are used only for computational purposes and should never be published or distributed to other users.

geoid, and the interpolation and correction grids are based on international standards. The RD correction grid is based on the Canadian NTv2 correction procedure (National Transformation version 2) and for the NLGEO2018 the VDatum format is used, both of which are better supported by existing softwares, including PROJ [77]. As shown in Figure 35.5, 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 35.5 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 centimeter 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 35.4. 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, using a pipeline of smaller transformation steps, each with its own `+proj` string (see Section 31.5).

Though many GIS packages support EPSG codes (see Section 31.4), and often use PROJ as the underlying mechanism for coordinates transformations, users should nevertheless be extremely careful with using the EPSG code EPSG:28992 for the Dutch RD coordinate system in coordinates transformations. The software you are using may not (yet) support the latest version of PROJ, have the correct parameters and/or downloaded the required correction grids. Besides, for accurate coordinates transformations it is equally important to specify the exact input coordinate system with the correct datum. The details do not matter for visualization on a map or computer screen, but should users wish to exchange RD coordinates, it is ad-

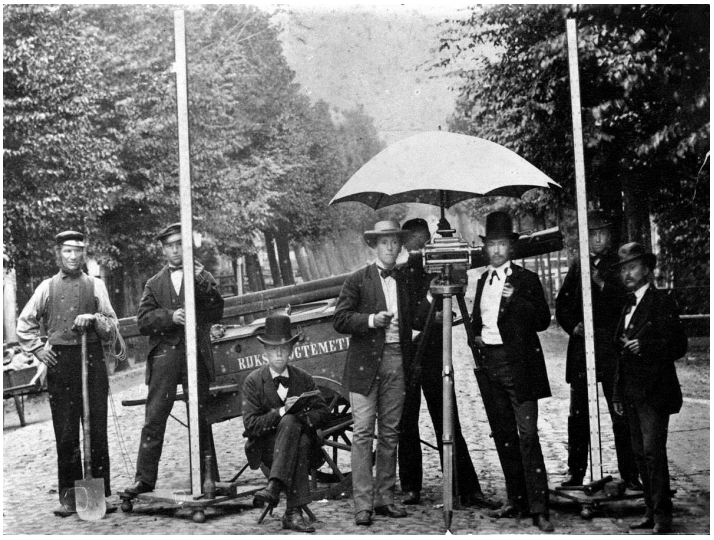


Figure 35.6: Team of surveyors posing for the camera during the first precise leveling, in Dutch *eerste nauwkeurigheidswaterpassing*, or *Rijkshoogtemeting*, probably in 1875 or 1876. The person with the white hat is Cornelis Lely (1854-1929), who just graduated as a civil engineer in Delft (1875). Later, as minister of infrastructure, he introduced the bill that resulted in the *Zuiderzee werken*, which comprised the construction of a 30 km *Afsluitdijk* dike forming the *IJselmeer* lake, and the creation of the two western polders in the former *Zuiderzee* sea. Photo taken from [93], see also [94]. Public Domain.

vised to use software that has been certified by the 'Nederlandse Samenwerking Geodetische Infrastructuur' [89] (NSGI) and carries the RDNAPTRANS™ trademark.

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

## 35.2. 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).

### 35.2.1. Precise first order levelings

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 levelings.

A series of five (first-order) precise leveling campaigns has been carried out to date. The 1st national precise leveling dates from the period 1875-1885, including 410 already existing points and 2100 km of continuous leveling lines. See also Figure 35.6. The datum was

Figure 35.7: Leveling lines of the 5th precise leveling in the Netherlands, 1996-1999 [91].

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 levelings 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 leveling, 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 precise leveling 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 now in a certain sense symbolic, as the height datum is nowadays defined based on the underground reference points in geologically more stable locations.

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 leveling 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 leveling was carried out between 1996-1999, see Figure 35.7. This was the first time that a combination of optical and hydrostatic leveling, satellite positioning (GPS) and gravity measurements, were used. It also include ice leveling measurements on the IJsselmeer and the Markermeer lakes. The leveling measurements still constitute the basis for the primary NAP grid. The gravity measurements constituted the 2nd measurement epoch of the Dutch gravity grid. They served to get an independent insight into subterranean movements. The GPS measurements served to enhance the leveling net over greater distances, and to connect the leveled NAP heights to ETRS89. The network of the 5th precise leveling was also connected to the German and Bel-

Figure 35.8: NAP (Amsterdam Ordnance Datum) 'datum point' at the Stopera, Amsterdam. Photo by M. Minderhoud - own work, July 2005, taken from Wikimedia Commons [9] under CC BY license.

gian networks. These connections play 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 Figure35.8.


Figure 35.9: Bronze NAP bolt, in a pillar (back row pillar, in the perspective of this photo under the 'C' of Civil Engineering ...) , near the B-entrance of the Faculty of Civil Engineering and Geosciences building in Delft. The NAP marker (peilmerk) number is 037E0612, and the given height in NAP is $-0.359$ m, thus below the NAP reference surface.

### 35.2.2. NAP Benchmarks

The primary NAP grid is comprised of about 300 underground points and 70 posts (in Dutch: 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. Figure 35.9 shows one such bronze NAP bolt near the entrance of the building for Civil Engineering and Geosciences in Delft.

The heights of the bronze NAP bolts have been determined by leveling loops with an average length of 2 km for each edge, with a precision better than 1 mm/km. A bronze bolt is installed after every kilometer. There are about 35,000 of these bronze bolts (peilmerken) installed in the Netherlands. The heights of the bolts are published by RWS in NAPinfo [95]. These bolts serve as the basis for height determination by consulting engineers, water boards, municipalities, provinces, state, and other authorities, whereby one of these bolts can almost always be found within a distance of 1 km.

GPS has not replaced leveling 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 33 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.

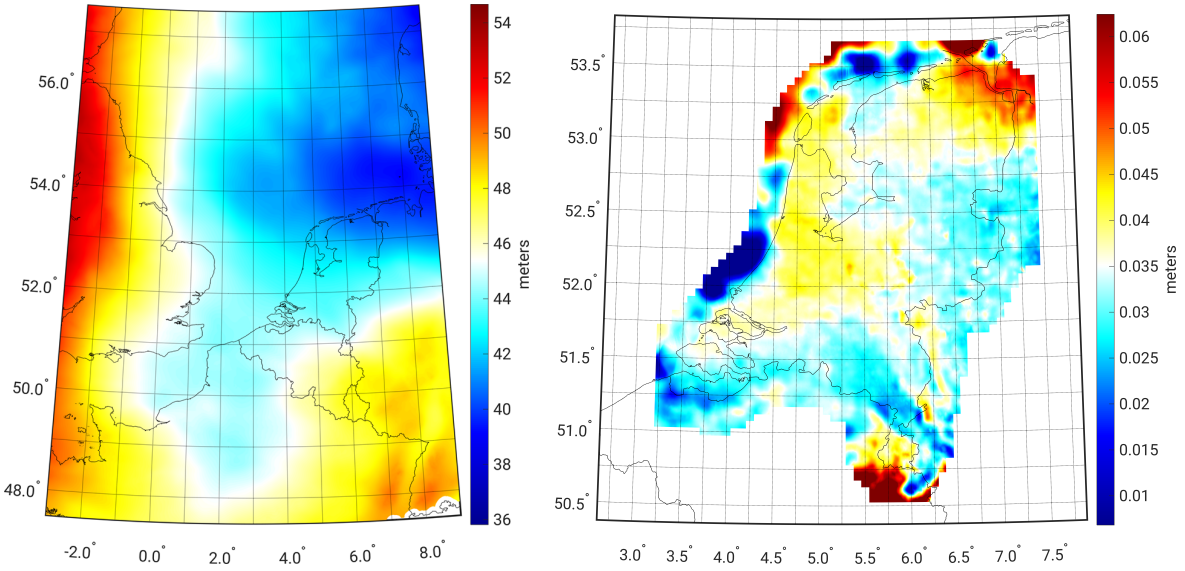



Figure 35.10: NLGEO2018 quasi-geoid for the Netherlands on the left, with geoid height in [m] with respect to the GRS80 ellipsoid in ETRS89, with on the right the differences with NLGEO2004. Clearly visible the much larger domain over which the NLGEO2018 quasi–geoid is computed, though only the values between $2^\circ - 8^\circ$ East longitude and $50^\circ - 56^\circ$ North latitude are published. Image by Cornelis Slobbe [96].

## 35.3. Geoid models – NLGEO2004 and NLGEO2018

Although it is unlikely that GPS will replace leveling alltogether, GPS can be used to obtain heights with an accuracy of about 1-2 cm using the RDNAPTRANS procedure, as outlined in Figure 35.5. The transformation from ellipsoidal height to NAP height, and vice versa, requires a correction for the geoid height.

Calculation of a geoid requires gravity measurements over - in principle - the entire Earth, cf. Chapter 32. 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 gravity measurements were carried out in a grid of almost 8,000 points (1 point per 5 km$^2$) 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. The improvements resulted from using additional gravity measurements on Belgian and German territory and a set of 84 GPS / leveling points from the 5th precise leveling 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 [91]. 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 [96]. 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 35.10 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 leveling 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 a maximum height difference of about 2.5 cm.

## 35.4. 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), see Section 33.4. Tide 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 Hydrographic Service of the Royal Netherlands Navy (in Dutch: Dienst der Hydrografie

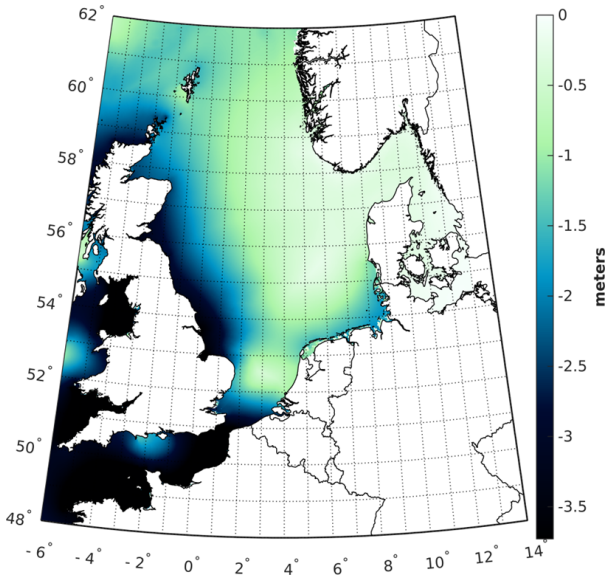



Figure 35.11: NLLAT2018 with respect to the NLGEO2018 quasi-geoid. Image by Cornelis Slobbe [96].

van de Koninklijke Marine) is responsible for the survey of the Netherlands Continental Shelf (in Dutch: Nederlands Continentaal Plat). In hydrographic maps the chartered depth of the seafloor is reported with respect to 0-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 35.11. The LAT reference surface, 0-LAT, in NLLAT2018 is given with respect to the GRS80 ellipsoid in ETRS89, shown by $L$ in Figure 33.4. It is available as correction grid with respect to the GRS80 ellipsoid in ETRS89 and as correction grid with respect to NLGEO2018. The accuracy of the LAT reference surface is about 1 decimeter.

## 35.5. Exercises and worked examples

Below is a simple exercise on retrieving the height of a benchmark from the NAP-database.

**Question 1** The heights of the NAP bolts are published by RWS in the NAPinfo-database [95]. Look up the height of NAP bolt (peilmerk) 037E0612 that was shown in Figure 35.9. The position of the bolt on the wall is given by *x-muur (cm)* and *y-muur (cm)* in the meta-data (Peilmerkinformatie). Y-muur is the height above the ground. This information is intended to find the bolt more easily, but we can also do the reverse. What is the height of the pavement (according to the meta-data)?

**Answer 1** From the NAPinfo website we find that the height of the bolt in NAP is -0.359 m and that the bolt was last measured on 2018-08-01. The x-muur is 25 cm, and y-muur is 20 cm. Thus the height of the pavement is -0.359-0.20 $\approx$ -0.56 m, thus approximatelty 56 cm below NAP. Of course, this is not very accurate or particularly useful, the purpose of this exercise is just to familiarize yourself with the NAPinfo-database.

# VI

# Mapping

# 36

# Introduction

This part is intended to provide a quick tour of the subjects of maps and geographic information. It is not intended to provide a comprehensive coverage of all topics in depth — this part provides just a first impression of the field of working with geospatial data in the context of surveying and mapping. Many books have already been written on the subject, and surely many more will be written still. This is with good reason: the scope of this subject is very broad and begs in-depth study.

As an aspiring engineer, you will hopefully see how important good practice is in cartography and in visual design in general. Communicating a message requires careful thoughts and an appreciation of the tools we have available for the task. For the end users of maps, this aspect of communication helps to understand the world around us.

## Dutch historical perspective on mapping

Cartography has a very long and rich history, going back thousands of years. Cartography is particularly relevant to the painting by the famous artist Johannes Vermeer (1632-1675) [97], shown in Figure 36.1. Johannes Vermeer worked and lived in Delft, and is particularly renowned for his masterly treatment and use of light in his work. The man in the painting, acting as a geographer, most likely is the Dutch scientist Anthonie van Leeuwenhoek, a contemporary of Vermeer, also born in Delft. Note the sea chart on the wall in the back, covering the coasts of Europe, and the globe on the cupboard.

In the days of Vermeer — the Dutch Golden Age — Dutch cartographers, such as Abraham Ortelius, Jacob van Deventer, Willem Blaeu and Lucas Waghenaer, to name a few, played influential roles in the domain of mapping.

## Overview of this part

While earlier parts of this book focussed on the acquisition of geospatial data and the processing of the measurements, this last part is on storing, working with and presenting geospatial data. Recognizing that maps are valuable and effective tools for communicating spatial information to colleagues, customers and the public, the communication process is briefly outlined in the next chapter. Next an overview of different types of maps is presented, and we proceed to the art and skill of working with visual variables, the knowledge of which is applicable to a wide range of presenting information in graphical form. The last chapter in this part is on Geographic Information Systems (GIS), which allow us to store, access and maintain geospatial data, as well as offer spatial analysis and interpretation functionality.

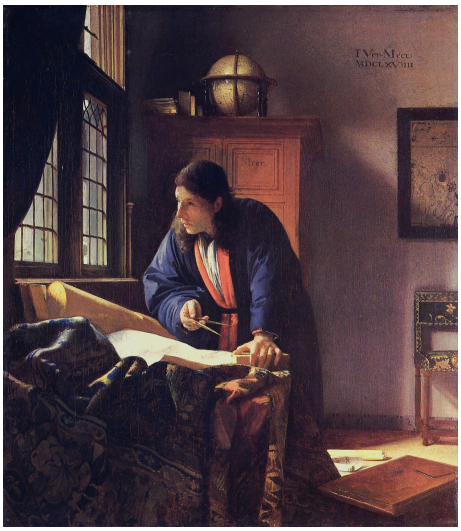


Figure 36.1: Mapping in Delft: The Geographer, painted by Johannes Vermeer in 1669. Städel Museum - Frankfurt am Main. Image taken from Wikimedia Commons [9]. Public Domain.

In three appendices, frequently used map services are presented: PDOK specifically for the Netherlands, Open Street Map and Google Earth, in Appendix I, J and K respectively.

# 37

# Communicating spatial information

Maps are tools for effectively communicating spatial information to a reader, organising it and representing it using a visual medium. According to the International Cartographic Association (ICA): 'a map is a representation, normally to scale and on a flat medium, of a selection of material or abstract features on, or in relation to, the surface of the Earth' [98].

The final goal is to eventually observe and understand geospatial relationships, and analyze spatial patterns. If the reader is unable to do so, then the map creator has failed. To this end, there is a communication process that cartographers must follow. Observe Figure 37.1. As we see there are four sequential stages, through which the cartographer's message is 'streamlined', improving the design of the map.

Selection and generalization are the interim steps between the real world, and the map. These are the processes which happen regardless, since maps will inevitably have less information and detail than reality. Selection, the first process, involves deliberately choosing which elements of reality are relevant to your message, and which elements can be left out. In the second process, generalization, the level of detail in the map is reduced to a lower level, such that the information is reasonably faithful to reality, while still being practical.

A picture is worth a thousand words, so examples of these two concepts will be demonstrated in Section 37.3.

## 37.1. What to communicate?

The modern day engineer has access to a vast amount of data, collected using methods our predecessors never would have dreamed of. To start with, humanity has succeeded in launching vehicles into the space outside the Earth's atmosphere. We have Global Navigation

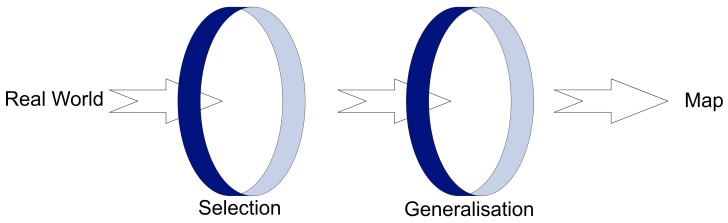



Figure 37.1: Cartographic communication process (diagram after [4])

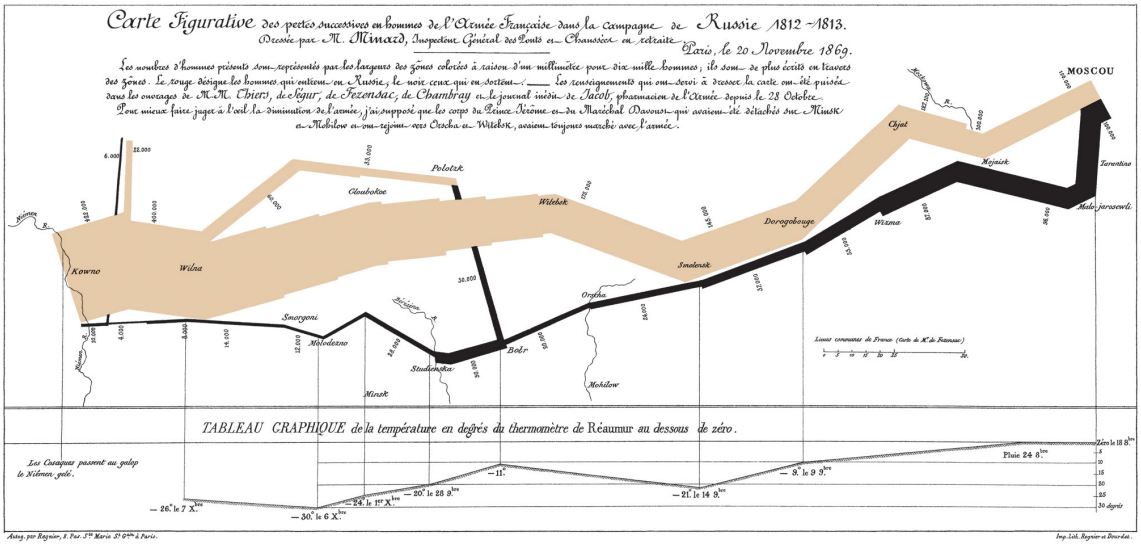



Figure 37.2: French civil engineer Charles Minard's map of Napoleon Bonaparte's Russian campaign is impressively dense. The 'Carte Figurative' contains six types of data: the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel; and location relative to specific dates. Charles Joseph Minard, 1869 - Mapping Napoleon's March. Carte figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813. Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite. Image taken from Wikimedia Commons [9]. Public Domain.

Satellite Systems (GNSS) freely available, that allow for precise positioning nearly anywhere on the globe. For a large engineering or scientific project, GNSS-equipped platforms may be used together with data collected using traditional methods such as tide gauges and leveling. This information may then yet again be combined with a third dataset, like a geological map. Space agencies such as ESA and NASA have launched dozens of imaging remote sensing satellites, regularly capturing images of the Earth's surface.

With such a broad range of data for an engineering project, which can involve hundreds of people, it is now more important than ever for geospatial data to be visualized in a way that is clear, concise, and complete.

As a case study, let us have a look at the chart in Figure 37.2 by civil engineer Charles Minard (1781-1870) [99]. Minard is noted for his representation of numerical data on geographic maps. The chart in Figure 37.2 - created in 1869, after the event - shows the status and surrounding conditions of Napoleon's army throughout his 1812 invasion of Russia. The figure is famously dense. It contains six types of data represented in two dimensions: but it may lack some *clarity*. There is simply too much information, the annotations are all oriented differently, and the legend is too small and poorly organised. Its compact form is great if you are the messenger who needs to deliver the map on horseback, but not so great for the eventual reader who has to read it by candlelight.

So there is a limit to the amount of information our soft squishy brains can absorb at once. But there is also an immensely complex and rich world outside our front doors. So how do we select key features of that world, and then somehow display it on a two-dimensional computer or smartphone screen or on a piece of paper?

Fortunately the modern day engineer(-ing student) has powerful software at his or her disposal, and an entire library of textbooks. With some key guidelines that are discussed in this chapter, you can learn how to effectively communicate important geospatial data from one squishy brain to another.

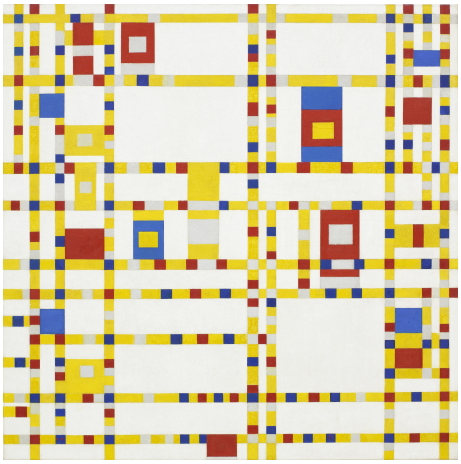


Figure 37.3: Inspired by the grid-like layout of Manhattan's streets, Mondrian takes an abstract approach to mapping. This is his *subjective* visualisation of a *real* neighborhood located in New York City. Broadway Boogie Woogie (1942-1943) painted by Piet Mondrian (1872-1944). Museum of Modern Art (MoMA), Manhattan, New York. Image taken from Wikimedia Commons [9]. Public Domain.

## 37.2. Modeling reality

The goal is to turn the world around us into a map or a model. As discussed in the introduction, the world is complex. Because it is bumpy and uneven, and changing in time, reality is hard to grasp and *subjectively* interpreted by humans. For impressionist art, the painter attempts to convey a certain subjective interpretation to the viewer, a certain 'model' of the world if you will. This is a visualization of the reality around the painter, but only the aspects that the painter considered relevant - for example Piet Mondrian's [100] impressionist painting 'Broadway Boogie Woogie' in Figure 37.3. The area around Broadway in Manhattan is reduced (simplified) to a couple of basic geometric entities, like lines, squares and rectangles, in a few elementary colors.

The approach to take for engineers is to define reality using *objective* models, so the aspects of reality that *we* find relevant can become workable. Reality is then reduced to a collection of points, lines, polygons/areas (geometric entities). In this respect, remember the simple example of Figure 1.1, and the discussion with Figure 9.4, and also see Section 39.2 later on. So, for both the young engineer and Mondrian, we notice the concept that Paul Klee (1920) described: *'Art does not reproduce the visible, but rather makes it visible'*. Swiss artist Paul Klee is known for his paintings which straddled reality and the abstract [101]. A cartographer or engineer has to make visible, out of a complex reality, the features which are relevant to his or her message.

## 37.3. Selection and generalization

As future engineers, we love to simplify and generalize cases so that we can more easily see what matters most for the problem we want to solve. But what is it that matters? What is our message? As we saw in the previous section, we want to reduce and simplify the case to the relevant entities and attributes, which comes to selection and generalization.

Take for example the map in Figure 37.4 on top, of the Rotterdam metro station network [102]. The different tracks and stations are located throughout Rotterdam and its surrounding areas, at different horizontal and vertical positions in this two-dimensional view, with a topographic map as background, cf. Section 38.1. The first metro network maps were created (of the extensive network of the London Underground), geographically correct, in this style, and


Figure 37.4: On top: geographic map of Rotterdam area with the Rotterdam Metro stations indicated, at bottom: map of Rotterdam Metro network. This style of transport map (at bottom), originally designed by Harry Beck for the London Underground, may have distorted the relative geographic positions of the individual stations in the network (some stations may appear closer on the map, than they are in reality, and the other way around), but has made it way easier to find the route from A to B, and in particular how to get there; this map comprises information about links and transfers. On top, BRT Achtergrondkaart by Kadaster, taken from PDOK [59] under CC BY 4.0 license. At bottom, own work by MichaBen, image taken from Wikimedia Commons [9], dated Nov. 16th, 2009. Public Domain.

they were confusing and difficult to interpret.

But, if your purpose is to inform travellers Bram and Neelie, who just need to get to the other side of the city as fast as possible, you may need to simplify the map (generalization) and omit some information (selection). It would be smart to focus on, and select the *connections* between the lines of the metro network, because that is the most important thing to know when you are on the metro: when to get off and change trains! That was exactly the type of innovative thought that Harry Beck had when, in 1931, he, as an electrical engineer, redesigned the London Underground map in a way that it resembled an electrical circuit diagram [103], and which since has been used for many metro networks around the world, also for the metro in Rotterdam, see Figure 37.4 at bottom. Rail tracks run horizontally, vertically or under 45 degrees. We leave it to the reader to decide whether it is a diagram or a map. Beck ignored — to some extent — the geographic locations of the stations, and simply presented the user with the information he or she needed. His rethinking of the map is still used today. Looking at the two different maps (Figure 37.4 on top and at bottom), which are on the same subject, we see that in the one inspired by Beck, the geometry of features may be much different, but the *topology* in fact remains intact! The subject of topology is covered in Section 39.4. Obviously, the map of Figure 37.4 at bottom does not have a uniform scale across the area. In the diagram at bottom, nearly all stations are positioned equidistantly along the tracks, while this is certainly not the case in reality.

Along the same lines, let us think of an imaginary, millimeter-accurate, map of the Netherlands. There would be a staggering amount of detail: every leaf, cobblestone and roof tile would be drawn out extremely precisely (by the way, if you have such an detailed map, there are many organisations that would pay you very well for it!). The Netherlands covers a surface area of about 41,500 $\mathrm{km}^2$. That is approximately 41.5 billion square meter. A typical laptop screen, however, has a surface area of less than 0.1 $\mathrm{m}^2$. So if you want to create a map of the Netherlands that can actually be viewed on a laptop screen, you would have to scale down your imaginary, enormous map to make it usable. The leaves and cobblestones in your original map will be several orders of magnitude smaller than a single pixel on your screen - hence the necessity for generalization!

The issue of overly-abundant detail is demonstrated in the TOP25NL-map in Figure 37.5, which is unusable at the smaller scale of the map at right, where a lower detail-level would be more practical. The need for choosing an appropriate level of generalisation is again made obvious in Figure 37.6 at right, which uses the less detailed raster map TOP500NL; the map at right no longer displays individual buildings in the Delft, it just shows the built-up area, and only main-roads are kept. That is all. Hypothetically, you could construct a computer screen the size of the Netherlands so you can display everything in full detail - but likely you see now the practical benefit of generalization and scaling.

While looking at the Figures 37.5, 37.6, and also 37.7, you observe different levels of *scale*. A scale of 1:$x$ simply expresses that 1 unit in the map (e.g. 1 cm) corresponds to $x$ units in reality (e.g. $x$ cm). For example, the map of Figure 37.5 at left, is displayed here at 1:25,000, meaning that 1 cm in the map, corresponds to 250 m in the terrain. The scale is $\frac{1}{x}$, meaning that a smaller scale map is able to show less detail. A larger scale map can show much more detail.

With dynamic maps visualized on a screen, rather than old days paper maps, the user can smoothly, and virtually endlessly, zoom-in and zoom-out, see also Figure 37.10.

So simplification and generalization are unavoidable, and can make your map more readable and effective. Your map is a medium for conveying the important data and features about reality; what is relevant to the theme, and what is not? Generally, 'less is more'. Keep in mind however, as the often quoted Albert Einstein once said, 'A problem [e.g. modeling reality]

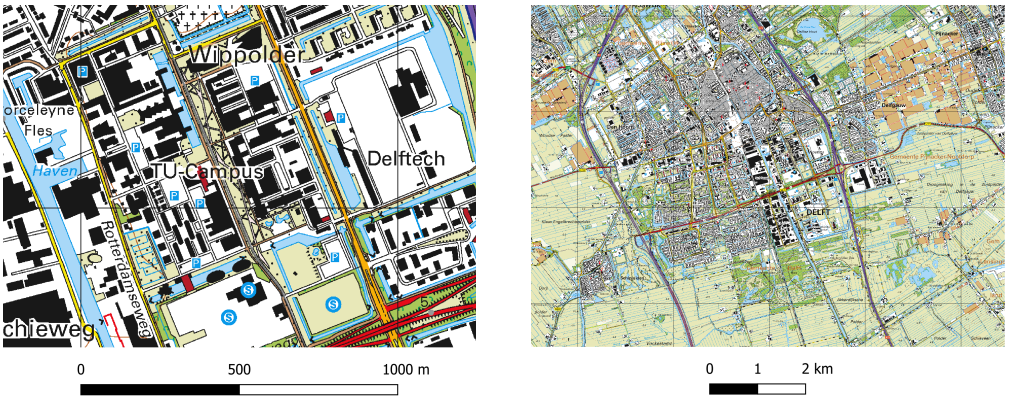



Figure 37.5: The city of Delft as mapped by TOP25NL, a topographic raster map created for a target scale of 1:25,000, containing detailed features such as roads, buildings and waterways. TOP25NL shown at a 1:25,000 scale, at left, and at (approx) a 1:165,000 scale, at right (stated scales refer to print on A4-paper). In the map at right, there is a far too high level of detail in relation to the scale. Basisregistratie Topografie (BRT) maps by Kadaster, taken from PDOK [59] under CC BY 4.0 license.


Figure 37.6: The city of Delft as mapped by TOP25NL (left) and TOP500NL (right), both shown here at the *same* scale, (approx) a 1:165,000 scale. Compared to TOP25NL, TOP500NL is produced for a target scale of 1:500,000 and includes far less detail, representing Delft by only the boundaries of its built-up area. Basisregistratie Topografie (BRT) maps by Kadaster, taken from PDOK [59] under CC BY 4.0 license.

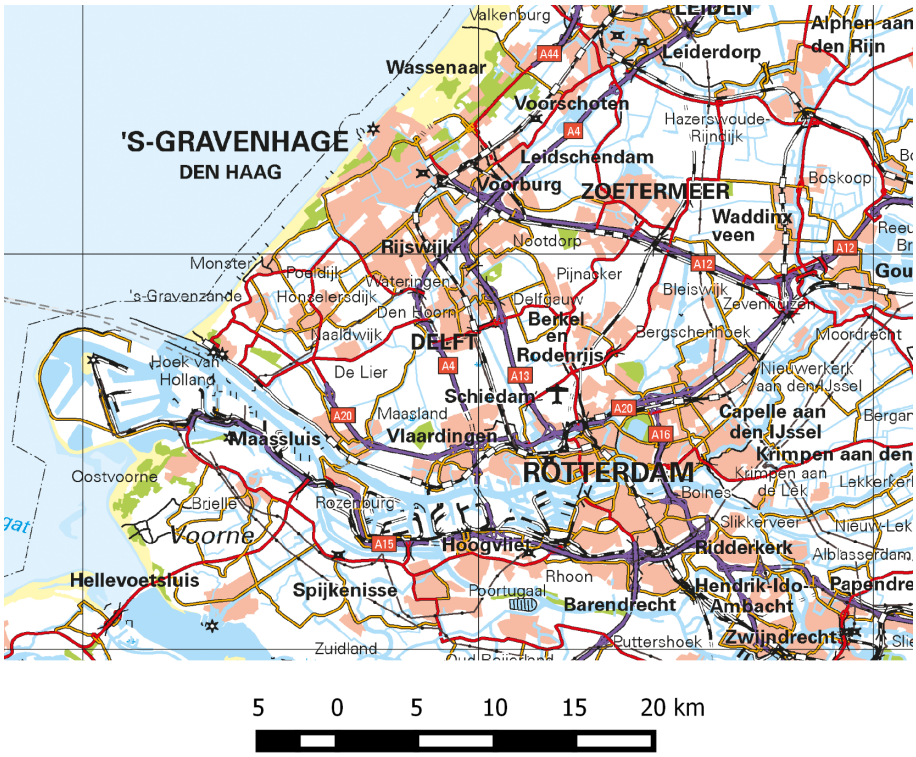



Figure 37.7: Same source as Figure 37.6 at right (TOP500NL), but at a smaller scale (here 1:500,000). We see that the lower level of detail is appropriate for this scale, as opposed to using TOP25NL. Basisregistratie Topografie (BRT) map by Kadaster, taken from PDOK [59] under CC BY 4.0 license.


Figure 37.8: Ptolemy's world map, from his book *The Geography* written in the 15th century [105], originally written by Claudius Ptolemy around 150 AD. Inset on lower right is a zoom-in on Western Europe. In Ptolemy's map we see the world as it was known at the time by the Roman Empire, which, as we can see, did *not* include the Americas, most of Africa and Asia, Oceania, and the Poles. Image taken from Wikimedia Commons [9], work credited to Francesco di Antonio del Chierico (1450-1475). Public Domain.

should be made as simple as possible, but no simpler'.

## 37.4. History of cartography

It would be appropriate at this time to reflect on the history of cartography. It was not so long ago that accurate maps were worth their weight in gold (well, more than their own weight even!) to traders, explorers and military leaders. As such, cartography has a long and rich history — of which we will highlight here just a few topics.

Claudius Ptolemaeus (c. AD 100 - c. AD 170) was an accomplished mathematician, astronomer, and geographer who lived in Alexandria, which was in the Egyptian part of the Roman empire at the time [104]. One of his main works is *The Geography* (Figure 37.8), a compilation of geographical data known to him and his contemporaries in the Roman Empire — roughly 8,000 different locations around the world [104]!

In the first part of *The Geography* Ptolemy discussed the data and methods he used — and a key innovation was that he recorded latitudes and longitudes of locations on Earth for the first time. He also devised two methods for representing these circles of latitude and longitude on a flat map, using a grid of lines [104]. Ptolemy's map was of course distorted, not accurately representing the size and orientation of all features. This distortion stems from the fact that the scale of landmasses and water bodies was determined in a very *subjective* manner — mostly based on the experience of the explorer as he travelled across it. From our own experience, we can recall how a few hours of time can seem very short when we are

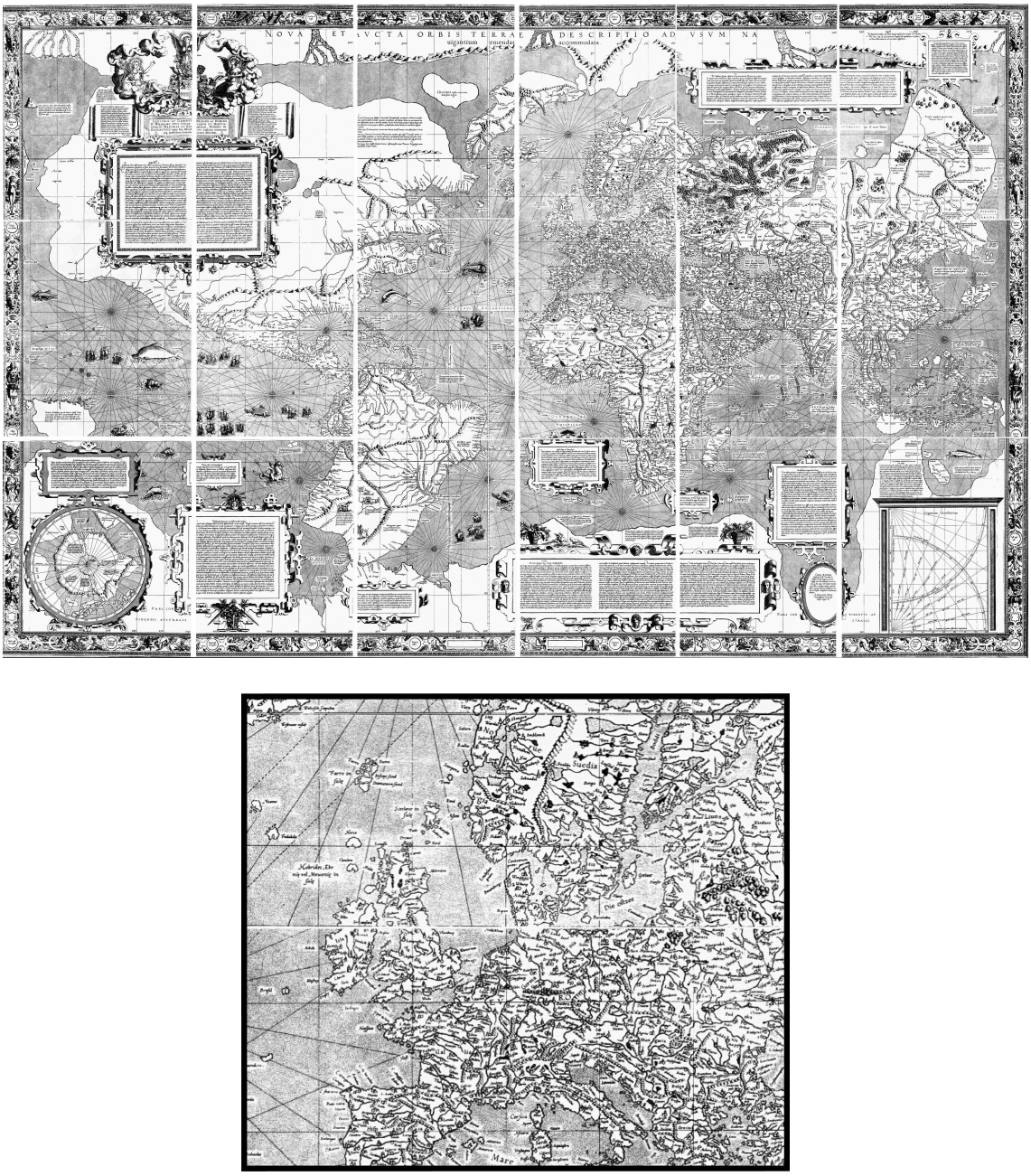


Figure 37.9: Mercator world map of 1569 - *'Nova et Aucta Orbis Terrae Descriptio ad Usum Navigantium Emendate Accommodata'* [106]. This map gives an overview of the Earth that comes close to our modern day knowledge - except for Australia, which was only suspected to exist as a continent, and Tasmania, which was entirely unknown - until it was discovered by Dutch explorer Abel Tasman in 1642. The Western Europe part of the map is also shown in more detail (at bottom). Image taken from Wikimedia Commons [9]. Public Domain.

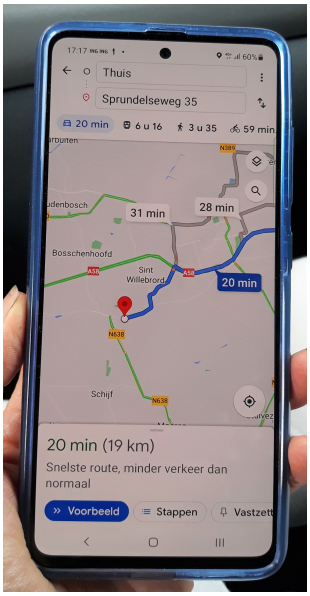



Figure 37.10: Map supported directions to your destination by means of Google Maps on your smartphone.

having fun, but very long when we are bored without entertainment. Nevertheless his map was a relatively clear and detailed illustration of the Roman Empire at the time.

Another main character, whose name you might encounter often, is Flemish cartographer and geographer Gerardus Mercator (1512 - 1594). Mercator is most famous for the world map he created based on the projection system he proposed — the Mercator projection — which he introduced to the world in 1569 [107], see the map in Figure 37.9. Its legend contains various texts: copyright claims, a message of thanks to his employer, and explanations of concepts. The Mercator projection is a a conformal map projection, meaning that it preserves the angle of intersection of any two curves, see Chapter 30.

The latitude lines are equally spaced, straight, and parallel, while the longitudinal lines are parallel and straight, but at increasingly larger distances closer to the poles. With this projection, any straight line in the map is a line of constant bearing on the globe, see Section 30.4.2 — this was very useful for sailors, who could then navigate a straight-line course. Hence the name of the map *Nova et Aucta Orbis Terrae Descriptio ad Usum Navigantium Emendate Accommodata*, which is Renaissance Latin for 'New and more complete representation of the terrestrial globe properly adapted for use in navigation' [106]. Mercator's map forms the basis of navigation until this very day — a milestone in the history of mapping and navigation.

Speaking of navigation, if we skip ahead a couple of hundred years, we find ourselves in the 21st century. Navigation and cartography have made great advances, and have become freely accessible to anyone with an Internet connection. In many countries, the geographic coordinates of nearly every building is known, and navigation is a breeze - like finding the fastest route to your destination using your smartphone (Figure 37.10) and based on GPS satellite navigation. Cartography has extended to animated and interactive maps, with personalized content.

There is much more to say about the rich and vast history of cartography, but the scope of this book is limited to this brief section. The interested reader is referred to e.g. a series of six volumes by Harley and Woodward: [108], [109], [110], [111], [112], and [113].

# 38

# Maps

As we have seen in the previous chapter, a map is a model, or visualization of reality, in our case, specifically the Earth's surface. Turning a spherical Earth into a flat piece of paper or screen, is covered in Chapter 30 on map projections.

In this chapter we present a concise overview of both topographic maps and thematic maps. Topographic maps are an indispensable resource of geographic information for building and construction, during the design, deployment, and maintenance of civil engineering projects. Thematic maps are an often used tool to communicate the results of a survey or enquiry, or the results of an impact study to customers and the public.

The second part of this chapter is devoted to guidelines for working with visual variables, and eventually producing a proper map — a map which is clear, attractive and well understandable, i.e. which serves its purpose of transferring the right message.

## 38.1. Topographic maps

One of the most commonly seen types of maps is the topographic map. The distinctive feature of a topographic map is that it shows the natural shape of the Earth's surface, and its topography. The focus is on *geometry*, with typically both natural and man-made features being shown in large-scale detail. A topographic map is a *general purpose* map, and thereby also often used as a background for other types of maps.

In this section we will cover a few different topographic maps which are (freely) available online, to give you an idea of the cartographic resources that exist out there.

### 38.1.1. Basisregistratie Grootschalig Topografie (BGT)

The Basisregistratie Grootschalige Topografie (BGT) is a detailed, large-scale digital base-map of the Netherlands. It is a key tool for civil engineering projects, as it maps the location of all permanent physical objects such as buildings, roads, water bodies, railway tracks, and (agricultural) plots of land — that is topography. A sample screenshot can be seen in Figure 38.1, and the data can be downloaded in both vector and raster format from the PDOK website [59] (see Appendix I).

The BGT has a long list of applications. It can answer questions like: What is the shape and direction of this sidewalk? What objects are located around the building of interest? As you can imagine, having an overview of a site by means of a base-map is exceptionally useful when carrying out building and construction works.

Figure 38.1: Sample of the Basisregistratie Grootschalige Topografie (BGT) Achtergrondkaart, for the area around the Stevinweg in Delft. In this map you can view every individual object registered in the BGT, even the small plot of your grass in front of your house, managed by the government, and obtain administrative information about the object. BGT Achtergrondkaart by Kadaster, taken from PDOK [59] under CC BY 4.0 license.

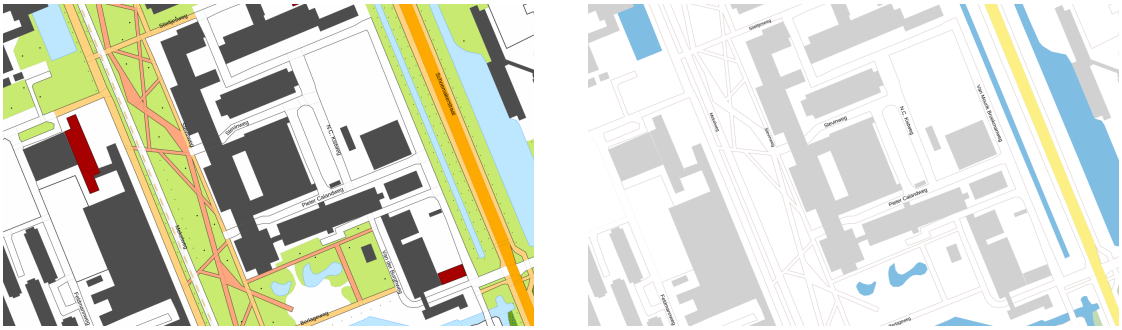


Figure 38.2: Sample of TOP10NL map for the area (of Figure 38.1), of about 330 m x 550 m, around the Stevinweg in Delft. At left the TOP10NL-map, and at right, much similar, the BRT Achtergrondkaart (BRT-A); they share the same geometry. Basisregistratie Topografie (BRT) by Kadaster, taken from PDOK [59] under CC BY 4.0 license.

### 38.1.2. Basisregistratie Topografie (BRT) - TOP10NL

The Dutch Kadaster (the Netherlands' Cadastre, Land Registry and Mapping Agency) publishes TOP10NL, the digital topographical base map of the Netherlands. It is the most detailed topographic product within what is known as the Basisregistratie Topografie (BRT) — and useful for creating maps with scales in the range between 1:5000 and 1:25.000. The target scale of this map, obviously is 1:10.000. A sample can be seen in Figure 38.2, and be compared with the BGT-map in Figure 38.1, the latter showing even a higher level of detail.

The TOP10NL is produced in Geography Markup Language (GML) format, a common format for geospatial data. You can download map sheets freely from the PDOK website [59], see also Appendix I, and use AutoCAD or a DWG viewer, or QGIS [5], for these maps.

The TOP25NL and TOP500NL maps of the Basisregistratie Topografie (BRT) were shown already in Section 37.3.

### 38.1.3. Underground topography

With civil engineering projects and construction-works, an important part of the infrastructure is located underground. By the WION (Wet Informatie-uitwisseling Ondergrondse Netwerken), and from July 2018 on, covering also 'bovengrondse netten' (WIBON), cable- and pipeline-infrastructure needs to be mapped. One can think of telecommunication infrastructure, and public utilities like gas, water, electricity and sewage, as well as privately owned and operated pipeline-infrastructure (e.g. for petro-chemical industry). The registration and mapping, on a

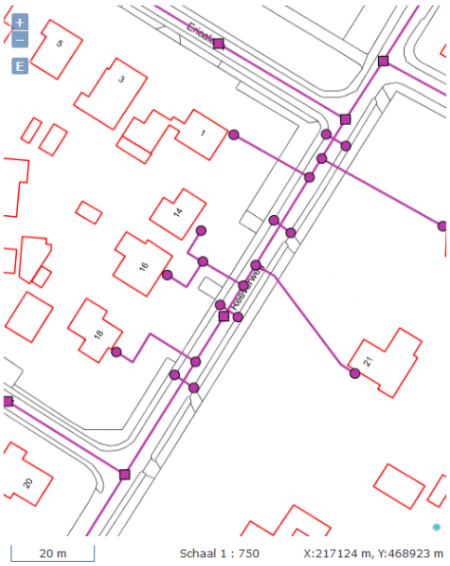



Figure 38.3: Example of underground infrastructure map: fragment of sewerage network around the Reeverweg in Harfsen. Image courtesy of Stichting RIONED [114], dated August 16th, 2018.

scale of 1:500, is organized in the Netherlands through the Kabels en Leidingen Informatie Centrum (KLIC) hosted by the Dutch Cadastre. Figure 38.3 shows an example of an underground infrastructure map. The purple line in the middle of the street is the combined sewer, for both storm water and sanitary/waste-water. The purple squares are manholes, and the purple circles are storm drains in the pavement or house lateral connections.

### 38.1.4. Actueel Hoogtebestand Nederland (AHN)

The Netherlands has an incredibly detailed and up-to-date raster dataset for heights with a horizontal grid resolution up to 0.5 meters, known as the Actueel Hoogtebestand Nederland (AHN). Chapter 22 mentions that the Actueel Hoogtebestand Nederland (AHN) has been collected with airborne laser scanning. The AHN-3 has been flown from 2014 to 2019, for all of the country. The result is a Digital Elevation Model (DEM) of all of the Netherlands. AHN3 can be downloaded per tile from PDOK [59], see also Appendix I.

The elevations pertain to the terrain ground level ('maaiveld'), buildings and infrastructure, in the so-called filtered version. In the unfiltered version also vegetation is present. The elevations in this model have a precision of about 5 cm (standard deviation), with at most a 5 cm of systematic offset. An example of the TU Delft campus area is shown in Figure 38.4.

### 38.1.5. 3D Maps

Traditionally, topographic maps have been two-dimensional. Now, with three-dimensional surface model data available, the visualization on a two-dimensional map or screen, poses challenges. One can use shadow-effects to indicate relief (shading), or use perspective viewing, see Figure 38.5. This is often done when illustrating Digital Terrain Models (DTM) and Digital Surface Models (DSM) to make them look more realistic. The different Level of Details (LODs), as conceptual model of buildings, are shown in Figure 38.6.

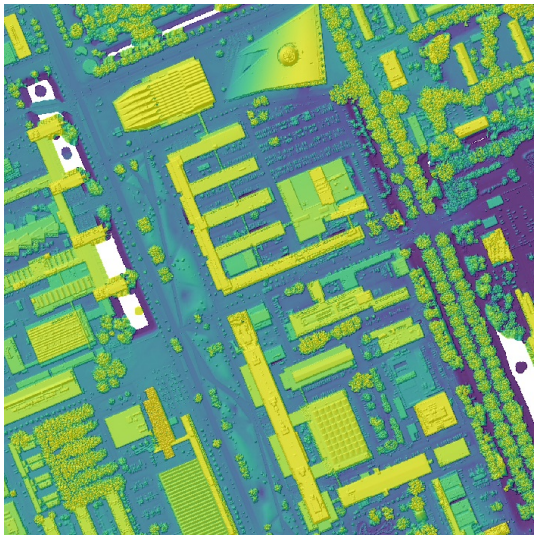


Figure 38.4: Example of Actueel Hoogtebestand Nederland (AHN-3), unfiltered (with vegetation and buildings present) - Digital Surface Model (DSM), shaded relief. The height (in NAP) per pixel (0.5 m x 0.5 m) is shown in color-scale from blue through green and yellow to red. The image shows a detailed scene of the TU Delft campus, with the Aula and library on top. AHN by Rijkswaterstaat; data retrieved from PDOK [59] under CC0 license.

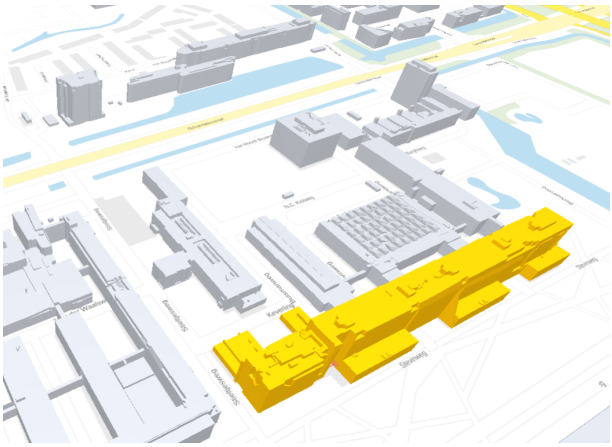


Figure 38.5: View on TU Delft campus around Stevinweg from 3D BAG 2.0 service; Level of Detail (LOD) models reconstructed from buildings in Basisregistratie Adressen en Gebouwen (BAG) and AHN-3 point clouds. Shown is LOD 2.2 on BRT Achtergrondkaart. Image taken from 3D BAG by 3D Geoinformation Research group at TU Delft, under CC BY 4.0 license. Also available as 3D Basisvoorziening by Kadaster through PDOK [59].

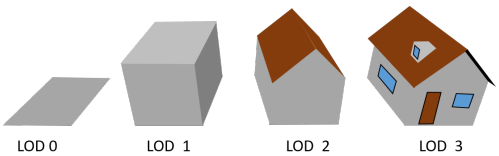



Figure 38.6: Level of Detail (LOD) for visualization of buildings and Digital Surface Models (DSM) in 3D maps, image after [115].

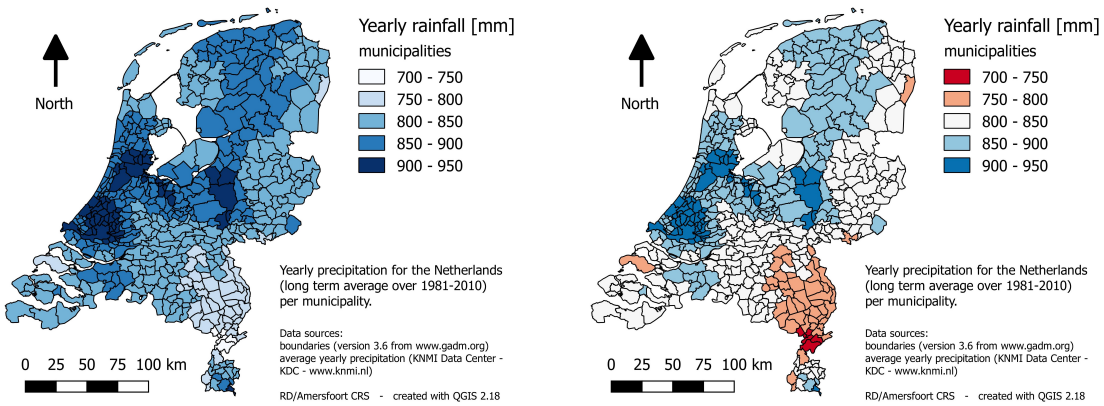



Figure 38.7: Chloropleth map of yearly rainfall in the Netherlands, per municipality, as an average over 1981-2010. At left, using an appropriate visual variable (value/lightness), the darker the blue-color, the more intense the phenomenon (rainfall), and at right, using an *in*-appropriate visual variable, the use of different colors (hue) is *not* suited to present hierarchy or rank - why would red imply less rainfall than blue, or the other way around? Municipality geometry data from GADM maps and data [118] and precipitation data from KNMI Dataplatform (KDP) [119] as Open Data CC0 1.0.

## 38.2. Thematic maps

Thematic maps, as opposed to general purpose topographic maps, focus on specific subjects, aiming to show a particular theme linked to a geographic area of interest. Themes can be land cover, traffic density, flooding risk, and for instance income, age and religion as census data, and many, many more. The two most common types of thematic maps are the choropleth map and the chorochromatic map.

### 38.2.1. Choropleth map

Chloropleth maps are used to show statistical information that is aggregated by geographic area. It was described by American geographer John Kirtland Wright to mean 'quantity in area' [116]. Additionally, a hint to the meaning of choloropleth can be found in its Greek root words, *choros* (meaning 'area') and *plethos* ('value') [117].

The magnitude of the variable of interest (attribute) may be represented using the saturation of a color or the lightness/darkness (value) of a color, or just shades of gray in a black and white figure. The differences in color-depth or darkness denote differences in the intensity of a phenomenon, for instance the amount of rainfall in millimeter over a period of one year in Figure 38.7 at left (created from spatially interpolated precipitation measurements). This allows the viewer to easily see how the phenomenon varies spatially; which areas do get most percipitation, and which areas are relatively dry?

Also Figures 38.11 and 38.12 later on, show choropleth maps.

### 38.2.2. Chorochromatic map

For qualitative data, as opposed to statistical data, we use another type of thematic map: a chorochromatic map. It is composed of the Greek root words *choros* (meaning 'area', as we just saw) and *chroma* ('color') [117]. The original meaning was to render nominal values for areas by using different colors (hue). As we can see in Figure 38.8, it is very useful for mapping descriptive data.

Using different colors or patterns, the distribution of a qualitative characteristic over a region is illustrated. Note that since chorochromatic maps visualize qualitative data, it is very

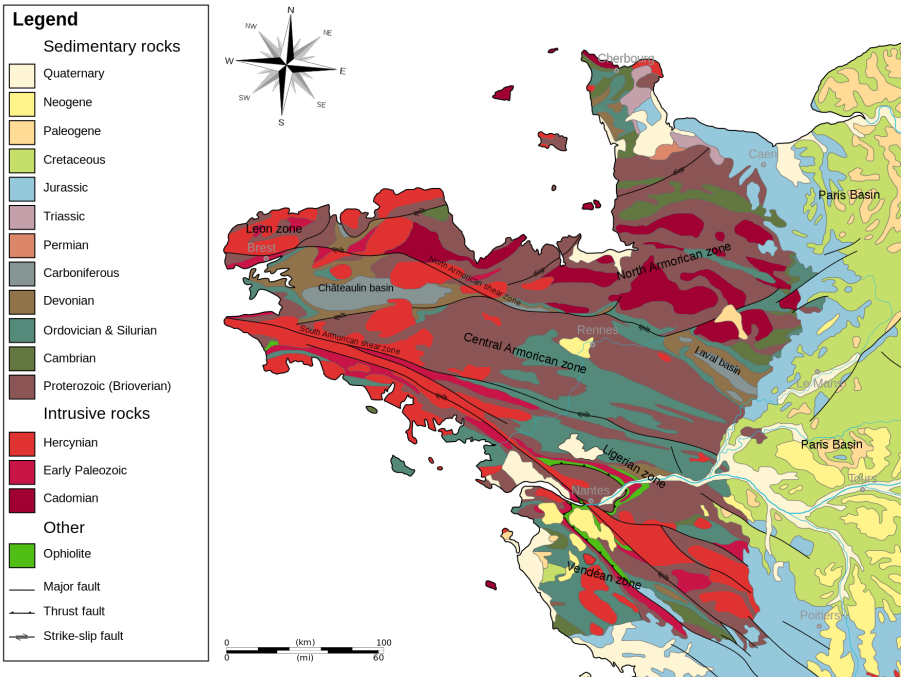



Figure 38.8: Chorochromatic map: geological map of the Armorican Massif in France (Bretagne). Map by Woudloper, own work, January 2009, taken from Wikimedia Commons [9] under CC BY-SA 3.0 license.

important that the choice of colors or patterns do *not* suggest a hierarchy or order of the classes, like quaternary, jurassic and cambrian in Figure 38.8. They imply just different types of sedimentary rocks, and cannot be mutually compared. Cambrian is neither less, nor more important than quaternary — it is just different from.

### 38.2.3. Types of map data

Map data are stored in a Geographic Information System (GIS), see Chapter 39, and prior to storing features and their attributes, it is worth thinking about how to 'code' them. An attribute can be defined by a character string (e.g. a street-name), or a number (e.g. 42), with the latter being an integer or real (float) number.

Another consideration is what is known as the 'levels of measurement' scale. This classification was developed by an American psychologist, Stanley S. Stevens, and it can be useful to conceptualize the differences between types of data values, in order to handle them in map-making. To clarify, Table 38.1 shows an example of data from the results of a marathon race, with the different types of data values indicated.

The four 'levels' are [4]:

- *Nominal* data (nominal as in 'to name') identifies or categorises data items. There is no indication of a relative value or ranking. Can we say that the name 'Caroline' is better than 'Dennis'? 'No'! Its purpose is only to *identify*, like a name, or a telephone number. An example of *unique* labels are parcel-numbers as used for land registration, e.g. 'DEL00-D-429' for parcel number 429 in section D of Delft. Examples of categories, or classes of road surface are tarmac, concrete, brick and cobblestone. The nominal data type is typically coded as a character string.

- *Ordinal* (as in 'to order') *does* indicate a type of order ranking. The intervals between ranks (such as 'great' or 'good') are not known, but the order is known ('great' is better

| Nominal | Ordinal | Interval | Ratio |
|---|---|---|---|
| participant | ranking | finish time | race time |
| Mick | 1 | 11:10 | 2:30 |
| Tycho | 2 | 11:15 | 2:35 |
| Bob | 3 | 11:17 | 2:37 |
| . | . | . | . |
| . | . | . | . |
| Patrick | 450 | 19:30 | 10:50 |

Table 38.1: An example of the four levels of measurement of data, using data from a marathon race as an example (after [4]).

or more than 'good'). This data type is often coded with characters (and you cannot apply math to them). Only relational operators as '<', '>', and '=' apply to them.

- *Interval* data, is purely numerical. The difference, or 'interval' between numbers is meaningful (arithmetic operations like '+' and '−'), but there is no meaning to dividing or multiplying the numbers. As an example, we can consider temperature data. $30°$ Celsius is $10°$ warmer than $20°$ Celsius: an interval of $10°$. However, we cannot say that $30°$ is twice as warm as $15°$, because interval data does not make sense when multiplied, as it does not start at some 'true zero' value. It is the *difference* that matters (and that can be quantified).

- *Ratio* data, however, does have an absolute zero, unlike interval data. Think of time spent, height, and weight. It makes sense to say that a race time of 4 hours, is twice the race time of 2 hours. Addition, subtraction, multiplication and division of ratio values make sense ('+', '−', '×', and ':'). Other examples of ratio data include the price of real-estate (in Euros), the amount of rainfall (in mm/year) and the percentage of unemployment (in percent). Obviously, ratio data is numerical.

### 38.2.4. Other types of thematic maps

A contour plot is a technique for representing the variation of an attribute over a geographical area of interest. Over a 2D map, we draw contour lines, also referred to as iso-lines, which each connect locations with the same data value. Figure 38.9 shows a contour map of rainfall data, based on measurements of over 300 meteo-stations at discrete locations throughout the Netherlands, interpolated in a raster to cover all of the country. In Figure 38.7, the interpolated raster map was aggregated to municipality areas (using the mean value).

## 38.3. Cartographic rules and guidelines

How to design a good map yourself? And, are there any useful guidelines for doing this? In this section we present a few basic principles of visualization. This field even touches on perception psychology.

### 38.3.1. Principles of information visualisation (Bertin)

In this section we discuss a crucial aspect of cartography: how do we encode information into visual variables and geometric shapes, and how does the reader decode that information? For this reason, Jacques Bertin [120] proposed a systematic set of rules for cartography. Part of his system can be seen in the annotated table of Figure 38.10.

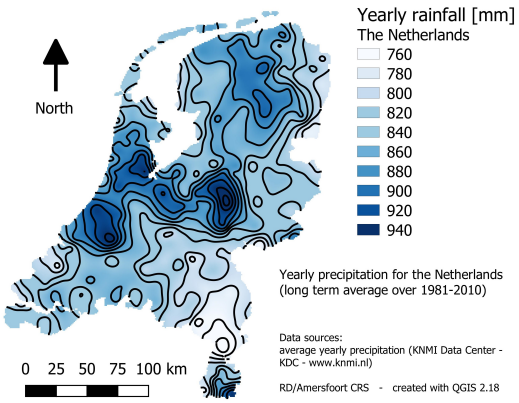



Figure 38.9: Annual precipitation interpolated over the Netherlands (average over 1981-2010) shown as a contour map. Precipitation data from KNMI Dataplatform (KDP) [119] as Open Data CC0 1.0.

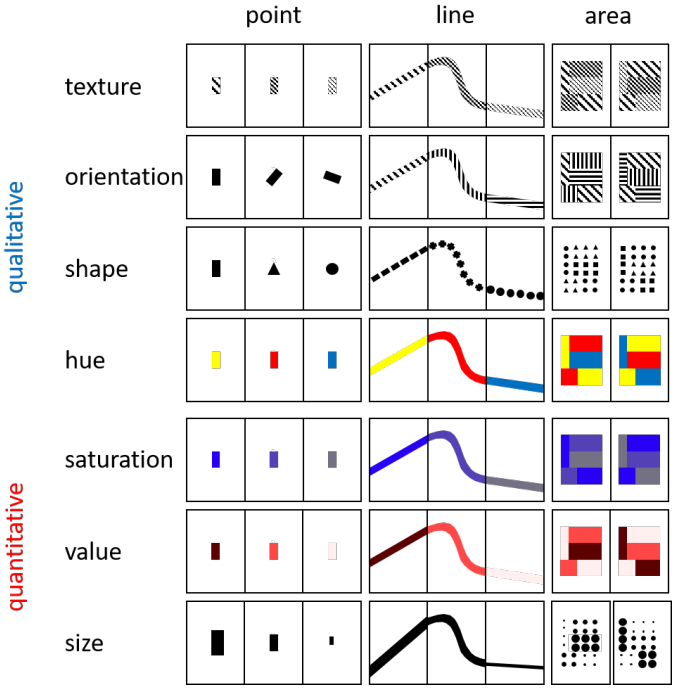



Figure 38.10: Table of qualitative and quantitative visual variables, listed vertically, for three geometric entities in maps: points, lines and areas, listed horizontally. This table can help you select an appropriate visual variable for your map (after [120]).

As we can see, many different types of visual variables exist and can be used, and depending on their purpose they may be more, or less appropriate. The first four rows are visual variables best suited for *qualitative* data (descriptive data, e.g. nationality, and land cover), whereas the remaining three are more appropriate for mapping *quantitative* data (data that can be quantified in numbers). Remember that in Section 38.2.3, we presented different types of data values! Things should come together now.

In the sequel, some examples will demonstrate how Bertin's visual variables are used in practice - and also abused, leading to unclear and misleading figures.

One example of a potentially misleading figure was shown already in Figure 38.7 at right. The visual variable 'hue' can*not* be used to present quantitative data.

The item 'value' (or similarly 'lightness') shows that differences in gray (ranging from black to white) show differences in density, e.g. the percentage of unemployed people (ratio data value, hence quantitative). An area which is darker is associated with a higher percentage of unemployment. There clearly is a hierarchy or order. When correctly applied, percentages or densities that are twice as high are represented by a gray value that is twice as dark.

Though, it is not always that easy — one should be aware of potential ambiguous interpretation of the map. Indeed, generally, the darker the gray values, the more intense or the higher the densities of the phenomenon. But, another interpretation could be that the darker the area tints, the less favourable the conditions of the phenomenon are. Then it might be difficult to combine these two rules and interpretations. Literacy might be taken as an example: to render increasing literacy percentages on a global map through tints that increase in value maybe interpreted as the less favourable condition (il-literacy) being represented by higher tints [117].

In Figure 38.10 we use the HSV (Hue, Saturation, Value) color model (or similarly HSL, with L for Lightness), as an alternative to the well-known Red-Green-Blue (RGB) color model. It describes 'colors' by using three variables. Individually, each of these can be used as a visual variable [120], with hue being appropriate for *qualitative* data, and saturation and value (lightness) for *quantitative* data.

When describing sediment rock type, we use descriptive words: 'jurassic', 'cambrian', and 'quartenary', etc. There is no particular numerical value to these classifications, and therefore no hierarchy either. Therefore a *qualitative* visual variable should be used to illustrate the data, as was done in Figure 38.8, namely hue! In this map, we can immediately and clearly see the distribution of sediment rock type over the area. In this context one can think also of a land cover map and use appropriate colors: when the reader sees blue areas on the map, they are quick to associate it with water — hence a good choice to use the color blue to represent water bodies, and not for grassland, for instance.

For quantitative data, it is also appropriate to use colors as a visual variable. However, it must be done differently than in the previous example of Figure 38.8. For the *quantitative* data in Figure 38.11 at right, using hue as a visual variable leads to an extremely confusing rainbow of polygons. Why are the more populated areas yellow, and the less populated ones blue or red?

So, this is a good example of a poor choice of visual variables! When displaying quantitative data, variations in color value (lightness) and color saturation allow the reader to quickly and clearly see how the phenomenon is spatially distributed, and which areas have the higher magnitude. This is what we see in Figure 38.11 at left, where color value/lightness is used as a visual variable.

As we have seen, the same dataset can look very differently depending on the choice of visual variables. It can sometimes be misleading, by implying a certain hierarchy in the data categories where there is not one.

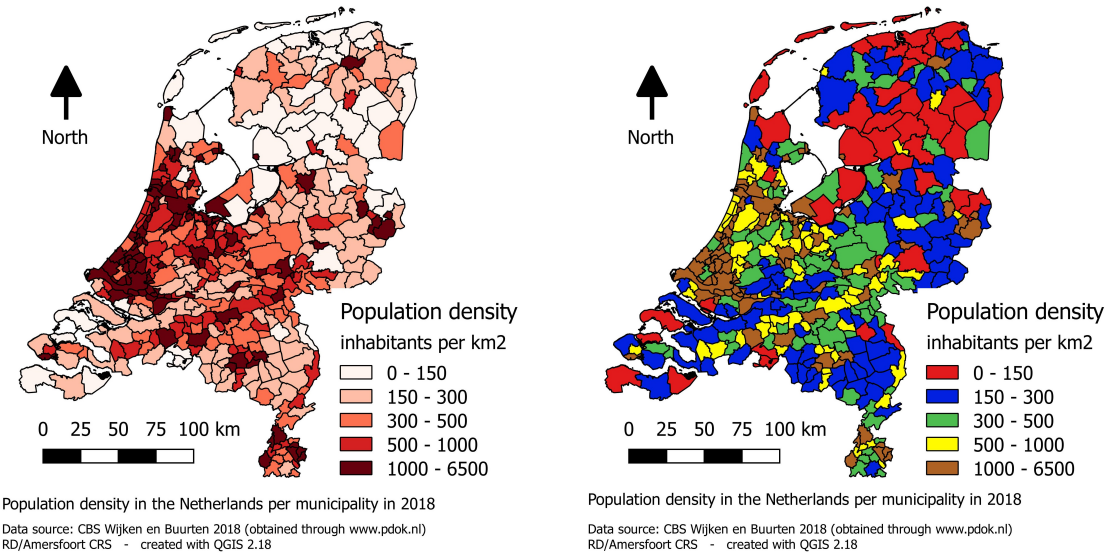



Figure 38.11: Population density in the Netherlands, per municipality, visualised using two different visual variables, namely value and hue, resulting in two different color maps. At left a clear and indicative choice of color: value (lightness), and at right a poor map with a confusing choice of color: hue. Data from CBS Wijken en Buurten [121], obtained through PDOK [59] as Open Data CC0 1.0, Public Domain.

Knowing that a biased map can strongly influence the viewer, it is important to realize that the cartographer's role is primarily to inform, not to influence!

### 38.3.2. Good practice

Data can be powerful, but they have no 'wind in their sails' until they get well visualized or otherwise presented. In this section some of the key points are presented to creating a good map that communicates well. This section by no means provides a full guide to create a good map — just a couple of major guidelines are given.

Consider the feature type you would like to map, cf. Section 39.2, and choose an appropriate mapping technique for the data type in Section 38.2.3, following the guidelines in Figure 38.10.

Keep in mind that using absolute data over percentages may give a skewed impression of the message you want to convey, in this respect also compare Figure 38.11 at left, showing density (inhabitants per square kilometer), and Figure 38.12 at right, showing absolute counts (number of inhabitants). An other example would be the absolute count of unemployed people versus percentage of unemployed people (of the total working population).

The legend should be clear, and also specify the data shown, as well as mention the source of the data. The legend may also contain a date, and mention the coordinate reference system. Make sure to use text with a proper font. The map should contain a North arrow for spatial orientation, and a scale-bar for proper reference.

The layout, and relative sizes of the different elements of the map (like title, legend, etc.) should be arranged such that first and foremost the reader's attention is drawn to the data ('what is the message?'). Do not overload the reader with information, focus on your message. Balance the shape and size of elements in your map. Distribute the elements of the map evenly (geographic map itself, legend, title, further annotation). Do not leave a large white unused space somewhere. Does the image look good overall?

Keep in mind that the use of (some of the) colors can attract more attention than it deserves, so balance color-usage. A map is a model of reality, and hence implies simplifications.

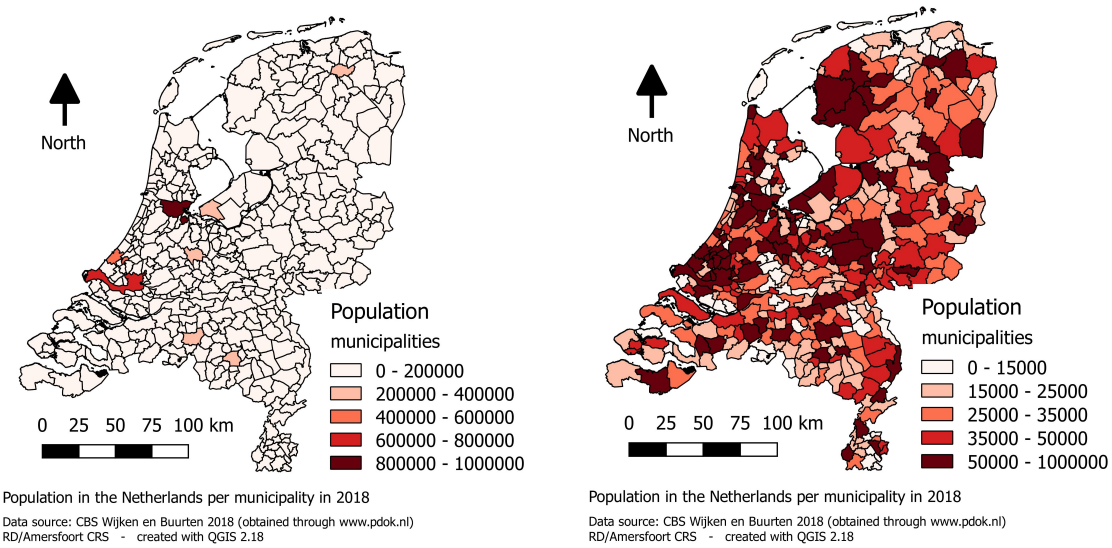



Figure 38.12: Population map of the Netherlands, with number of inhabitants per municipality. At left shown with *equal-interval* classes, and at right with *quantile-based* classes. At left, there are five classes, each covering an interval of 200.000 inhabitants. At right, the class boundaries are certainly *not* equidistant, but chosen such that each class covers (about) the same number of municipalities (the 'data points' are equally distributed over the five classes for the visualization). Data from CBS Wijken en Buurten [121], obtained through PDOK [59] as Open Data CC0 1.0, Public Domain.

The attribute you would like to present should be categorized, or classified into a limited number of categories or classes. One can think of a limited number of different rock types, as in the map of Figure 38.8, rather than showing 238 different ones .... A quantitative attribute may cover, as a real number, an infinite amount of different values, and this needs to be reduced to a limited number of bins, or classes for a clear interpretation, see the map of Figure 38.7. There shall be not too few, nor too many categories or classes. With quantitative data, generally five or six classes suffice.

There are several, or even many ways to set the classes. One can use equal intervals, just by cutting the minimum-maximum range of the attribute into five or six equal length classes, as used in Figure 38.12 at left, but this simple choice may not always be the best choice. An alternative is to use quantiles to set the class-boundaries. In this way, the total amount of data is distributed evenly over the classes. Each class contains the same number of 'data points'. One could first create a histogram of the data, to get a first impression of the distribution of the data.

Figure 38.12 shows a population map of the Netherlands, based on data of the Centraal Bureau voor de Statistiek (CBS), [121]. On the left with (default) equal interval on the population count, which results in pretty much a 'flat' map — there is not much to see, only the real big cities stand out, and for the rest the Netherlands is 'pretty empty'. At right with quantile-based classes, in order to emphasize the spatial variation of population in the Netherlands, also across the country. The two maps give a really different impression, and the only cause for this lies in a different definition of the classes for the population attribute. Showing population *density* as in Figure 38.11, will actually give a more trustworthy impression, instead of absolute values like in Figure 38.12. Showing absolute counts rather than density may distort or obscure the message. In the map of Figure 38.12 at right, you see high population values also in the provinces of Friesland, Groningen and Drenthe, whereas we do not see high population density values there in Figure 38.11 at left. The municipalities do not all have equal

size (area) and this may corrupt the message of the map. Large area municipalities may have a reasonable number of inhabitants, but due to the large area, the density is still low, and, as they cover a large area, their fairly large absolute counts may visually dominate the map of Figure 38.12 at right. The question again is: what is your message?

# 39

# Geographic Information System (GIS)

Geographic Information Systems (GIS) are everywhere. You may not have realized, but likely you already often used a GIS. You are strolling through town and consulting a service or app on your smartphone, in order to find the nearest pizza-restaurant. A GIS can answer questions like 'What is located here?', and 'Where is the TU Delft Aula (or another object of interest)?', and also 'Where is the nearest ATM?'. As soon as spatial, geographic information comes into play, a GIS enters the scene.

A GIS may be a valuable tool in analyzing the noise zones around a newly built track of railway. Town-planners may use a GIS, based on satellite images, to see where informal settlements are, and consequently to help in planning infrastructure in big cities for instance in Africa, Asia and South America. A GIS is an indispensable ingredient of a digital twin of built-area, a town or a city. Rescue-teams may use a GIS, with a recent satellite image, to see where impacted areas are of a natural disaster, such as a flood, and where people may need rescueing. And, there are many, many more uses.

This chapter provides a brief introduction to the subject of Geographic Information Systems.

## 39.1. Geographic information: early trace

In presenting the early traces of Geographic Information Systems, often the example of John Snow is brought up (also in [4]). John Snow was an English physician, and considered to be one of the fathers of modern epidemiology [122]. He traced the source of a cholera outbreak in the Soho-district in London, in 1854 [123]. How did he do this? Exactly, by using geographic information in his analyses!

Snow identified the source of the outbreak of this epidemy as the public water pump in Broad Street, see Figure 39.1. All pumps are indicated on this map by a rectangle in red, and the one on Broad Street, in the middle of the map, also by the arrow. The spatial analysis Snow carried out was convincing enough to persuade the local council to disable this well pump by removing its handle in order to end the outbreak of this disease.

Snow indicated the houses with occurrences of cholera. Rather than producing an administrative list or table (with addresses of the patients and fatalities), he marked them in black on the map in Figure 39.1. And soon a *cluster* of the occurrences appeared (of black houses). Snow concluded that nearly all deaths had taken place within a short distance of the pump in Broad Street. He also observed that there had been no particular outbreak or prevalence of

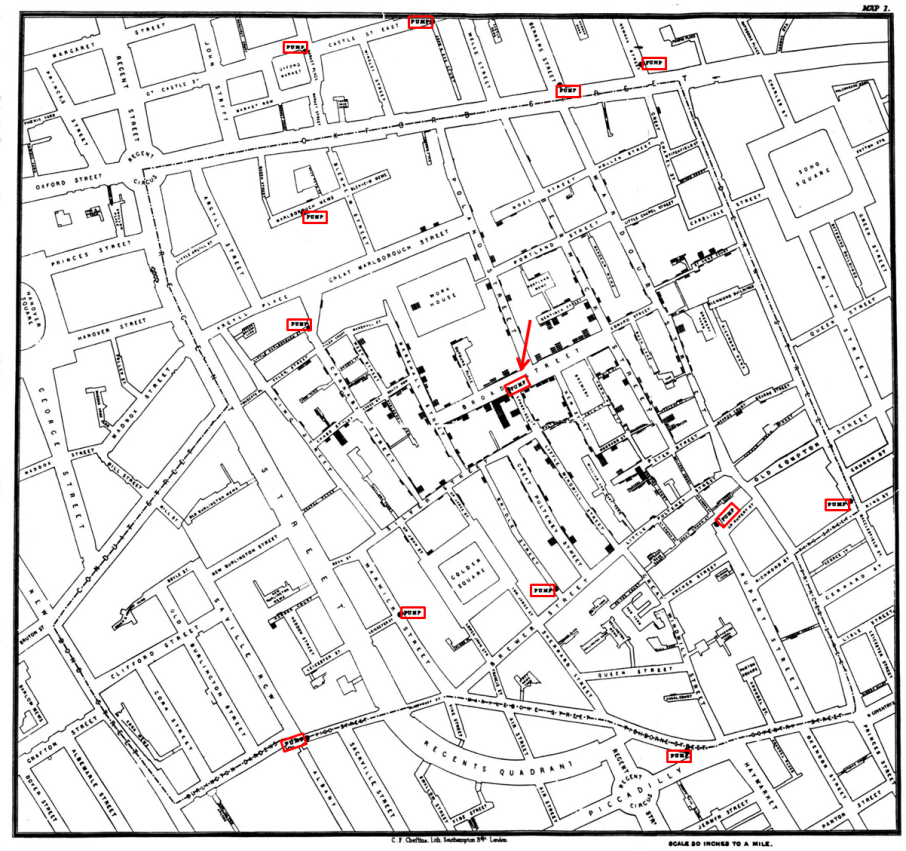



Figure 39.1: Original map of Soho-district in London made by John Snow in 1854, with 'a scale of 30 inches to a mile'. Cholera cases, from 19th August to 30th September 1854, are highlighted in black. The red rectangles have been added, in order to clearly indicate the locations of the pumps in this area. Published by C.F. Cheffins, Lith, Southhampton Buildings, London, England, 1854. On the Mode of Communication of Cholera, by John Snow [123]. Image taken from Wikimedia Commons [9]. Public Domain.

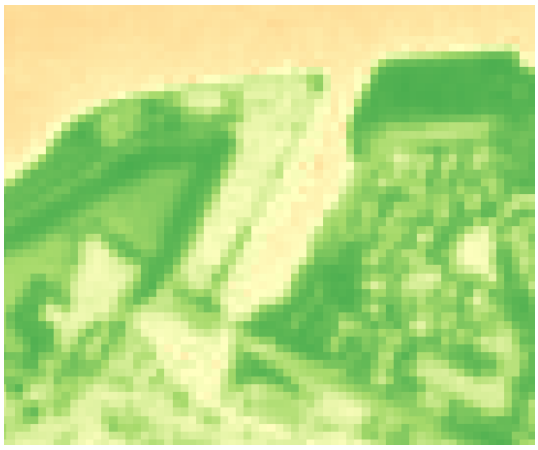


Figure 39.2: Zoom-in on raster data (of Figure 25.10): the individual pixels can be seen, representing each a 10 x 10 meter area in the terrain.

cholera in this part of London, except among the persons who were in the habit of drinking the water of the above-mentioned pump well.

Snow was right in pointing to the water of the pump as the culprit, though the sanitary mechanism by which the disease was transmitted (bacteria in the water) was not fully understood yet, at that time.

## 39.2. Vector and raster data

There are two fundamental data models for recording and storing geographical data in a GIS: vector data and raster data.

For a description in terms of *vector data*, one uses points, lines and areas to model the real world. The basic entity in vector data is a *point*, of which the position coordinate pair $(x, y)$, or the coordinate triplet $(X, Y, Z)$ in 3D, is stored. A line is represented by its start- and end-point, and an area in its turn by a series of line-segments. GPS-positioning and tachymetry with a total station naturally deliver vector data. Also stereo photogrammetric measurements result in vector data (cf. Chapter 19).

For a description in terms of *raster data*, the whole area of interest is divided in a regular grid, and a value, of for instance the average or total amount of Sun-light reflected to our sensor, in a specific optical frequency band, is stored for each cell. The cell or *pixel* is the basic geometric entity with raster data. Remote sensing naturally delivers raster data (think of satellite imagery and (digital) aerial photography), see Figure 25.1. The example of Figure 25.10, on the Normalized Difference Vegetation Index (NDVI), is also based on remote sensing data (from the Sentinel-2 satellite mission). Zooming-in a lot on this image, on the small harbor and the residential area, yields Figure 39.2, which is a bit 'blocky'; we can observe the individual pixels, which, in this case, correspond to a 10 x 10 meter foot-print on the Earth's surface.

### 39.2.1. Vector data

Identifiable objects on the Earth's surface are referred to as *features*. One can think of buildings, roads, rivers, orchards and islands. Points can be used to represent objects like a railway station or a residential house. Lines can be used to represent linear features such as roads, railways and rivers, see Figure 39.3. And areas can be used to represent parcels, lakes and forests.

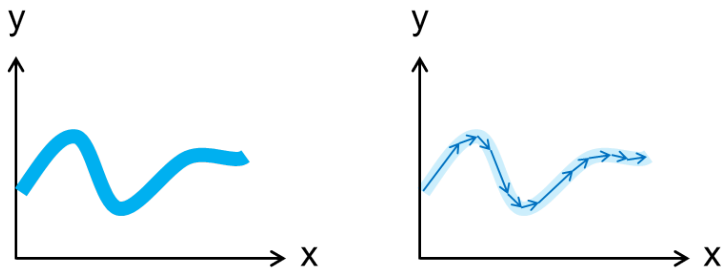



Figure 39.3: A (curved) feature on the Earth's surface, like a river, is captured (digitized) by a series of straight-line segments, delivering vector-data.

Figure 39.4: Small example of a road network stored as vector data. Both sides of the road are mapped and road segments are actually stored as closed polygons (areas) in this large-scale topographic map. On the left a 750 meter wide area of the road network is shown, while the 'zoom-in' on the right covers only 100 meter. One can clearly see from the curved road that the geometry of the road is described by a series of points, connected by straight lines. Basisregistratie Topografie (BRT) TOP10NL data by Kadaster, taken from PDOK [59] under CC BY 4.0 license.

Modeling a curved road by straight line segments implies an approximation, see Figure 39.4. The graph at left shows a part of a road network in a rural area along a highway (taken from the TOP10NL-map), and the graph at right shows a 'zoom-in'. The curved road gets slightly jagged.

With vector data the point is the basic geometric entity. A linear feature, such as a river, is described by a polyline (a series of vectors or line segments), as shown in Figure 39.3, and in principle by just the series of vertices (as start- and end-points of the line segments). Two successive vertices are connected by a so-called edge. The edges together form the polyline. When a polyline is used to describe an area, it becomes an (enclosed) polygon (and the last vertex is equal to the first one). The elementary geometric entities are shown in Figure 39.5. In three-dimensional mapping one may add the tetrahedron as a volume element, which is a polyhedron with four sides (four 'triangular faces'); it has four corners (vertices) and six straight edges. Maps on flat paper or screen are two-dimensional, though a third dimension can be added by means of perspective view. An actual three-dimensional map or model can be realized by means of a maquette.

A raster map, representing for instance terrain elevation, continously over the area of interest, can be converted into a vector map, by means of *contouring*. The value range in the raster map is divided into a set of distinct classes. Linear features are created as the boundaries between classes (such that pixels with a larger value are on one side, and pixels with a lower value on the other). The result is a map with polygons and polylines as so-called iso-lines, connecting points with equal value (e.g. elevation or amount of rainfall), see also Figure 38.9.

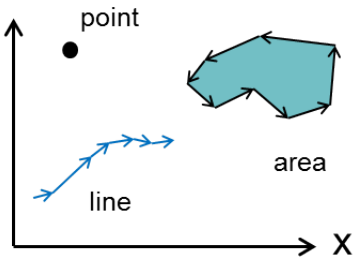



Figure 39.5: Vector based data: point, polyline, and polygon.

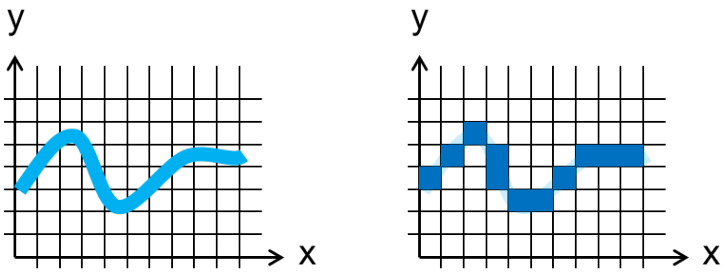



Figure 39.6: A (curved) feature on the Earth's surface, like a river, is captured (digitized) by means of a set of pixels, delivering raster-data.

### 39.2.2. Raster data

With raster data the area of interest is divided in a *regular* grid, and the data stored can be regarded as a matrix with rows and columns, and each element representing the value for that small area. The value represents the condition of that specific part of the Earth's surface and can be the amount of Sun-light reflected to our sensor by that area, or the (average) height of the terrain surface of that area. The matrix is a rectangular array of cells or pixels (short for picture elements), see Figure 39.6.

The entire area is covered, so 'there is a value everywhere'. Hence, raster data are particularly suited when the attribute of interest (e.g. elevation of the terrain surface, amount of vegetation, or concentration nitrogen dioxide $NO_2$ in the air) is continuous in space.

When original measurements are taken at discrete points (locations), for instance air temperature at meteo-stations, and you though want to have a continuous representation over the whole area of interest, you can use spatial *interpolation*, see Chapter 11. The result is then typically available as raster data, and the temperature at any location has been predicted based on measurements from neighboring meteo-stations, see also Figure 38.9, with the blue color scheme originally varying gradually.

### 39.2.3. Pros and cons

In this section we briefly review the major pros and cons of the use of vector and raster data. Typically a pro of one, is a con of the other.

With vector data we are able to maintain the original resolution (and precision) of the measurements underlying the spatial position information (resolution with which the position coordinates of the points or vertices are stored). A map based on vector data has 'infinite' resolution.

The use of vector data is particularly suited for linear features (roads, rivers) and fea-

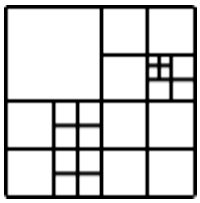


Figure 39.7: Example of a so-called (hierarchical) quadtree data structure.

tures with boundaries (parcels, orchards, lakes). Vector data allows us to correctly maintain topology, cf. Section 39.4.

Vector data allows for efficient storage of so-called sparse data. We need to store only the features present. Empty areas do not need to be stored.

The data structure of raster data is very easy to understand, and straightforward to work with. Spatial analyses are simple and easy (simple math), think for instance of overlaying different data sets / layers (just find the corresponding pixels, whereas with vector data intersections of polygons need to be computed). Similarly, raster data are easily and directly obtained from imagery, whereas vector data, obtained for instance from tachymetry, involves computations. With raster data, there is data for every part in the area covered (a full matrix); thereby raster data is particularly suited to represent so-called continuous data (for instance terrain elevation or amount of air pollution).

With raster data the resolution of spatial information, e.g. the ability to pin-point an object, is by default limited to the pixel size.

Raster data requires typically large amounts of storage space; in principle we need to store a value for every pixel, no matter whether a feature is present or not. Also, as shown in Figure 39.6, raster data deals poorly with linear features, and topological issues may arise (the river may no longer be a nice continuous feature, as river pixels may not be directly adjacent).

### 39.2.4. Raster with adaptive grid

The raster data structure copes with spatially continuous phenomena by means of a 'one-size fits all'-fixed cell-size — all grid cells have equal spatial size. This drawback can be overcome by the use of an *adaptive* grid, to say, a vario-scale cell size. The cell size is made dependent on for instance the spatial gradient of the parameter or phenomenon of interest. A smaller cell is used when there is a lot of change or variation, and a larger cell is used when 'there is not happening much'. Figure 39.7 shows an example of a so-called (hierarchical) quadtree structure.

A quadtree starts from a regular partitioning of the two-dimensional (typical horizontal) domain. Starting from a set of discrete observation points irregularly positioned, one could go for an *irregular* partioning as shown in Figure 11.3.

## 39.3. GIS structure

As a Geographic Information System (GIS) may house data sets from many different sources, the data are organized in *layers*. With raster data, different images are stored on different layers. With vector data, it is common to group features into layers. One layer then contains features of the same geometric type, and with the same kind of attributes. Figure 39.8 shows an example, with the complete, real world at bottom. There is a layer specifically for roads (containing roads of different categories, such as provincial roads and local streets), one layer for railways, another for buildings (containing houses, schools, railway stations, hospitals and churches), and one for (linear) water bodies (like rivers and canals). Together, the layers build

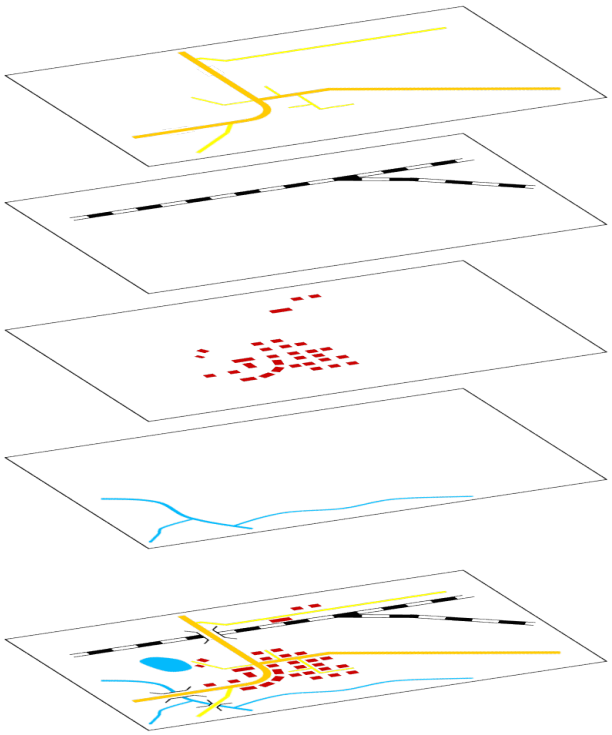


Figure 39.8: Vector data in a GIS, organized in different layers. The top layer contains roads, the second one railways, the third one buildings and the fourth one waterways. Layers with vector data are typically stored in so-called *shape* files.

| road ID | class | surface | road number | street name |
|---|---|---|---|---|
| 0001 | provincial road | tarmac | 266 | Bundesstrasse |
| 0002 | local street | brick | - | Schulstrasse |
| 0003 | local street | concrete | - | Weiherhof |
| 0004 | local street | cobblestone | - | Kirchplatz |

Table 39.1: Example of a simple attribute table, of road features.

a model of reality.

In a GIS one typically would like to store also non-geometric information, associated to the objects (features). This alpha-numerical type of information, or descriptive data, is referred to as *attributes*. The layer of roads may be associated to a table, containing the road class (highway, provincial road, local street, private road etc.), the type of surface (tarmac, concrete, gravel, or unpaved), directionality (two-way or one-way), the number of lanes, street name (if applicable), etc. Road-class, type of surface, and street name etc. are the attributes of the roads.

An example of an attribute table is shown in Table 39.1. Each row — a record — refers to a feature, and each column — a field — covers one attribute of the features.

With a GIS you can make queries on attributes in the table. For instance, show all tarmac roads. Or show all residential houses with a floor surface exceeding 250 square meter.

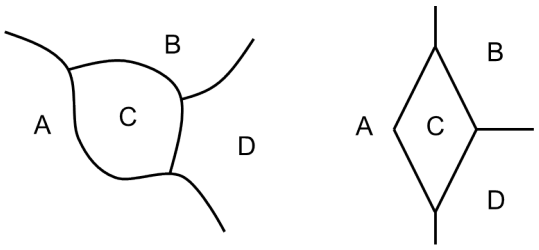



Figure 39.9: Two spatial structures which are topologically identical.

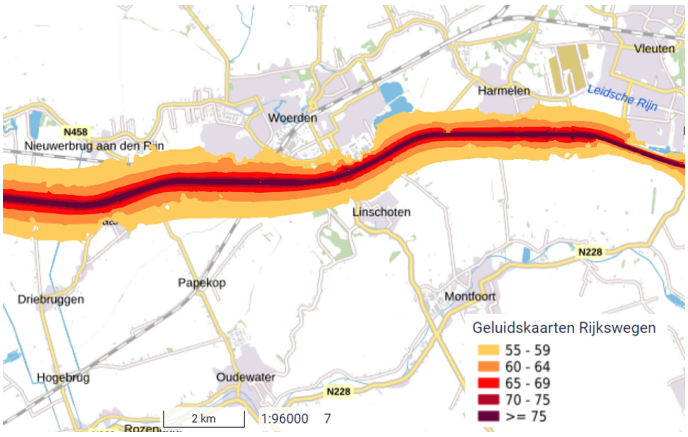



Figure 39.10: Example of a highway-noise map - 24 hour average noise in dB in 2016, around the A12 highway near Woerden. Data by Rijkswaterstaat on BRT Achtergrondkaart by Kadaster, map obtained through PDOK [59], under CC0 1.0 license, Public Domain.

## 39.4. Topology

Topology expresses the spatial relationships between vector features. Two line segments in a road network may not meet perfectly at an intersection. They are not connected, while they should be, and this poses an topological issue. Parcels as registered by the Cadastre (national land registration authority) shall not overlap (otherwise that certain piece of land would have two owners at the same time). Adjacent parcels shall have common edges. Land property boundaries of one parcel shall not run into a neighboring parcel.

Figure 39.9 shows an example of spatial relationships. Areas A and C share a common boundary, and so do B and C, and D and C. This is the case for both the model at left, and the one at right. Adjacency relations are maintained, e.g. area B being adjacent to area D.

A correct topology is crucial for network analyses (connectivity). For instance for the Rotterdam Metro map of Figure 37.4: it matters whether vertices (or nodes) in the network are linked or not.

## 39.5. Spatial analysis

In the introduction of this chapter we mentioned that a GIS is able to answers questions like 'What is located here?', and 'Where is the TU Delft Aula?', and also 'Where is the nearest ATM?'. Similarly you may want to find (or select) all gas stations in a certain area.

But with a GIS, also more complex questions can be answered. A GIS enables us to carry out spatial analyses. We may discover and research spatial patterns, similar to the analysis of Dr. John Snow. For instance, is there a relation between soil type, yearly amount of rain fall and land cover (vegetation)? This analysis you would do by making an *overlay*, i.e. by stacking

various layers on top of each other. And next, what is the best place for agricultural activities? A GIS may also answer a 'What if' type of question. For instance, suppose a dike will break, what areas will be flooded? Related to this type of analysis, are so called proximity analyses. What areas will — in terms of noise — be impacted by the construction of a new highway or railway? It comes to determining whether features (buildings) are within or outside a noise buffer zone around the new railway, see also Figure 39.10. Creating a *buffer*, for instance according to distance to the object, is a common functionality of a GIS. And, as you likely are already aware of, a GIS can help you to find the best (e.g. shortest, or quickest) route in a network. The algorithm behind this function is often based on finding the shortest path between nodes in a so-called graph, for instance a road network, and conceived by Dutch computer scientist Edsger Dijkstra in 1956 [124].

# VII

# Appendices

# A

# Error sources in land-surveying [*]

## A.1. Atmospheric refraction

Figure 3.16 illustrated the Earth's atmosphere as spherical layers around a spherical Earth, with increasing density $\rho$ the closer one gets to the surface. The density of air drives the refractive index, and for standard atmospheric conditions at sea level the refractive index, for visible light, is about $n$=1.0003, as mentioned in Section 4.3 (the refractive index of water, at 20° C, for instance is way larger, $n$=1.333). Compared to vacuum, with $n$=1, the presence of air molecules in the Earth's atmosphere cause the electromagnetic waves to slow down - in the atmosphere they travel slower than in vacuum.

Pressure $p$, volume $V$ and temperature $T$ of an ideal gas are related, see e.g. [52], as $pV = NkT$, where $k$ is the Boltzmann constant, and $N$ is the number of gas molecules. Bringing the volume to the right-hand side leads to

$$p = R\rho T$$

with $R$ the specific gas constant. This equation shows how pressure $p$ and temperature $T$ of for instance dry air, are related to density $\rho$, and thereby to the refractive index $n$. In practice we do not deal with an ideal gas. The atmospheric composition may vary (level of carbon dioxide for instance), and it may contain less or more water vapor.

From the above expression we can see that a larger pressure leads to a higher density and thereby to a larger refractive index, and a higher temperature leads to a lower density and thereby to a smaller refractive index. These factors, pressure and temperature, are the main drivers in deviations of the refractive index of $n$=1.0003 in standard conditions. Rising the pressure by 100 mbar causes a $3 \cdot 10^{-5}$ increase of the refractive index (30 ppm), and rising the temperature by 30 degrees causes a $3 \cdot 10^{-5}$ decrease of the refractive index (30 ppm), hence both changes each lead to a 3 mm effect in measuring a 100 m distance (while using the default refractive index of $n$=1.0003). Changing the relative humidity by 50% has an impact which is smaller than the given pressure and temperature examples by one order of magnitude. As a conclusion we state that changes larger than these examples are only possible high up in mountainous areas.

A lot of research has been carried out on the physical background of the refractive index and lots of models have been developed, typically based on extensive laboratory experiments and measurement campaigns. For more details the reader is referred to [125].

So far in this appendix on atmospheric refraction we were concerned with the propagation *speed* of electromagnetic waves in relation to measuring distances. In a layered medium, the *signal path* will not be a geometrically straight line, but instead be (continuously) *curved* or

| $L$ | 100 m | 1000 m | 10000 m |
|---|---|---|---|
| $c$ | 0.8 mm | 7.8 cm | 7.8 m |
| $c'$ | 0.7 mm | 6.8 cm | 6.8 m |
| $L - L'$ | 1.5 nm | 1.5 $\mu$m | 1.5 mm |
| $L - L''$ | 1.5 nm | 1.5 $\mu$m | 1.5 mm |

Table A.1: Effects of Earth's curvature and atmospheric refraction on leveling and distance measurements, computed with $R$=6378 km and $k$=0.13.

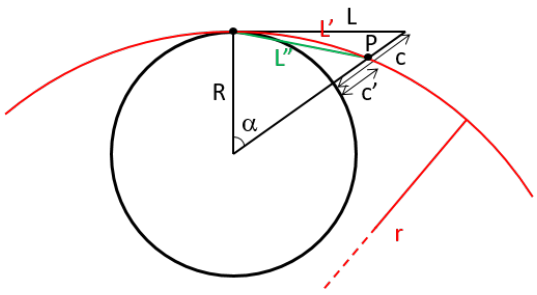



Figure A.1: The impact of Earth's curvature and atmospheric refraction on measuring distances, with the Earth shown in black. The geometric straight line distance (in green) between instrument and reflector at point P is $L''$. In practice the distance is measured along the curved path in red, resulting in distance $L'$.

*bended*, as a result of refraction. This is shown in Figure G.9, and discussed, for a horizontally layered medium, and in Figure 3.16 for a spherically layered medium.

As outlined in Section 3.4 this bending due to atmospheric refraction can be a serious effect in leveling, being expressed as $c - c'$. In measuring distances this bending-effect can be neglected in all practical circumstances, as shown in Table A.1.

$c$ shows the effect of the Earth's curvature on leveling, and $c'$ shows the combined effect of the Earth's curvature and the bending of the signal path due to a spherically layered atmosphere, as shown in Figure 3.16 at left. The effects of $c$ and $c'$ are computed according to the approximations given in Figure 3.16 at right.

For the measurement of distance the signal path is bended, due to atmospheric refraction, and the signal travels along the red curve in Figure A.1 to the reflector at point P - this is distance $L'$. In the table $L - L'$ shows by how much the distance measured along the actual (bended) signal path $L'$ is shorter than the assumed horizontal distance $L$. The straight line geometric distance from instrument to the target at P is $L''$. The difference $L' - L''$ is each time smaller by two orders of magnitude than the listed differences $L - L'$ and $L - L''$.

The conclusion is that, on account of signal path bending, the Earth's curvature and atmospheric refraction do play a role in leveling, see Sections 3.3 and 3.4, but not in (horizontal) distance measurements.

## A.2. Prism constant definition

In practice different conventions are used by different survey-equipment manufacturers in defining the prism constant $C_{\text{prism}}$, as introduced in Section 4.4. Usually, when using a brand X total station together with a brand X corner reflector, the software of the instrument (automatically) takes care of applying the correct prism constant. Problems may arise when mixing equipment from different brands, and in this appendix we review the definitions, such that

appropriate actions can be taken in practice to handle this issue correctly.

From Figure 4.14 we know that the observed one-way distance equals $d_{\text{correct}} = \frac{d_{\text{total}}}{2} = d + \frac{d_{\text{prism,air}}}{2} = d + w$, and refers to apparent reflection point So. In order to make the observed distance refer to the defined center of symmetry of the reflector, point Sc, one has to subtract the prism constant $C_{\text{prism}}$. It is custom practice, for manufacturers like Nikon, Pentax, Sokkia, Topcon and Trimble (with Geodimeter and Spectra Precision), to report actually the negative of the prism constant, hence in documents and manuals you will find the value for $-C_{\text{prism}}$. The only exception is Zeiss (now also part of Trimble) which directly reports the value for $C_{\text{prism}}$.

In general the standard prisms of these manufactures have a prism constant of $C_{\text{prism}}$ = 30 mm. The diameter of the frontal face of these prisms is 60 mm.

Leica defines the prism constant differently. The prism constant Leica is using is defined with reference to its standard round prism type (GPH1 and GPR1). These prisms have an absolute or true constant of $C_{\text{prism},o}$ = 34.4 mm. This is the standard reference value in Leica total stations. All other prism constants are *relative* to this reference value, hence $C_{\text{prism,rel}} = C_{\text{prism}} - C_{\text{prism},o}$. They work with a relative prism constant $C_{\text{prism,rel}}$. Obviously, when their standard prism is used, the relative prism constant equals zero $C_{\text{prism,rel}}$ = 0.

When for example a Topcon prism is used, with a given $-C_{\text{prism}}$ = -30 mm (mind reporting the negative of the prism constant), in combination with a Leica total station, then the correct input value for the (relative) prism constant to the Leica total station is $-C_{\text{prism,rel}}$ = –30 + 34.4 = +4.4 mm.

Or the other way around, when a Leica 360-degree prism type GRZ121 is used with a (relative) prism constant, as specified by the manufacturer, of $-C_{\text{prism,rel}}$ = 23.1 mm, in conjunction with a non-Leica total station, then the to be entered prism constant is: $-C_{\text{prism}} = -C_{\text{prism,rel}} - C_{\text{prism},o}$, hence $-C_{\text{prism}}$ = 23.1 - 34.4 = -11.3 mm.

# B

# Several mathematical proofs [*]

## B.1. Mean and variance of normal distribution

The Probability Density Function (PDF) of a normally distributed random variable $\underline{y}$ reads

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-x)^2}{2\sigma^2}} \tag{B.1}$$

and in this section we prove that $x$ is the mean, and $\sigma$ the standard deviation ($\sigma^2$ the variance).

Using (6.9) the mean reads

$$E(\underline{y}) = \int_{-\infty}^{+\infty} y \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{(y-x)}{\sigma})^2} dy \tag{B.2}$$

which can be split into

$$E(\underline{y}) = x \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{(y-x)}{\sigma})^2} dy + \int_{-\infty}^{+\infty} \frac{(y-x)}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{(y-x)}{\sigma})^2} dy$$

The first part yields just the mean $x$, and in the second part we apply the change of variable $z = \frac{y-x}{\sigma}$

$$E(\underline{y}) = x + \int_{-\infty}^{+\infty} \sigma \frac{z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

where the factor $\sigma$ enters in the second term as the change of variable implies $\sigma dz = dy$.

$$E(\underline{y}) = x + \frac{\sigma}{\sqrt{2\pi}} \left[ -e^{-\frac{z^2}{2}} \right]_{-\infty}^{\infty} = x$$

as the second term yields zero. The mean of the random variable $\underline{y}$ reads $x$.

Using (6.10) the variance reads

$$D(\underline{y}) = \int_{-\infty}^{+\infty} (y-x)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{(y-x)}{\sigma})^2} dy \tag{B.3}$$

where we used already $E(\underline{y}) = x$. Again we apply the change of variable $z = \frac{y-x}{\sigma}$

$$D(\underline{y}) = \int_{-\infty}^{+\infty} \sigma^2 \frac{z^2}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

which can be split into

$$D(\underline{y}) = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (ze^{-\frac{z^2}{2}})(z)\, dz$$

where we apply integration by parts

$$D(\underline{y}) = \frac{\sigma^2}{\sqrt{2\pi}} \left( -\left[ ze^{-\frac{z^2}{2}} \right]_{-\infty}^{\infty} + \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}}\, dz \right) = \sigma^2$$

where we used the definite integral $\int_0^{\infty} e^{-\frac{z^2}{(\sqrt{2})^2}}\, dz = \frac{\sqrt{2\pi}}{2}$. The variance of the random variable $\underline{y}$ reads $\sigma^2$.

## B.2. Mean and variance propagation laws

We consider the following linear transformation, as given by (7.7), of $m$-vector $\underline{y}$

$$\underline{v} = R\underline{y} + s$$

where vector $\underline{v}$ has $n$ elements, and consequently matrix $R$ has $n$ rows and $m$ columns, and vector $s$ is an $n$-vector.

In this section we prove that $E(\underline{v}) = RE(\underline{y}) + s$ (7.8) and that $Q_{vv} = RQ_{yy}R^T$ (7.9).

Using

$$E(G(\underline{y})) = \int_{-\infty}^{\infty} G(y) f(y)\, dy$$

we compute the mean of $\underline{v}$

$$E(\underline{v}) = E(R\underline{y} + s) = \int_{-\infty}^{\infty} (Ry + s) f(y)\, dy$$

$$= R \int_{-\infty}^{\infty} y f(y)\, dy + s \int_{-\infty}^{\infty} f(y)\, dy = RE(\underline{y}) + s$$

which proves (7.8) in Section 7.2.

With (7.4) applied to vector $\underline{v}$

$$D(\underline{v}) = E((\underline{v} - E(\underline{v}))(\underline{v} - E(\underline{v}))^T) = E((R\underline{y} + s - RE(\underline{y}) - s)(R\underline{y} + s - RE(\underline{y}) - s)^T)$$

$$= RE((\underline{y} - E(\underline{y}))(\underline{y} - E(\underline{y}))^T)R^T = RD(\underline{y})R^T = RQ_{yy}R^T = Q_{vv}$$

which proves (7.9).

## B.3. Non-linear mean and variance propagation laws

In Section 7.4 the mean and variance propagation laws were given for the non-linear transformation $\underline{v} = G(\underline{y})$ (7.10), where $G()$ represents $n$ non-linear functions of each time $m$ random variables (the elements of vector $\underline{y}$).

The Taylor series, up to second order term, of one of the $n$ non-linear functions $\underline{v}_i = G_i(\underline{y})$ with $i = 1, \ldots, n$, at $E(\underline{y})$, reads

$$G_i(\underline{y}) \approx G_i(E(\underline{y})) + \left.\frac{\partial G_i}{\partial y^T}\right|_{E(\underline{y})} (\underline{y} - E(\underline{y})) + \frac{1}{2}(\underline{y} - E(\underline{y}))^T \left.\frac{\partial^2 G_i}{\partial y y^T}\right|_{E(\underline{y})} (\underline{y} - E(\underline{y}))$$

where $\frac{\partial G_i(y)}{\partial y}$ is the $m \times 1$ gradient vector of $G_i(y)$, and $\frac{\partial^2 G_i(y)}{\partial y y^T}$ is the $m \times m$ Hessian matrix of $G_i(y)$ (matrix with second order partial derivatives).

The expectation of $\underline{v}_i$ in terms of $\underline{y}$ becomes

$$E(\underline{v}_i) = E(G_i(\underline{y})) \approx E(G_i(E(\underline{y}))) + E(\frac{\partial G_i}{\partial y^T}\Big|_{E(\underline{y})} (\underline{y} - E(\underline{y}))) + E(\frac{1}{2}(\underline{y} - E(\underline{y}))^T \frac{\partial^2 G_i}{\partial y y^T}\Big|_{E(\underline{y})} (\underline{y} - E(\underline{y})))$$

$$= G_i(E(\underline{y})) + \frac{1}{2}\text{trace}(\frac{\partial^2 G_i}{\partial y y^T}\Big|_{E(\underline{y})} E((\underline{y} - E(\underline{y}))(\underline{y} - E(\underline{y}))^T))$$

$$= G_i(E(\underline{y})) + \frac{1}{2}\text{trace}(\frac{\partial^2 G_i}{\partial y y^T}\Big|_{E(\underline{y})} Q_{yy})$$

as $E(\underline{y} - E(\underline{y})) = E(\underline{y}) - E(E(\underline{y})) = E(\underline{y}) - E(\underline{y}) = 0$, and where trace means taking the sum of the diagonal elements of the matrix, and, we used definitions (7.4) and (7.5) for $Q_{yy}$. This completes the proof of (7.11).

The $n \times n$ variance matrix of $\underline{v}$ is formally given by

$$Q_{vv} = E((\underline{v} - E(\underline{v}))(\underline{v} - E(\underline{v}))^T)$$

see (7.4) and (7.5). We consider one element $(i, j)$, which is $Q_{v_i v_j}$

$$Q_{v_i v_j} = E((\underline{v}_i - E(\underline{v}_i))(\underline{v}_j - E(\underline{v}_j))^T)$$

with $i = 1, \ldots, n$ and $j = 1, \ldots, n$. We supply the above Taylor series, up to second order term, of $\underline{v}_i = G_i(\underline{y})$, and of $\underline{v}_j = G_j(\underline{y})$, and we substitute the just obtained approximation for $E(\underline{v}_i)$, and for $E(\underline{v}_j)$. In the expansion we neglect all terms with cross products of first and second order derivatives of $G(y)$, as well as products of second order derivatives, and we are left with just one term, namely

$$Q_{v_i v_j} \approx E(\frac{\partial G_i}{\partial y^T}\Big|_{E(\underline{y})} (\underline{y} - E(\underline{y}))(\underline{y} - E(\underline{y}))^T \frac{\partial G_j}{\partial y^T}\Big|_{E(\underline{y})}^T)$$

$$= \frac{\partial G_i}{\partial y^T}\Big|_{E(\underline{y})} E((\underline{y} - E(\underline{y}))(\underline{y} - E(\underline{y}))^T) \frac{\partial G_j}{\partial y^T}\Big|_{E(\underline{y})}^T = \frac{\partial G_i}{\partial y^T}\Big|_{E(\underline{y})} Q_{yy} \frac{\partial G_j}{\partial y^T}\Big|_{E(\underline{y})}^T$$

where we used again definition (7.4). Finally the above expression can be used for any element $Q_{v_i v_j}$, hence, the full $n \times n$ variance matrix $Q_{vv}$ results as

$$Q_{vv} \approx \frac{\partial G}{\partial y^T}\Big|_{E(\underline{y})} Q_{yy} \frac{\partial G}{\partial y^T}\Big|_{E(\underline{y})}^T$$

This completes the proof of (7.12). The term $\frac{\partial G(y)}{\partial y^T}$ is an $n \times m$ matrix, containing, as rows, the gradient vectors of non-linear functions $G_i(y)$, with $i = 1, \ldots, n$, all evaluated at $E(\underline{y})$.

## B.4. Least-squares

In this section we prove that $\hat{x} = (A^T A)^{-1} A^T y$ (8.4) is the solution to $\min_x \|y - Ax\|^2$ (8.5).

For the least-squares solution we minimize, for $x$, the function $g(x) = \|y - Ax\|^2 = (y - Ax)^T (y - Ax)$. The gradient, the first order derivative, is set equal to zero, and the Hessian, the second order derivative, should be larger than zero (be a positive definite matrix), in order for the found (single) extremum to be a global minimizer.

$$g(x) = y^T y - y^T Ax - x^T A^T y + x^T A^T Ax = y^T y - 2x^T A^T y + x^T A^T Ax$$

as the innerproduct $y^T b = b^T y$. Setting $\frac{\partial g(x)}{\partial x} = 0$ yields

$$-2A^T y + 2A^T Ax = 0$$

where we used $\frac{\partial (x^T b)}{\partial x} = b$ and $\frac{\partial (x^T Mx)}{\partial x} = (M + M^T)x$ (which both can be verified by expansion), and recognize that $A^T A$ is a symmetric $n \times n$ matrix. Solving this for $x$ yields

$$A^T A\hat{x} = A^T y$$

which is referred to as the system of normal equations (it is a system with $n$ unknowns in $n$ equations), and inverting matrix $A^T A$ yields indeed (8.4). The Hessian reads

$$\frac{\partial^2 g(x)}{\partial x x^T} = 2A^T A > 0$$

which is indeed a positive definite matrix, and where we used $\frac{\partial (b^T x)}{\partial x^T} = b^T$.

## B.5. Concerns on non-linear estimation

For model (8.8) in Section 8.4, application of the least-squares criterion (8.5), taking into account the variance matrix $Q_{yy}$, yields

$$\underline{\hat{x}}' = \arg \min_{x \in R^n} \|y - F(x)\|^2_{Q_{yy}^{-1}}$$

the non-linear estimator $\underline{\hat{x}}' = G(\underline{y})$, which in most cases in practice can not be found in an analytical form. Also we mention that non-linear estimation is not a trivial subject. The propagation of the random characteristics of $\underline{y}$ into those of $\underline{\hat{x}}'$ is difficult. For instance the mean and variance can not be propagated through a non-linear relation in a straightforward way, e.g. $E(\underline{\hat{x}}') = E(G(\underline{y})) \neq G(E(\underline{y}))$, the non-linear estimator is biased $E(\underline{\hat{x}}') \neq x$.

Through the iterative procedure in Section 8.4 the estimate $\hat{x}$ is a numerical approximation of the realization of the non-linear estimator $\underline{\hat{x}}'$.

## B.6. Line-fitting with observed independent variable

Equation (8.2) $E(\underline{y}) = Ax; D(\underline{y}) = Q_{yy}$ presented the (linear) model of observation equations, with observations $y$, unknown parameters $x$, and known coefficients in $m$x$n$-matrix $A$. Example (10.2) was on line fitting, and positions were observed, with error, but timing, the independent variable, was assumed to be perfect (error-free), yielding the coefficients of matrix $A$.

If this assumption (on the independent variable) can*not* be made, the model for line fitting has to be set up differently. As an example we consider the measurement of the expansion of a steel bar due to a rise in temperature. Both the dependent variable (the length, observed,

in vector $y$; sometimes called the response), and the independent variable (temperature, occuring in matrix $A$), are now subject to observational error.

The length of the bar is measured $m$ times: $y_1, y_2, \ldots, y_m$ (and inevitably measurement errors are being made), and, at the same time also the (corresponding) temperature is *measured*: $T_1, T_2, \ldots, T_m$ (and also here inevitably with some error). The functional model (in absence of any error) reads $y_i = T_i x_1 + x_2$, for $i = 1, \ldots, m$, where $x_2$ is the length of the bar at reference temperature (e.g. $T = 0$), and $x_1$ is the coefficient of thermal expansion (in this case the increase in length per degree, for instance for steel approximately 10 $\mu$m per degree, for a 1 meter bar).

Trying to formulate the model as before yields

$$E\begin{pmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_m \end{pmatrix} = \underbrace{\begin{pmatrix} T_1 & 1 \\ T_2 & 1 \\ \vdots & \vdots \\ T_m & 1 \end{pmatrix}}_{A} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

but, this model now has observational error also in the independent variable $T_i$, and is referred to as an Error-in-Variables (EIV) model. This is an example of a so-called *total least squares* problem. There exist various approaches to solve such a problem, e.g. using singular value decomposition. In the sequel we describe another way.

Earlier, only $y$ was a random vector, namely $\underline{y}$, and $T$ was not, but now also $T$ is a random vector, namely $\underline{T}$, and with keeping just $x_1$ and $x_2$ as unknown parameters, we would have random variables in the design matrix $A$. Therefore the model is set up differently, namely by introducing also the temperatures as unknown parameters, next to $x_1$ and $x_2$. The actual temperatures are not known — they have been measured, but with error. The full model of observation equations becomes:

$$\begin{array}{l} E(\underline{T}_i) = T_i \\ E(\underline{y}_i) = T_i x_1 + x_2 \end{array} \quad \text{for} \quad i = 1, 2, \ldots, m \qquad ; \; D\begin{pmatrix} \underline{T} \\ \underline{y} \end{pmatrix} = \begin{pmatrix} Q_{TT} & 0 \\ 0 & Q_{yy} \end{pmatrix}$$

with in total $2m$ observations (which is exactly all we measured), $m$ temperatures and $m$ lengths, and $m + 2$ unknown parameters, namely $x_1$, $x_2$, and $T_1, T_2, \ldots, T_m$. The total $2m$ x $2m$ variance matrix will typically be a block-diagonal matrix, as there is generally no correlation between temperature and length readings.

The above extended model of observation equations is *non-linear*, notably by the product of $T_i$ and $x_1$, cf. Section 8.4 on linearization and iteration.

Figure B.1 shows an example *without* (at left), and *with* (at right) observational error in the independent variable. In the former case, the least-squares residuals (indicated by the dashed lines) are exactly along the vertical axis, whereas in the latter case both observed values $T_i$ and $y_i$ get 'corrected'.

## B.7. Ordinary Kriging

In this section we prove the key equation for Ordinary Kriging (11.9).

The requirement of *linear* interpolation is already implied in (11.2), $\hat{\underline{z}}_0 = w^T \underline{y}$.

The interpolation is also required to be *unbiased*. With just a constant unknown mean (11.7), with $A = l$, yields $E(\underline{y}) = lx$, where $l$ is a vector of all ones. Then with (11.2) and (7.8) $E(\hat{\underline{z}}_0) = w^T E(\underline{y}) = w^T lx$. Now with (11.8) for $\underline{z}(p_0) = \underline{z}_0$ we know that $E(\underline{z}_0) = x$, hence $E(\hat{\underline{z}}_0) = E(\underline{z}_0)$ yields the constraint $w^T l = 1$, or $\sum_{i=1}^{m} w_i = 1$.

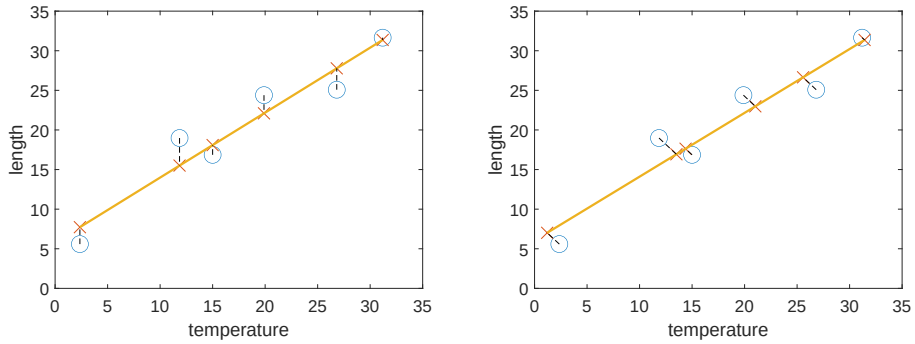



Figure B.1: Line fitting: the circles show the (same) original observations, pairs $(T_i, y_i)$, on the left the (ordinary) least-squares solution when temperature is error-free, and at right, when both length and temperature measurement are subject to observational error, and the alternative (non-linear) approach with temperatures as unknown parameters is used. The crosses show pairs $(T_i, \hat{y}_i)$ at left, for the line $\hat{x}_1$=0.82 and $\hat{x}_2$=5.79, and pairs $(\hat{T}_{i,NL}, \hat{y}_{i,NL})$ at right, for the line $\hat{x}_{1,NL}$=0.81 and $\hat{x}_{2,NL}$=6.05; crosses are all on the fitted line in both cases.

Eventually, to achieve the *best* interpolator $\underline{\hat{z}}_0$, we require minimum error variance with the error as $\underline{\hat{\epsilon}} = \underline{z}_0 - \underline{\hat{z}}_0$. Variance $\sigma^2_{\hat{\epsilon}}$ should be as small as possible, implying the largest probability on a small error at any position, and practically, the interpolated value being as close as possible to the actual water-depth. Hence, the goal is to determine the elements of vector $w$, such that

$$\min_w \sigma^2_{\hat{\epsilon}} \quad \text{subject to} \quad w^T l - 1 = 0 \tag{B.4}$$

The error variance is obtained through noting, with (11.2), that

$$\underline{\hat{\epsilon}} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \underline{z}_0 \\ \underline{\hat{z}}_0 \end{pmatrix} = \begin{pmatrix} 1 & -w^T \end{pmatrix} \begin{pmatrix} \underline{z}_0 \\ \underline{y} \end{pmatrix}$$

with the variance matrix of the vector on the right hand side as

$$\begin{pmatrix} Q_{z_0 z_0} & Q_{z_0 y} \\ Q_{y z_0} & Q_{yy} \end{pmatrix}$$

and $Q_{z_0 z_0} = \sigma^2_{z_0}$. Applying the variance propagation law (7.9) yields

$$\sigma^2_{\hat{\epsilon}} = \sigma^2_{z_0} + w^T Q_{yy} w - Q_{z_0 y} w - w^T Q_{y z_0} = \sigma^2_z + w^T Q_{yy} w - 2 w^T Q_{y z_0} \tag{B.5}$$

and the error variance depends on the weights in vector $w$.

The minimization (B.4), including the constraint, is solved using the Lagrange multiplier rule. The Lagrange function becomes

$$L(w, \lambda) = \sigma^2_{\hat{\epsilon}} + \lambda(w^T l - 1) = \sigma^2_z + w^T Q_{yy} w - 2 w^T Q_{y z_0} + \lambda(w^T l - 1)$$

where $\lambda$ is the Lagrange multiplier, and (B.5) has been substituted. Setting the partial derivatives of $L(w, \lambda)$ to zero yields

$$\frac{\partial L(w, \lambda)}{\partial w} = 2 Q_{yy} w - 2 Q_{y z_0} + \lambda l = 0$$

which is an $m$-vector, and with $Q_{yy}$ a symmetric matrix, and

$$\frac{\partial L(w, \lambda)}{\partial \lambda} = w^T l - 1 = 0$$

These $m+1$ equations, with $m+1$ unknowns can be cast in

$$\begin{pmatrix} Q_{yy} & l \\ l^T & 0 \end{pmatrix} \begin{pmatrix} w \\ \nu \end{pmatrix} = \begin{pmatrix} Q_{yz_0} \\ 1 \end{pmatrix} \tag{B.6}$$

with $\nu = \frac{\lambda}{2}$, see (11.9), from which the weights can be obtained through (11.10).

The inverse in (11.10) can be shown to read

$$\begin{pmatrix} Q_{yy} & l \\ l^T & 0 \end{pmatrix}^{-1} = \begin{pmatrix} Q_{yy}^{-1} - Q_{yy}^{-1} l (l^T Q_{yy}^{-1} l)^{-1} l^T Q_{yy}^{-1} & Q_{yy}^{-1} l (l^T Q_{yy}^{-1} l)^{-1} \\ (l^T Q_{yy}^{-1} l)^{-1} l^T Q_{yy}^{-1} & -(l^T Q_{yy}^{-1} l)^{-1} \end{pmatrix}$$

Once the values for vector $w$ have been obtained, the interpolation error variance can be evaluated using (B.5).

# C

# Normal distribution: table

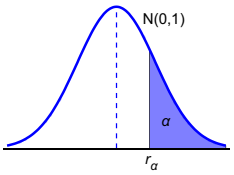


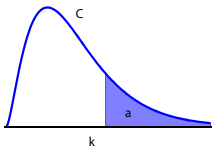

Figure C.1: Standard normal distribution $\mathrm{N}(0,1)$: one-sided level of significance $\alpha$ as function of the critical value $r_\alpha$, i.e. $\alpha = 1 - \Phi(r_\alpha)$.

| $r_\alpha$ | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.5000 | 0.4960 | 0.4920 | 0.4880 | 0.4840 | 0.4801 | 0.4761 | 0.4721 | 0.4681 | 0.4641 |
| 0.1 | 0.4602 | 0.4562 | 0.4522 | 0.4483 | 0.4443 | 0.4404 | 0.4364 | 0.4325 | 0.4286 | 0.4247 |
| 0.2 | 0.4207 | 0.4168 | 0.4129 | 0.4090 | 0.4052 | 0.4013 | 0.3974 | 0.3936 | 0.3897 | 0.3859 |
| 0.3 | 0.3821 | 0.3783 | 0.3745 | 0.3707 | 0.3669 | 0.3632 | 0.3594 | 0.3557 | 0.3520 | 0.3483 |
| 0.4 | 0.3446 | 0.3409 | 0.3372 | 0.3336 | 0.3300 | 0.3264 | 0.3228 | 0.3192 | 0.3156 | 0.3121 |
| 0.5 | 0.3085 | 0.3050 | 0.3015 | 0.2981 | 0.2946 | 0.2912 | 0.2877 | 0.2843 | 0.2810 | 0.2776 |
| 0.6 | 0.2743 | 0.2709 | 0.2676 | 0.2643 | 0.2611 | 0.2578 | 0.2546 | 0.2514 | 0.2483 | 0.2451 |
| 0.7 | 0.2420 | 0.2389 | 0.2358 | 0.2327 | 0.2296 | 0.2266 | 0.2236 | 0.2206 | 0.2177 | 0.2148 |
| 0.8 | 0.2119 | 0.2090 | 0.2061 | 0.2033 | 0.2005 | 0.1977 | 0.1949 | 0.1922 | 0.1894 | 0.1867 |
| 0.9 | 0.1841 | 0.1814 | 0.1788 | 0.1762 | 0.1736 | 0.1711 | 0.1685 | 0.1660 | 0.1635 | 0.1611 |
| 1.0 | 0.1587 | 0.1562 | 0.1539 | 0.1515 | 0.1492 | 0.1469 | 0.1446 | 0.1423 | 0.1401 | 0.1379 |
| 1.1 | 0.1357 | 0.1335 | 0.1314 | 0.1292 | 0.1271 | 0.1251 | 0.1230 | 0.1210 | 0.1190 | 0.1170 |
| 1.2 | 0.1151 | 0.1131 | 0.1112 | 0.1093 | 0.1075 | 0.1056 | 0.1038 | 0.1020 | 0.1003 | 0.0985 |
| 1.3 | 0.0968 | 0.0951 | 0.0934 | 0.0918 | 0.0901 | 0.0885 | 0.0869 | 0.0853 | 0.0838 | 0.0823 |
| 1.4 | 0.0808 | 0.0793 | 0.0778 | 0.0764 | 0.0749 | 0.0735 | 0.0721 | 0.0708 | 0.0694 | 0.0681 |
| 1.5 | 0.0668 | 0.0655 | 0.0643 | 0.0630 | 0.0618 | 0.0606 | 0.0594 | 0.0582 | 0.0571 | 0.0559 |
| 1.6 | 0.0548 | 0.0537 | 0.0526 | 0.0516 | 0.0505 | 0.0495 | 0.0485 | 0.0475 | 0.0465 | 0.0455 |
| 1.7 | 0.0446 | 0.0436 | 0.0427 | 0.0418 | 0.0409 | 0.0401 | 0.0392 | 0.0384 | 0.0375 | 0.0367 |
| 1.8 | 0.0359 | 0.0351 | 0.0344 | 0.0336 | 0.0329 | 0.0322 | 0.0314 | 0.0307 | 0.0301 | 0.0294 |
| 1.9 | 0.0287 | 0.0281 | 0.0274 | 0.0268 | 0.0262 | 0.0256 | 0.0250 | 0.0244 | 0.0239 | 0.0233 |
| 2.0 | 0.0228 | 0.0222 | 0.0217 | 0.0212 | 0.0207 | 0.0202 | 0.0197 | 0.0192 | 0.0188 | 0.0183 |
| 2.1 | 0.0179 | 0.0174 | 0.0170 | 0.0166 | 0.0162 | 0.0158 | 0.0154 | 0.0150 | 0.0146 | 0.0143 |
| 2.2 | 0.0139 | 0.0136 | 0.0132 | 0.0129 | 0.0125 | 0.0122 | 0.0119 | 0.0116 | 0.0113 | 0.0110 |
| 2.3 | 0.0107 | 0.0104 | 0.0102 | 0.0099 | 0.0096 | 0.0094 | 0.0091 | 0.0089 | 0.0087 | 0.0084 |
| 2.4 | 0.0082 | 0.0080 | 0.0078 | 0.0075 | 0.0073 | 0.0071 | 0.0069 | 0.0068 | 0.0066 | 0.0064 |
| 2.5 | 0.0062 | 0.0060 | 0.0059 | 0.0057 | 0.0055 | 0.0054 | 0.0052 | 0.0051 | 0.0049 | 0.0048 |
| 2.6 | 0.0047 | 0.0045 | 0.0044 | 0.0043 | 0.0041 | 0.0040 | 0.0039 | 0.0038 | 0.0037 | 0.0036 |
| 2.7 | 0.0035 | 0.0034 | 0.0033 | 0.0032 | 0.0031 | 0.0030 | 0.0029 | 0.0028 | 0.0027 | 0.0026 |
| 2.8 | 0.0026 | 0.0025 | 0.0024 | 0.0023 | 0.0023 | 0.0022 | 0.0021 | 0.0021 | 0.0020 | 0.0019 |
| 2.9 | 0.0019 | 0.0018 | 0.0018 | 0.0017 | 0.0016 | 0.0016 | 0.0015 | 0.0015 | 0.0014 | 0.0014 |
| 3.0 | 0.0013 | 0.0013 | 0.0013 | 0.0012 | 0.0012 | 0.0011 | 0.0011 | 0.0011 | 0.0010 | 0.0010 |
| 3.1 | 0.0010 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0007 | 0.0007 |
| 3.2 | 0.0007 | 0.0007 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0005 | 0.0005 | 0.0005 |
| 3.3 | 0.0005 | 0.0005 | 0.0005 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0003 |
| 3.4 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0002 |

Table C.1: Standard normal distribution $\mathrm{N}(0,1)$: one-sided level of significance $\alpha$ as function of the critical value $r_\alpha$, i.e. $\alpha = 1 - \Phi(r_\alpha)$. Values of $r_\alpha$ are given up to the first decimal in the first column, the second decimal in the first row. Example: $\alpha = 0.0250$ for $r_\alpha = 1.96$.

# D

# Chi-squared distribution: table



| $\alpha$ | 0.9990 | 0.9950 | 0.9900 | 0.9750 | 0.9500 | 0.9000 | 0.7500 | 0.5000 | 0.2500 | 0.1000 | 0.0500 | 0.0250 | 0.0100 | 0.0050 | 0.0010 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | | | | | | | | | | | | | | | |
| 1 | 0.0000 | 0.0000 | 0.0002 | 0.0010 | 0.0039 | 0.0158 | 0.1015 | 0.4549 | 1.3233 | 2.7055 | 3.8415 | 5.0239 | 6.6349 | 7.8794 | 10.8276 |
| 2 | 0.0020 | 0.0100 | 0.0201 | 0.0506 | 0.1026 | 0.2107 | 0.5754 | 1.3863 | 2.7726 | 4.6052 | 5.9915 | 7.3778 | 9.2103 | 10.5966 | 13.8155 |
| 3 | 0.0243 | 0.0717 | 0.1148 | 0.2158 | 0.3518 | 0.5844 | 1.2125 | 2.3660 | 4.1083 | 6.2514 | 7.8147 | 9.3484 | 11.3449 | 12.8382 | 16.2662 |
| 4 | 0.0908 | 0.2070 | 0.2971 | 0.4844 | 0.7107 | 1.0636 | 1.9226 | 3.3567 | 5.3853 | 7.7794 | 9.4877 | 11.1433 | 13.2767 | 14.8603 | 18.4668 |
| 5 | 0.2102 | 0.4117 | 0.5543 | 0.8312 | 1.1455 | 1.6103 | 2.6746 | 4.3515 | 6.6257 | 9.2364 | 11.0705 | 12.8325 | 15.0863 | 16.7496 | 20.5150 |
| 6 | 0.3811 | 0.6757 | 0.8721 | 1.2373 | 1.6354 | 2.2041 | 3.4546 | 5.3481 | 7.8408 | 10.6446 | 12.5916 | 14.4494 | 16.8119 | 18.5476 | 22.4577 |
| 7 | 0.5985 | 0.9893 | 1.2390 | 1.6899 | 2.1673 | 2.8331 | 4.2549 | 6.3458 | 9.0371 | 12.0170 | 14.0671 | 16.0128 | 18.4753 | 20.2777 | 24.3219 |
| 8 | 0.8571 | 1.3444 | 1.6465 | 2.1797 | 2.7326 | 3.4895 | 5.0706 | 7.3441 | 10.2189 | 13.3616 | 15.5073 | 17.5345 | 20.0902 | 21.9550 | 26.1245 |
| 9 | 1.1519 | 1.7349 | 2.0879 | 2.7004 | 3.3251 | 4.1682 | 5.8988 | 8.3428 | 11.3888 | 14.6837 | 16.9190 | 19.0228 | 21.6660 | 23.5894 | 27.8772 |
| 10 | 1.4787 | 2.1559 | 2.5582 | 3.2470 | 3.9403 | 4.8652 | 6.7372 | 9.3418 | 12.5489 | 15.9872 | 18.3070 | 20.4832 | 23.2093 | 25.1882 | 29.5883 |
| 11 | 1.8339 | 2.6032 | 3.0535 | 3.8157 | 4.5748 | 5.5778 | 7.5841 | 10.3410 | 13.7007 | 17.2750 | 19.6751 | 21.9200 | 24.7250 | 26.7568 | 31.2641 |
| 12 | 2.2142 | 3.0738 | 3.5706 | 4.4038 | 5.2260 | 6.3038 | 8.4384 | 11.3403 | 14.8454 | 18.5493 | 21.0261 | 23.3367 | 26.2170 | 28.2995 | 32.9095 |
| 13 | 2.6172 | 3.5650 | 4.1069 | 5.0088 | 5.8919 | 7.0415 | 9.2991 | 12.3398 | 15.9839 | 19.8119 | 22.3620 | 24.7356 | 27.6882 | 29.8195 | 34.5282 |
| 14 | 3.0407 | 4.0747 | 4.6604 | 5.6287 | 6.5706 | 7.7895 | 10.1653 | 13.3393 | 17.1169 | 21.0641 | 23.6848 | 26.1189 | 29.1412 | 31.3193 | 36.1233 |
| 15 | 3.4827 | 4.6009 | 5.2293 | 6.2621 | 7.2609 | 8.5468 | 11.0365 | 14.3389 | 18.2451 | 22.3071 | 24.9958 | 27.4884 | 30.5779 | 32.8013 | 37.6973 |
| 16 | 3.9416 | 5.1422 | 5.8122 | 6.9077 | 7.9616 | 9.3122 | 11.9122 | 15.3385 | 19.3689 | 23.5418 | 26.2962 | 28.8454 | 31.9999 | 34.2672 | 39.2524 |
| 17 | 4.4161 | 5.6972 | 6.4078 | 7.5642 | 8.6718 | 10.0852 | 12.7919 | 16.3382 | 20.4887 | 24.7690 | 27.5871 | 30.1910 | 33.4087 | 35.7185 | 40.7902 |
| 18 | 4.9048 | 6.2648 | 7.0149 | 8.2307 | 9.3905 | 10.8649 | 13.6753 | 17.3379 | 21.6049 | 25.9894 | 28.8693 | 31.5264 | 34.8053 | 37.1565 | 42.3124 |
| 19 | 5.4068 | 6.8440 | 7.6327 | 8.9065 | 10.1170 | 11.6509 | 14.5620 | 18.3377 | 22.7178 | 27.2036 | 30.1435 | 32.8523 | 36.1909 | 38.5823 | 43.8202 |
| 20 | 5.9210 | 7.4338 | 8.2604 | 9.5908 | 10.8508 | 12.4426 | 15.4518 | 19.3374 | 23.8277 | 28.4120 | 31.4104 | 34.1696 | 37.5662 | 39.9968 | 45.3147 |
| 21 | 6.4467 | 8.0337 | 8.8972 | 10.2829 | 11.5913 | 13.2396 | 16.3444 | 20.3372 | 24.9348 | 29.6151 | 32.6706 | 35.4789 | 38.9322 | 41.4011 | 46.7970 |
| 22 | 6.9830 | 8.6427 | 9.5425 | 10.9823 | 12.3380 | 14.0415 | 17.2396 | 21.3370 | 26.0393 | 30.8133 | 33.9244 | 36.7807 | 40.2894 | 42.7957 | 48.2679 |
| 23 | 7.5292 | 9.2604 | 10.1957 | 11.6886 | 13.0905 | 14.8480 | 18.1373 | 22.3369 | 27.1413 | 32.0069 | 35.1725 | 38.0756 | 41.6384 | 44.1813 | 49.7282 |
| 24 | 8.0849 | 9.8862 | 10.8564 | 12.4012 | 13.8484 | 15.6587 | 19.0373 | 23.3367 | 28.2412 | 33.1962 | 36.4150 | 39.3641 | 42.9798 | 45.5585 | 51.1786 |
| 25 | 8.6493 | 10.5197 | 11.5240 | 13.1197 | 14.6114 | 16.4734 | 19.9393 | 24.3366 | 29.3389 | 34.3816 | 37.6525 | 40.6465 | 44.3141 | 46.9279 | 52.6197 |
| 26 | 9.2221 | 11.1602 | 12.1981 | 13.8439 | 15.3792 | 17.2919 | 20.8434 | 25.3365 | 30.4346 | 35.5632 | 38.8851 | 41.9232 | 45.6417 | 48.2899 | 54.0520 |
| 27 | 9.8028 | 11.8076 | 12.8785 | 14.5734 | 16.1514 | 18.1139 | 21.7494 | 26.3363 | 31.5284 | 36.7412 | 40.1133 | 43.1945 | 46.9629 | 49.6449 | 55.4760 |
| 28 | 10.3909 | 12.4613 | 13.5647 | 15.3079 | 16.9279 | 18.9392 | 22.6572 | 27.3362 | 32.6205 | 37.9159 | 41.3371 | 44.4608 | 48.2782 | 50.9934 | 56.8923 |
| 29 | 10.9861 | 13.1211 | 14.2565 | 16.0471 | 17.7084 | 19.7677 | 23.5666 | 28.3361 | 33.7109 | 39.0875 | 42.5570 | 45.7223 | 49.5879 | 52.3356 | 58.3012 |
| 30 | 11.5880 | 13.7867 | 14.9535 | 16.7908 | 18.4927 | 20.5992 | 24.4776 | 29.3360 | 34.7997 | 40.2560 | 43.7730 | 46.9792 | 50.8922 | 53.6720 | 59.7031 |
| 40 | 17.9164 | 20.7065 | 22.1643 | 24.4330 | 26.5093 | 29.0505 | 33.6603 | 39.3353 | 45.6160 | 51.8051 | 55.7585 | 59.3417 | 63.6907 | 66.7660 | 73.4020 |
| 50 | 24.6739 | 27.9907 | 29.7067 | 32.3574 | 34.7643 | 37.6886 | 42.9421 | 49.3349 | 56.3336 | 63.1671 | 67.5048 | 71.4202 | 76.1539 | 79.4900 | 86.6608 |
| 60 | 31.7383 | 35.5345 | 37.4849 | 40.4817 | 43.1880 | 46.4589 | 52.2938 | 59.3347 | 66.9815 | 74.3970 | 79.0819 | 83.2977 | 88.3794 | 91.9517 | 99.6072 |
| 70 | 39.0364 | 43.2752 | 45.4417 | 48.7576 | 51.7393 | 55.3289 | 61.6983 | 69.3345 | 77.5767 | 85.5270 | 90.5312 | 95.0232 | 100.4252 | 104.2149 | 112.3169 |
| 80 | 46.5199 | 51.1719 | 53.5401 | 57.1532 | 60.3915 | 64.2778 | 71.1445 | 79.3343 | 88.1303 | 96.5782 | 101.8795 | 106.6286 | 112.3288 | 116.3211 | 124.8392 |
| 90 | 54.1552 | 59.1963 | 61.7541 | 65.6466 | 69.1260 | 73.2911 | 80.6247 | 89.3342 | 98.6499 | 107.5650 | 113.1453 | 118.1359 | 124.1163 | 128.2989 | 137.2084 |
| 100 | 61.9179 | 67.3276 | 70.0649 | 74.2219 | 77.9295 | 82.3581 | 90.1332 | 99.3341 | 109.1412 | 118.4980 | 124.3421 | 129.5612 | 135.8067 | 140.1695 | 149.4493 |

Table D.1: Central $\chi^2$ distribution: critical value $\chi^2_\alpha(n, 0)$ as function of one-sided level of significance $\alpha$ (top row) and degrees of freedom $n$ (left column). Given is one-minus-the CDF, hence $\alpha$ represents the right-tail probability. Example: $\alpha = 0.010$ and $n = 10$ yield $\chi^2_\alpha(n, 0) = 23.2093$.

# E

# NMEA [*]

## E.1. Introduction

NMEA 0183 is a data format, a communication protocol and an electrical interface, for the Position, Velocity and Time (PVT) output of a GNSS receiver, next to for instance an echo sounder, and an anemometer.

This standard has been initiated and maintained by the National Marine Electronics Association (NMEA) [126]. It is around already for a long time, several decades, and still often used to communicate for instance *position solutions* of a GNSS receiver.

The NMEA 0183 Interface Standard is a copyrighted document and available only from the NMEA. The latest version, as of November 2018, is version 4.11, which accomodates multi-GNSS (GPS, Glonass, Galileo, BeiDou and QZSS).

The NMEA 0183 standard is based on ASCII character encoding (American Standard Code for Information Interchange - ASCII; for instance the bit-sequence 1000001 standing for 'A', and 1100001 for 'a'), and a serial communication protocol. The data are communicated in records or sentences, also called messages, starting with the character '$', and ending with a checksum of two characters, preceeded by a '*'.

## E.2. Example

Below is shown an example of NMEA 0183 sentences, for two seconds.

```
$GNRMC,220332.00,A,5137.3391118,N,00443.2462641,E,0.017,,240320,,,A,V*15
$GNVTG,,T,,M,0.017,N,0.032,K,A*3A
$GNGGA,220332.00,5137.3391118,N,00443.2462641,E,1,12,0.67,-0.390,M,46.060,M,,*63
$GNGSA,A,3,01,03,08,11,22,14,28,32,27,,,,1.22,0.67,1.02,1*03
$GNGSA,A,3,88,65,87,72,,,,,,,,,1.22,0.67,1.02,2*08
$GNGSA,A,3,03,08,15,13,21,,,,,,,,1.22,0.67,1.02,3*0E
$GNGSA,A,3,27,09,28,14,,,,,,,,,1.22,0.67,1.02,4*04
$GPGSV,3,1,09,01,63,288,47,03,25,224,41,08,49,168,47,11,82,163,44,1*68
$GPGSV,3,2,09,14,35,113,43,22,51,221,47,27,19,152,39,28,27,304,38,1*6F
$GPGSV,3,3,09,32,38,082,47,1*5E
$GPGSV,3,1,09,01,63,288,42,03,25,224,35,08,49,168,41,11,82,163,,6*6F
$GPGSV,3,2,09,14,35,113,,22,51,221,,27,19,152,38,28,27,304,,6*66
$GPGSV,3,3,09,32,38,082,40,6*5E
$GLGSV,2,1,06,65,64,158,43,66,08,196,37,71,09,029,,72,54,052,44,1*72
$GLGSV,2,2,06,87,53,110,47,88,71,335,42,1*71
$GLGSV,2,1,06,65,64,158,36,66,08,196,40,71,09,029,18,72,54,052,42,3*7D
$GLGSV,2,2,06,87,53,110,36,88,71,335,29,3*78
```

```
$GAGSV,2,1,06,03,50,087,45,08,37,163,43,13,68,298,45,15,56,106,44,7*7F
$GAGSV,2,2,06,18,67,242,48,21,16,260,37,7*71
$GAGSV,2,1,06,03,50,087,44,08,37,163,44,13,68,298,45,15,56,106,46,2*7E
$GAGSV,2,2,06,18,67,242,48,21,16,260,38,2*7B
$GBGSV,1,1,04,09,35,064,37,14,26,049,30,27,77,272,46,28,43,136,47,1*7E
$GBGSV,1,1,04,09,35,064,43,14,26,049,33,27,77,272,,28,43,136,,3*7D
$GNGLL,5137.3391118,N,00443.2462641,E,220332.00,A,A*77
$GNRMC,220333.00,A,5137.3391090,N,00443.2462609,E,0.022,,240320,,,A,V*1F
$GNVTG,,T,,M,0.022,N,0.041,K,A*38
$GNGGA,220333.00,5137.3391090,N,00443.2462609,E,1,12,0.67,-0.394,M,46.060,M,,*6B
$GNGSA,A,3,01,03,08,11,22,14,28,32,27,,,,1.22,0.67,1.02,1*03
$GNGSA,A,3,88,65,87,72,,,,,,,,,1.22,0.67,1.02,2*08
$GNGSA,A,3,03,08,15,13,21,,,,,,,,1.22,0.67,1.02,3*0E
$GNGSA,A,3,27,09,28,14,,,,,,,,,1.22,0.67,1.02,4*04
$GPGSV,3,1,09,01,63,288,47,03,25,224,41,08,49,168,47,11,82,163,44,1*68
$GPGSV,3,2,09,14,35,113,43,22,51,221,47,27,19,152,39,28,27,304,38,1*6F
$GPGSV,3,3,09,32,38,082,47,1*5E
$GPGSV,3,1,09,01,63,288,42,03,25,224,35,08,49,168,41,11,82,163,,6*6F
$GPGSV,3,2,09,14,35,113,,22,51,221,,27,19,152,38,28,27,304,,6*66
$GPGSV,3,3,09,32,38,082,40,6*5E
$GLGSV,2,1,06,65,64,158,43,66,08,196,37,71,09,029,,72,54,052,44,1*72
$GLGSV,2,2,06,87,53,110,47,88,71,335,42,1*71
$GLGSV,2,1,06,65,64,158,35,66,08,196,40,71,09,029,17,72,54,052,42,3*71
$GLGSV,2,2,06,87,53,110,36,88,71,335,29,3*78
$GAGSV,2,1,06,03,50,087,45,08,37,163,43,13,68,298,45,15,56,106,44,7*7F
$GAGSV,2,2,06,18,67,241,48,21,16,260,37,7*72
$GAGSV,2,1,06,03,50,087,44,08,37,163,44,13,68,298,45,15,56,106,46,2*7E
$GAGSV,2,2,06,18,67,241,48,21,16,260,38,2*78
$GBGSV,1,1,04,09,35,064,37,14,26,049,30,27,77,272,47,28,43,136,47,1*7F
$GBGSV,1,1,04,09,35,064,43,14,26,049,33,27,77,272,,28,43,136,,3*7D
$GNGLL,5137.3391090,N,00443.2462609,E,220333.00,A,A*7B
```

## E.3. Position output

Sentences starting with GN are generic for Global Navigation Satellite Systems (GNSS), or refer to a combination of GNSSes. The sentence GNRMC contains the time (in UTC) 22:03:32 (with seconds given with two decimals), the latitude 51 deg and 37.3391 min, North (N), and longitude 4 deg and 43.2462 min, East (E), and the date 24 March 2020. The latitude and longitude are given in degrees and decimal minutes. The latitude in this example equals 51.622318 deg. This sentence, GNRMC, contains the so-called recommended minimum data.

The sentence GNGGA presents the position (fix) solution, with again latitude and longitude, and, height above Mean Sea Level (MSL), -0.390 m (altitude), and the geoid height above the ellipsoid, 46.060 m (computed using the Earth Gravitational Model (EGM), see Section 32.4). The ellipsoidal height is reconstructed as -0.390 + 46.060 = 45.67 m. There are 12 GNSS satellites used for this position fix (in this case a standalone or single point position solution).

An alternative data format for position information (e.g. tracks) is KML, which is described in Appendix K.

## E.4. GNSS

The GPGSV sentence shows the GPS satellites in view, in this case 9 satellites (09), listing four per sentence, with the satellite ID (PRN-number), 01, the elevation angle (63 deg; from 0 to 90), the azimuth angle (288 deg; from 0 to 359), and the C/N0 as a measure for the signal-strength (47 dB-Hz), and then on for the next satellite. There are similar sentences for

Glonass, `GLGSV`, Galileo, `GAGSV`, and BeiDou, `GBGSV`. The first two characters, like `GP`, `GN`, are also referred to as the talker ID.

The above example serves the purpose of giving a flavour of how NMEA looks like, and only a few of the basic sentences are discussed. There are many more, and often a GNSS equipment manufacturer even adds some additional, proprietary, sentences (for which the standard actually allows).

# F

# RINEX [*]

## F.1. Introduction

RINEX stands for Receiver Independent Exchange Format (RINEX) [127]. It is the standard format for the exchange of GNSS *measurement data* and is supported by almost all GNSS processing software packages, and many GNSS equipment manufacturers.

As opposed to proprietary format, binary data in the GNSS receiver, you can directly view, read and understand a GNSS measurement data file in RINEX format. In the next section, with Figures F.1 and F.2 two examples are presented of such GNSS measurement data files in RINEX format.

## F.2. Examples

RINEX files are ASCII files that can be viewed and edited by any editor. RINEX *Observation* files contain the pseudo-range (code) and carrier-phase measurements (the observed distances to the satellites), while RINEX *Navigation* files contain the decoded data message from the satellites including the broadcast ephemeris.

RINEX observation files consist of a block with meta-data followed by several data blocks. Each data block then contains the data (measurements) for one measurement epoch. A data block with epoch data consist of an epoch header, with the epoch time, number of satellites and a list with satellite ID's, in subsequent lines followed by the measurements for each of the satellites. The epoch time (in the red box) presents year-month-day (in the example of Figure F.1 2013, September 20th), and next, the time in hour-minutes-seconds, the first record being 14:06:35. The time reported is GPS system time, as kept locally by the receiver clock (and typically accurate to 1 millisecond). At the start of GPS, in January 1980, GPS time was equal to UTC, but GPS time does not introduce leap-seconds as UTC does (currently the applicable number of leap-seconds is 18, so, GPS time is 18 seconds ahead of UTC). Apart from a number of leap-seconds, GPS system time is, at the 10 nanosecond level, aligned with UTC. In the Netherlands local time — in the GMT+1 time zone — is one hour ahead of UTC, during Winter, and in Summer, with day-light saving time, two hours ahead of UTC.

In Figure F.1, at 14:06:35 (GPS time), 8 GPS satellites were observed, and for example for GPS satellite 16 (G16), the pseudorange measurement is 22141508.477 m (this is the observed distance to the satellite, listed in the first column). The order of the observation types is presented in the orange box. Observation-type 'C1' refers to the pseudorange code measurement on the L1-frequency (1575.42 MHz), and 'L1' is the corresponding carrier phase measurement. 'P2' and 'L2' are the pseudorange code and carrier phase measurement on the L2-frequency (1227.60 MHz).

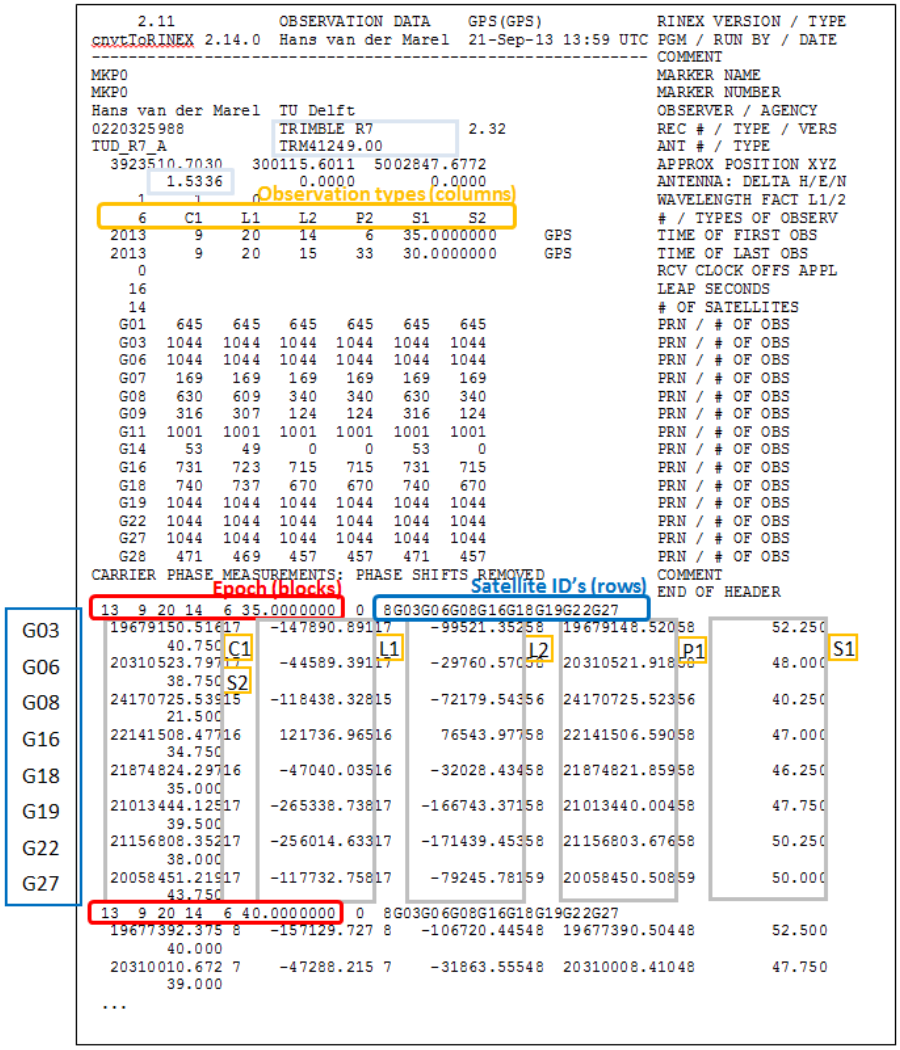
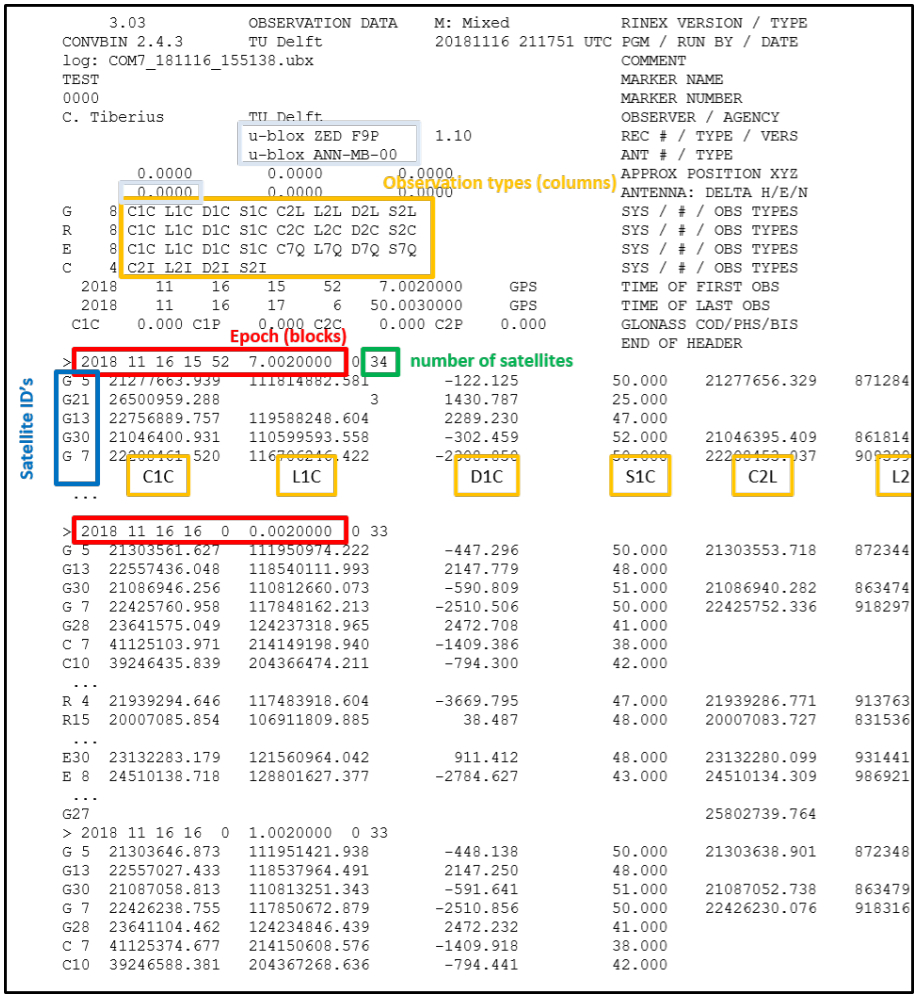



Figure F.1: Example of a GNSS measurement file in RINEX version 2, specifically a RINEX observation file.




```
     3.03           OBSERVATION DATA    M: Mixed            RINEX VERSION / TYPE
CONVBIN 2.4.3       TU Delft            20181116 211751 UTC PGM / RUN BY / DATE
log: COM7_181116_155138.ubx                                 COMMENT
TEST                                                        MARKER NAME
0000                                                        MARKER NUMBER
C. Tiberius         TU Delft                                OBSERVER / AGENCY
                    u-blox ZED F9P      1.10                REC # / TYPE / VERS
                    u-blox ANN-MB-00                        ANT # / TYPE
        0.0000          0.0000          0.0000              APPROX POSITION XYZ
        0.0000          0.0000          0.0000              ANTENNA: DELTA H/E/N
G    8 C1C L1C D1C S1C C2L L2L D2L S2L                      SYS / # / OBS TYPES
R    8 C1C L1C D1C S1C C2C L2C D2C S2C                      SYS / # / OBS TYPES
E    8 C1C L1C D1C S1C C7Q L7Q D7Q S7Q                      SYS / # / OBS TYPES
C    4 C2I L2I D2I S2I                                      SYS / # / OBS TYPES
  2018    11    16    15    52    7.0020000     GPS         TIME OF FIRST OBS
  2018    11    16    17     6   50.0030000     GPS         TIME OF LAST OBS
 C1C    0.000 C1P    0.000 C2C    0.000 C2P    0.000        GLONASS COD/PHS/BIS
                                                            END OF HEADER
> 2018 11 16 15 52  7.0020000  0 34
G 5  21277663.939   111814882.581        -122.125           50.000    21277656.329    871284
G21  26500959.288                3       1430.787           25.000
G13  22756889.757   119588248.604        2289.230           47.000
G30  21046400.931   110599593.558        -302.459           52.000    21046395.409    861814
G 7  22000461.520   116706216.422       -2000.050           50.000    22000453.037    900000
 ...

> 2018 11 16 16  0  0.0020000  0 33
G 5  21303561.627   111950974.222        -447.296           50.000    21303553.718    872344
G13  22557436.048   118540111.993        2147.779           48.000
G30  21086946.256   110812660.073        -590.809           51.000    21086940.282    863474
G 7  22425760.958   117848162.213       -2510.506           50.000    22425752.336    918297
G28  23641575.049   124237318.965        2472.708           41.000
C 7  41125103.971   214149198.940       -1409.386           38.000
C10  39246435.839   204366474.211        -794.300           42.000
 ...
R 4  21939294.646   117483918.604       -3669.795           47.000    21939286.771    913763
R15  20007085.854   106911809.885          38.487           48.000    20007083.727    831536
 ...
E30  23132283.179   121560964.042         911.412           48.000    23132280.099    931441
E 8  24510138.718   128801627.377       -2784.627           43.000    24510134.309    986921
 ...
G27                                                                   25802739.764
> 2018 11 16 16  0  1.0020000  0 33
G 5  21303646.873   111951421.938        -448.138           50.000    21303638.901    872348
G13  22557027.433   118537964.491        2147.250           48.000
G30  21087058.813   110813251.343        -591.641           51.000    21087052.738    863479
G 7  22426238.755   117850672.879       -2510.856           50.000    22426230.076    918316
G28  23641104.462   124234846.439        2472.232           41.000
C 7  41125374.677   214150608.576       -1409.918           38.000
C10  39246588.381   204367268.636        -794.441           42.000
```

Figure F.2: Example of a GNSS measurement file in RINEX version 3, specifically a RINEX observation file.

A full discussion of the RINEX format is beyond the scope of this appendix, and the reader is referred to the IGS-website (International GNSS Service (IGS)) [46].

More recent is RINEX version 3, as shown in Figure F.2. Where RINEX version 2 is limited to the measurements of GPS and Glonass, version 3 supports all GNSSes, including Galileo and BeiDou. The first letter of the satellite IDs in the blue box refers to the constellation, with 'G' for GPS, 'R' for Glonass (Russian), 'E' for Galileo (Europe) and 'C' for BeiDou (China).

The example in Figure F.2 contains GPS measurements on L1 (1575.42 MHz) and L2 (1227.60 MHz), respectively referred to as 'C1' and 'C2'. For Galileo, there are measurements on E1 (1575.42 MHz) and E5b (1207.14 MHz), referred to as 'C1' and 'C7'.

RINEX version 4.00 was released in December 2021. The new version primarily implied a major revision of the navigation message files to accomodate new nav-messages from all the GNSS constellations and system data messages.

Finally we note that an important field in the header of the file is the antenna height ('ANTENNA: DELTA H/E/N'). According to conventions, this must be the vertical height from the marker (or object of interest) to the antenna reference point (ARP), typically the bottom of the antenna housing (center of the bottom-plane). In the example of Figure F.1 it reads 1.5336 m (DELTA H). The DELTA E and N (East and North) are generally zero (the antenna is centered exactly above the marker).

# G

# Signal propagation [*]

In this appendix we deal with electromagnetic signals, though many of the phenomena are similar for acoustic signals, and we consider what happens to these signals on the way from transmitter to receiver. As shown in Figure G.1 there is a propagation medium in between the transmitter and receiver, e.g. air in the Earth's atmosphere.

In this appendix we first consider signal propagation at a global level, and we describe how signal power gets spread as the signal moves away from the transmitter, even when an electromagnetic signal travels through vacuum (a lossless, homogeneous medium), and next, we consider, more on a local scale, propagation effects when the medium is not homogeneous, and at the interface between two media.

Vector quantities are explicitly denoted in bold in this appendix, hence the electric field vector is $\mathbf{E}$, and the magnetic field vector $\mathbf{H}$.

## G.1. Signal spreading

For signal propagation at a global level, one might imagine a *point source*, that radiates a (perfectly) spherically symmetric electromagnetic field, see Figure G.2. A fictitious antenna which radiates uniformly in all directions is called an *isotropic* antenna. Strictly, this can not be accomplished in practice, see Section 21.3. The isotropic antenna is nevertheless commonly used as a reference.

An electromagnetic field can be used to transfer energy. Energy becomes available, or is dissipated, with the combustion of petrol fuel and with the passage of an electric current through a resistance. Operation of a measurement system is based on the transfer of electromagnetic energy from transmitter to receiver. As an electromagnetic wave propagates

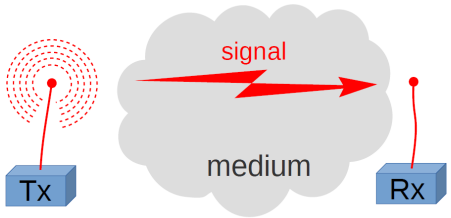



Figure G.1: A signal leaves transmitter Tx, propagates through a medium, and arrives at receiver Rx.

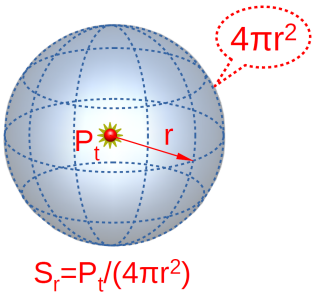



Figure G.2: An isotropic point source antenna, which radiates a perfectly spherical electromagnetic field.

through space, energy is transferred. The Poynting vector, see [128], is defined as

$$\mathbf{S} = \mathbf{E} \times \mathbf{H} \quad [\frac{\text{VA}}{\text{m}^2} = \frac{\text{W}}{\text{m}^2}] \tag{G.1}$$

where $\mathbf{E}$ is the electric field vector [V/m] and $\mathbf{H}$ the magnetic field vector [A/m]; it results as the outer, or cross-product of $\mathbf{E}$ and $\mathbf{H}$. The Poynting vector is interpreted as giving the direction, and quantifying the rate of energy transfer per unit area (expressed in Watt per squared meter), which is a power density (formally, it is the directional energy flux density). It is a vector quantity, in general depending on position and time, hence $\mathbf{S}(\mathbf{x}, t)$.

Power is energy per unit time. Energy is expressed in Joule [J], and power consequently in Joule/second, which is Watt [W] (=[Nm/s]).

A point-source antenna allows for a simple analysis of the *spreading* of the power. The power per unit area decreases as $1/r^2$ as the field propagates radially outward, with distance $r$, away from the source as shown in Figure G.2. The total power transmitted by the antenna is $P_t$ [W] (typically time-averaged). The *area* of a sphere with radius $r$ in the three-dimensional space equals $4\pi r^2$ $[\text{m}^2]$. The sphere is centered at the source, and the power *density* at a distance $r$ from the source is consequently

$$S_r = \frac{P_t}{(4\pi r^2)} \quad [\frac{\text{W}}{\text{m}^2}] \tag{G.2}$$

The transmitted signal power is distributed over the full surface (skin) of the sphere. The decrease in power density with increasing distance to the transmitting antenna is commonly referred to as *free space loss*, although strictly it is not a loss, but just a spreading of the power. It is not a loss of energy; in a strict sense, loss only occurs when a medium really takes away part of the signal energy (e.g. due to interaction with particles in the medium - for instance by absorption).

### G.1.1. Link budget

Analysing the link budget yields an answer to the question — given the amount of signal power made available at the source by the transmitter — how much signal power eventually arrives at the target (the receiver). The free space loss typically forms a big part of the link budget, but there is also the antenna gain, both for transmitter and receiver.

The amplification of a transmitted signal is referred to as *gain*. The gain $G$, in terms of power, is the ratio of the received signal power and the transmitted signal power:

$$\frac{P_{\text{Rx}}}{P_{\text{Tx}}} = G$$

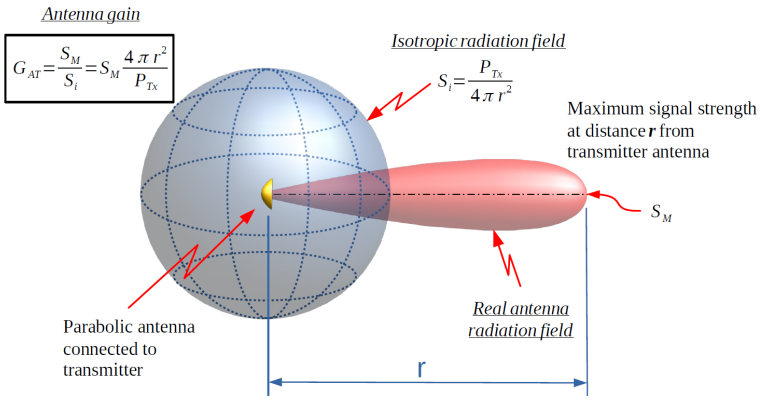



Figure G.3: The power transmitted by an isotropic antenna is represented by the sphere in blue. The radiation pattern of the actual antenna under consideration is shown in red; the direction with maximum signal power density is indicated by $S_M$; in that direction the antenna gain equals $G_{AT} = \frac{S_M}{S_i}$.

or

$$P_{\mathrm{Rx}} = P_{\mathrm{Tx}}\, G$$

which demonstrates explicitly that the gain is a factor; the transmitted signal power is multiplied by the gain $G$ to arrive at the received signal power. When the transmitted signal gets attenuated, e.g. due to the 'gain' of a cable, the received signal power is less than the transmitted signal power, and $G < 1$. Typically the gain of free-space propagation is (much) smaller than one ($G_{\mathrm{FS}} \ll 1$), due to *spreading* of the signal in space.

The gain is decomposed into several contributing factors — in this appendix the transmitter antenna gain, the 'gain' due to signal spreading, and the gain of the receiver antenna.

An antenna converts an electric signal into an electromagnetic wave; it is the interface between the wired and wireless transmission. From a mathematical point of view, an isotropic antenna is the most simple antenna. An isotropic antenna is a hypothetical antenna which radiates signal power uniformly in all directions (in three dimensions), cf. Figure G.2. The gain of an isotropic antenna is $G = 1$. With isotropic antennas, we have $G_{\mathrm{AT}} = G_{\mathrm{AR}} = 1$ (for transmit and receiver antenna respectively).

In practice, an antenna will concentrate (focus) the transmitted power in certain directions; the antenna gain is elevation and azimuth dependent. Typically the gain of the transmitter antenna is larger than one $G_{\mathrm{AT}} > 1$, when the receiver is located within the beam of the transmitter antenna (and vice versa), cf. Figure G.3. The gain actually expresses how well the antenna can concentrate (focus) the power in a specific direction. The Effective Isotropic Radiated Power (EIRP) is defined as

$$P_{\mathrm{EIRP}} = P_{\mathrm{Tx}} G_{\mathrm{AT}}$$

The receiver — located in the transmitter antenna beam — may believe that the transmitter is transmitting isotropically (in all directions), and then the total radiated power would be $P_{\mathrm{EIRP}}$. The transmitter is providing power $P_{\mathrm{Tx}}$, but to the receiver it looks like an isotropic antenna providing power $P_{\mathrm{EIRP}}$.

Given the power density, the power intercepted by the receiving antenna depends on the (effective) area of the antenna $A_e$. The effective area is generally not the actual (physical) area, but rather a modeling parameter of the antenna. The total received power $P_{\mathrm{Rx}}$ simply

equals the product of the power density at the receiver and the effective area of the receiving antenna

$$P_{\mathrm{Rx}} = \frac{P_{\mathrm{Tx}} G_{\mathrm{AT}}}{4\pi r^2} A_e$$

The gain $G_{\mathrm{A}}$ and the effective area $A_e$ of an antenna are related ([129]) by

$$G_{\mathrm{A}} = \frac{4\pi}{\lambda^2} A_e$$

with $\lambda$ the wavelength of the signal (related to the frequency $f$ through the speed of light $c$, by $\lambda = c/f$). Applying this to the receiving antenna, the total link budget equation becomes

$$P_{\mathrm{Rx}} = \underbrace{P_{\mathrm{Tx}} G_{\mathrm{AT}}}_{P_{\mathrm{EIRP}}} \underbrace{\left(\frac{\lambda}{4\pi r}\right)^2}_{G_{\mathrm{FS}}} G_{\mathrm{AR}}$$

For an isotropic antenna ($G = 1$), the effective area equals $A_e = \frac{\lambda^2}{4\pi}$. There exists a rich variety of different antennas, think of dipole antennas, horn and parabola antennas — discussing their specifics is beyond the scope of this book. In its most simple form, the above link budget consists of two isotropic antennas ($G_{\mathrm{AT}} = G_{\mathrm{AR}} = 1$) and we just have the free-space gain $G_{\mathrm{FS}}$.

Note that further details can be added in to the above link budget, such as some loss due to the atmosphere (in a satellite – Earth link), and so-called line-losses, which are losses due to feed-cables, connecting the transmitter and receiver with their respective antennas.

### G.1.2. Example: radar equation

In radar remote sensing the transmitter and receiver are co-located and share the same antenna. The transmitter sends a radio pulse, and listens for, and times the reflection from some topography on the Earth's surface Consequently, the signal has travelled distance $d$ twice, it is a two-way range. The link budget for this two-way ranging is set up in two parts: the signal is originally transmitted by the transmitter and captured by the object on the ground (area), and consequently 're-transmitted' (reflected) back to the transmitter. The Radar Cross Section (RCS) is a measure of the electric/reflective area; it may, or may not correspond with the actual physical area of the object. The RCS is denoted by $\sigma_t$ and expressed in $[\mathrm{m}^2]$. The RCS is sometimes decomposed into $\sigma_t = \sigma_o A$, where $A$ is the area intercepting the transmitted signal (in $[\mathrm{m}^2]$), and $\sigma_o$ presents the fraction of power which is reflected, and which is a dimensionless quantity, similar to the antenna gain.

The total power (in W) 'transmitted' by the reflecting object is obtained similarly as before, but instead of multiplication by the effective area $A_e$, we multiply by the Radar Cross Section, and equals

$$P_{\mathrm{reflect}} = \frac{P_{\mathrm{Tx}} G_{\mathrm{AT}}}{4\pi d^2} \sigma_t$$

The radar transceiver is located in the beam of the re-transmission, and considers the reflecting object to transmit isotropically with $P_{\mathrm{reflect}}$. To conclude we have to apply again the spreading loss along the path from reflecting object back to transmitter, and the effective area of the receiving antenna.

$$P_{\mathrm{Rx}} = \frac{P_{\mathrm{Tx}} G_{\mathrm{AT}}}{4\pi d^2} \sigma_t \frac{1}{4\pi d^2} \frac{G_{\mathrm{AR}} \lambda^2}{4\pi} = \frac{P_{\mathrm{Tx}} G_{\mathrm{A}}^2 \sigma_t \lambda^2}{(4\pi)^3 d^4}$$

As the transmitter and receiver antenna are identical, we have $G_{\mathrm{AT}} = G_{\mathrm{AR}} = G_{\mathrm{A}}$. The distance $d$ between transceiver and reflecting object appears in the link budget with power 4. In one way ranging this would be only a power of 2.

### G.1.3. Noise, and Signal-to-Noise ratio

The presence of noise may obscure proper reception and processing of the desired signal. Noise signals may originate from (processes in) the measurement equipment itself (circuitry and transmission lines) or be picked up by the receiving antenna, next to the desired signal. Noise around in the atmosphere can come from outer space (galactic, cosmic noise), be caused by physical processes in atmosphere and Earth, or by human activities.

The receiver shall gather sufficient (desired) signal power to properly function. The Signal-to-Noise Ratio (SNR) is a key parameter to successful operation of a communication or measurement system. This ratio compares the power of the desired electromagnetic signal $P_s$, with the power of (always present) electromagnetic background noise $P_n$, simply as $\frac{P_s}{P_n}$.

In an attempt to bridge large power-level differences, it is common in electrical engineering and communications technology to employ the following logarithm of the ratio

$$\mathrm{SNR} = 10 \ {}^{10}\log\left(\frac{P_s}{P_n}\right) \ \ [\mathrm{dB}] \tag{G.3}$$

that has been assigned to the 'unit' decibel, denoted by dB. The above ratio is a general *relative* measure of power, but in practice commonly used to compare signal power with noise power. When $P_s = P_n$, the SNR is zero. When the desired signal is 100 times more powerful than the noise, the Signal-to-Noise Ratio equals 20 dB. When the signal is 100 times less powerful than the noise, the SNR is -20 dB.

One decibel is one tenth of one bel, named in honour of Alexander Graham Bell, but the bel is seldom used without the deci-prefix.

When the power of a signal is compared to a standard power level of for instance 1 Watt ($= P_n$), the SNR is said to have unit decibel-Watt [dBW], or with $10^{-3}$ Watt, unit decibel-milliwatt [dBm].

## G.2. Propagation effects

The Maxwell equations govern the behaviour of an electromagnetic field. Solutions to these equations give the electric and magnetic field strengths as functions of time and position, and consequently can describe the propagation of electromagnetic waves through space. For a introduction to these equations, see e.g. [128], as this textbook offers an introduction to the theoretical concepts of electromagnetic waves, and contains the basic material on time-varying wavefields and their applications in electrical engineering, communication and remote sensing.

According to the Maxwell equations, spatial variations of the electric and magnetic field components are related to temporal variations of the magnetic and electric field components respectively. The variations of the electromagnetic field can be perceived as an *electromagnetic wave* propagating through space as time passes by. The topic of particular interest is, once the wave has been excited by an electromagnetic source, to study and describe the wave *propagation*, based on the Maxwell equations for electromagnetic fields.

A basic wave type is the *plane wave*. The electric and magnetic fields are perpendicular to each other, and to the direction of propagation, see Figure G.4. The electric and magnetic field vectors $\mathbf{E}$ and $\mathbf{H}$, and the propagation direction form a right handed triad, and have been chosen to lie along respectively the third ($z$), first ($x$) and second ($y$) axis.

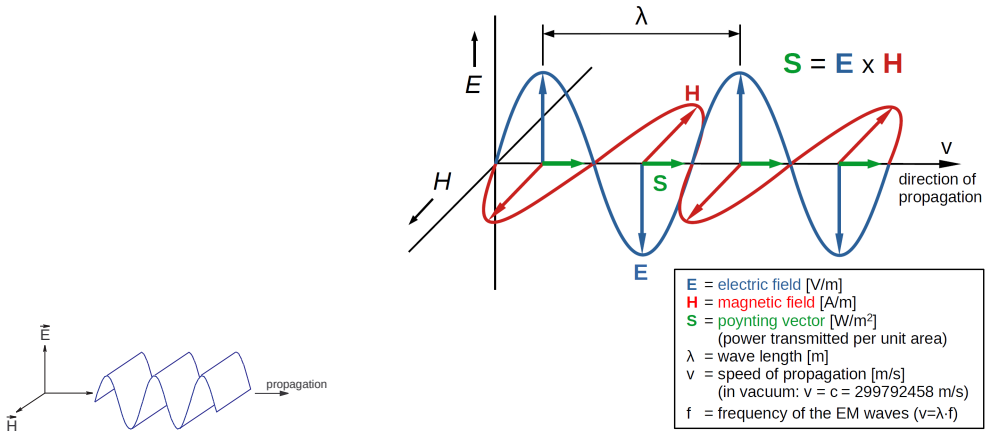



Figure G.4: The relation between the electric $\mathbf{E}$ and magnetic $\mathbf{H}$ field vector and the propagation direction of a plane wave. The Poynting vector lies along the propagation direction, $\mathbf{S} = \mathbf{E} \times \mathbf{H}$, and power transfer, by this plane wave, takes place in the propagation direction.

In practice a plane wave does not exist, as it can not be excited by any finite-sized antenna. It may serve however, as a useful approximation for wave propagation, for instance at larger distances from the source, studying local and regional effects. The reader is referred to the textbook [128] for an in-depth treatment.

The plane wave exists in *homogeneous* medium, implying that the propagation medium parameters are constants, and our discussion will be restricted to this type of medium for the moment. The study of electromagnetic waves is, as done here as well, often restricted to *steady state* analysis, in which the electromagnetic field quantities are taken to depend sinusoidally on time (a sinusoidal wave is called a *harmonic* wave). In addition to the above restrictions, only lossless media are considered here (in some cases the Earth's atmosphere may be approximated as a lossless medium), and vacuum is an example of lossless medium.

The electric field vector, as shown in Figure G.4, is given by

$$\mathbf{E}(\mathbf{x}, t) = \begin{pmatrix} 0 \\ 0 \\ E_o \cos(\omega t - ky) \end{pmatrix}$$

where the wave propagation takes place in vacuum, which is a lossless medium. In the above equations $k$ is the so-called wave number, and in vacuum defined as $k = \frac{\omega}{c}$ or $k = \frac{2\pi}{\lambda}$ in radians per meter (it is the number of cycles in a distance of $2\pi$ m). The magnetic field vector reads

$$\mathbf{H}(\mathbf{x}, t) = \begin{pmatrix} H_o \cos(\omega t - ky) \\ 0 \\ 0 \end{pmatrix}$$

The propagation 'velocity' $v$ of a wave in this type of medium correspondingly becomes $v^2 = \frac{1}{\varepsilon\mu}$. Relating this 'velocity' to the propagation speed in vacuum $c$

$$n = \frac{c}{v} = \frac{\sqrt{\varepsilon\mu}}{\sqrt{\varepsilon_o \mu_o}} = \sqrt{\varepsilon_r \mu_r} \tag{G.4}$$

yields the (real-valued) index of refraction $n$. The refractive index is usually larger than one, and the propagation speed in most media is less than the speed of light in vacuum, $v < c$, causing a *delay* in the signal travel time.

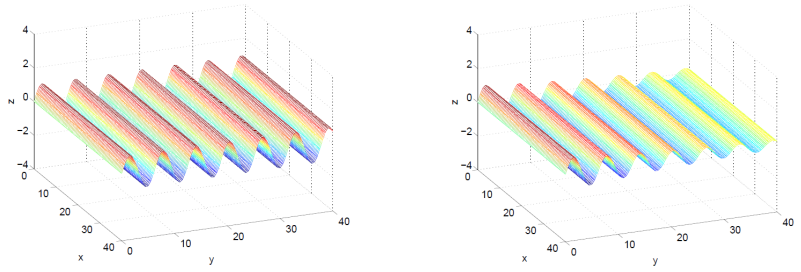


Figure G.5: Uniform two-dimensional plane wave, at left in lossless medium, at right in lossy medium. The wave, shown in the three-dimensional space at a particular epoch in time, propagates along the $y$-axis to the right.

In the above equations is $\varepsilon_o$ the electric permittivity in vacuum in farad per meter [F/m], and $\mu_o$ the magnetic permeability in vacuum in henry per meter [H/m]. This last quantity is fixed by the SI International System of Units to $\mu_o = 4\pi.10^{-7}$ H/m. The speed of light in vacuum is $c = 299792458$ m/s and by the relation $c^2 = \frac{1}{\varepsilon_o \mu_o}$, the permittivity in vacuum follows as $\varepsilon_o \approx 8.85 \cdot 10^{-12}$ F/m. The null-index refers to vacuum. The permittivity and permeability of a medium are related to those of vacuum

$$\varepsilon_r = \frac{\varepsilon}{\varepsilon_o} \text{ and } \mu_r = \frac{\mu}{\mu_o}$$

these are the relative permittivity and relative permeability. Both are dimensionless quantities.

### G.2.1. Example: two dimensional wave

In this section the two dimensional electromagnetic wave is introduced, merely by means of a few graphical examples. The electromagnetic field vectors $\mathbf{E}$ and $\mathbf{H}$ do depend on two spatial coordinates, on $x$ and $y$, but still not on $z$.

Figure G.5 presents two examples of the *uniform* plane wave: the planes of equal phase and those of equal amplitude are parallel. The graph at left shows the plane wave in lossless medium, whereas the graph at right pertains to lossy medium, in which the wave gets 'damped' as it propagates (the wave gets attenuated, by absorption). In the following we focus on lossless media.

The wave at left is strictly periodic in space, with period $\lambda$, the wavelength. A unit amplitude was taken (initially). Note that the example is actually still a one dimensional wave as the $x$-coordinate is still not involved in the field vectors, $\mathbf{E}$ is along the $z$-axis, and $\mathbf{H}$ along the $x$-axis.

When the wave hits an object or medium which it can not penetrate, the wave gets reflected. The angle of incidence $\Theta_i$ then equals the angle of reflection $\Theta_r$, cf. Figure G.6 at left; the arrowed lines represent the propagation directions.

When a wave is incident upon the plane boundary between two different media, the wave is partly reflected and partly transmitted through. Figure G.7 shows, beside the incident wave, only the transmitted wave. The reflected wave will interfere with the incident one. If the incident wave is plane and uniform, the reflected and transmitted ones are as well, in lossless media. For the reflected wave holds again $\Theta_i = \Theta_r$, *Snell's law of reflection*, see Figure G.6 at right. The transmission angle $\Theta_t$ follows from *Snell's law of refraction*

$$n_1 \sin \Theta_i = n_2 \sin \Theta_t \tag{G.5}$$

where $n_1$ and $n_2$ are the indices of refraction for medium 1 and 2 respectively, functions of $\varepsilon$ and $\mu$ of the two media. Both laws are known from geometric optics.

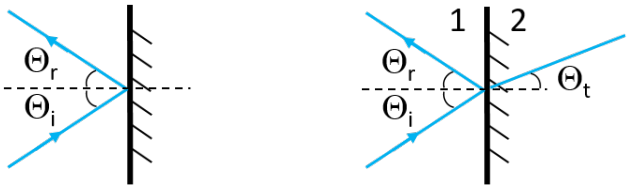



Figure G.6: At left, reflection: for a uniform plane wave the angle of incidence $\Theta_i$ equals the angle of reflection $\Theta_r$, both measured with respect to the normal on the reflecting plane surface. At right, the angle of transmission $\Theta_t$ is according to Snell's law of refraction; the transmitted wave is travelling at a different speed in medium 2, than in medium 1 (in this example slower; $n_2 > n_1$), and in a different direction - the wave propagation direction gets bended.

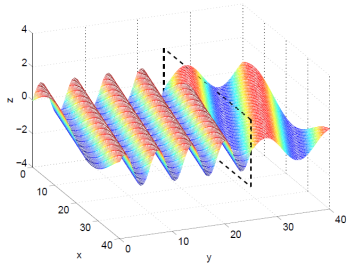


Figure G.7: Transmission of a uniform two-dimensional plane wave. The (incident) wave propagates in the $x$-$y$-plane, to the front at an angle of $10°$ with the (positive) $y$-axis. The transmission is slightly refracted to $20°$ ($n_1 > n_2$).

## G.2.2. Electromagnetic rays

So far we dealt with plane electromagnetic waves, which can exist in homogeneous media. In general, electromagnetic waves are however not plane, for instance when they propagate through an inhomogeneous medium. The propagation of an electromagnetic field through a weakly *inhomogeneous* medium is described in terms of rays. The theory of geometric, or ray optics can be derived from the Maxwell equations as an asymptotic solution in the limit as the frequency approaches infinity. Geometric optics is generally a valid and useful approximation when the medium parameters change very little over a distance that is large compared with the wavelength.

The electromagnetic ray is introduced, as an approximate solution to the Maxwell equations in inhomogeneous medium. The *wavefront* is formed by all points or locations in space where the electromagnetic wave, transmitted by the source, has just arrived, see also Figure G.8.

The positions traced by a particular 'point' or spot on the wavefront, during the propaga-

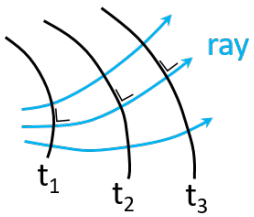



Figure G.8: Electromagnetic wavefront at times $t_1$, $t_2$ and $t_3$. The curve connecting corresponding points on the fronts represents a ray trajectory.

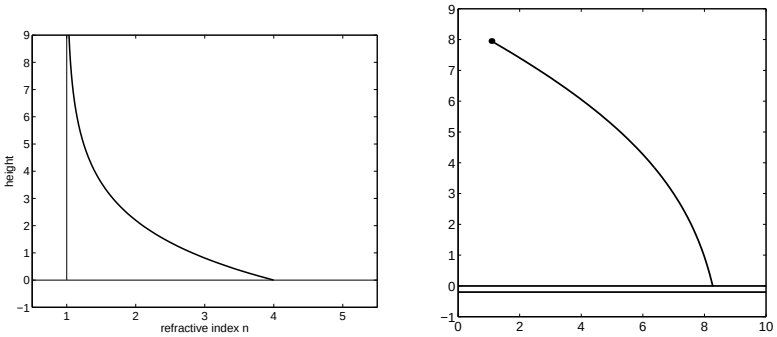



Figure G.9: Refractive index $n$ as function of height $z$ [km], at left, and resulting ray trajectory in horizontally layered medium (e.g. cross-section of the Earth's atmosphere, with the Earth's surface shown by the double line at bottom), at right. The effect has been largely exagerated for illustrative purpose.

tion of the wave through space, form a *ray* trajectory, see Figure G.8. Wavefronts and ray trajectories are generally curved in inhomogeneous media. For general *uniform* electromagnetic rays, they (fronts and rays) are — as long as the medium is isotropic — perpendicular to each other. The tangent to a ray trajectory coincides in each point with the direction of wave propagation. The key parameter for propagation can then be shown to be the *index of refraction* $n$:

$$n = c\sqrt{\varepsilon\mu} \tag{G.6}$$

that relates the propagation speed $v$ in the medium to the one in vacuum $c$, $n = \frac{c}{v}$, as for the plane wave before. In vacuum the index of refraction equals $n = 1$.

At this point wave propagation has become very similar to the propagation of rays in geometrical optics. The trajectory of a uniform electromagnetic ray can be described by second order differential equations in the coordinates $x$, $y$, $z$, in which only the index of refraction $n$ is involved. Once initial position and direction have been specified, the ray trajectory can be determined, for instance numerically, which is referred to as ray tracing. Finally we give an example for the trajectory in a horizontally layered medium.

In a horizontally layered medium, the refractive index varies only (but continuously) in the vertical direction, in this case the $z$-coordinate: $n = n(z)$, not necessarily being a monotonic in- or decreasing function.

With $\Theta$ the angle between the vertical axis (z) and the propagation direction, Snell's law for horizontally layered medium states that $n(z)\sin\Theta$ is constant along the ray trajectory, hence

$$n(z_1)\sin\Theta_1 = n(z_2)\sin\Theta_2 \tag{G.7}$$

Figure G.9 shows an example of a ray trajectory in a horizontally layered medium. The graph at left gives the refractive index as function of height $z$. It increases monotonically with decreasing height, as is common when an electromagnetic wave approaches the Earth's surface from space. The amount of increase and the eventual value at the Earth's surface have however been largely exagerated in this example, for illustrative purpose.

Propagation through a horizontally layered medium can apply to electromagnetic waves traveling through a local part of the Earth's atmosphere, that is, over a relatively small distance so that the curvature of Earth (and its atmosphere) can be neglected.

The signal path gets more and more bended (refracted), all the time, as shown in Figure G.6 at right, towards the Earth's surface, as the refractive index keeps on increasing. And the signal

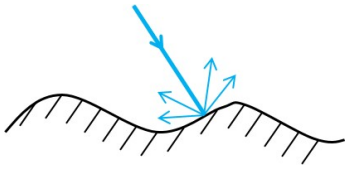


Figure G.10: Diffuse reflection. The surface is not a 'perfect' mirror, and the electromagnetic wave is reflected in multiple directions. Diffuse reflection occurs when the surface irregularities are comparable to, and larger than the wavelength. For specular reflection, surface imperfections/irregularities have to be smaller than the wavelength.

gets delayed, as the refractive index is larger than one (and keeps on increasing), and hence, the propagation speed gets smaller and smaller, than the speed of light in vacuum.

Not just any electromagnetic wave will pass through the Earth's atmosphere; visible light and microwaves will pass, as well as certain bands in the infrared-part of the spectrum.

In the lower part of the Earth's atmosphere (generally the lower 10 km of the atmosphere), the so-called troposphere, the refractive index $n$ practically depends on the temperature, the atmospheric pressure, and the amount of water vapour (humidity). This holds for the radio-spectrum, frequencies below 20 GHz; the troposphere is non-dispersive at these frequencies, that is, the refractive index is the same for all frequencies. For signals in the visible and near-visible part of the spectrum (see also Appendix A.1), the refractive index does depend also on the frequency.

At frequencies exceeding a few GHz, up to the visible and near-visible spectrum, rain (percipitation), clouds and fog may cause significant absorption and scattering of radio waves. Visible and near-visible signals for sure do not allow all weather operations.

Finally we note that in this section, we briefly reviewed the subject of propagation of electromagnetic waves, and we covered, next to attenuation and refraction, also reflection, thereby assuming a perfect, flat boundary between the two media, resulting in a so-called *specular* reflection (as in Figure G.6). In practice the boundary surface may not be perfectly flat and smooth, and cause actually reflection of the wave into multiple directions, as illustrated in Figure G.10. This is an example of *diffuse* reflection, also refered to as *scattering*.

## G.3. Acoustic waves

In acoustics, energy is transferred through a medium, by means of a vibrational wave. For instance, molecules in a fluid are displaced from their normal configuration, and this displacement (for instance compression) causes an internal elastic restoring force. As a sound wave travels through a medium, the particles of the medium vibrate and produce density and pressure changes along the path of motion of the wave. These compressions and dilations propagate from transmitter to receiver, as a longitudinal wave.

There are various types of transducers which convert electric energy into sound pressure, or vice versa, for instance by means of the piezo-electric effect. An electric potential difference may arise between the different sides of a certain crystal when undergoing mechnical strain, and vice versa. In hydrography, the transmitter is typically referred to as the source or the projector, and a hydrophone is the receiving unit.

### G.3.1. Acoustic wave propagation

The propagation speed $v$ of acoustic waves in a medium is given by

$$v = \sqrt{\frac{E}{\rho}}$$

where $E$ is the elasticity modulus (also known as bulk modulus), the relative variation of volume due to pressure in [Pa], which is $[\mathrm{kg/ms^2}]$, and $\rho$ the density in $[\mathrm{kg/m^3}]$, both are properties of the medium. For water the elasticity modulus is $E \approx 2.2 \cdot 10^9$ Pa, and the density is $\rho \approx 1.0 \cdot 10^3$ $\mathrm{kg/m^3}$, and this yields $v \approx 1.5 \cdot 10^3$ m/s. In practice the speed of sound in (sea)-water depends on temperature, pressure, and salinity.

The pressure obviously depends on the water depth. With an atmospheric pressure of 101325 Pa (1013.25 mbar) at the surface, the pressure increases roughly by the same amount every 10 m of depth, due to the density of water of about $\rho \approx 1.0 \cdot 10^3$ $\mathrm{kg/m^3}$. The temperature may range from up to 20° - 30° in the top layer, to close to zero at large depths. The salinity is roughly at the 0.35‰-level (per mille). As a result the speed of sound may vary, as a function of depth, by several percents, and range from about 1475 m/s to 1520 m/s. Of course, local variations in pressure, temperature and salinity translate in variations of the speed of sound.

Similar to the Poynting vector (G.1) with electromagnetic waves, the rate of energy transfer (or flow) per unit area is defined for sound waves as well, and referred to as *intensity* $I$. It is a function of the amplitude of the pressure changes, (ambient) density, and sound propagation speed. The unit of intensity $I$ is $[\mathrm{W/m^2}]$.

Most of the concepts covered in this appendix on electromagnetic waves apply to acoustic waves as well, like the Signal-to-Noise ratio and propagation effects such as attenuation, reflection and refraction.

When going from one medium to another, an acoustic wave may propagate (penetrate) and/or get reflected. The acoustic impedance determines this behaviour. The larger the difference in impedance between the two media, the larger the amount of wave energy that is reflected back. The acoustic *impedance* is a measure of opposition that a system or medium presents to the acoustic flow resulting of an acoustic pressure being applied to the system or medium. For example, upon an air - brick wall interface, the wall has a (much) higher impedance than air, and most of the sound is reflected back.

# H

# 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 meters, means that it is 25 times the length of the unit meter. A quantity shall be used (defined), with appropriate unit, that provides a reproducible standard.

| quantity | dimension | unit |
|---|---|---|
| distance | length | meter [m] |
| time duration | time | second [s] |
| mass | mass | kilogram [kg] |

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

As listed in Table H.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, see e.g. [52]. This system was first established in 1889 by the Bureau International des Poids et Mesures (BIPM) [130]. The official *definitions* read:

- *the meter* 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 meter, 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 H.1. A supplementary unit (actually the quantity is dimension- and unitless) is the radian [rad] for plane angles. A full circle corresponds to $2\pi$ rad. Other units for angles are grades or gradians (also known as gon), a

| 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 | $\mu$ | $10^{-6}$ |
| nano | n | $10^{-9}$ |
| pico | p | $10^{-12}$ |
| femto | n | $10^{-15}$ |
| atto | a | $10^{-18}$ |

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

circle is $400$ grad (the centesimal system), and degrees, a circle is $360$ deg, or $360^\circ$, 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, the sexagesimal system, a numeral system with sixty as its base. As an example $52^\circ 03' 44''$ means $52 \cdot 60^2 + 3 \cdot 60^1 + 44 \cdot 60^0$ seconds of arc.

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 H.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$^2$].

# I

# Publieke Dienstverlening Op de Kaart (PDOK)

PDOK, a collection of public map services, enables users to access digital geographical data for the Netherlands using Open Source Geospatial (OSG) compliant web services. Over 200 datasets and over 500 view- and download services (aimed at digital mapping) are available to the general public, private companies, organisations and the public sector. The use of PDOK is for free. For details and the website, see [59].

## I.1. Introduction

PDOK unlocks digital geospatial data from the Dutch government. This unlocking is done through a central facility: PDOK, [59]. The data are available first of all through a viewer (PDOK Viewer) in your web-browser in order to get familiar with the available data — to say *browse* through the available data, and secondly, as a web service or as a downloadable file (then you use your GIS-software, for instance QGIS [5], to analyze and visualize the data). In the sequel we discuss the web service in more detail.

In PDOK you can also find: 'Basisregistratie Topografie Achtergrondkaarten (BRT-A)', as discussed in Section 38.1. You can use this map for your own website or publication with the advantage that it is free of charge, and free of advertisements. In this book we frequently use maps and data from PDOK, and refer to the website as a single point of entry, though indicating each time the keyword of the dataset or map.

Where necessary, PDOK ensures that the services and files comply with the INSPIRE directive on improving the exchange of spatial data in Europe[1]. PDOK also manages the Nationaal Georegister (NGR). This online catalogue contains a lot of links to Dutch spatial datasets including a lot of web services from different providers.

Almost all PDOK files and data services are available under an 'open' license. This means that everyone is free to use the data in accordance to a Fair Use Policy. PDOK offers an overview of all freely available services and datasets. In some cases the access policy of the holders of the data source determines that the data are not accessible to everyone.

The available map information is easy to reuse in your own applications. The use of the web service is done by copying and pasting a URL from the overview, or from the Nationaal Georegister (NGR). If your software supports open standards that apply to PDOK, then the

[1] INSPIRE, which stands for Infrastructure for Spatial Information in Europe, is a directive, since 2007, leading to standardization of describing and sharing environmental spatial data. INSPIRE defines common standards for 34 spatial data themes (such as transport, energy, population and natural risk) [131].

| abbreviation | description | type |
|---|---|---|
| WMS | Web Map Service | raster data (images) |
| WMTS | Web Map Tile Service | raster data(images) |
| WFS | Web Feature Service | vector data (lines, points, etc.) |
| WCS | Web Coverage Service | coverage |

Table I.1: Open Geospatial Consortium (OGC) web service protocols.

requested map layer is shown. For example, to get a background map into your GIS application, you should link the URL from the overview into your GIS application. In QGIS [5] this is done by opening a WMTS layer (raster) using this URL. In QGIS there is also a plugin that can do this for you: the PDOK plugin. You can install this plugin though the plugin menu in QGIS.

## I.2. Protocols

The use of the PDOK data web services [59] requires certain knowledge of geotools and the various protocols used by PDOK and GIS applications, see Table I.1. The *Open Geospatial Consortium* (OGC) standards are used by PDOK to the maximum extent possible. The standard being used is shown in the PDOK overview. Sometimes the same product is available in multiple standards and you need to understand what each standard can offer and more importantly what not.

A *Web Map Service* (WMS) or *Web Map Tile Service* (WMTS) server acts upon requests by the client (e.g. you, using QGIS) for a *raster* map with a *given* extent, set of layers, symbolization style, and transparency. The WMS server then consults its local data sources, (if needed) rasterizes the map, and sends it back to the client in a raster format. For QGIS, this format would typically be JPEG or PNG. WMTS is used for distributing *tile* sets of geospatial data. The dataset is geographically cut into small parts, to enable manageable operation. This is a faster and more efficient way of distributing data than WMS, because with WMTS, the tile sets are pre-generated (and stored at the server), and the client only requests the transmission of the tiles, not their production. A WMS request typically involves both the generation and transmission of the data. In order to display the data at a variety of scales close to what the user might want, the WMTS tile sets are produced at several different scale levels and are made available for the GIS client to request them. Only the resulting ‘‘image’ is sent to the user — the original data remain at the PDOK-server. Also Google Maps and Google Earth work this way, see Appendix K.

PDOK adopts WMTS tiling services (and not the WMS protocol) for popular data sets because the tiling services are more efficient than the WMS protocol. With WMS, the map is recreated every time and this is a heavy burden on the PDOK servers. With *tiling* services *images* are *pre-stored* and these images are ready for loading from the cache at all times.

It is important to understand that WMS/WMTS-serves images (PNG/JPEG) that are rendered by the WMS/WMTS server. Therefore, you, as a user, can*not* change style information. You basically get the image as-is. QGIS is able to display a legend graphic in the table of contents’ layer list and in the map composer only if the WMS/WMTS-server has GetLegendGraphic-capability, and the layer has the getCapability-url specified.

In case you want to apply your own styling, you should use WFS or WCS protocols instead (if available).

A *Web Coverage Service* (WCS) provides access to raster data in forms that are useful for client-side rendering, as input into scientific models, and for other clients. WCS allows

clients to choose portions of a server's information holdings based on spatial constraints and other query criteria. QGIS has a native WCS provider which handles all network requests and uses all standard QGIS network settings. It is also possible to select cache mode ('always cache', 'prefer cache', 'prefer network', 'always network'), and the QGIS provider also supports selection of time position if a temporal domain is offered by the server.

A *Web Feature Service* (WFS) in QGIS behaves pretty much like any other vector layer. You can identify and select features, and view the attribute table. In general, adding a WFS layer is very similar to the procedure used with WMS. Though with WFS you can still customize, in your GIS software, the visualization to your own needs and wishes — there is no pre-created 'image'. Practically spoken, with a Web Feature Service (WFS), the server primarily acts like a geographical database.

# J

# OpenStreetMap (OSM)

OpenStreetMap is a free, editable, open-source map of the world. From the website [51]: 'OpenStreetMap is built by a community of mappers that contribute and maintain data about roads, trails, cafés, railway stations, and much more, all over the world'. OpenStreetMap (OSM) is a collaborative project (collaborative mapping). Rather than the map itself, the data generated by the project are considered its primary output.


Figure J.1: Sample map of the Delft area from OpenStreetMap [51]. ©OpenStreetMap contributors, data available under the Open Data Commons Open Database License (ODbL).

The key-assets of OpenStreetMap are, quoted from [51]:

- OpenStreetMap emphasizes *local knowledge*; contributors use aerial imagery, GPS devices, and low-tech field maps to verify that OSM is accurate and up to date
- OpenStreetMap's *community* is diverse, passionate, and growing every day; the contributors include enthusiast mappers, GIS professionals, engineers running the OSM servers, humanitarians mapping disaster-affected areas, and many more
- OpenStreetMap is *open data*: you are free to use it for any purpose as long as you credit OpenStreetMap and its contributors; if you alter or build upon the data in certain ways, you may distribute the result only under the same license

# K

# Google Earth

## K.1. Introduction

Google Earth, [132], features an interactive 3D representation of the Earth, based, in the first place on satellite imagery, covering the entire Earth, supplemented by aerial photography and geographic information. In many cities Google Earth can show 3D building models with photo-realistic 3D imagery.

A three-dimensional representation is, for much of the Earth, enabled by using radar and optical remote sensing measurements, resulting in a Digital Elevation Model (DEM). At present, Google Earth relies on radar measurements to the Earth's surface, taken from two locations (for instance two passes of the same space-vehicle or two antennae on the same space-vehicle), according to the principle of Interferometric Synthetic Aperture Radar (InSAR), see Section 23.2. Google Earth allows a user to go below the surface of the ocean, and view, similarly to land surface, also the seafloor.

Google uses a WGS84 datum as its reference system for Google Earth, and coordinates in KML (see next section) are longitude, latitude and altitude. Longitude and latitude values are in decimal degrees, and negative for West and South, respectively. The vertical component, altitude, is expressed, in meters, with respect to the Earth Gravitional Model 1996 (EGM96), which is a geopotential model of the Earth, see Section 32.4 and Chapter 33. When altitude is omitted, then the default value of zero is substituted, or, the corresponding height of the terrain is taken at that location ('clamped to ground'). Instead of altitude, ellipsoidal height may be included in KML.

For visualization, Google Earth uses a simple *cylindrical* map-projection (web Mercator projection), which means that meridians and parallels are straight lines and perpendicular to each other, though differently from the Mercator map-projection, both the meridians and parallels are *equidistant*; both are discussed in Section 30.4.

Google Earth can run as an application on the user computer and as a Web Map Service (WMS) client, see Appendix I.

In the domain of civil engineering, Google Earth is useful for visualization and presentation, and to obtain a first impression of the landscape and the area of interest. It should be kept in mind that, at present, the geographic information in Google Earth at best has a meter-accuracy, and certainly not centimeter-accuracy required for building and infrastructure-construction.

## K.2. KML

A user may add, to Google Earth, his/her own data, using Keyhole Markup Language (KML). KML, used by Earth browsers like Google Earth for the visualization of geographic information

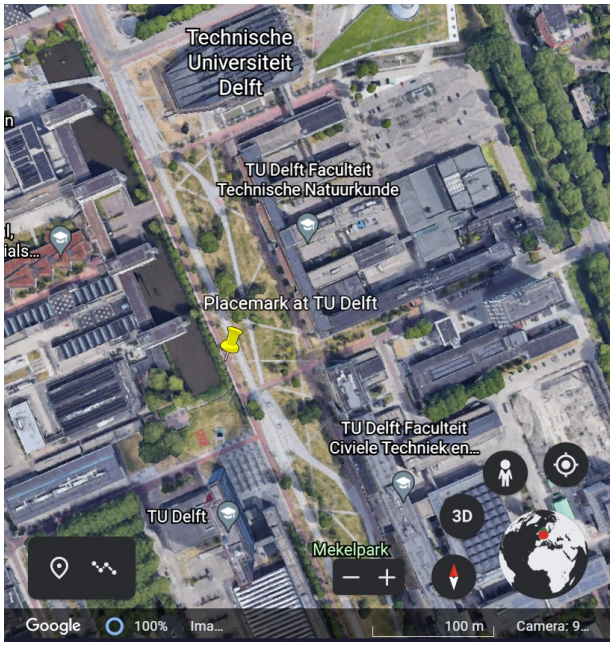



Figure K.1: Result of the simple example KML-file in Google Earth [132], showing the TU Delft campus. Image from Google Earth (imagery date 31/08/2019).

on Earth, is the equivalent of HyperText Markup Language (HTML) for web browsers. KML, an eXtendable Markup Language (XML), was created by Keyhole, Inc., specifically for, at that time, Keyhole EarthViewer. The company was acquired by Google in 2004, leading to Google Earth. KML became an international standard of the Open Geospatial Consortium (OGC) in 2008.

KML specifies a set of features, like placemarks, images, polygons, 3D models and textual descriptions, for display in an Earth browser. A very simple example of a KML-file, producing the yellow push-pin at the TU Delft campus, see Figure K.1, is shown below. The placemark is at 52 degrees latitude North, cf. Figure 26.1.

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://www.opengis.net/kml/2.2">
      <Placemark>
        <name>Placemark at TU Delft</name>
        <description>Clamped to ground</description>
        <Point>
          <coordinates>4.3734,52.000,0</coordinates>
        </Point>
      </Placemark>
</kml>
```

## K.3. Google Earth Engine

Google Earth Engine is a cloud computing platform for processing primarily satellite imagery [133]. It provides, through a web-interface, access to a large database of satellite imagery (Landsat, MODIS, and Sentinel, see Table 25.1), and the computational power needed to analyze those images. Using repeated satellite imagery (time-series of images), a user can analyze — using ready-to-use functions and his/her own scripts — changes on the Earth's surface, for instance

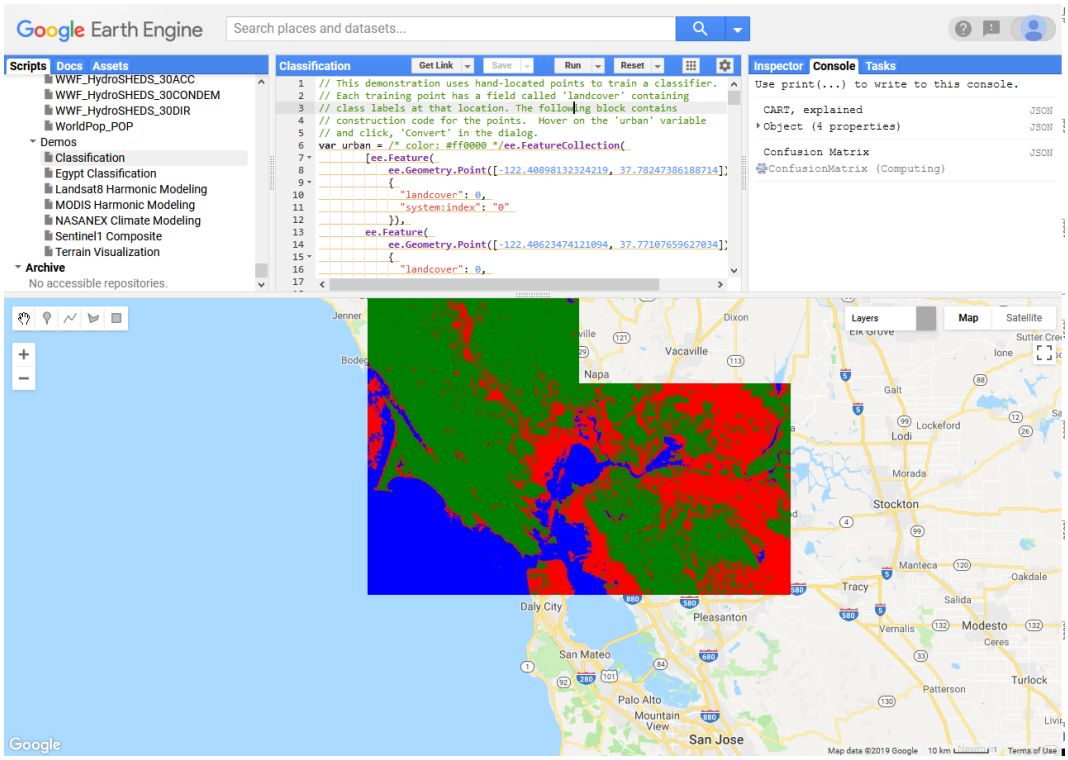



Figure K.2: Screenshot of a land cover example in the San-Francisco area, using Google Earth Engine (a demo classification script is used). Map data ©2019 Google, [133].

urbanization and deforestation. An example of land cover classification in Google Earth Engine is shown in Figure K.2.

## K.4. Data storage and processing architecture

These days many new remote sensing satellite missions are being launched, think for instance of the ESA Sentinel program. These missions deliver a huge amount of Earth observation data, in the order of Petabytes of data. The use of these data, with many users simultaneously, requires a different architecture for storage and processing.

Traditionally a user would *download* data from a server to his own computer system (client), store it locally, and process and analyze the data with software installed and running on his computer system.

Today, there is a trend, with the use of big (geo) data, to leave the data on the server, which is possibly largely distributed, meaning to consist of an array of physical machines. The user instead *uploads* to the server, instructions and operations, for instance in a Python script, and has them run on the server, and eventually gets back his/her result, typically in the form of an image or a map. This is the so-called virtual machine approach.

The architecture of Google Earth Engine (GEE) is more diffuse, in the sense that you interact within the cloud, both for storage and computing [133]. The data are stored on a network of computers, and the processing is carried out on a network of computers as well (distributed storage, and distributed computing).

# VIII

# Bibliography

# Bibliography


[1] Sinergise Laboratory for geographical information systems, Ltd., *Sentinel Hub,* (n.d.).

[2] F. Dekking, C. Kraaikamp, H. Lopuhaa, and L. Meester, *A modern introduction to probability and statistics — understanding why and how* (Springer Verlag, 2005).

[3] P. de Bakker, *GPS Positioning,* (2017), reader for CTB3310 Surveying and Mapping, 15 pages, Delft University of Technology.

[4] M. Schmandt, *GIS Commons: an introductory textbook on Geographic Information Systems,* (n.d.), available under CC BY-SA 3.0 license.

[5] *QGIS - A Free and Open Source Geographic Information System,* (n.d.), website.

[6] J. E. Alberda, *Inleiding Landmeetkunde*, 3rd ed. (Delftse Uitgevers Maatschappij (DUM) - VSSD, 1983) in Dutch.

[7] H. C. Pouls, *De landmeter Jan Pietersz. Dou en de Hollandse Cirkel*, Green series, Vol. 41 (Nederlandse Commissie voor Geodesie (NCG), Delft, 2004) in Dutch.

[8] J. P. Dou, *Tractaet Vant Maken Ende Gebruycken Eens Nieu Gheordonneerden Mathematischen Instruments: In Welcke Verscheyden Konstighe Stucken (de Geometrie Betreffende) Vervatet Ende Begrepen Zijn* (Willem Jansz Blaeu, Amsterdam, 1620).

[9] *Wikimedia Commons,* (n.d.), media file repository, for public domain and freely licensed educational media content.

[10] Anonymous, *Hans Lipperhey,* Wikipedia The Free Encyclopedia (2022), dated Jun 20.

[11] S. K. Roy, *Fundamentals of surveying*, 2nd ed. (Prentice Hall, 2011).

[12] Anonymous, *Hero of Alexandria,* Wikipedia The Free Encyclopedia (2021), dated Oct 19.

[13] Kotsanas, *Photo of dioptra,* Katakolo, Ilia, Greece (n.d.), Kotsanas Museum of Ancient Greek Technology.

[14] W. Leybourn, *The compleat surveyor: or, the whole art of surveying of land, by a new instrument lately invented*, 5th ed. (Samuel Ballard at the Blue-Ball in London (and others), 1722).

[15] Rijksmuseum Boerhaave, *Photo of Hollandse Cirkel manufactured by Kleman en Zoon, Amsterdam (1825-1830),* Leiden, The Netherlands (2017).

[16] Museum of Lands, Mapping and Surveying, *Photo of spirit level used in the 1840s, artefact gallery, heighting instruments,* Queensland, Australia (2018).

[17] Johnson Level and Tool Mfg Company, *Photo of 22x builder's level, model 40-6960,* Mequon, Wisconsin (2021).

[18] Smithsonian National Museum of American History, *Photo of theodolite manufactured by J. Gilbert, Tower Hill, London,* Behring Center, Washington D.C. (n.d.).

[19] South Surveying and Mapping Technology Co., Ltd., *Photo of South 6N+ total station and photo of South DT-02L electronic theodolite,* Guangzhou, China (2021).

[20] *Leica TPS-System 1000,* Leica Geosystems AG, Heerbrugg, Switzerland (1998), user manual.

[21] *Leica TC605/TC805/TC905/L,* Leica Geosystems AG, Heerbrugg, Switzerland (1998), user manual.

[22] *Surveying made easy,* Leica Geosystems AG, Heerbrugg, Switzerland (2013).

[23] A. Hald, *A history of mathematical statistics from 1750 to 1930* (Wiley-Interscience, 1998).

[24] G. Székely, *Paradoxa: Klassische und neue Uberraschungen aus Wahrscheinlichkeit-srechnung and mathematischer Statistik* (Akadémiai Kiadó, Budapest, 1990) translated in German.

[25] A. Smits, *Portrait photo of Willem Baarda,* (n.d.), Delft University of Technology.

[26] *Accuracy,* Oxford Learner's Dictionaries (2021), Oxford University Press.

[27] D. Lay, *Linear algebra and its applications*, 4th ed. (Pearson, 2014).

[28] Rijkswaterstaat, *Beeldbank,* (n.d.), website, Ministerie van Infrastructuur en Waterstaat.

[29] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars, *Computational Geometry: Algorithms and Application*, 3rd ed. (Springer Verlag, 2008).

[30] M. Oliver and R. Webster, *Basic Steps in Geostatistics: The Variogram and Kriging* (SpringerBriefs in Agriculture, 2015).

[31] P. Teunissen and O. Montenbruck, eds., *Springer handbook of Global Navigation Satellite Systems*, Springer Handbooks (Springer Verlag, 2017).

[32] J. Morton, F. van Diggelen, J. Spilker, and B. Parkinson, eds., *Position, Navigation, and Timing Technologies in the 21st Century: Integrated Satellite Navigation, Sensor Systems, and Civil Applications, Volume 1 and 2* (Wiley - IEEE Press, 2021).

[33] Boeing, *Image of GPS block IIF satellite,* Chicago, Illinois (n.d.).

[34] *GSA GNSS Market Report*, Tech. Rep. (European Union Agency for the Space Programme (EUSPA), 2019).

[35] H. Verhoef, *Photo of Peter Teunissen,* (n.d.).

[36] R. Odolinski, *Graph of 3D GNSS position solutions in terms of East-North scatter, and Up time-series,* University of Otago, New Zealand (n.d.).

[37] A. Smits, *Visualization of GPS positioning,* (n.d.), Delft University of Technology.

[38] 06-GPS, *Image of network of permanent GPS tracking stations in the Netherlands, Belgium and Luxemburg,* Sliedrecht, The Netherlands (2021).

[39] Heijmans N.V., *Photo of excavator in the process of constructing a motorway embankment, guided by RTK-GPS,* Rosmalen, The Netherlands (n.d.).

[40] T. Takasu, *RTKLIB: An Open Source Program Package for GNSS Positioning,* (n.d.), website.

[41] Bundesamt für Kartographie und Geodäsie (BKG), *The BKG GNSS Data Center — BKG Ntrip Client (BNC),* (n.d.), website of the German Federal Agency for Cartography and Geodesy.

[42] M. de Schipper, *Photo of RTK-GPS positioning with a quad on shore, and a jet-ski in the water, at the Zandmotor,* (n.d.).

[43] United States Coast Guard, *Navigation Center,* (n.d.), website of the United States Coast Guard (USCG), U.S. Department of Homeland Security, Navigation Center.

[44] U.S. government, *Official U.S. government information about the Global Positioning System (GPS) and related topics,* (n.d.), GPS.gov website.

[45] *European Union Agency for the Space Programme (EUSPA),* (n.d.), website.

[46] *International GNSS Service (IGS),* (n.d.), website.

[47] Anonymous, *Willebrord Snellius,* Wikipedia The Free Encyclopedia (2021), dated Jul 18.

[48] Anonymous, *Christiaan Huygens,* Wikipedia The Free Encyclopedia (2021), dated Sep 24.

[49] Beeldmateriaal Nederland, *Aerial photograph of part of the TU Delft campus,* Amersfoort, The Netherlands (2021).

[50] W. Förstner and B. Wrobel, *Photogrammetric Computer Vision - Statistics, Geometry, Orientation and Reconstruction*, Geometry and Computing book series, Vol. 11 (Springer, 2016).

[51] *OpenStreetMap,* (n.d.), available under the Open Database License, ©OpenStreetMap contributors.

[52] P. Tipler, *Physics for scientists and engineers* (Freeman-Worth Publishers Inc., New York, 1991).

[53] *Deutsches Zentrum für Luft- und Raumfahrt (DLR),* (n.d.), German Aerospace Center - Oberpfaffenhofen.

[54] C. Audoin and B. Guinot, *The measurement of time - time, frequency and the atomic clock* (Cambridge, 2001).

[55] National Metrology Institute van Swinden Laboratorium (VSL), *Photo of Netherlands national time standard control room,* Delft, The Netherlands (2020).

[56] L. Truong-Hong, *Photo of de Trambrug in Schipluiden,* (2019).

[57] L. Truong-Hong, *Image of point cloud obtained with laser scanning, as 3D-model of de Trambrug in Schipluiden,* (2019).

[58] R. Lindenbergh, *Image of point cloud obtained with laser scanning, of the Symbio-bridge in the Delft Technopolis Science Park area,* (2018).

[59] Kadaster, *Publieke Dienstverlening Op de Kaart (PDOK),* (n.d.), data sets and maps available under CC BY 4.0 license.

[60] Rijkswaterstaat, *Actueel Hoogtebestand Nederland (AHN),* (n.d.), website.

[61] R. Hanssen, *Radar interferometry - data interpretation and error analysis* (Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001).

[62] F. van Leijen, *Persistent scatterer interferometry based on geodetic estimation theory*, Publications on Geodesy, Vol. 86 (Nederlands Centrum voor Geodesie en Geo-informatica (NCG), 2014).

[63] Nederlands Centrum voor Geodesie en Geo-informatica (NCG), *Bodemdalingskaart,* (2021), website.

[64] Nederlands Centrum voor Geodesie en Geo-informatica (NCG), *Image of Bodemdalingskaart Nederland,* (n.d.).

[65] R. Jaarsma, *Photo of EA400 and AA400 echo-sounder (Kongsberg), and screenshot of depth measurements in Stevin III-lab,* (2019), in context of bachelor final thesis project.

[66] Rijkswaterstaat, *Morfologie - Waterinfo Extra - Monitoring,* (n.d.), website.

[67] E. Peeters, *Photo of calibration dock for acoustic surveys, and screenshot of acoustic survey results,* (2012), Van Oord Dredging and Marine Contractors, Rotterdam, The Netherlands.

[68] European Space Agency (ESA), *Sentinel Online,* (2021), website.

[69] W. Rees, *Physical principles of remote sensing*, 3rd ed. (Cambridge University Press, 2013).

[70] C. van Uffelen, *Photo of 52 degrees North latitude marker-line on TU Delft campus,* (2018), Delta 27 August.

[71] C. van Uffelen, *Why is a blue line running across campus?* (2018), Delta 27 August.

[72] B. Bowring, *Transformation from spatial to geographical coordinates,* Survey Review **23**, 323 (1976).

[73] C. de Jong, G. Lachapelle, S. Skone, and I. Elema, *Hydrography*, 2nd ed., Series on Mathematical Geodesy and Positioning (VSSD, Delft, 2010).

[74] *OGP publication 373-7-1 – Guidance Note 7 Part 1: Using the EPSG geodetic parameter dataset,* (2012), version 8.

[75] *OGP publication 373-7-2 - Guidance Note 7 Part 2: Coordinate conversions and transformations including formulas,* (2021), version 61.

[76] Anonymous, *Well-known text representation of coordinate reference systems,* Wikipedia The Free Encyclopedia (2021), dated Aug 29.

[77] PROJ contributors, *PROJ coordinate transformation software library,* (2021), Open Source Geospatial Foundation.

[78] *European Space Agency (ESA),* (2004), image of the Earth's gravity field (geoid) as it will be seen by GOCE.

[79] National Geospatial-Intelligence Agency, *WGS 84 DATA & APPS,* (2021), webpage.

[80] W. Heiskanen and H. Moritz, *Physical geodesy* (W.H. Freeman, 1967).

[81] *Bundesamt für Kartographie und Geodäsie (BKG),* (2021), website of the German Federal Agency for Cartography and Geodesy, image of transformation parameters from national heights in Europe to EVRF2019 in cm and reference tide gauges.

[82] Z. Altamimi, X. Collilieux, and L. Métivier, *ITRF2008: an improved solution of the International Terrestrial Reference Frame,* Journal of Geodesy **85**, 457 (2011).

[83] *International Earth Rotation and Reference Systems Service (IERS),* (2013), website.

[84] Z. Altamimi, X. Collilieux, and L. Métivier, *Analysis and results of ITRF2008*, Technical Note 37 (International Earth Rotation and Reference Systems Service (IERS) - BKG, 2012).

[85] G. Petit and B. Luzum, eds., *IERS Conventions (2010)* (Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt am Main, 2010) IERS Technical Note 36, all software and material associated with IERS Conventions (2010) can be found at the IERS Conventions website.

[86] *EUREF Permanent GNSS Network (EPN) - EPN Central Bureau,* (2021), Royal Observatory of Belgium (ROB).

[87] *EuroGeographics,* (2021), representing Europe's National Mapping, Cadastral and Land Registration Authorities, website.

[88] H. van der Marel, *Dutch Permanent GNSS Array (DPGA),* (n.d.), DPGA website, Delft University of Technology.

[89] *Nederlandse Samenwerking Geodetische Infrastructuur (NSGI),* (2021), website.

[90] Z. Altamimi, *EUREF Technical Note 1 : Relationship and transformation between the International and the European Terrestrial Reference Systems,* (2018), version June 28, 2018.

[91] 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*, Green series, Vol. 43 (Netherlands Geodetic Commission (NCG), Delft, 2005) also available in Dutch.

[92] H. Heuvelink, *Stereographic projection and its application in the Rijksdriehoeksmeting* (Rijkscommissie voor Graadmeting en Waterpassing, Delft, 1918) in Dutch.

[93] P. Wijnterp, *De eerste nauwkeurigheidswaterpassing van Nederland (1875-1885)*, Tech. Rep. MDTNO-R-9231 (Meetkundige Dienst, Directoraat-Generaal Rijkswaterstaat, Ministerie van Verkeer en Waterstaat, 1993) available through Rijkswaterstaat Rapportendatabank, in Dutch.

[94] W. Ankersmit, R. Hagman, H. Heijmans, G. Olsder, G. van de Schootbrugge, and P. van Woerkom, eds., *175 jaar TU Delft: Erfgoed in 33 verhalen* (Histechnica, 2017) chapter by W.A. van Beusekom 'Het waterpas instrument van Caminada: ontwerp van L. Cohen Stuart' (in Dutch).

[95] Rijkswaterstaat, *Normaal Amsterdams Peil (NAP),* (2021), webpage.

[96] C. Slobbe, *NLGEO2018 & NLLAT2018 – berekeningsstrategie en nauwkeurigheid,* (2018), Presentation at Vertical reference frame for the Netherlands mainland, Wadden islands and continental shelf (NEVREF) workshop, Rijkswaterstaat, Utrecht, June 27.

[97] Anonymous, *Johannes Vermeer,* Wikipedia The Free Encyclopedia (2021), dated Sep 26.

[98] Anonymous, *Multilingual dictionary of technical terms in cartography,* International Cartographic Association (ICA), Steiner, Wiesbaden (1973).

[99] Anonymous, *Charles Joseph Minard,* Wikipedia The Free Encyclopedia (2021), dated Mar 10.

[100] Anonymous, *Piet Mondrian,* Wikipedia The Free Encyclopedia (2021), dated Sep 21.

[101] J. Woodall, *Reality, distorted,* Prospect (2013).

[102] Anonymous, *Rotterdamse metro,* Wikipedia The Free Encyclopedia (2021), dated Sep 18 (in Dutch).

[103] K. Garland and H. Beck, *Mr Beck's Underground Map* (Capital Transport, 1994).

[104] A. R. Jones, *Ptolemy,* Encyclopedia Britannica (2021), dated Jan 20.

[105] C. Ptolemy, *Ptolemy World Map,* (1450–1475), translated by Emanuel Chrysoloras and Jacobus Angelus, housed by British Library.

[106] Anonymous, *Mercator 1569 world map,* Wikipedia The Free Encyclopedia (2021), dated Sep 18.

[107] Anonymous, *Gerardus Mercator,* Encyclopedia Britannica (2021), dated Mar 1.

[108] J. B. Harley and D. Woodward, eds., *Cartography in Prehistoric, Ancient, and Medieval Europe and the Mediterranean*, The History of Cartography, Vol. 1 (Chicago and London: The University of Chicago Press, 1987).

[109] J. B. Harley and D. Woodward, eds., *Cartography in the Traditional Islamic and South Asian Societies*, The History of Cartography, Vol. 2 (Chicago and London: The University of Chicago Press, 1992).

[110] J. B. Harley and D. Woodward, eds., *Cartography in the Traditional East and Southeast Asian Societies*, The History of Cartography, Vol. 2 (Chicago and London: The University of Chicago Press, 1995).

[111] J. B. Harley and D. Woodward, eds., *Cartography in the Traditional African, American, Arctic, Australian, and Pacific Societies*, The History of Cartography, Vol. 2 (Chicago and London: The University of Chicago Press, 1998).

[112] J. B. Harley and D. Woodward, eds., *Cartography in the European Renaissance*, The History of Cartography, Vol. 3 (Chicago and London: The University of Chicago Press, 2007).

[113] M. Monmonier, ed., *Cartography in the Twentieth Century*, The History of Cartography, Vol. 6 (Chicago and London: The University of Chicago Press, 2015).

[114] Stichting RIONED, *Underground infrastructure map,* Ede, The Netherlands (2018), fragment of sewerage network around the Reeverweg in Harfsen.

[115] F. Biljecki, *Level of detail in 3D city models*, Ph.D. thesis, Delft University of Technology, Delft, The Netherlands (2017).

[116] J. K. Wright, *Notes on statistical mapping: with special reference to the mapping of population phenomena,* (American Geographical Society, 1938) p. 37.

[117] M. Kraak and F. Ormeling, *Cartography: Visualization of Geospatial Data*, 3rd ed. (Routledge Taylors and Francis Group (CRC Press) - London and New York, 2010).

[118] GADM, *Boundary geometry for the Netherlands (data-set),* (2021), maps and data.

[119] Koninklijk Nederlands Meteorologisch Instituut (KNMI), *Precipitation - long term average 1981-2010 - average yearly precipitation (data-set),* De Bilt, The Netherlands (n.d.), KNMI Data Platform (KDP).

[120] J. Bertin, *Sémiologie graphique: les diagrammes, les réseaux, les cartes*, 4th ed. (Editions de l'Ecole des hautes études en sciences sociales Paris (EHESS), 2013) first edition in 1967 (in French).

[121] Centraal Bureau voor de Statistiek (CBS), *CBS Wijken en Buurten (data-set),* (2018), Statistics Netherlands, Heerlen, The Netherlands.

[122] Anonymous, *John Snow,* Wikipedia The Free Encyclopedia (2021), dated Sep 13.

[123] J. Snow, *On the mode of communication of cholera*, 2nd ed. (John Churchill, New Burlington Street, London, England, 1855).

[124] E. Dijkstra, *A note on two problems in connexion with graphs,* Numerische Mathematik **1**, 269 (1959).

[125] J. Rüeger, *Electronic Distance Measurement - an introduction*, 4th ed. (Springer-Verlag, 2012).

[126] *National Marine Electronics Association (NMEA),* (n.d.), website.

[127] W. Gurtner and L. Estey, *RINEX: The Receiver Independent Exchange Format Version 2.11*, Tech. Rep. (Astronomical Institute, University of Bern, 2007).

[128] H. Blok and P. van den Berg, *Electromagnetic waves - an introductory course* (Delft University Press, 1999).

[129] L. Couch, *Digital and analog communication systems*, seventh ed. (Pearson Prentice Hall, Upper Saddle River, NJ, 2007).

[130] *Bureau International des Poids et Mesures (BIPM),* (n.d.), website.

[131] *INSPIRE Infrastructure for Spatial Information in Europe,* (2021), website.

[132] Google LLC, *Google Earth,* (n.d.), Mountain View, California.

[133] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, *Google Earth Engine: Planetary-scale geospatial analysis for everyone,* Remote Sensing of Environment (2017), 10.1016/j.rse.2017.06.031.

# Surveying and Mapping

*Christian Tiberius, Hans van der Marel, René Reudink & Freek van Leijen*

This book provides an introduction, at academic level, into the field of surveying and mapping. The book has been compiled based on hand-outs and readers written for the third-year course Surveying and Mapping, in the bachelor program Civil Engineering at Delft University of Technology. This book covers a wide range of measurement techniques, from land surveying, GPS/GNSS and remote sensing to the associated data processing, the underlying coordinate reference systems, as well as the analysis and visualization of the acquired geospatial information.

TU Delft | Civil Engineering & Geosciences

*Christian Tiberius has over 20 years of teaching experience at Delft University of Technology in the field of mathematical geodesy and positioning. His research interests are high-accuracy, high-integrity positioning by GNSS satellite navigation and terrestrial radio positioning.*

*Hans van der Marel is known for his work on permanent GNSS networks for accurate positioning, deformation monitoring and meteorology, and has over 30 years' experience in geometric infrastructure and reference systems. Currently he focuses on data processing and integration of GNSS and InSAR for land subsidence monitoring.*

*René Reudink is a researcher in the field of gravimetry as a cornerstone for vertical reference. He also has expertise in land surveying education and a keen interest in the development of measurement sensors and instruments.*

*Freek van Leijen is specialized in time series analysis of InSAR remote sensing. Applications of his research lie in the monitoring of infrastructure and analysis of surface motion due to mining.*

TU Delft


ISBN 978-94-6366-489-9
DOI: https://doi.org/10.5074/T.2021.007

**textbooks.open.tudelft.nl**

Cover image shows a `true color' remote sensing image of the South-Western part of the Netherlands. Data from ESA Sentinel-2 satellite (Copernicus Sentinel data 2019) obtained through Sentinel-Hub [1], CC BY-NC 4.0, taken on August 24th, 2019 (10:56:43 UTC)