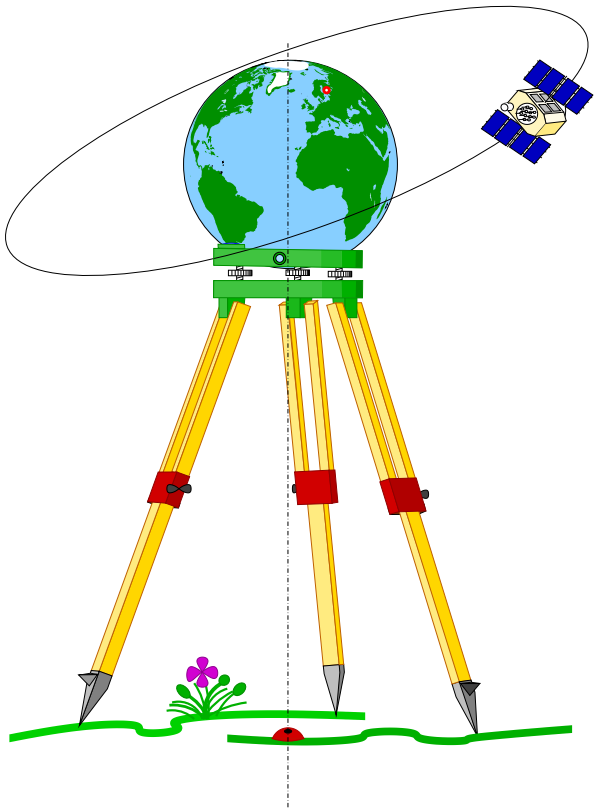
Geodesy



Martin Vermeer
martin.vermeer@aalto.fi

Preface

Although Finland is, and has been since independence, a superpower in the field of geodesy, there doesn't seem to be a modern geodesy textbook in the Finnish language area. Finnish-language textbooks, and popular books as well, do exist, but they are either badly outdated or treat only a certain sub-area of geodesy. Of these, we may mention the writings by Martti Tikka on measurement and instrument techniques and geodetic computation (Tikka, 1991, 1985), now in large part obsolete, and Salmenperä (1998). The work on satellite positioning by Poutanen (1998) has now been updated (2016) and is very useful. The book by Kallio (1998) explains the least-squares statistical computation technique used in geodesy. All these sources have been helpful in writing this book.

Internationally, clearly more geodesy textbooks are on offer, and we have benefited from Torge (2001), Vaníček and Krakiwsky (1986), in measurement and instrumental techniques Kahmen and Faig (1988), in physical geodesy Heiskanen and Moritz (1967), and with satellite geodesy, Hofmann-Wellenhof et al. (2001).

The material in this book divides naturally into two parts: classical geodesy, and modern geodesy. Each could be the textbook of its own course, which would each be worth three ECTS points.

The subjects discussed in the *classical geodesy* part (chapters 1 – 9) are the history of geodesy, the figure of the Earth and gravity, the reference ellipsoid, co-ordinates and heights, basics of geodetic measurements, units of measurement, uncertainty of measurement; Helmert transformations, the direct and inverse geodetic problems; levels and levelling, height systems, the geoid; theodolites and total stations, angle measurements; distance measurement using electromagnetic radiation and propagation of the measurement ray in the atmosphere; geodetic networks, measurement classes, network hierarchy; base and mapping measurements; area and volume calculations.

In the part on *modern geodesy* (chapters 10 – 19) we concentrate instead on the development, during the past century, from two-dimensional geo-

desy on the Earth's surface to genuinely three-dimensional geodesy, comprising space and satellite geodesy and truly three-dimensional positioning methods based on electromagnetism. We discuss the basics of three-dimensional reference systems, hyperbolic positioning systems and the global positioning system GPS; GPS satellites, orbits, signals, receivers; measurements of pseudo-range and carrier phase, measurement geometry, differencing of observations, integer-valued ambiguitues and their fixing; processing GPS observations, relative and differential as well as real-time positioning. We dive also deeper into the statistical foundations of geodesy, like the least squares method, residuals, statistical testing, outlier detection, reliability, planning of measurement networks. Finally, we look at the borderlands between geodesy and geophysics, comprising the gravity field of the Earth and the gravimetric geoid; space geodesy, the rotational and orbital motions and deformations of the Earth; satellite orbits and the role of geodesy in geophysical research.

We have chosen in this text to concentrate on conceptual and fundamental matters. That also means describing the internal workings of instruments and processes which are in today's systems handled automatically by smart software – even if to some, this may feel like we are teaching outdated skills.

Helsinki, February 12, 2018,

Martin Vermeer
martin.vermeer@aalto.fi

Acknowledgements

In drafting the various manuscript versions, the lecturing transparencies and other materials by professors Teuvo Parm and Martti Martikainen were a great help. Assistant professor Jaakko Santala, teaching assistants Mauri Väisänen, Panu Salo and Henri Turto, professor Markku Poutanen, National Land Survey chief expert Marko Ollikainen, as well as students in many years of teaching are gratefully acknowledged for many useful remarks, corrections, and pieces of information.

Nicolàs de Hilster is gratefully acknowledged for contributing figure 8.8.

Contents

List of Figures

List of Tables

Abbreviations

AGU	American Geophysical Union
ANNA	("Army, Navy, NASA, Air Force"), 1962-060A, geodetic satellite
APPS	Automatic Precise Positioning Service (JPL)
ARIA	Advanced Rapid Imaging and Analysis (JPL and Caltech)
ARP	antenna reference point
ATR	Automatic Target Recognition
AUSPOS	Australian on-line GPS processing service
BGI	Bureau Gravimétrique International, International Gravimetric Bureau
BIFROST	Baseline Inferences for Fennoscandian Rebound, Sea level, and Tectonics
BIH	Bureau International de l'Heure, International Time Bureau
CAD	computer-aided design
CalTech	California Institute of Technology
CCD	charge-coupled device, image sensor type
CDMA	code division multiple access
CHAMP	200039B, Challenging Minisatellite Payload, German satellite
CIO	Conventional International Origin, reference pole for polar motion
Decca	marine navigation system
DEM	digital elevation model
DGPS	differential GPS
DHM	digital height model
DOP	dilution of precision, a measure for the geometric strength of satellite positioning
DORIS	Doppler Orbitography and Radiopositioning Integrated by Satellite, a French satellite positioning system
DTM	digital terrain model
DWT	discrete wavelet transform
ED50	European Datum 1950

EET Eastern European Time
EEST Eastern European Summer Time
ECEF Earth-centred, Earth-fixed
EGM2008 Earth Gravity Model 2008
EGM96 Earth Gravity Model 1996
EGNOS European Geostationary Navigation Overlay System
EOP Earth orientation parameters
ETRF European Terrestrial Reference Frame
ETRS, ETRS89
European Terrestrial Reference System. Coincides with ITRS for the epoch 1989.0
ERTS-GK*n*
Gauss-Krüger projection for Finland, central meridian n degrees East
ETRS-TM35FIN
UTM projection for Finland, zone 35
EUREF IAG Reference Frame Subcommission for Europe
EUREF89
First European realization of ETRS89
EUREF-FIN
Finnish national realization of ETRS89
FAT file allocation table (file system)
FDMA frequency division multiple access
FGI Finnish Geodetic Institute, 1918 – 2015, Finnish Geospatial Research Institute, 2015 –
FIN2000 Finnish geoid model
FIN2005N00
Finnish geoid model
FRS fellow of the Royal Society (of London)
FRSE fellow of the Royal Society of Edinburgh
USGS US Geological Survey
GAST Greenwich Apparent Sidereal Time
GCP ground control point
GDGPS Global Differential GPS (JPL)
GIA glacial isostatic adjustment
GKT geodeettinen käänteistehtävä, inverse geodetic problem
GLONASS
Global Navigation Satellite System (Russian)
GMT Greenwich Mean Time
GNSS Global Navigation Satellite Systems, generic name
GOCE Gravity Field and Steady-State Ocean Circulation Explorer
GPS Global Positioning System
GPT geodeettinen päätehtävä, forward geodetic problem
GPU Geopotential Unit, $10\,\mathrm{m}^2/\mathrm{s}^2$

GPS/MET
95017C, GPS radio occultation satellite mission

GPU geopotential unit

GRACE 0201201+0201202, Gravity Recovery and Climate Experiment. Satellite pair

GRS80 Geodetic Reference System 1980

GSI Geospatial Information Authority (Japan)

IAG International Association of Geodesy

IB inverted barometer

IERS International Earth Rotation and Reference Systems Service

InSAR interferometric SAR

IGS International GNSS Service

ITRF International Terrestrial Reference Frame

ITRS International Terrestrial Reference System

IONEX Ionosphere Map Exchange Format

Jason 1 – 3
radar altimetric satellites

JHS julkisen hallinnon suositukset, Recommendations for Public Administration

JPL Jet Propulsion Laboratory, NASA

JPEG 2000
image format

JUHTA julkisen hallinnon tietohallinnon neuvottelukunta, the Advisory Board on Data Management in Public Administration

KKJ Kartastokoordinaattijärjestelmä, the Finnish National Map Grid Co-ordinate System

KM2, KM10
Finnish national terrain models

LAGEOS 1 – 2
Laser Geodynamics Satellite.

LAST local apparent sidereal time

LoD Length of Day

LORAN-C
marine navigation system

MHD magnetohydrodynamics

MIF member of the French Academy of Sciences (technically, of the *Institut de France*)

MRI magnetic resonance imaging

MSAS Multi-functional Satellite Augmentation System

N60 Finnish height system, epoch 1960.0

N2000 Finnish height system, epoch 2000.0

NASA National Aeronautics and Space Administration, US

NLS National Land Survey of Finland

NOAA National Oceanic and Atmospheric Administration, US

NTRIP Networked Transport of RTCM via Internet Protocol
NUVEL-1A
global plate-motion model
NWP numerical weather prediction
Omega marine navigation system
PAGEOS 1966-056A, Passive Geodetic Earth Orbiting Satellite
PCMCIA memory card bus standard
RINEX Receiver-Independent Exchange Format
RTK real-time kinematic positioning
SAR synthetic-aperture radar
SBAS satellite based augmentation systems
Seasat 1978-064A, radar altimetric satellite
SLR satellite laser ranging
SP3 Standard Product 3, precise ephemeris data format
SRTM Shuttle Radar Topography Mission
TAI International Atomic Time
TEC total electron content
TIN Triangulated Irregular Network
TOPEX/Poseidon
92052A, radar altimetric satellite
USB Universal Serial Bus
UTC Universal Time Co-ordinated
UTM Universal Transverse Mercator (map projection)
VLBI very long baseline interferometry
VRS-RTK virtual reference station RTK
WAAS Wide Area Augmentation System
WADGPS wide area differential GPS
WGS84 World Geodetic System 1984
YKJ Yhtenäiskoordinaatisto, Uniform Co-ordinate System

1. The history and societal status of geodesy

1.1 The figure of the Earth, early conceptions

In traditional societies, undoubtedly the most common conception of the figure of the Earth was, that the Earth is a flat disc extending to the horizon, with the sky doming over her. On the inner surface of the dome, the celestial bodies describe their complicated orbits. Also children have generally the same conception. Only with formal education does this "naïve world model" give way. Psychologically, from the viewpoint of childhood development, this is by no means an easy process, surely as difficult as it was back in time for all of society, historically speaking.

However, already the antique Hellenes were aware of the spherical shape of the Earth. Free of preconceptions, they had observed how, during a lunar eclipse, the Earth cast her shadow on the surface of the Moon. They also observed that a lunar eclipse that was high in the sky on one end of the Mediterranean, happened near the horizon at the other end. Assuming that this was one and the same event, this could only mean that the Earth's surface must be curved at least in the East-West direction.

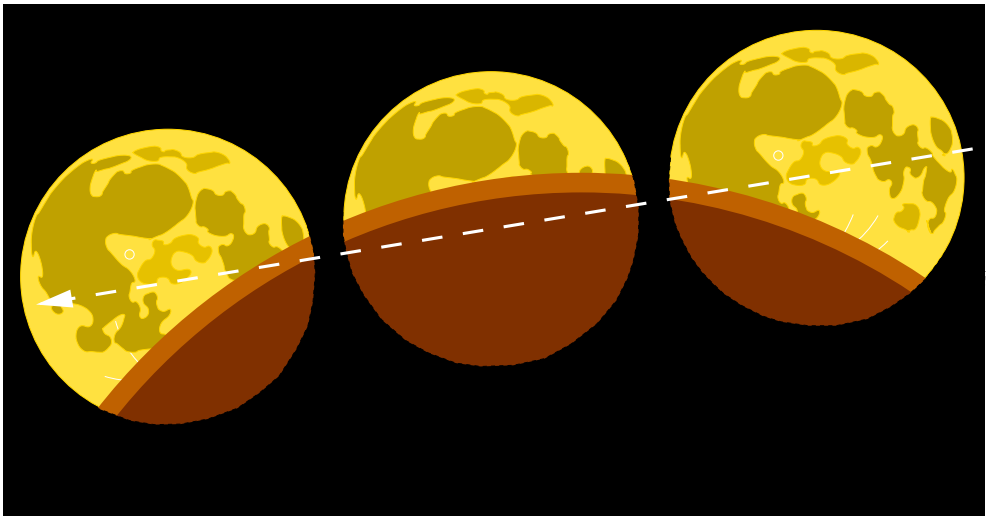


Figure 1.1. A lunar eclipse. The shape of the shadow, always circular, shows that the Earth must be a sphere.

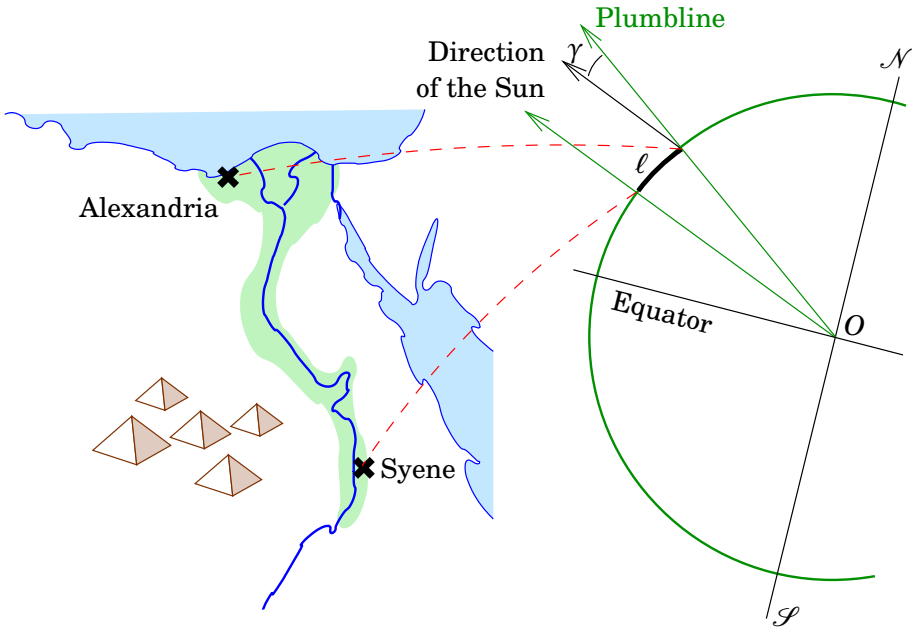


Figure 1.2. The grade measurement of Eratosthenes.

Eratosthenes, the "father of geography", lived 276 – 195 BCE[1]. He was the first to measure the size, or radius, of this spherical Earth. The measurement was the same in principle as the later grade measurements: measure the length of an arc on the surface of the Earth by geodetic means, and the difference in direction between the plumb lines at the ends of the arc by astronomical means. By combining the length ℓ of the arc and the difference in plumb-line directions γ one obtains for the radius of the Earth:

$$R = \frac{\ell}{\gamma}.$$

See figure 1.2.

The information on the directions of the plumb lines was obtained from the midsummer Sun, which in Syene (today's Assuan) did not throw any shadows at all[2]. In Alexandria, on the other hand, the Sun was not in the zenith but, based on the lengths of shadows, some 1/50th part of a circle more to the South. Eratosthenes obtained for the radius of the Earth[3] 6317 – 7420 km – pretty close to the current best value 6371 km.

More information can be found in Torge (2001, pages 5 – 6).

The principle of *triangulation*, so important in geodesy, that, in a network consisting of triangles, the geometry may be uniquely determined,

[1] "Before the common (or Christian) era".

[2] The story that he used the circumstance that the Sun illuminated the bottom of a well, is apparently a misunderstanding (Dreyer, 1914).

[3] In fact he obtained his results in a unit called the *stadium*, the length of which varies. The length used by Eratosthenes is controversial.

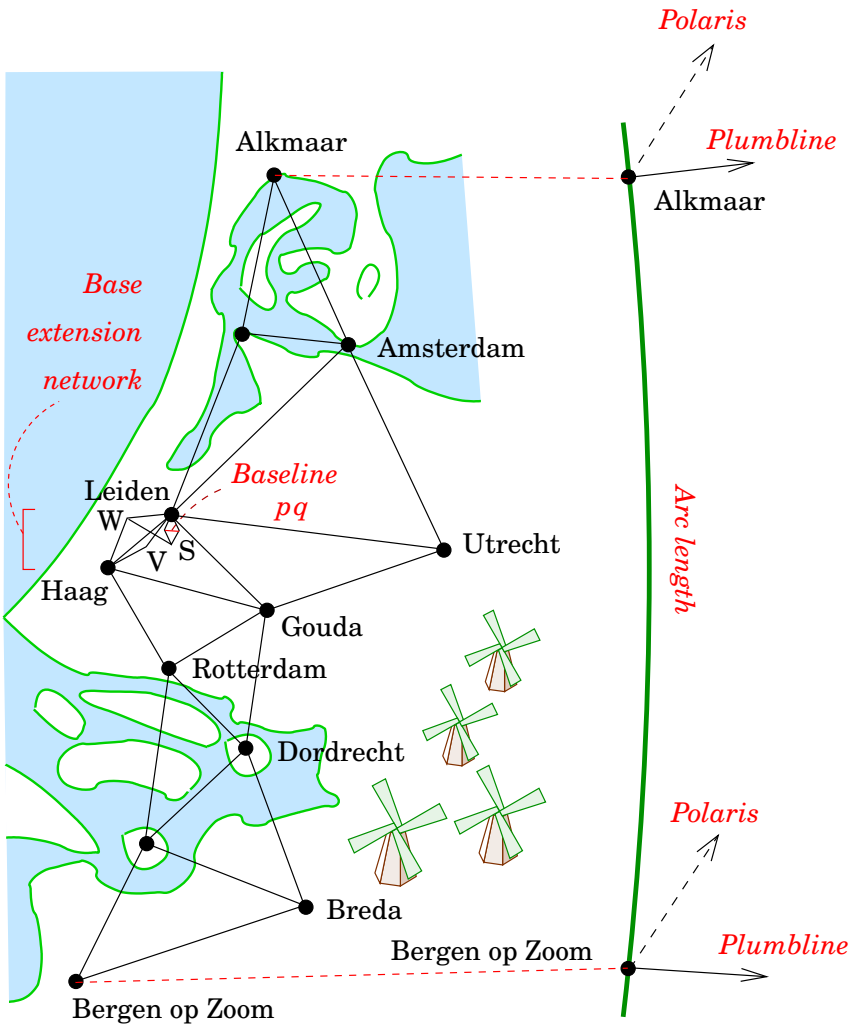


Figure 1.3. The Snellius grade measurement. The length of baseline pq is 326.45 *roeden* (1229 m) . This length was derived through a local base extension network from the only measured length, the original baseline of length 87.05 *roeden* (328 m) (personal comm. L. Aardoom).

□

if, in addition to the angles of the triangle, only *one distance* is measured, was presumably discovered by Gemma Frisius[4] in 1533 (Crane, 2002, pages 56 – 57).

The use of the method for grade measurement happened also for the first time in the Netherlands, using the numerous church towers dotting the prosperous but flat country. *Snellius*[5] was among the first who used triangulation to determine the length of an arc ℓ. By measuring in the network one length, and otherwise only angles, he managed to determine the distance between two cities, Bergen op Zoom and Alkmaar, although

[4]Gemma Frisius (1508 – 1555) was a Dutch polymath.

[5]Willebrord Snell van Rooyen (1580 – 1626) was a Dutch astronomer and mathematician.

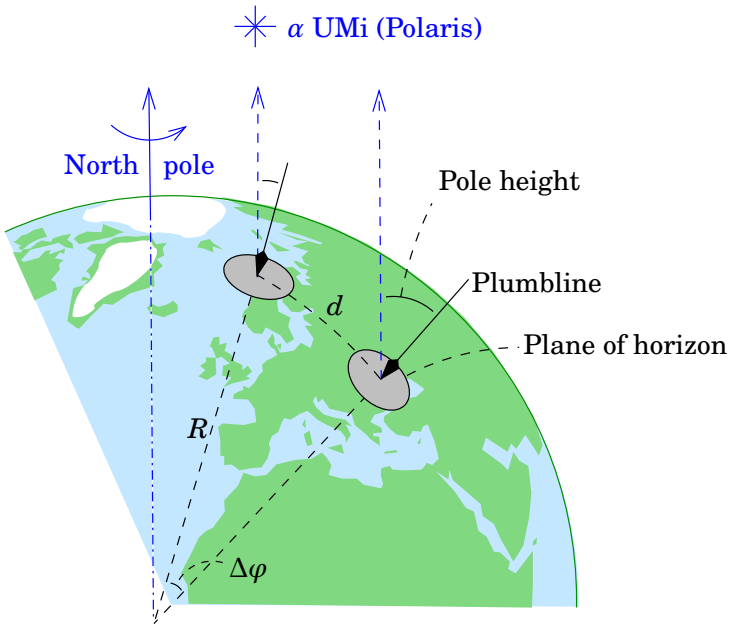


Figure 1.4. Astronomically determining the difference of plumb-line directions. From the direction difference and the metric distance d on can determine the radius of curvature of the Earth $R = d/\Delta\varphi$.

the cities are separated by the broad river branches of the Rhine delta. See figure 1.3.

The secret of triangulation is that with the aid of angle measurements one can build, either computationally or graphically, a *scale model* of the whole measurement network, where all proportions are correct. To determine the true scale it suffices to measure just *one* distance in the model also in reality. In the case of Snellius, this was the distance pq, in the meadow by Leiden, a *baseline* of only 326 "roeden".

With the aid of *astronomical position determination* one may measure the difference between the directions of the plumb line in two locations, see figure 1.4. When we travel along the meridian in the North-South direction, the *absolute direction* of the local plumb line, the direction with respect to the stars, changes. Also the local *plane of the horizon*, which is always perpendicular to the plumb line, the local direction of gravity, turns by the same amount when we travel in the North-South direction – or in whatever direction.

The direction in space of the rotation axis of the Earth is very stable due to the gyroscope phenomenon. It points to a place in the sky near the star α Ursae Minoris, or Polaris. With the aid of this star, the North direction can be found. The *latitude* Φ of a location is obtained by determining astronomically the height of the so-called *celestial pole* above the horizon. This is easiest to do using Polaris, though a precise determination is a bit more involved.

By thus measuring astronomically the difference in plumb-line direc-

tions between Alkmaar and Bergen of Zoom, and combining this with the metric distance obtained by triangulation, Snellius managed to determine the radius of curvature of the Earth. The method is referred to as *grade measurement*.

1.2 Newton's laws and the figure of the Earth

The understanding of the figure of the Earth made a great leap forward, when Newton[6] published in 1687 his main work, the *Principian* (*Philosophiæ Naturalis Principia Mathematica*, "The mathematical principles of natural philosophy [physics]"). In this opus he created the foundations of the whole of classical mechanics, including celestial mechanics.

The universal law of gravitation: Between two masses m_1, m_2 acts an *attraction* of size

$$F = G\frac{m_1 m_2}{r_{12}^2},$$

in which r_{12} is the distance separating the masses. The constant G is Newton's universal gravitational constant, the value of which is $6.672 \cdot 10^{-11}\,\mathrm{kg}^{-1}\mathrm{m}^3\mathrm{s}^{-2}$.

This attraction acts between *all* pairs of masses. So, not only does the Earth's attraction act on the Moon and the Sun's attraction on the Earth, but the Moon's attraction affects also the Earth, etc. In geophysics again we know that the attraction works between all *parts* of the Earth: sea, atmosphere, mountains all affect the gravitational field surrounding the Earth. And, because our Earth consists of materials that – however more of less reluctantly – deform under the influence of external force, gravitation also shapes the physical figure of the Earth.

In the *Principia* Newton calculated, using his famous laws, that a homogeneous, liquid Earth, in equilibrium and rotating once in 24 hours, with gravitation acting between its elements of liquid, would be *flattened* at the poles by the centrifugal force (figure 1.6). The definition of the flattening (oblateness) is

$$f = \frac{a-b}{a}, \tag{1.1}$$

in which a and b are the semi-major and semi-minor axes of the Earth ellipsoid, in other words, the equatorial and polar radii.

The theoretical flattening calculated by Newton was $f = 1/230$.

The assumption that the Earth is of homogeneous density is not correct. Christiaan Huygens calculated in 1690, by assuming that all the Earth's mass was concentrated in her centre, that the flattening would only be

[6] Sir Isaac Newton (1642 – 1727) was an English phyisicist and mathematician, the father of classical mechanics.

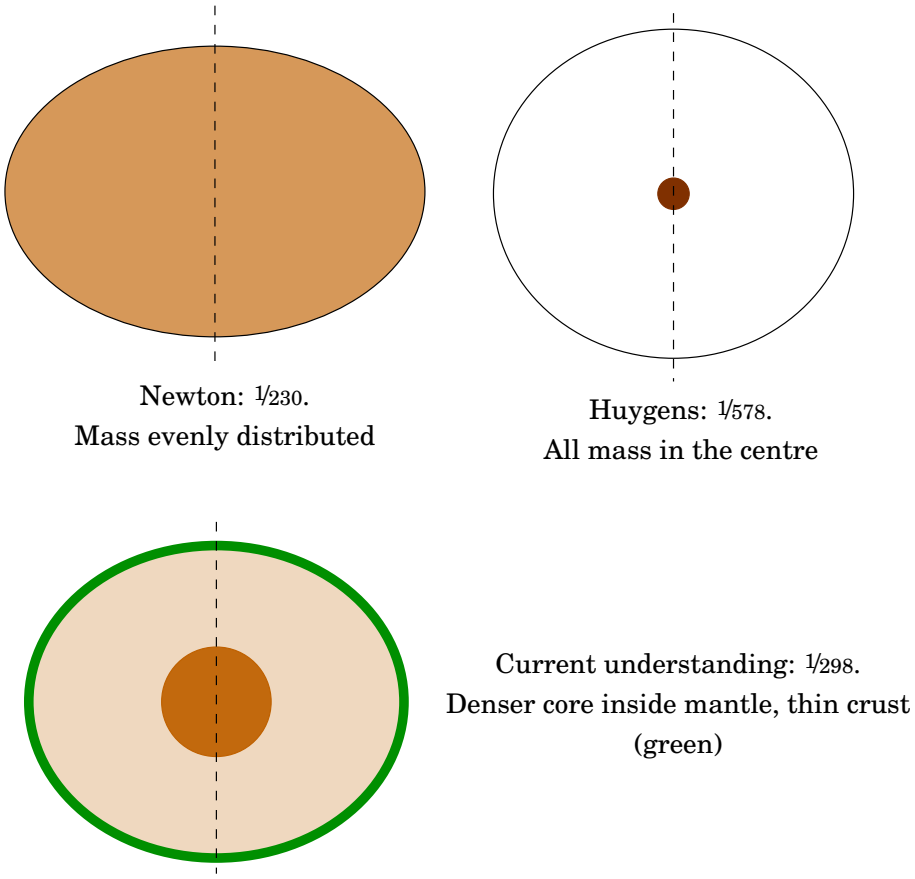


Figure 1.5. Different mass distribution models for the Earth, and their theoretical flattening values.

$f = 1/578$. As we know today, the truth lies between these two extremes: the density of the Earth's crust is about $2.7\,\mathrm{g/cm^3}$, that of the underlying mantle is $3.0-5.4\,\mathrm{g/cm^3}$, and the density of the iron core of the Earth is $10-13\,\mathrm{g/cm^3}$. The average density of the whole Earth is about $5.4\,\mathrm{g/cm^3}$. So, while the density increases a lot toward the centre of the Earth, a large part of the Earth's mass is nevertheless far from her centre.

In Newton's days there were influential scientists, like the astronomer Cassini[7], who believed that the Earth was in fact elongated like a rugby ball, $b > a$, and not flattened. An empirical answer to the question was needed!

The flattening issue remained unsolved until half a century later, when the French Academy of Sciences organized two expeditions, one to Finnish Lapland – then part of the Swedish empire – (1736 – 1737), the other to Peru, South America (1735 – 1744). The goal of the expeditions was to measure, by geodetic and astronomical means, the *length of a meridian arc of one degree* on two different latitudes, one close to the equator in

[7] Jean Dominique (Giovanni Domenico) Cassini (1625 – 1712) was an Italian-French astronomer, mathematician and engineer.

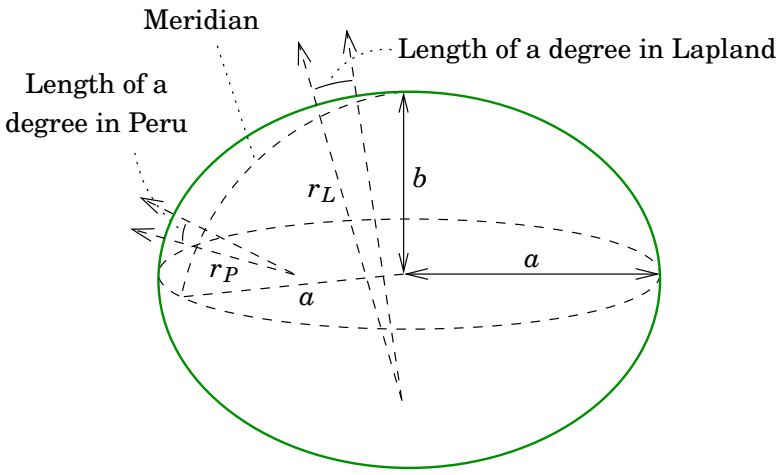


Figure 1.6. Parameters of an ellipsoid of revolution.

□

Peru, the other close to the North pole in Lapland in the Torne river valley. This was thus a similar *grade measurement* as Snellius had carried out over a century earlier... but this measurement took place far away from the home country, in strange lands in different climate zones, one of them even beyond the ocean.

The idea of the measurement is illustrated in figure 1.6. By astronomical measurements, a baseline is established in the North-South direction, in the end points of which the directions of the plumb line differ from each other *by one degree*. Over land, it is measured how much is the distance between the points in metres[8]. If Newton was right, the length of a degree close to the North pole would be greater than close to the equator, in other words, the *radius of curvature* of the Earth would, at the poles, be longer than at the equator:

$$r_L > r_P.$$

The joint result of both expeditions was an empirical flattening of $f = 1/210$. For comparison, the current best value for the flattening of the Earth is $f = 1/298.257$.

A lot has been written about the adventures of the expedition led by Pierre L. M. de Maupertuis in the Torne river valley 1736 – 1737[9].

[8] In reality, the French Academy measurements used the *toise* as the unit of length, as the metre hadn't been invented yet.

[9] Mostly in Finnish,
http://lapinkavijat.rovaniemi.fi/maupertuis/mittausretki.html,
http://www.maupertuis.fi/tieteelliset-mittaukset/astemittaus/,
and in older writings in French, e.g.,
http://www.doria.fi/bitstream/handle/10024/69470/001.pdf, not so much in English.

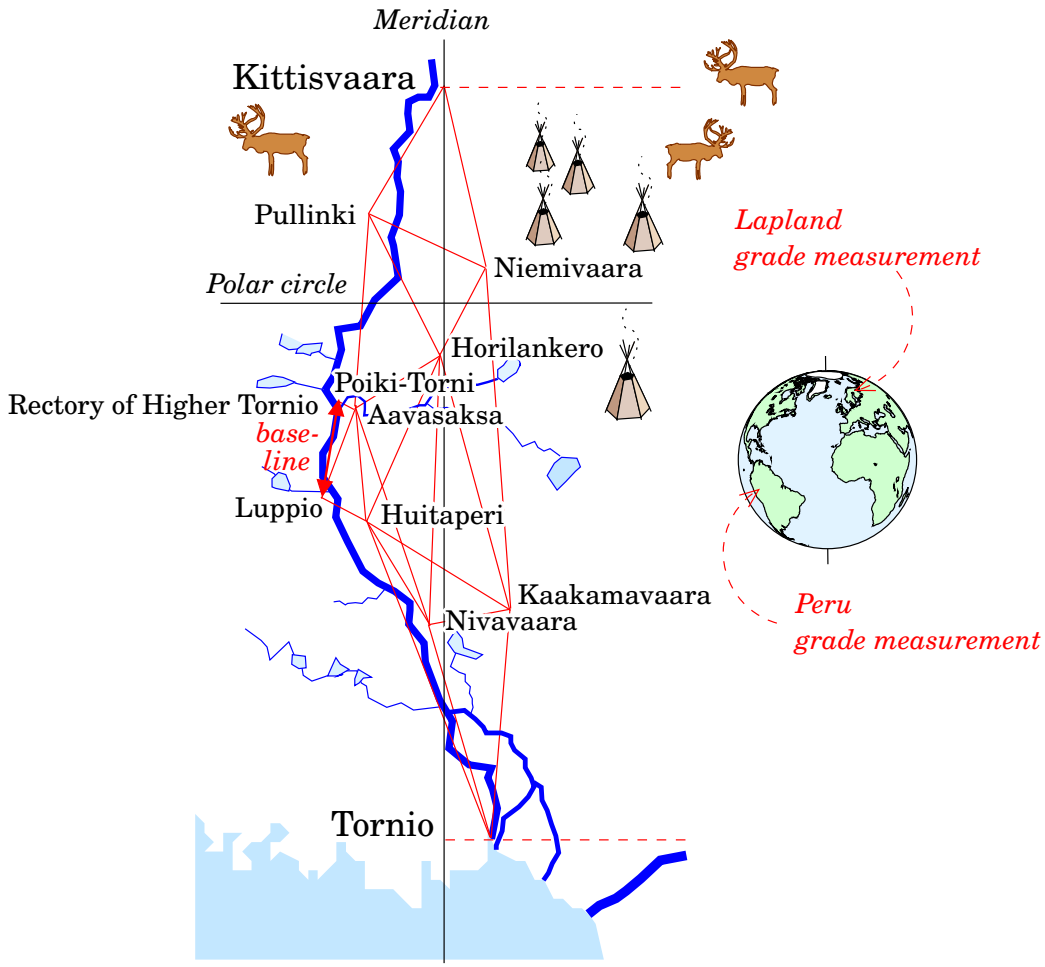


Figure 1.7. The grade measurement project of the French Academy of Sciences: the Lapland grade measurement network.

Of the later grade measurements we may mention Struve[10]'s Russian-Nordic grade measurement ("Struve chain") 1816 – 1855 which extended from Norway's Atlantic coast all the way to the Black Sea[11]. Some points of the chain have been preserved also on Finnish territory.

1.3 The mathematical figure of the Earth or *geoid*

The change from place to place in the direction of the plumb line along an arc on the Earth's surface are thus a means to find out about the true figure of the Earth. In the previous section we described how the grade measurement project of the French Academy of Sciences exploited this phenomenon in order to determine the figure of the Earth, *assuming* that the Earth had the figure of an ellipsoid of revolution.

[10] Friedrich Georg Wilhelm von Struve (1793 – 1864) was a Russian astronomer and geodesist.

[11] https://en.wikipedia.org/wiki/Struve_Geodetic_Arc.

Figure 1.8. The Northernmost point of the Struve chain in Fuglenes, Norway, © Wikimedia Commons.

□

With the aid of more precise geodetic measurements it was noticed that this assumption *doesn't precisely apply*. Already in the context of the Peru grade measurement Pierre Bouguer[12] noticed that the direction of the plumb line on both sides of the Andes had a tendency to deflect toward the mountain range, and interpreted this correctly as an expression of Newtonian *gravitation* or attraction. George Everest[13] in India noticed the same phenomenon near the Himalayas. As geodetic measurements, especially astronomical determinations of the direction of the plumb line, progressed, the understanding spread that the figure of the Earth is irregular.

One started to speak about the "mathematical figure of the Earth" or *geoid* (J.B. Listing[14], 1873), the continuation of mean sea level under the continental masses, a surface that is everywhere perpendicular to plumb lines, and along which a fluid at rest – like sea water – would settle. See figure 1.9.

---

[12]Pierre Bouguer (1698 – 1758) was a French polymath, mostly a geophysicist and shipbuilder.

[13]Sir George Everest (1790 – 1866) was a geodesist and geographer born in Wales, director-general of the Survey of India. In 1865 Mount Everest was named after him against his protestations.

[14]Johann Benedict Listing (1808 – 1882) was a German mathematician, the inventor of *topology*.

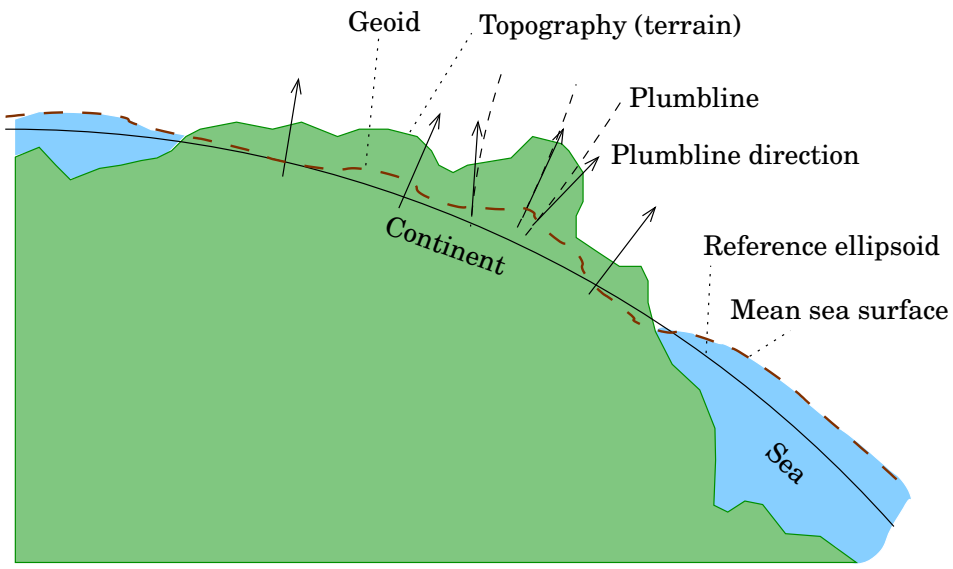


Figure 1.9. Deviations of the plumb line and the shape of the geoid.

In 1862, under the leadership of the Prussian J. J. Baeyer[15], the "*Mitteleuropäische Gradmessung*" ("Central European Grade Measurement") was established, which later developed into the global organization IAG, *International Association of Geodesy*. Its task was the determination of the figure of the Earth or geoid, especially on the European territory, and unite the geodetic networks of Europe into a single network. This objective wasn't properly achieved until 1950, when the first common European network adjustment ED50, "European Datum 1950", was completed, even though initially only on the Western European territory.

Also elsewhere, like in North America, continental-scale triangulation networks were being measured, to determine the figure and flattening of the Earth as well as the locations of points on the Earth's surface in support of map making. Determining the general figure of the Earth is however difficult from the Earth's surface using classical geodetic techniques, because extended networks on the Earth's surface are not geometrically strong, and their unification across oceans is impossible. Satellites have fundamentally changed this picture: satellite techniques have provided precise data on, e.g., the flattening of the Earth by exploiting the rapid precession of the satellite orbital plane it causes. Already several weeks after the launch of Sputnik, much better values became available for the flattening, and the American Vanguard 1 satellite showed the Earth to be "pear shaped" – though only very, very slightly.

[15]Johann Jacob Baeyer (1794 – 1885) was a Prussian officer and a geodesist.

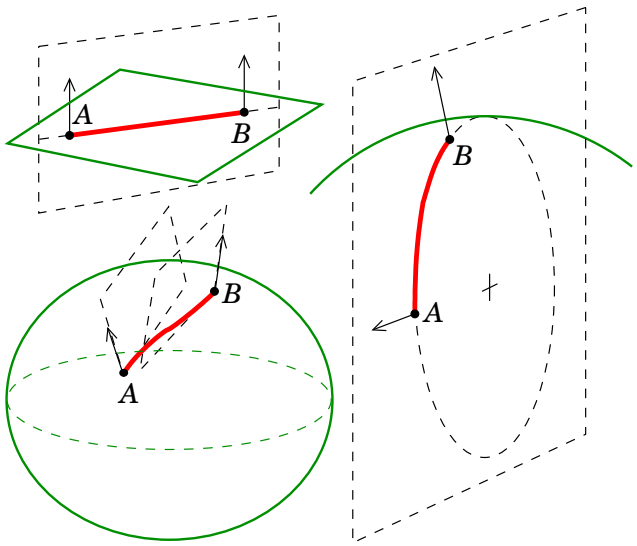


Figure 1.10. The geodesic in the plane, on the sphere and on the ellipsoid of revolution. The arrows depict the local normal to the surface. The plane of the normal section is not uniquely defined for the ellipsoid of revolution.

1.4 The geodesic

In the plane, the shortest path between two points is the *straight line*. On the curved surface of a sphere, or of an ellipsoid of revolution, the shortest path is a curve. In the case of the sphere it is an *arc of a great circle*[16]; in the case of the ellipsoid it is a surface curve that is ever so slightly S-shaped.

geodeettinen viiva

Figure 1.10 shows this general concept of "the shortest path within a surface" or *geodesic*.

The figure shows another curiosity: both in the plane and on the sphere, the plumb lines, or *normals*, to the surface at the end points A and B lie in the same plane, together with the connecting line or curve itself. This plane is called the *normal section*. In the case of an ellipsoid of revolution this is however not the case: in the general case *there is no normal section containing both normals*. This effect is indeed extremely small and it can be ignored in all but the most precise calculations.

Traditional geodetic measurement networks on the Earth's surface, especially triangulation networks, can be considered to consist of *measurement lines* that are geodesics on some reference surface, generally a reference ellipsoid.

In practice the geodetic instruments and signals are never precisely on the reference surface, but some distance above or, more rarely, below it.

[16] Between any two points there are two great-circle arcs, one the shortest and the other the longest between the points. Only in case of antipodes are there an infinity of great-circle arcs, all 180° long.

Then, a *reduction* of the raw observations to this reference surface must be done. This applies to both angle and distance observations. Generally, the corrections needed are small.

1.5 The flattening of the Earth and gravity

As we descibed above, already the Greeks were well aware of the sphericity of the Earth and even her approximate size. During the 17th and 18th century, the idea developed of the flattening of the Earth, i.e., the *ellipsoid of revolution* as descriptive of the figure of the Earth. Astronomers observed the flattening of Jupiter, describing it correctly as a dynamic phenomenon caused by the planet's fast rotation.

On a flattened Earth, of course also *gravity* changes with latitude; this was observed with a pendulum clock, the pendulum of which had to be shortened when travelling to Cayenne in French Guyana (Jean Richer[17], 1672), so the clock would run on time again. Gravity is weaker near the equator than in France. Upon returning to France, Richer had again to lengthen his pendulum, so it would swing on time.

Newton and Huygens calculated theoretically a value for the Earth's flattening $f = \dfrac{a-b}{a}$, equation 1.1. A. C. Clairaut[18] again derived his famous equation giving the relationship between flattening f and "gravity flattening" $\beta = \dfrac{\gamma_b - \gamma_a}{\gamma_a}$. Here, γ_a and γ_b are the accelerations of gravity on the equator and on the poles, respectively. Clairaut's theorem's approximate but elegant form is

$$f + \beta = \frac{\omega^2 a}{\gamma_a},$$

in which ω is the rotation rate or angular velocity of the Earth.

1.6 Reference surfaces and reference systems

As a reference surface for heights, the *geoid* is used, that level or *equipotential surface* of the gravity field of the Earth, which is on average on the same level as mean sea level. This is thus the surface, the minimum state of potential energy, which sea water would reach if there were no currents in the ocean, no salinity or temperature differences, above it no variations in air pressure, etc. In reality all these disturbing factors exist and the global mean sea level deviates as much as over one metre from this equipotential surface, both above and below it. Of this devia-

[17] Jean Richer (1630 – 1696) was a French astronomer.

[18] Alexis Clairaut (1713 – 1765) was a French mathematician and astronomer.

tion, part is varying in time, like, e.g., the tides, part is permanent, the so-called *sea-surface topography*.

The *levelled heights* of countries may be understood as heights above this geoid surface. In practice the heights are nevertheless tied to sea-level observations by tide gauges or *mareographs* operating on the coast, and the reference level is transferred inland by levelling, thus creating a *height system*.

More about the geoid and heights in chapter 4.

A space geodetic measurement method, like GPS, provides a way to determine the height of a point from the reference ellipsoid, because the ellipsoid is a simple mathematical surface in space. The reference ellipsoid is also otherwise a good approximation to the true figure of the Earth, which has also traditionally been used as a reference surface for the adjustment of national or continental triangulation networks.

In a very small area and for many purposes, the "flat Earth approximation", the assumption that the curvature of the Earth may be neglected, continues to serve us. The location of points may be described by two plane co-ordinates and the height by one height co-ordinate, the vertical distance in metres from the reference surface. Depending on the application, a "small area" may be a city or all of Finland – or, in special situations, taking special care, even an area the size of Europe, e.g., Strang van Hees (1990).

1.7 The sub-fields of geodesy

Geodesy is defined as "the science of measuring and mapping the Earth's surface" (Helmert, 1880; see Torge (2001, page 1). This definition continues to hold today: geodesy includes the mapping of the sea floor as well as the determination of the gravity field of the Earth, or *geopotential*. More and more also the study of *changes* in the shape of the Earth and of the physical mechanisms causing these, *geodynamics*, has become a part of the geodetic research field.

Thus, geodesy, in particular physical geodesy, belongs to the Earth sciences. However, geodesy belongs clearly also to the engineering sciences. In Finland, geodesy is currently (2017) being taught both at the University of Helsinki (two docents, external lecturers) and at Aalto University (a professor, docents, an external lecturer).

According to Torge, geodesy may be divided into three sub-fields:

- "global geodesy, also "measuring the Earth" (*geomensuration*, German *Erdmessung*). More precisely:

  > "The task of geodesy is to determine shape and size of the Earth and other celestial bodies, and study their changes in time; as well as to

determine the mean reference ellipsoid of the Earth from parameters observed on and outside the Earth's surface."

Figure of the Earth:

1. the *physical figure* of the Earth: the solid surface of the Earth, i.e., the interface between solid and gaseous or liquid matter – atmosphere, ocean – with all its mountains and depths
2. the *mathematical figure* of the Earth or *geoid*: the equipotential surface of the Earth's gravity field that on average coincides with the mean sea surface, and that may be considered the continuation of the mean sea surface under the land masses. See section 1.3

- geodetic surveying, surveying science
- ordinary surveying ("plane surveying"). To this belong, e.g., topographic surveying and engineering surveying measurements. These measurement are of such areal extent and accuracy class, that in all calculations the curvature of the Earth may be neglected or taken into account by simple correction formulas.

  These measurements are made, besides by geodetic means, often also photogrammetrically by means of aerial photography. The objective of the work is always the production of geometrically precise and correct mapping material for use by society.

  Computations may be done using plane co-ordinates x and y, maps for small areas may be drawn without the use of a map projection, and the separate height co-ordinate H may be assumed to be purely metric ("metres above sea level") without unpleasant consequences.

1.8 Topographic surveying: from terrain to map

In the booklet "*Maastomittaus ja Kartoitus*" ("*Topographic Surveying and Mapping*") (Heiskanen and Härmälä, 1963) it is stated:

maastomittaus

> "Topographic surveying and the mapping that goes with it most often serve to provide of a larger or smaller area a depiction that is as correct and precise as possible, in the form of a *map*."

Society needs maps and location based information for many purposes. In today's society, the ownership of real estate, its buying and selling, and expecially its use as collateral (mortgage) for loans aimed at maintaining and developing the property's value, are foundational for a modern society with a high level of investment. To this end exists the *cadastral system*, which registers as reliably as possible the state of properties and the rights in them. There are millions of real-estate properties and

parcels and their collective monetary value is astronomical[19].

Another use of maps and geospatial information that is important to society, is the *planning and construction of infrastructure*. Roads, railroads, bridges, tunnels, canals, airports, harbours, power plants, water works, telephone and data networks etc. etc. This is a vital public service from the viewpoint of economic productivity.

In every developed society, construction is limited in some sense. You may not build what and how you wish, even on land that you own. *Zoning* regulates in a co-ordinated way, for what purposes land may and may not be used. These regulations are contained in so-called *zoning plans*, the approval process of which is statutorily prescribed, public and consisting of several stages. The reason for this is, that zoning affects the value of property, so the legal status of property owners requires that the democratic approval process of zoning plans contains sufficient instruments of appeal. Maps and other surveying-based information sources are essential for this process.

kaavoitus

In Finland, the planning of land use and the associated local infrastructure construction happens for the most part in the public administration, most often in municipalities. We speak of *spatial planning*.

yhdyskunta-suunnittelu

Topographic surveying is, in terms of volume, the overwhelmingly largest field of application of land surveying.

1.8.1 Spatial planning

Spatial planning is a continuous activity, to which belongs zoning, planning of land use and the built environment, and managing building activity. In spatial planning a major role falls to the municipality. City planning and regional planning are forms of spatial planning.

The technical planning and building of a community require reliable information on the environment which is being planned and built. Figure 1.11 gives an idea of all the places where geospatial information plays a role in the continuous process of spatial planning and construction.

[19] According to the report (RAKLI ry, 2014) “All of Finland’s building stock together has a value, including parcels, of some 480 billion euros”. This means almost a hundred thousand euros for every Finnish man, woman and child.

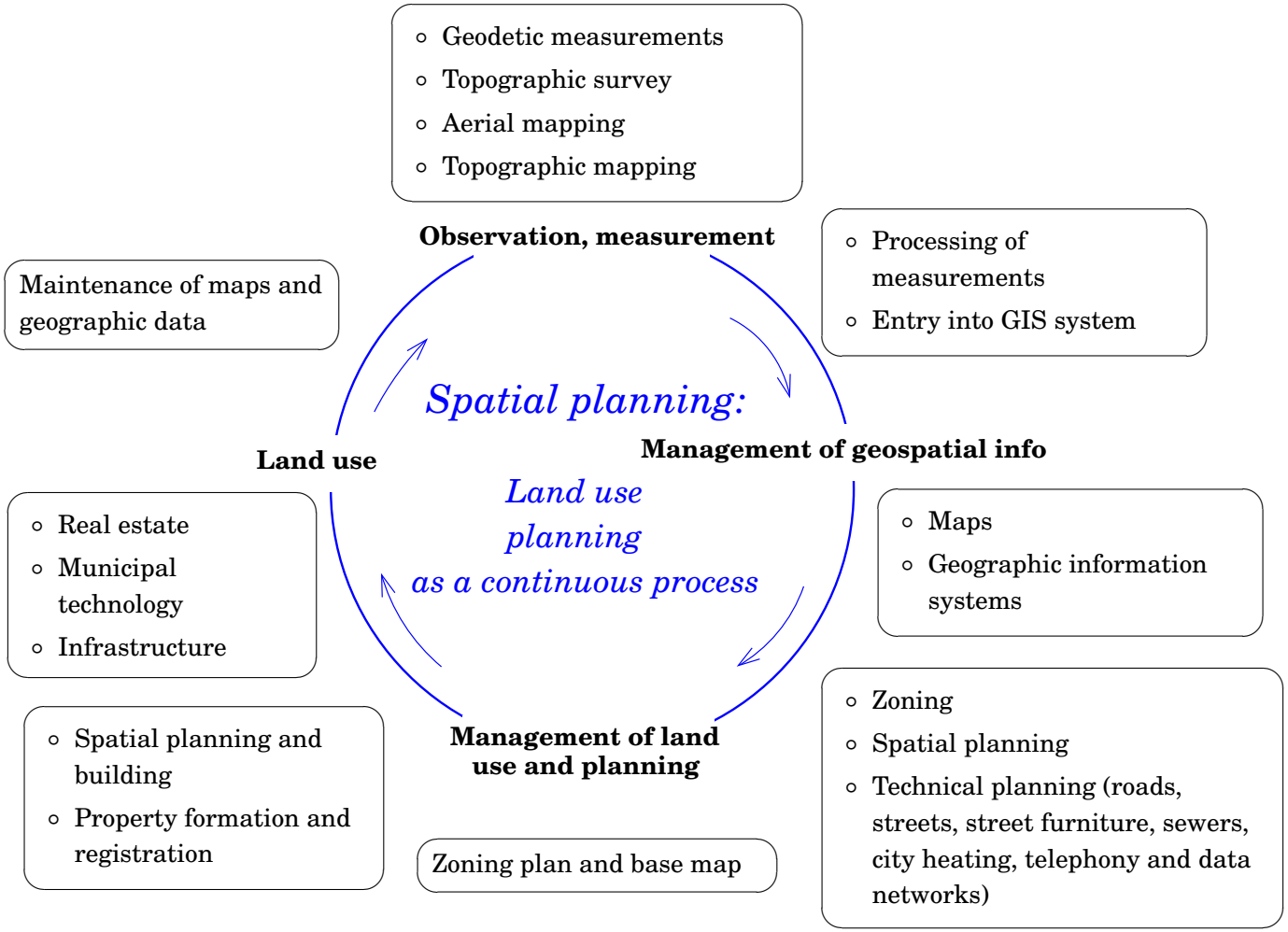


Figure 1.11. The roles of the map and topographic surveying in spatial planning and construction.

Topographic surveying is along in the whole process:

- Topographic information must be measured onto the map in a certain co-ordinate reference frame.
- When the zoning plan is ready, it has to be set out in the terrain.
- The properties have to be measured and mapped.
- The technical structures must be placed into the terrain.

Topographic surveying has the following *goals* and *tasks*:

runkoverkko

- creating a *base network* for mapping: getting the measurements into a certain, known co-ordinate frame
- mapping terrain details while using base network points: *detail survey*

kartoitusmittaus

- *setting out*: the transfer of plans into the terrain, to be realized in their right locations. Setting out is, in a way, *mapping's inverse problem*.

1.8.2 Carrying out the tasks of topographic surveying

Successfully performing the tasks of topographic surveying requires that the surveyor is acquainted with the following things:

- geodetic instruments
- measurement technique, the planning of measurement, measurement conditions
- geodetic computation: co-ordinates and transformations, derived quantities, accuracy
- zoning-plan calculations, setting out
- geographic information systems
- maps, printing, reporting.

Professor Matti Martikainen drafted a pretty tableau, table 1.1, on the role of the measurement plan and the place of topographic surveying in the whole of the measurement and mapping process. We present it here, slightly modernized. The various parts of the table belong to the sub-fields of geodesy, photogrammetry, and cartography.

Tableau 1.1. Topographic surveying as a part of the whole measurement and mapping process.

Measurement plan					
Base network measurement (GNSS), computation					
Collection of topographic data	Mapping surveying (often GNSS-RTK)	Existing maps	Digital imagery	Airborne laser scanning	Photographs
Processing of topographic data	Geodetic computation	Digitization	Image processing	Processing	Stereo photo-grammetry
Data integration	Geographic information systems				
Products	*Graphical*	*Numerical*	*Textual*		*Metadata*
Data presentation to the end user	◦ Topographic maps ◦ Thematic maps ◦ Special maps ◦ Customer printouts ◦ 3D-visualization	◦ Point data bases ◦ Elevation models ◦ Digital CAD models	Reports		

1.8.3 The end product of topographic surveying

The visible end product of topographic surveying is a *map*. A map must be, before all else, *correct*, but also clearly drawn, give all relevant information, and it would be nice if it were pretty as well (Heiskanen and Härmälä, 1963).

A suitable scale is chosen for the map, which defines the level of accuracy of the information presented. The scale is chosen to suit the purpose of use of the map. The objects to be presented need to be *generalized* in a suitable way: too small details should be taken out, however, essential details should be made clearer[20].

Nowadays maps can also be in digital form. In that case the significance of the scale is not quite as clear.

paikkatieto
ominaisuustieto

Additionally there is a lot of location-related information in numerical form (*geospatial information*, consisting of location and attribute data) and "metadata", data describing other data, e.g., map information. The legend in a paper map is an example of metadata.

The following items of information are or can be part of the final product:

- plane co-ordinates, these state the location inside a municipality, a country, the world. On the map, co-ordinate curves, a co-ordinate grid
- height information, e.g., height contours, height values of points, possibly profiles
- the forms of the physical Earth's surface, presented in various ways
- attribute data. The measured objects are shown on the map in accordance with an agreed presentational style. To every piece of information, a suitable identifier or symbol is given
- The information shown on the map may also be divided into natural and cultural data. See table 7.4.

Self-test questions

1. You have a car with an FM radio, a clock, a passport, a few weeks of free time, and money for food and fuel. How do you establish to your own satisfaction that the Earth is round, not flat? https://www.salon.com/2017/09/25/b-o-b-flat-earth/.
2. How does Newton's universal law of gravitation explain that larger celestial bodies are approximately spherical? Why are they not

[20] E.g., on a road map, the widths of the drawn roads have no relationship whatever with the widths of the roads in the terrain! The drawn width expresses the *importance* of the road to traffic. This is how generalization works.

precisely spherical?

3. Describe the hypotheses of Newton and Huygens on the interior mass distribution of the Earth, and its effect on the Earth's flattening.
4. What is the "mathematical figure of the Earth", and how is it related to the direction of the plumb line?
5. Clairaut's theorem gives the relationship between the rotation of the Earth, her flattening, and her "gravity flattening". Intuitively, give two reasons why gravity on the equator should be weaker than at the poles, and one reason why it should be stronger.
6. What is the geoid, and what is the reference ellipsoid?
7. What is the task of topographic surveying according to Heiskanen and Härmälä?
8. Name the three main fields of application of topographic surveying in society.
9. What are the three goals and tasks of topographic surveying?

2. Geodetic measurements and co-ordinates

2.1 Units of measurement

2.1.1 Definitions

When we talk in geodesy, like more generally in physics, about measurement units to be used, we distinguish between *units* and *quantities*. E.g., length is a *quantity*, the *unit* of which may be, e.g., the metre [m][1]. So, for example:

suure

> *The length of the distance AB is* 15 *metres, i.e.,* 15 m.

Quantity	Value	Unit	Symbol
Length	15	metres	m

dimensio

In the literature, also the term *dimension* is used, e.g., the dimension of volume is length^3, the dimension of acceleration is length $\times$ time^{-2}. In this way is expressed how the definition of a certain quantity depends on the definitions of other quantities. E.g., if one wants to precisely measure accelerations, one has to precisely measure both lengths and time intervals. This belongs to the field of *metrology*, the science of measurement.

In Finland, like in most countries of the world, the SI system or International System of Units (SI = *Système International d'Unités*) is used.

The system consists of *base units* and *derived units*. There are seven base units[2]. See table 2.1.

[1] The official symbol of a unit according to the SI system is always written in upright letters, *not in italics*! Italics is used for mathematical symbols. So: $E = mc^2$, but $\mathrm{J} = \mathrm{kg\,m^2\,s^{-2}}$.

[2] http://www.bipm.org/en/measurement-units/base-units.html.

Figure 2.1. A public metre in Paris, at 36 rue de Vaugirard, sixth arrondissement © Wikimedia Commons.

2.1.2 Prefixes

To the SI units (but not to the additional units[3]!) one may add a prefix indicating order of magnitude according to table 2.2 on the facing page. The table is not complete.

Example: $1\,\mathrm{MHz} = 10^{+6}\,\mathrm{Hz}$.

[3]This is nevertheless widely done, e.g., kcal means kilocalorie, a traditional unit of energy content of chemical substances and food. In the ICT field again, the prefix k, or sometimes K, expresses the non-standard binary order of magnitude 1024× (http://en.wikipedia.org/wiki/Kilobyte). And also the monetary unit k$ is being used!

Table 2.1. Measured quantities, units and their symbols.

	Quantity	Unit	Symbol	How derived
Base units	Length	Metre	m	
	Mass	Kilogram	kg	
	Time	Second	s	
	Electric current	Ampere	A	
	Temperature	Kelvin	K	
	Luminous intensity	Candela	cd	
	Amount of substance	Mole	mol	
Derived units	Plane angle	Radian	rad	
	Solid angle	Steradian	sr	
	Frequency	Hertz	Hz	s^{-1}
	Force	Newton	N	$\mathrm{kg\,m\,s}^{-2}$
	Pressure	Pascal	Pa	Nm^{-2}
	Energy	Joule	J	Nm
	Power	Watt	W	Js^{-1}
	Electric tension	Volt	V	WA^{-1}
	Electrical resistance	Ohm	Ω	VA^{-1}

Table 2.2. Prefixes indicating order of magnitude in the SI system.

Value	Prefix	Symbol	Value	Prefix	Symbol
+1	deca	da	+6	mega	M
+2	hecto	h	+9	giga	G
+3	kilo	k	+12	tera	T
−1	deci	d	−6	micro	μ
−2	centi	c	−9	nano	n
−3	milli	m	−12	pico	p

2.1.3 Non-SI units accepted for use with SI

In daily life, and in many scientific disciplines, a large number of non-SI units are in widespead use and are not going away soon. E.g., a calendar, or a clock, would not be very practical in kilo- and megaseconds. The international metrological community, realizing this, has created a category for these units, "Non-SI units accepted for use with the SI[4].".

lisäyksikkö

Table 2.3 gives some often used additional units "accepted for use with the SI":

Celsius temperature is obtained from Kelvin temperature by subtracting 273.15 K from it. The temperature *differences* are the same in the Celcius and Kelvin scales, i.e., 1 °C = 1 K.

$$0\,^{\circ}\mathrm{C} \sim 273.15\,\mathrm{K},$$
$$0\,\mathrm{K} \sim -273.15\,^{\circ}\mathrm{C}.$$

2.1.4 Units of angle

The units *radian* and *steradian* are dimensionless numbers ("bare numbers") because they are ratios. E.g., the radian is the ratio between the

[4] https://en.wikipedia.org/wiki/Non-SI_units_mentioned_in_the_SI

Table 2.3. Non-SI units accepted for use with the SI.

Quantity	Base unit	Accepted unit	Symbol
Time	s	Hour, minute, second	h m s
		Day	d
		Year	a
Plane angle	rad	Degree, minute, second	°′″
		Gon	gon
Temperature	K	Degree Celsius	°C
Volume		Litre	l
Mass	kg	Tonne	t

length of a circular arc and its radius, and thus dimensionless. There exist however other units of angle, like the degree and the gon. Therefore it make sense, in a way, to treat all of these also as units.

A somewhat similar situation exists for logarithmic scales: Richter (total energy of earthquakes), the magnitude scale of stars, the decibel (dB) scale, the °DIN scale for the light sensitivity of photographic emulsions, and the pH, the degree of acidity of a solution.

2.1.5 Length and time

Of the base units, both length and time are based on atomic phenomena.

The metre is the distance travelled by light in vacuum in 1/299,792,458 seconds. In practice, the metre is realized by a so-called iodine stabilized helium-neon laser, the wavelength of which is very precisely (2.5 parts in 10^{11}) known[5].

The second again is "the duration of 9,192,631,770[6] oscillation periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the ^{133}Cs atom" [official SI definition]: it is based on the use of a *caesium clock*.

2.1.6 Angular units

2.1.6.1 Degrees

In ordinary life we use as unit of angle the *degree*, symbol °. Also geographical latitude and longitude as read from a map are commonly given in degrees. A *right angle* is 90°, and a *straight angle* 180°. In addition to degrees, we have as traditional units the *minute* (′) and the second (″). These behave like their namesakes in time measurement: one degree is 60 minutes and one minute 60 seconds.

2.1.6.2 Calculation example

Convert degrees, minutes, seconds to degrees and decimals:

$$\begin{aligned} 56°47'33'' &= 56° + \left(\frac{47}{60}\right)^\circ + \left(\frac{33}{60\times 60}\right)^\circ = \\ &= 56° + 0°.783333\ldots + 0°.0091666\ldots = \\ &= 56°.7924999\ldots \end{aligned}$$

Convert back in the other direction (Note the rounding error!):

$$\begin{aligned} 56°.7925 &= 56° + (60\times 0.7925)' = \\ &= 56° + 47'.55 = \\ &= 56°47' + 0'.55 = \\ &= 56°47' + (60\times 0.55)'' = \\ &= 56°47'33''.0. \end{aligned}$$

[5] http://museum.nist.gov/object.asp?ObjID=50.

[6] https://www.quora.com/Why-is-a-second-defined-as-the-duration-of...

2.1.6.3 Gon

In geodesy and geodetic instrumentation, often the *gon* is used as a measurement unit[7]. Sometimes the name "new degree" is used. A new minute, or centigon, is 0.01 gon, a new second, or decimilligon, 0.0001 gon. One notation is $1.2345\,\text{gon} = 1^{\text{g}}23^{\text{c}}45^{\text{cc}}$. In units of gon, a right angle is 100^{g}.

2.1.6.4 Radians

A full circle contains 2π radians, 360 degrees (360°) and 400 gon (400^{g}). A right angle has thus $2\pi/4 = \pi/2$ radians.

Here are some formulas to convert an angle given in radians to gon, degrees. See Kahmen and Faig (1988):

A full circle is

$$
\begin{aligned}
2\pi\,\text{rad} &= 400^{\text{g}} = 360^{\circ}.\\
1\,\text{rad} &= \left(\frac{360}{2\pi}\right)^{\circ} \approx 57^{\circ}.2957795;\\
&= \left(\frac{400}{2\pi}\right)^{\text{g}} \approx 63^{\text{g}}.6619772.\\
\alpha\,\text{rad} &= \left(\frac{360}{2\pi}\alpha\right)^{\circ} = \left(\frac{400}{2\pi}\alpha\right)^{\text{g}}.
\end{aligned}
$$

And

$$
\begin{aligned}
\alpha^{\circ} &= \left(\frac{400}{360}\alpha\right)^{\text{g}} = 1.1111\cdots\times\alpha^{\text{g}},\\
\alpha^{\text{g}} &= \left(\frac{360}{400}\alpha\right)^{\circ} = 0.9\times\alpha^{\circ}.
\end{aligned}
$$

2.2 Measurement error and uncertainty

No measurement is perfectly accurate. Land surveying is also, being a human activity, prone to error. Just like in computer programming, one shouldn't even try to measure completely errorlessly. A more realistic goal is, in addition to the usual carefulness, to develop methods by which

1. Errors of a certain size can be noticed and removed from the observational material (statistical testing) ;
2. The impact of not noticed errors on the end result can be evaluated and minimized (adjustment calculus).

With these methods, measurement errors may be taken into account and measurement results as correct as possible be produced, the quality of

[7] Another used name is *grad*. The unit was taken into use by the French in connection with the revolution and introduction of the metric system. Today, it is only used in the land surveying field.

which, i.e., their accuracy, is known, and the size of any errors possibly still be hiding out in them is known or can at least be judged.

The processing of the measurements, the *adjustment*, will yield the following *results*: tasoitus

1. The "best" value for the unknown quantities, based on the measurements. This may be
   (a) the most likely value
   (b) the statistical expectancy odotusarvo
   (c) a value with respect to which the deviations are as little damaging as possible.
2. A judgment on the "goodness" of the original measurements (i.e., their *precision*[8]). We speak of standard deviation or *mean error*. Nowadays the official term is *standard uncertainty*.
3. Similarly, a judgment on the "goodness" or precision of the *end results*, i.e., the computed values (estimates) of the unknown quantities.
4. A judgment on the possible occurrence of gross errors and their maximal magnitude after statistical testing has been done: *reliability*.
5. A judgment on the possible presence of systematic errors and their magnitude.

Measurement error is the difference between the value to be measured[9] and the measured value. The *measured value* is the result of an often complex[10] measurement process.

2.2.1 Types of errors

We may divide the errors coming out of this measurement and reduction process into the following categories:

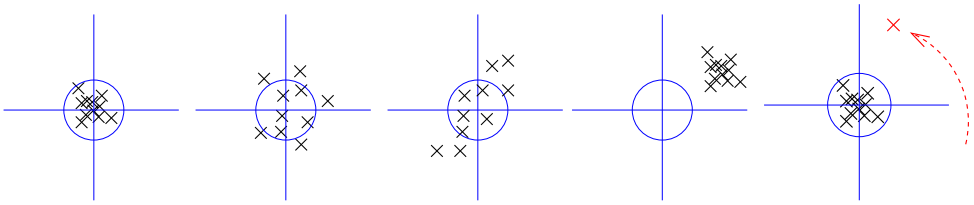
2.2.1.1 Random errors

Generally it is assumed that the *random errors* that represent the natural uncertainty of the observation process are *normally distributed*, i.e., the distribution of errors is a pretty Gaussian bell curve or normal distribution – on which more later on.

[8] In English the terms "precision" and "accuracy" are used. *Precision* refers only to statistical spread, the variability of measurement values, whereas *accuracy* refers to the deviancy of the observation values from the "correct values". The latter concept thus includes also systematic errors.

[9] The "true value", though one may philosophically ask if it even exists.

[10] Sometimes the measurement process contains a pretty complex reduction chain or modelling effort.



(a) Small random **(b)** Large random **(c)** Correlated **(d)** Systematic **(e)** Gross

Figure 2.2. Examples of different error types. Random precise, random imprecise, correlated, systematic, gross.

If this is so – and one shouldn't assume this without further testing! – one has available the so-called least-squares adjustment method, which minimizes the joint impact of random errors in the end result.

2.2.1.2 *Gross errors*

Gross errors are caused by human mistakes or measurement device malfunctions. They happen only now and then and they cannot be described by a statistical distribution. An example of a gross error is writing a digit wrong into the observation notebook.

One strives to eliminate gross errors by statistical testing. Statistical testing is a broad scientific subject; as a rule of thumb one may say, that if a function of a measurement differs from its known expectancy (e.g., zero) by more than three times its known, own standard deviation (mean error) σ, there is reason to suspect that some of those measurements are in error. This is criterion is called the *three-sigma rule*.

2.2.1.3 *Systematic errors*

Systematic errors are a sign that there are deficiencies in the theories used to describe the measurement process, observation geometry and physics of the observational situation. The occurrence of significant systematic errors should lead to a study and re-assessment of the theoretical assumptions. An example: we know that the sum of the angles of a triangle must be 180°: the triangle condition. If the sum of the measured angles of a plane triangle differs significantly and consistently from this value, one must suspect

- lateral refraction in the atmosphere
- that the triangle is so large, that on the curved surface of the Earth we no longer have a plane triangle and the theorem no longer applies
- other things.

Also *uncertainty*, the knowledge on how large errors could possibly be, can be divided into different categories:

Table 2.4. Dice throwing statistics.

Number of throws	Value 1	2	3	4	5	6	Mean ± standard deviation
60	14	8	9	12	9	8	10 ± 2.4
%	13.33	23.33	13.33	15.00	20.00	15.00	16.67 ± 4.08
600	103	91	106	114	87	99	100 ± 9.9
%	16.50	17.17	15.17	17.67	19.00	14.50	16.67 ± 1.65
6000	973	1007	1003	962	1015	1040	1000 ± 28.5
%	17.33	16.22	16.78	16.72	16.03	16.92	16.67 ± 0.47
60,000	10,138	9936	10,057	10,029	9925	9915	$10{,}000 \pm 89.5$
%	16.52	16.90	16.56	16.76	16.72	16.54	16.67 ± 0.15

2.2.1.4 Type A uncertainty

Type A uncertainty is standard uncertainty or standard error calculated, e.g., from repeated measurements. This type of uncertainty can be handled by means of statistical theory, i.e., it corresponds to *random errors*.

2.2.1.5 Type B uncertainty

Type B uncertainty is uncertainty that can *not* be determined by statistical means, but needs to be assessed in some other way. E.g., *systematic errors* can be determined by calibration, from a statement by the device manufacturer, etc.

There is no uncertainty concept corresponding to gross errors. They should be avoided or eliminated by statistical testing.

2.3 Stochastic quantities

2.3.1 Discrete stochastic quantities

A measurement process is a so-called random or *stochastic quantity*. A stochastic quantity is obtained by doing something the outcome of which is random. E.g., casting a die creates a stochastic quantity $\underline{n}$ the realizations ("throws") of which are $n_1, n_2, n_3, \ldots$.

noppa

The possible values of a die throw are the integer numbers $\{1,2,3,4,5,6\}$. The same with a coin flip, if we take as the values heads = 0, tails = 1. We say that the value set or *codomain* is $\{0,1\}$, a discrete value set.

kruuna
klaava
arvojoukko

If we perform a die throw again and again – i.e., we collect realizations $n_i, i = 1,2,3,\ldots$ —, we may always after a certain total number of throws tabulate the results. We obtain the example table 2.4, which expresses how many throws of the total number were ones, how many twos, etc.

According to experience, the greater the number of dice throws, the smaller appears to be the deviation of the end result from the ideal outcome, in which the frequency of occurrence of every value in percents would be 16.666…%, i.e., 1/6. This empirical result is called the *law of large numbers*.

2.3.2 Expectancy

Based on this, we may assign to the outcomes of die thows a theoretical *probability value*, which expresses how often, "in the long term", a certain value will happen. For a balanced ("fair") die, the probability values are:

$$p(1) = p(2) = \ldots = p(6) = \frac{1}{6}.$$

In the discrete case the *expectancy* – the value around which the throws group, their "centre of mass" – is computed by the formula[11]

$$E\{\underline{n}\} \stackrel{\text{def}}{=} \sum_{i=1}^{N} i \cdot p(i) \tag{2.1}$$

which yields $2\frac{1}{6} = 3.5$. Here, N is the number of alternatives. (Note that 3.5 isn't even a possible throw value!)

In case of a "fair" coin we have

$$p(0) = p(1) = 0.5,$$

and the expected value is, by the same formula, 0.5.

2.4 Statistical distributions

2.4.1 Real-valued quantities and density distributions

Geodetic measurements are generally stochastic quantities, the value set or *codomain* of which is a sub-set of the real numbers $\mathbb{R}$, i.e., *continuous*. E.g., the result of a distance measurement is a distance in metres, and it is a real[12] number. The same for angle measurements. The value space may be bounded, $\alpha \in [0, 360^\circ)$, but it is in any case continuous.

In the case of a continuous value space we speak of a *probability density distribution*, or a *distribution* in short. If we make a large number of measurements of the same object, we may draw a so-called *histogram*,

[11] More generally,

$$E\{\underline{n}\} \stackrel{\text{def}}{=} \sum_{i=1}^{N} v_i \cdot p(i),$$

in which v_i is the value alternative with serial number i. For a die, $v_i = i$.

[12] Well, actually a rational number $\in \mathbb{Q}$.

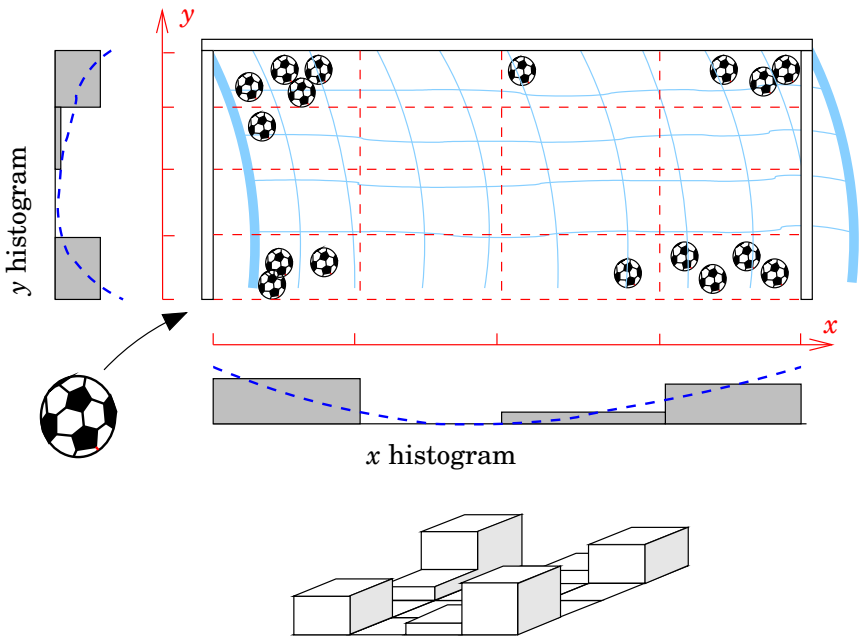


Figure 2.3. An example of a stochastic quantity on a continuous (two-dimensional) value set. Drawn are also the histograms separately according to arguments x and y, and according to both together. The blue dashed line is a possible density distribution function.

which shows the numbers of measurement results that fall between certain values in the value space.

As an example of a continuous statistical distribution may serve the places of the goals scored in the goal of a football match (figure 2.3).

The more measurements, the more bars we may draw, and the narrower we may make them. In the limit for an infinite number of measurements we obtain a continuous curve, the probability density distribution, which expresses the probability with which a measurement result will fall within some value interval $[x_1, x_2]$. See figure 2.4 for an example.

If on a certain interval $[x_1, x_2]$ the integral

$$\int_{x_1}^{x_2} p(x)\,dx$$

is zero, we say that it is *impossible* for a value x in this interval to occur as a realization of the stochastic quantity. If the integral is 1, we say that the occurrence of $x \in [x_1, x_2]$ is *certain*. In all other cases, the value of the integral gives the *probability* that the realization of the stochastic quantity would hit inside this interval. The value of the probability is thus always between zero and one[13].

[13] Mathematically, probability is a *measure*, like, e.g., surface area or volume.

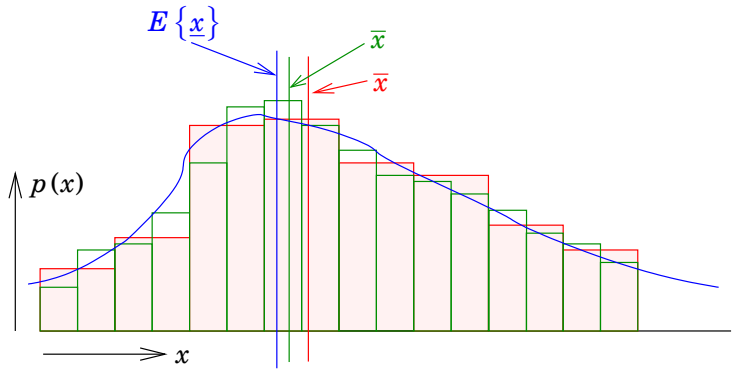


Figure 2.4. The probability density distribution $p(x)$ of a real-valued stochastic quantity as the limit of histograms. Also the expectancy $E\{\underline{x}\}$ and the averages of the two measurement series are drawn.

Note that always

$$\int_{-\infty}^{+\infty} p(x)\,dx = 1,$$

because the joint probability that the measurement value will be *any* real number, is 1, i.e., 100%: it is *certain*.

Also for a continuous stochastic quantity one can compute an *expectancy*. The formula is an integral which is similar looking to the discrete counterpart 2.1:

$$E\{\underline{x}\} = \int_{-\infty}^{+\infty} x \cdot p(x)\,dx.$$

The expectancy is the "centre of mass" of the area under the density distribution curve $p(x)$.

2.4.2 The Gaussian bell curve

In geodesy we use almost always the *normal* or Gaussian *distribution* ("bell curve").

In figure 2.5 we see the *expectancy* $\mu = E\{\underline{x}\}$. The expectancy may be understood as the "true value" of the observed quantity $\underline{x}$, which is not itself measurable, but around which the measurement values group under the influence of their random errors. In the figure we see also the *mean error*[14] σ, which describes the tendency of an individual measurement to differ from the expectancy, In the case of the normal distribution, the deviations equal in magnitude to the mean error $\mu \pm \sigma$ are located in the curve's *inflection points*.

taivutuspiste

Also a concept by the name of *variance* is used, the square of the mean error:

$$\mathrm{Var}\{\underline{x}\} = \sigma^2,$$

[14] . . . or standard deviation, or *standard uncertainty*.

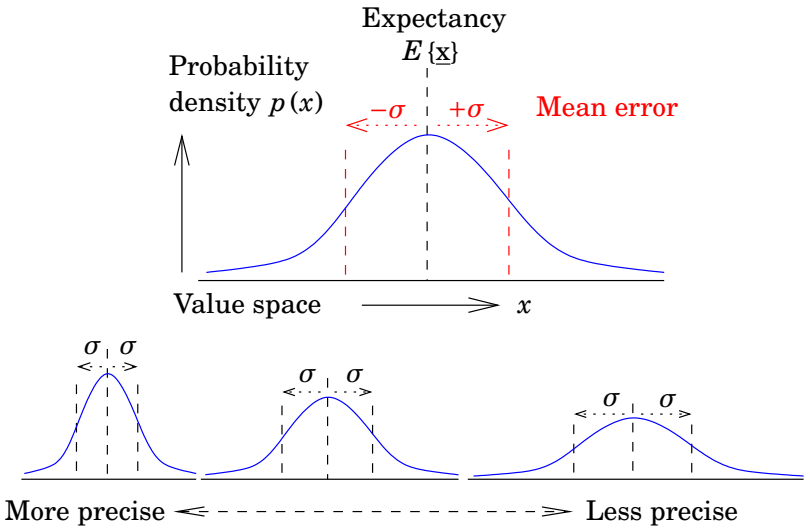


Figure 2.5. Properties of the normal distribution.

the formal definition of which is[15]:

$$\mathrm{Var}\{\underline{x}\} \stackrel{\mathrm{def}}{=} E\left\{\left(\underline{x} - E\{\underline{x}\}\right)^2\right\}. \tag{2.2}$$

Here $E\{\cdot\}$ is the *expectancy* operator as defined above.

Like the expectancy, also the variance is the kind of "true value" that theoretically exists, but that we never actually really measure. When we have available a finite number of observations, or *sample*, we may calculate the *sample mean* and *sample variance*, that, for the number of observations increasing, go ever closer and closer to expectancy and variance. Also this phenomenon is called the *law of large numbers*.

The mathematical expression $p(x)$ for the normal distribution – which we do not present or discuss here – gives through integration the following probability values:

- The probability that x deviates from $E\{\underline{x}\}$ more than an amount σ (in either direction): 32%.

[15] I.e., the expectancy of the square of the deviation of a stochastic quantity from its own expectancy. This is a kind of cost function: if the "cost" of the "error" $\underline{x} - E\{\underline{x}\}$ is proportional to its square, then $\mathrm{Var}\{\underline{x}\}$ is the expected cost of the error.

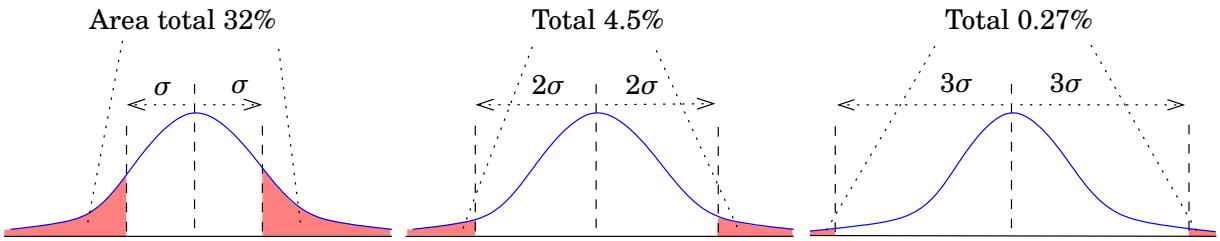


Figure 2.6. Probability values for the normal distribution.

- The probability that x deviates from $E\{\underline{x}\}$ more than an amount 2σ: 4.5%.
- The probability that x deviates from $E\{\underline{x}\}$ more than an amount 3σ: 0.27%.

> *In practice a normally distributed stochastic quantity really never deviates from its expectancy by more than three times its mean error.*

This rule of thumb, the *three-sigma rule*, is made use of in statistical testing.

The properties of a number of statistical distributions are tabulated on this web page:
http://www.stats.gla.ac.uk/steps/glossary/probability_distributions.html.

2.4.3 Covariance and correlation

When there are two stochastic quantities $\underline{x}$ and $\underline{y}$, one can, besides their individual behaviours, also study how thet behave *together*.

Let us call *covariance* the expression

$$\mathrm{Cov}\{\underline{x},\underline{y}\} \stackrel{\mathrm{def}}{=} E\{(\underline{x}-E\{\underline{x}\})(\underline{y}-E\{\underline{y}\})\}. \tag{2.3}$$

This definition is analoguous to the similar one (equation 2.2) for variances, but describes the "behaving in the same way" property of the quantities $\underline{x}$ and $\underline{y}$, i.e., the similarity in their random behaviours.

Often it makes sense to *scale* this covariance relative to the variances of the quantities $\underline{x}$ and $\underline{y}$ in the following way:

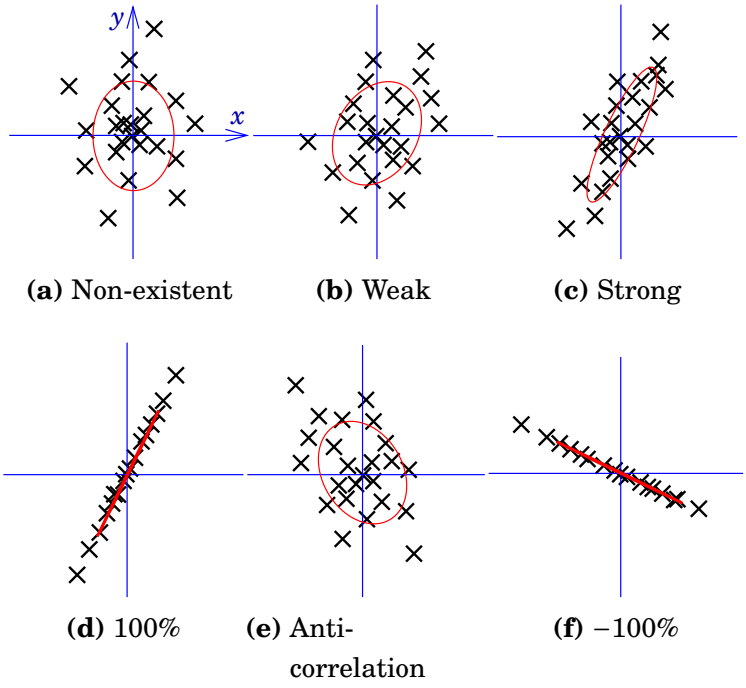
$$\mathrm{Corr}\{\underline{x},\underline{y}\} \stackrel{\mathrm{def}}{=} \frac{\mathrm{Cov}\{\underline{x},\underline{y}\}}{\sqrt{\mathrm{Var}\{\underline{x}\}\,\mathrm{Var}\{\underline{y}\}}}.$$

This is how *correlation* is defined between the quantities $\underline{x}$ and $\underline{y}$. Correlation lies always in the interval $[-1,1]$, or correspondingly in the interval $[-100\%,100\%]$.

The correlation (and covariance) between statistically independent quantities $\underline{x}$ and $\underline{y}$ equals 0. The non-vanishing of correlation is a sign of a cause-and-effect relationship – but one cannot say that $\underline{x}$ is the cause and $\underline{y}$ the effect! $\underline{y}$ may be the cause of $\underline{x}$, or $\underline{x}$ and $\underline{y}$ may have a common cause $\underline{z}$. If the correlation is 1, i.e., 100%, we speak of *perfect correlation*. In this case there exists an exact functional relationship between $\underline{x}$ and $\underline{y}$. If the realization value of one is given, the corresponding realization value of the other can be calculated precisely.

If the correlation is negative, we speak of *anticorrelation*. If the correlation is -1, i.e., -100%, there exists a perfect correlation between the quantities $\underline{x}$ and $-\underline{y}$, and we speak of *perfect anticorrelation*.

In figure 2.7 some examples of correlation are shown.



(a) Non-existent **(b)** Weak **(c)** Strong

(d) 100% **(e)** Anti-correlation **(f)** −100%

Figure 2.7. Some examples of correlation.

2.4.4 Propagation law of errors

An important property of mean errors is the *propagation law of errors*. kasautumislaki

If a stochastic quantity $\underline{z}$ is a linear combination of two other stochastic quantities $\underline{x}$ and $\underline{y}$:

$$\underline{z} = a\underline{x} + b\underline{y},$$

we may also write

$$E\{\underline{z}\} = aE\{\underline{x}\} + bE\{\underline{y}\},$$

and thus

$$\left[\underline{z} - E\{\underline{z}\}\right] = a\left[\underline{x} - E\{\underline{x}\}\right] + b\left[\underline{y} - E\{\underline{y}\}\right].$$

This is probably intuitively clear. Now according to the above equation 2.2 it follows that

$$\begin{aligned}\mathrm{Var}\{\underline{z}\} &= E\left\{\left(\underline{z} - E\{\underline{z}\}\right)^2\right\} = \\ &= a^2 E\left\{\left(\underline{x} - E\{\underline{x}\}\right)^2\right\} + b^2 E\left\{\left(\underline{y} - E\{\underline{y}\}\right)^2\right\} + \\ &\quad + 2ab E\left\{\left(\underline{x} - E\{\underline{x}\}\right)\left(\underline{y} - E\{\underline{y}\}\right)\right\}.\end{aligned}$$

If now the expression

$$\mathrm{Cov}\{\underline{x},\underline{y}\} = E\left\{\left(\underline{x} - E\{\underline{x}\}\right)\left(\underline{y} - E\{\underline{y}\}\right)\right\}$$

is called, like in the previous subsection, the *covariance*, we may write

$$\mathrm{Var}\{\underline{z}\} = a^2\,\mathrm{Var}\{\underline{x}\} + b^2\,\mathrm{Var}\{\underline{y}\} + 2ab\,\mathrm{Cov}\{\underline{x},\underline{y}\}. \qquad (2.4)$$

This equation is called the *law of propagation of variances*.

In case $\mathrm{Cov}\{\underline{x},\underline{y}\} = 0$ (i.e., the quantities $\underline{x}$ and $\underline{y}$ *are uncorrelated*), we obtain

$$\mathrm{Var}\{\underline{z}\} = a^2\,\mathrm{Var}\{\underline{x}\} + b^2\,\mathrm{Var}\{\underline{y}\},$$

or similarly

$$\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2.$$

Special case: the situation in which z is the sum or difference of x and y, i.e., $a = \pm 1, b = \pm 1$, and x and y are statistically independent of each other. Then we get a simple, Pythagoras-like, much used equation:

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2.$$

Often, this special case is referred to as the law of error propagation.

2.4.5 Multi-dimensional distributions

Often, like in the above football example, we speak of stochastic quantities existing of several components. E.g., the co-ordinates (x, y) of a point in the plane. In the same way as in the above case, we may draw two-dimensional histograms and speak about the probability density distribution $p(x, y)$.

Figure 2.8 shows a two-dimensional normal density distribution, and how its *standard ellipse* or *error ellipse* is defined. It is the two-dimensional counterpart to the mean error. A cross section through the error ellipse produces again always a one-dimensional Gaussian bell curve, the mean error points of which are precisely intersections of the cutting line and the ellipse.

Whereas the mean-error points of the one-dimensional distribution are at $x = \mu \pm \sigma$, the formula for the error ellipse is more complicated. The expectancy ("true value") of the stochastic co-ordinate pair $(\underline{x},\underline{y})$ is itself a co-ordinate pair, (μ_x, μ_y), the centre of the error ellipse.

In the general case where the axes of the error ellipse are not parallel to the co-ordinate axes (and of different lengths!), the stochastic quantities $\underline{x}$ and $\underline{y}$ can no longer be treated as independent variables: as one says, they *correlate* with each other. This means that the most probable value of x will depend on the actual value of y, and vice versa. In adjustment calculus, this statistical dependence must be taken into account in order to arrive at the best possible (i.e., optimal) estimates both of μ_x and of μ_y.

Correlation is a curious thing. It is always good to remember, that statistical dependence does *not* necessarily mean that there is a direct cause-and-effect relationship, only that the observations have something in common. The connection may be indirect, like in the famous case in

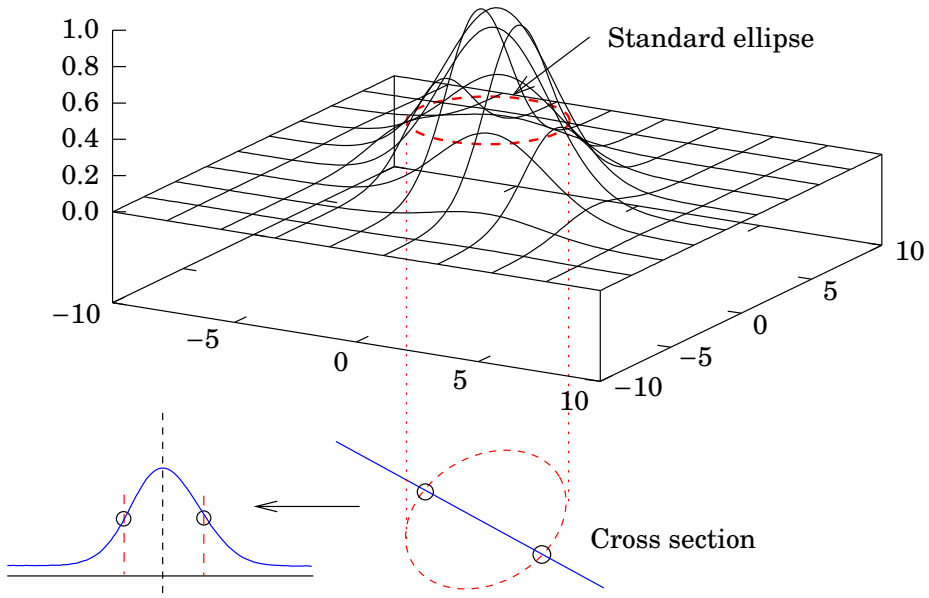


Figure 2.8. A two-dimensional probability density distribution.

which the icecream sales over summer correlated with the drowning statistics.

In three-dimensional space there exist probability density distributions of three arguments $p(x,y,z)$, the visual representation of which is a triaxial *error ellipsoid*.

2.4.6 The variance matrix

The variances of multi-dimensional stochastic quantities become themselves also multi-dimensional, i.e., they become *variance matrices*: let[16]

$$\underline{\mathbf{x}} \stackrel{\text{def}}{=} \begin{bmatrix} \underline{x} \\ \underline{y} \end{bmatrix}.$$

Then the variance matrix is

$$\mathrm{Var}\{\underline{\mathbf{x}}\} \stackrel{\text{def}}{=} \begin{bmatrix} \mathrm{Var}\{\underline{x}\} & \mathrm{Cov}\{\underline{x},\underline{y}\} \\ \mathrm{Cov}\{\underline{x},\underline{y}\} & \mathrm{Var}\{\underline{y}\} \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \stackrel{\text{def}}{=} \Sigma_{\mathbf{xx}},$$

in which again apply the definitions

$$\mathrm{Var}\{\underline{x}\} \stackrel{\text{def}}{=} E\left\{\left(\underline{x} - E\{\underline{x}\}\right)^2\right\},$$

$$\mathrm{Cov}\{\underline{x},\underline{y}\} \stackrel{\text{def}}{=} E\left\{\left(\underline{x} - E\{\underline{x}\}\right)\left(\underline{y} - E\{\underline{y}\}\right)\right\},$$

and in which we see the often used notation[17]

$$\mathrm{Var}\{\underline{x}\} = \sigma_x^2,$$

$$\mathrm{Cov}\{\underline{x},\underline{y}\} = \sigma_{xy}.$$

---

[16] ... and of course also the expectancy is a vector: $E\{\underline{\mathbf{x}}\} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$!

[17] Sometimes we see in older texts $\mathrm{Var}\{\underline{x}\} = m_x^2, \mathrm{Cov}\{\underline{x},\underline{y}\} = m_{xy}$ as back then, typewriting Greek letters was difficult.

In appendix A a short explanation is given of the basics of matrix calculus.

In fact, the *error ellipse* is a graphical representation of this 2×2 variance matrix. If $\sigma_{xy} = 0$, we say that $\underline{x}$ and $\underline{y}$ do not correlate with each other, or, carelessly[18], that they are statistically independent of each other.

Also the propagation law of variances, equation 2.4:

$$\mathrm{Var}\left\{\underline{z}\right\} = a^2\,\mathrm{Var}\left\{\underline{x}\right\} + b^2\,\mathrm{Var}\left\{\underline{y}\right\} + 2ab\,\mathrm{Cov}\left\{\underline{x},\underline{y}\right\},$$

may be written into a new, more general form: if we form the row vector

$$\mathrm{a} \overset{\mathrm{def}}{=} \begin{bmatrix} a & b \end{bmatrix},$$

and the corresponding column vector[19]

$$\mathrm{a}^\mathsf{T} = \begin{bmatrix} a & b \end{bmatrix}^\mathsf{T} = \begin{bmatrix} a \\ b \end{bmatrix},$$

we obtain

$$\mathrm{Var}\left(\underline{z}\right) = \begin{bmatrix} a & b \end{bmatrix} \cdot \begin{bmatrix} \mathrm{Var}\left\{\underline{x}\right\} & \mathrm{Cov}\left\{\underline{x},\underline{y}\right\} \\ \mathrm{Cov}\left\{\underline{x},\underline{y}\right\} & \mathrm{Var}\left\{\underline{y}\right\} \end{bmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix} = \mathrm{a} \cdot \mathrm{Var}\{\underline{\mathrm{x}}\} \cdot \mathrm{a}^\mathsf{T}.$$

Understanding this requires knowing how to multiply matrices (row × column). Writing it out gives us back the original equation 2.4.

This is presented in the literature in somewhat varying notations:

$$\sigma_z^2 = \mathrm{a}\,\mathrm{Var}\{\underline{\mathrm{x}}\}\,\mathrm{a}^\mathsf{T} = \mathrm{a}\Sigma_{\mathrm{xx}}\mathrm{a}^\mathsf{T}.$$

We will discuss more generally the propagation of variances in linear models of many variables in section 13.7 on page 336.

2.5 Geodetic observables

2.5.1 Angles, directions

Traditionally, the large national or continental geodetic networks have always been *triangulation networks*. Between the network stations, *directions* are measured, and the measurements in the network are adjusted: from the measured values, small contradictions caused by the uncertainty of measurement are computationally removed, leading to the solution of the *triangulation*. Back then, distance measurement was

[18] . . . because there may be a more complicated statistical dependence that doesn't show as correlation.

[19] We often use the transpose notation to make it easier to write a vector in running text.

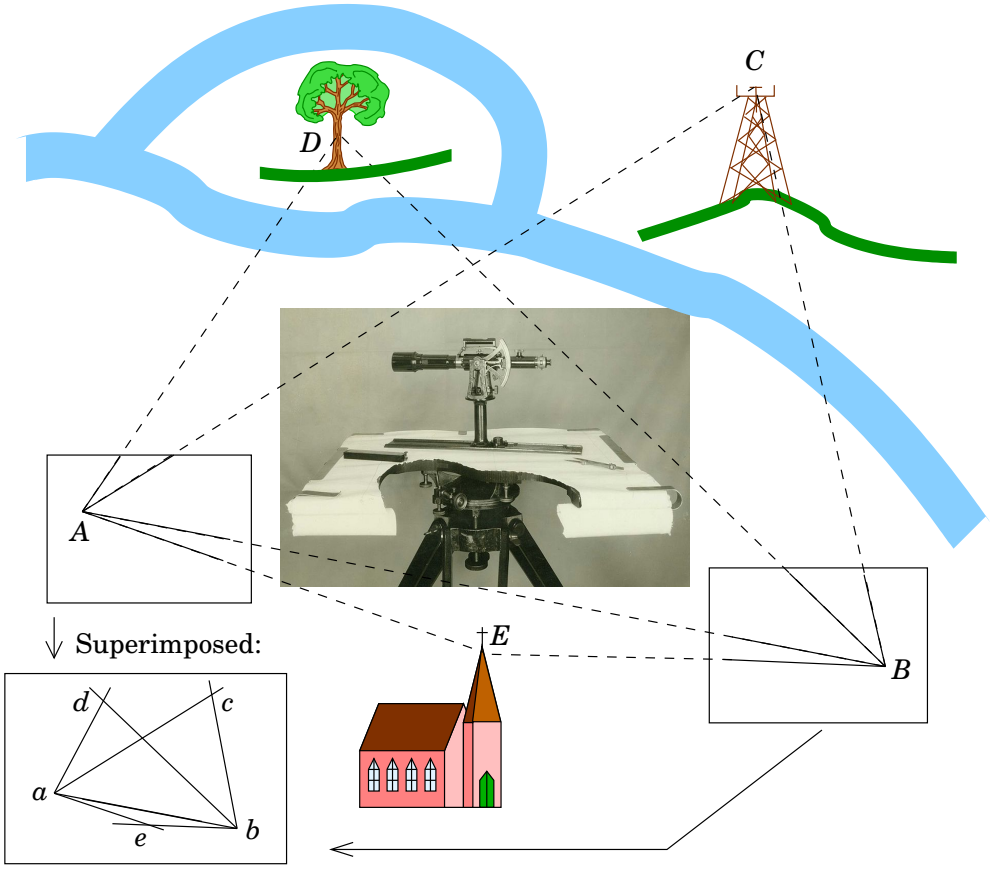


Figure 2.9. Triangulation by means of a plane table and alidade. Photo © Wikimedia Commons.

□

difficult over larger distances, as the only working technique was measurement using mechanical means such as measuring tapes and rods.

In triangulation, the procedure is that in a network only all directions are measured, and in addition *one single distance*, and from these, all other distances, point locations, etc. are computed. This is how Snellius carried out his famous grade measurement between Bergen op Zoom and Alkmaar, cities 100 km apart, by measuring directly only the baseline of 326 *roeden* long, built by him in a meadow! Similarly the Lapland grade measurement: the baseline was built during winter 1736 on the ice of the Torne river, all other measurements were directions between the points of the triangulation network.

The basic idea of triangulation can be explained using the *plane table*[20]. mittapöytä
In point A in the terrain a plane table is erected, a sheet of transparent paper placed on it, and all directions to the terrain points B, C, D and E[21] are drawn. Move to point B, and draw a similar rose of directions to

[20] German *Messtisch*, French *planchette*, Swedish *lantmätartavla*, Danish, Norwegian *målebord*.

[21] The drawing is made easier by using an *alidade*, German *kippregel*.

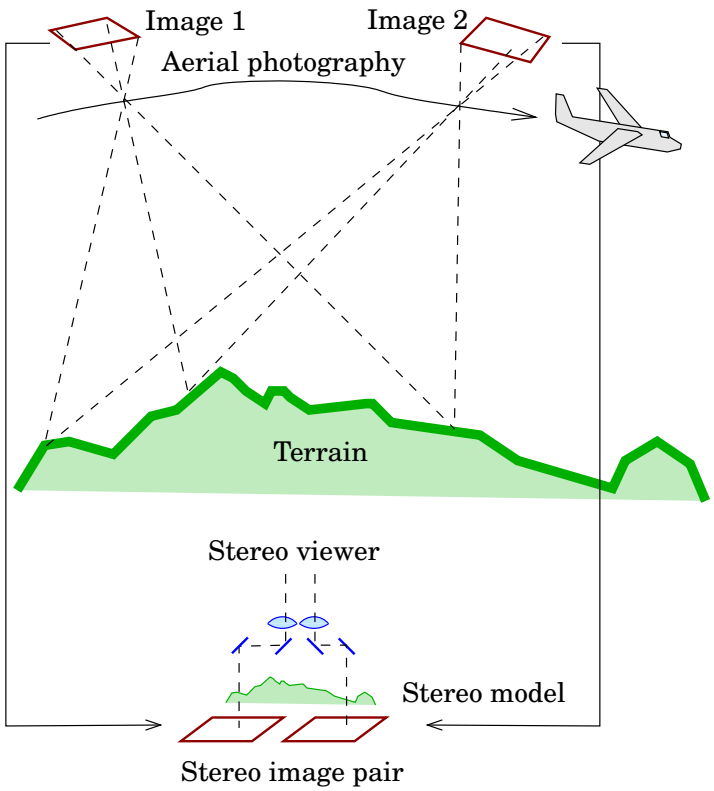


Figure 2.10. Forming a stereo model in photogrammetry.

□

targets A, C, D and E. In the office, the sheets of paper are placed on top of each other, and the result is a miniature image of the true geometry of the terrain – at a scale $m = ab : AB$! So, if the distance AB in the terrain is 3 km, and on the superposed papers $ab = 30$ cm, then the scale of the thus obtained "map" is 1 : 10,000.

Nowadays directions are measured by *theodolite*, a precise measurement instrument for horizontal and vertical angles, and creating the map takes place computationally.

In photogrammetry there is a three-dimensional analogue to the plane table method, *stereo model restitution*, which however nowadays is also invariably realized fully digitally. Figure 2.10 shows, how from two aerial images a stereo model is formed using a viewing device. In a stereo restitution instrument similarly a model is formed, inside which a *floating mark* may be moved around three-dimensionally. The coordinates of the mark are continuously output to a computer hooked up with the instrument, and maps can be drafted immediately.

avaruusmittamerkki

ilmakolmiointi

Also *aerotriangulation* is based on forming stereo models.

A fairly modern geodetic technique based on the idea of the plane table but in three dimensions, is Yrjö Väisälä[22]'s *stellar triangulation*. This

[22] Yrjö Väisälä (1891–1971), "the Wizard of Tuorla", was a Finnish astronomer, physicist, geodesist, metrologist, builder of telescopes, finder of comets and as-

technique, in which the stellar background is used for measuring directions, is discussed in section 16.3.

2.5.2 Distances, distance differences

In the past, distances could only be measured mechanically, using measuring wires, tapes or rods. The precision achieved in the measurements was often impressive thanks to careful procedures.

In local surveying, steel measuring tapes continue to be used, as they are inexpensive and easy to use and carry along. One should only remember to clean and grease them after use. The lengths are 20 – 60 m. The precision measuring wires of old were made of invar, a steel alloy with a very low thermal expansion coefficient.

Today, distance measurements are done electronically or electro-optically. Only a straight line of sight between points is needed. The devices may use microwaves (Tellurometer) or more often visible light (Kern Mekometer), laser light (Geodimeter[23]), or, in the case of most modern range finders (distance measurement devices) and total stations, infrared light produced by a light-emitting diode (LED). The effect of atmospheric delays on signal propagation must always be carefully taken into account in reducing the observations.

hohtodiodi

Terrestrial geodetic measurement instruments are able to measure both horizontal and vertical angles as well as slant ranges. They are called *electronic tacheometers*[24] [25], and they thus combine the properties of theodolite and range finder (distance measuring device) in one integrated, fully electronic, highly automated device.

Also satellite geodesy uses electronic distance measurement. The GNSS, Global Navigation Satellite Systems – like the Global Positioning System GPS – in broad use today are based on distance measurements made by microwaves – more precisely, distance *difference* measurements[26]. Satellite lasers on the other hand (Metsähovi!) measure the travelling time from the observation station to a satellite reflecting the light, and back.

The advantage of using electronics is, that the incredibly high precision of frequency measurement is thus harnessed as distance measurement

---

teroids, recreational sailor and Esperanto practitioner.

[23] Geodimeter™ is a Swedish invention and the name is a trademark of Trimble AB, Danderyd, Sweden.

[24] Or more often, *total stations*.

[25] Tacheometer, Greek *rapid measurer*.

[26] Many so-called *hyperbolic positioning systems*, like Decca, Transit/Doppler, and also GPS and the other satellite positioning systems, are based on distance difference measurement.

precision. When the precision of frequency measurement can well be $1:10^{12}$, it is understandable that with GPS, we measure intercontinental distances with a relative precision of even $1:10^{9}$. The satellite techniques are even more precise than the terrestrial ones, because such a large part of signal propagation takes place outside the denser parts of the atmosphere.

2.5.3 Potential differences, levelling

Measurement of geopotential differences is traditionally done by *levelling*: one measures the height difference between two points in metres, and converts this difference to a potential difference using local gravity g. If the height difference is ΔH and the difference in geopotential is ΔC, we use the equation[27]

$$\Delta C = g\Delta H.$$

The measurements are done along a line, and the differences ΔC_i are summed up. From the lines, a closed *levelling network* is built, in which the potential differences calculated around loops must sum to zero. The network is adjusted using this condition.

In subsection 4.3.1 a few more exotic height determination alternatives to traditional levelling are presented, many of which are based on the direct physical comparison of geopotential values, e.g., using a fluid surface.

2.6 About co-ordinates

In geodesy we use, in order to describe the figure and size of the Earth and to determine the locations of points on or near the Earth's surface, *co-ordinates*.

The co-ordinate reference systems used in geodesy are generally *three-dimensional*, as the Earth is a three-dimensional object in a three-dimensional space. As an example we may mention latitude, longitude and height (φ, λ, h), which let us describe the location of a point in an intuitive way.

"Two-dimensional" co-ordinate reference systems are really *map projection co-ordinates*, i.e., derived quantities, and not directly interesting to geodesy. They belong more to the field of cartography, though they are used very widely in applied surveying. E.g., on older Finnish topographic maps one encounters KKJ co-ordinates, which are map-plane

[27] This is an example of "work is force times path". Work (per unit of mass) is ΔC, path is ΔH, and force (per unit of mass, i.e., according to the law $F = ma$, *acceleration*) is g.

co-ordinates, i.e., the kind of co-ordinates (x, y) used on a map sheet, directly measurable by ruler.

In addition to spatial co-ordinates, of course also used are *time*, for describing processes of change, and in physical geodesy, as a kind of co-ordinate, the *geopotential*, the potential of the gravity field of the Earth. See subsection 4.1.1.

2.7 Why plane co-ordinates

Because the Earth is a three-dimensional object, geodesy is a three-dimensional science. The Earth and points connected to her are located in three-dimensional space, and the task of geodesy is to describe their location with three-dimensional co-ordinates, e.g., (X, Y, Z). Modern measurement systems, like the global positioning system GPS, are able to directly measure three-dimensional co-ordinates.

Geodesy, however, is also an Earth science, an applied branch of science serving people. Mankind lives, forced by gravity, in a quasi two-dimensional "subspace", in which there is freedom of movement almost only in the horizontal direction, along the surface of the Earth. Furthermore, the important medium of communication called *paper* is unconditionally two-dimensional, and maps are commonly drawn on paper!

For this reason, *plane co-ordinates* are very generally used in geodesy and surveying, rectangular, two-dimensional co-ordinates in the horizontal plane.

There exist a multitude of practical plane co-ordinate reference systems suitable for surveying work. The main differences between them are:

1. the location of the *origin* and the *orientation* of the axes. The origin, the starting point where $x = y = 0$, must be known. Generally the axes are x to the North and y to the East, but not always, and not necessarily accurately.
2. the *technology of determination*, i.e., geocentricity: modern plane co-ordinates are obtained from geocentric, three-dimensional co-ordinates produced using GNSS.

Geodetic plane co-ordinates are, in fact, *map projection co-ordinates*: they have been calculated using map projection formulas from originally three-dimensional co-ordinates through geodetic latitude and longitude (φ, λ).

In a very small area, like a building site, no proper map projection is needed. In those, plane co-ordinates may be understood as a special, rectangular case of *topocentric* (observation-site centric) *co-ordinates*.

2.8 Co-ordinates of location in three dimensions

Three-dimensional co-ordinates can be of different types, e.g., *rectangular* or *geodetic*, also called geographical. The most common co-ordinate reference system used is the *three-dimensional rectangular or Cartesian* (X,Y,Z) *system*. Often it is also *geocentric*.

Geocentricity means literally, that the origin is in the centre of mass of the Earth (to a certain accuracy). Additionally, the Z axis of a geocentric system points along the Earth's axis of rotation, i.e., to the celestial (North) pole.

In a geocentric co-ordinate reference system the direction of the X axis is in principle *arbitrary*, i.e., conventionally agreed. For a geocentric system on the Earth's surface, the *meridian which contains the direction of the plumb line at the Royal Observatory Greenwich* serves as an international standard[28] for fixing the direction of the X axis.

The X axis lies both in the plane of the Greenwich meridian and in the equatorial plane, and is thus perpendicular to the Z axis. The Y axis in turn is perpendicular to both the Z and X axes, so that all three axes are mutually perpendicular. See figure 2.12.

Three-dimensional rectangular co-ordinates X,Y and Z may be generally applicable, but they are not particularly intuitive. E.g., the co-ordinates of the GPS antenna at the Metsähovi research station are, expressed in the Finnish EUREF-FIN reference frame (see subsection 3.2.3):

$$X = 2{,}892{,}571.1204\,\mathrm{m},$$
$$Y = 1{,}311{,}843.2621\,\mathrm{m},$$
$$Z = 5{,}512{,}633.9521\,\mathrm{m}.$$

The numbers are interesting looking, but don't give a very enlightening or easy to grasp answer to the question "where is Metsähovi?"...

As a first step to more practical co-ordinates we construct *geodetic co-ordinates*. First we construct mathematically a reference ellipsoid, a suitably flattened ellipsoid of revolution the measures of which are reasonably close to those of the real Earth[29].

[28] The treaty of Washington DC of 1884 made the Greenwich meridian the world's zero or reference meridian. At the same time, a "world time" or universal time was agreed: Greenwich Mean Time, GMT. The basic idea was, that civil times of countries would differ from GMT by a whole number of hours, for Finland +2 h in winter (EET) and +3 h in summer (EEST). Without this time zone system, international traffic (by sea, air or telephone) would be cumbersome.

[29] E.g., the GRS80 reference ellipsoid: equatorial radius 6,378,137.0 m, polar radius 6,356,752.3141 m (some 21 km shorter), and flattening 1 : 298.257,222,101.

Figure 2.11. The Greenwich meridian. Image © Wikimedia Commons.

A point is projected along the ellipsoidal *normal* onto the surface; the distance of projection, h, is the ellipsoidal height, the *direction angles* of the projection line or ellipsoidal normal[30] φ, the geodetic latitude reckoned from the equator, and λ, the geodetic longitude, reckoned from the Greenwich meridian. The triplet (φ, λ, h) – or often only the pair (φ, λ) – is referred to as *geodetic co-ordinates*. The geodetic co-ordinates of Metsähovi are, e.g.:

$$\begin{aligned} \varphi &= 60^\circ 13' 2''.89046, \\ \lambda &= 24^\circ 23' 43''.13336, \\ h &= 94.568\,\mathrm{m}, \end{aligned} \tag{2.5}$$

... and this tells people already a whole lot more about where this point is located!

There is a simple mathematical relationship between rectangular and geodetic co-ordinates: each can be *converted* into the other without losing precision. They are equivalent[31] presentations of the location of a point:

$$(X, Y, Z) \quad \begin{array}{c} \text{equation} \\ \Longleftrightarrow \\ \text{equation}^{-1} \end{array} \quad (\varphi, \lambda, h)$$

[30] Note that the ellipsoidal normal generally does *not* go through the centre of the ellipsoid! See figure 2.12.

[31] So, if (X, Y, Z) is given, then (φ, λ, h) can be calculated, and also the reverse is true.

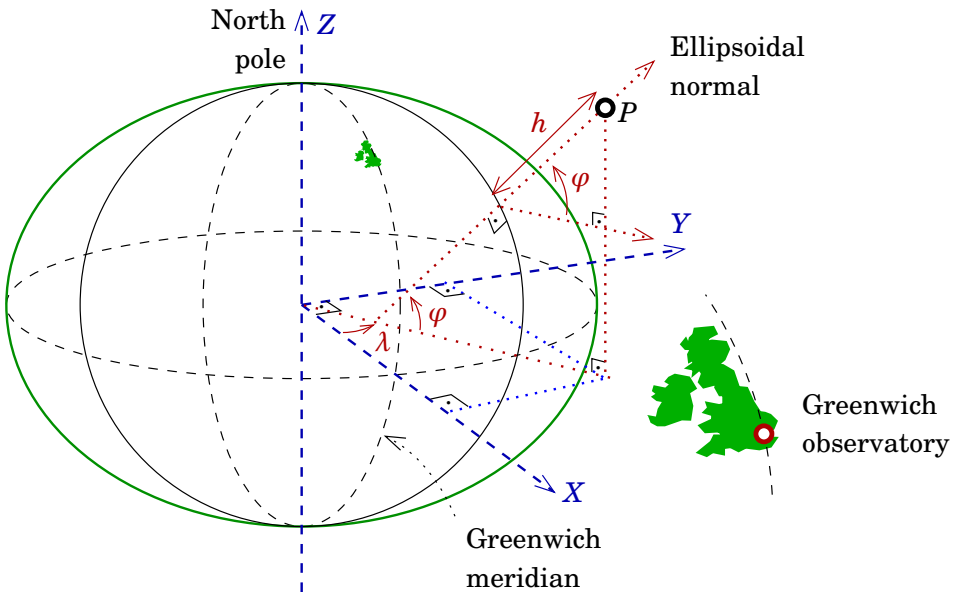


Figure 2.12. Rectangular and geodetic co-ordinates.

□

... and both are geocentric and three-dimensional. The differences between them are:

- rectangular co-ordinates are more convenient to handle in numerical work, but
- geodetic co-ordinates are more intuitive, closer to the human understanding.

In addition to geocentric co-ordinates, often *topocentric* co-ordinates are encountered; co-ordinates reckoned relative to some local origin. Topocentric co-ordinates are naturally produced by measurements in which the location of the measurement instrument becomes the origin of the co-ordinate system: one measures a direction (*azimuth* or horizontal direction angle, and *elevation* angle above the horizon) and a *distance* from instrument location to the points to be measured.

korkeuskulma

□

Self-test questions

1. What is a measurement unit, a measured quantity, and its dimension? Describe and explain the differences.
2. "Three score years and ten". How much is this in gigaseconds?
3. The mean temperature of the Earth for 2016 was[32] 14.8 °C. How much is this in kelvin?
4. Explain random errors, gross errors, systematic errors.
5. Consider a pack of 52 playing cards, with the numbers 2 – 10, a

[32] Measured at a height of 2 m following standard procedure.

jack counted as 11, a queen as 12, a king as 13, and an ace as 1. What is the expected value for a card drawn at random from the pack?

6. Look up a table of the normal distribution's *probability density* $p(x)$. What is the function value $p(x)$ for $x = \mu \pm \sigma$? And for $x = \mu$? And what is the ratio $\frac{p(\mu \pm \sigma)}{p(\mu)}$? (Be aware that tabulations of the normal distribution are for the distribution for which $\mu = 0$ and $\sigma = 1$, the *standard normal distribution*.)
7. Would you expect that the incidence of bush fires in Australia and the thickness of snow cover in Lapland would be correlated? Anti-correlated? Uncorrelated? Why?
8. Why do we use two-dimensional (plane) co-ordinates in geodesy?
9. How is geocentricity defined?
10. What is the advantage of using geodetic co-ordinates (φ, λ, h) over rectangular co-ordinates (X, Y, Z)?

Exercise 2–1: Co-ordinates and a street address

A riddle for seafarers

1. To what street address do the geodetic co-ordinates in the accompanying picture refer?

2. What is the nature of the object at the address? Picture!
3. Are the co-ordinates geocentric? On which reference ellipsoid?
4. Find the "Easter egg" in the index of this text.

3. Map projections, datums and transformations

3.1 Map projections

Although geodetic co-ordinates are already a whole lot more practical than rectangular ones from the viewpoint of the end user, they can certainly be improved still. The method for this is called *map projection*.

The practice of depicting the Earth's surface on a two-dimensional plane or map is an old one. *Cartography* is the science that has grown around this art. If the area to be depicted is small, the method of depiction is straightforward and error-free: local horizontal rectangular landscape co-ordinates (x, y) can be depicted, through a *scale*, onto the paper map plane. We speak of a *plan*[1]. In most countries, also Finland, the x axis points North and the y axis East.

In a larger area we use a *map projection*, a mathematical method to map the location, or latitude and longitude (φ, λ), of a point to the map plane. Thus we can draw a graphical representation of the Earth's surface onto paper, a *map*. Many digital applications – like CAD software – that don't actually even require the use of a paper map, nevertheless are based on the intuitive use of the map plane.

In the simplest approach we use the co-ordinate pair (φ, λ) directly as map co-ordinates x and y, i.e., $x = S\varphi$, $y = S\lambda$, S being the nominal scale. This is a pathetic solution, because

- The co-ordinates φ and λ are in degrees, angular units, whereas the map co-ordinates have to be in metric units.
- One degree of longitude λ expressed in kilometres diminishes toward the poles. At the latitude of Helsinki, a degree of longitude is only 55 km, when at the equator it is 111 km. A slightly better approach is to use the co-ordinate pair $(\varphi, \lambda \cos\varphi)$.

karttaprojektio-oppi

Better solutions are offered by *map projection science*. Thus we may map the parameter pair (φ, λ) on the surface of the reference ellipsoid onto

[1] Swedish *plankarta* (archaic; nowadays the term refers to zoning maps), Dutch *plattegrond*.

(x, y) in the map plane in a sensible way. Unfortunately a method that depicts everything exactly correctly doesn't exist. *Something is always distorted*[2].

In map projections we always make approximations. As we already mentioned in section 3.1, we always lose something. There does not exist projections that would't distort anything, or the scale of which would be the same thoughout the map plane. A projection is chosen *according to the purpose of use*, so that something that is important to the user is preserved: the shapes of objects, their surface area, some distances, or the compass direction.

- If angles and ratios of distances are preserved, we speak of a *conformal* or angle-preserving projection. In this case both the linear scale (distances) and the scales of surface areas are distorted, except in some special points of the map. kulmatarkka
- If surface area is preserved, we speak of an *equivalent* or equal-area projection. Here, angles and shapes are distorted, again with the exception of special points or lines. pinta-alatarkka
- If distances are preserved, we speak of an *equidistant* projection. A projection can be equidistant only along certain lines, never everywhere. etäisyystarkka
- In navigation it is important that compass directions are preserved. In that case, *loxodromes*[3] or rhumb lines, lines of constant compass direction, are straight. The Mercator projection has this property.
- If great circles are depicted as straight lines, we have the *gnomonic* projection. It is used in aviation and in connection with meteor observations – the track of an aircraft on the Earth's surface or the track of a meteor on the sky is mapped to a straight line, which may be drawn with a ruler.

It is important to understand that *there is no "correct" projection*! The choice is dictated by the purpose of use and the distortions considered acceptable. In a way they are all "wrong". On the other hand, they are also all *useful*[4].

Philosophically one may be of the view, that map co-ordinates (x, y) *aren't actually geodetic quantities*. Genuinely geodetic quantities are always three-dimensional. Map projection co-ordinates are *derived quantities*.

[2] https://en.wikipedia.org/wiki/Theorema_Egregium.

[3] https://en.wikipedia.org/wiki/Rhumb_line.

[4] "Essentially, all models are wrong, but some are useful" – George E. P. Box (1919 – 2013), statistician.

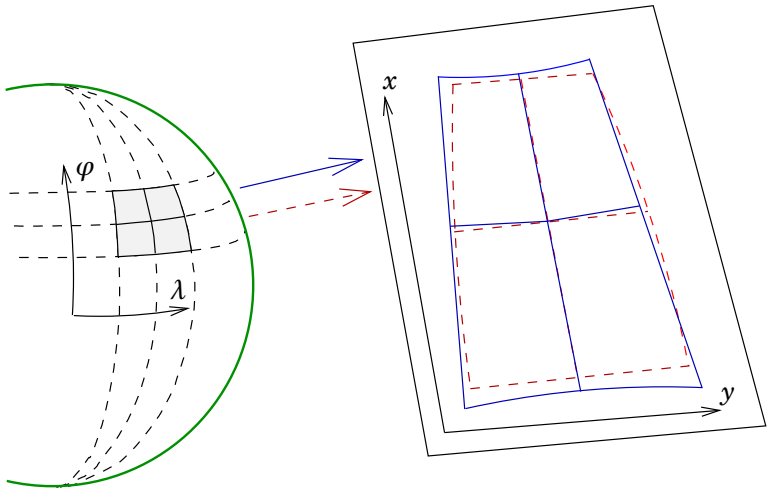


Figure 3.1. Depicting the curved surface of the Earth to the map plane using different projections. Something will always be distorted!

3.2 The various co-ordinate solutions used in Finland

In Finland, like everywhere on Earth, several different co-ordinate reference frames[5] are in use for historical reasons. This complicates the use of geographic information and co-ordinate materials. It is a precondition for use of co-ordinate data, that we know in which co-ordinate reference frame the material is. If necessary, one *transforms* from one frame to another before use.

The most fundamental distiction between different co-ordinate reference frames is, whether we have a "traditional" one, i.e., created before the satellite era using traditional geodetic measurement methods, or "modern", i.e., geocentric, created using satellite positioning technology. In the sequel we tell more about co-oordinate reference frame alternatives and the transformations between them.

3.2.1 National Map Grid Co-ordinate System (KKJ)

In Finland, the old, in its time official, and now largely obsolete, co-ordinate reference frame is KKJ, the *National Map Grid Co-ordinate System*, (Parm, 1988). The system was created based on the results of the national triangulation effort. It is also a good example of the use of a *map projection* for mapping a national territory onto a two-dimensional map plane. More about the map projection in subsection 3.3.1.

[5]Here, we don't yet make any clear distinction between co-ordinate reference *systems* and reference *frames*. A bit more on this in section 3.7.

Figure 3.2. The systematic shift between the road network and the aerial photograph base could have something to do with the use of different co-ordinate reference frames. Google Earth™, © see text in image.

3.2.2 Helsinki system (VVJ)

The Helsinki system, or VVJ, was the predecessor of KKJ. The national triangulation was a huge project[6], lasting for over half a century. The users of accurate co-ordinates, however, could not wait. Thus, VVJ was created "on the fly" in support of practical mapping work. The co-ordinates were calculated stage after stage, as the Finnish triangulation progressed from Southern Finland toward the North. Not until the 1970s, when the national triangulation was almost completed and its complete adjustment carried out, was the KKJ system formed as a nationally unified system.

The co-ordinate differences between the Helsinki system and KKJ are up to a few metres (Tikka, 1991, page 193).

Ancient as the Helsinki system may be, it was in use in many municipalities still in the 2010s: e.g., Tuusula moved away from it (directly to EUREF-FIN) not until 2013...

3.2.3 The EUREF-FIN reference frame

With the spread of satellite positioning, already in the 1990s it was possible to determine everywhere on Earth three-dimensional co-ordinates, in a system that was on the several-centimetre level *geocentric*: the *origin* of three-dimensional co-ordinates (X, Y, Z) is that close to the centre of mass of the Earth. In addition to this, the Z axis of the co-ordinate system points in the direction of the Earth's axis of rotation, whereas the X and Y axes lie within the equatorial plane.

Geocentric systems are *global*. In Europe, the International Association

[6] https://www.fig.net/resources/proceedings/fig_proceedings/fig2008/papers/hs05/hs05_02_puupponen_jarvinen_2934.pdf.

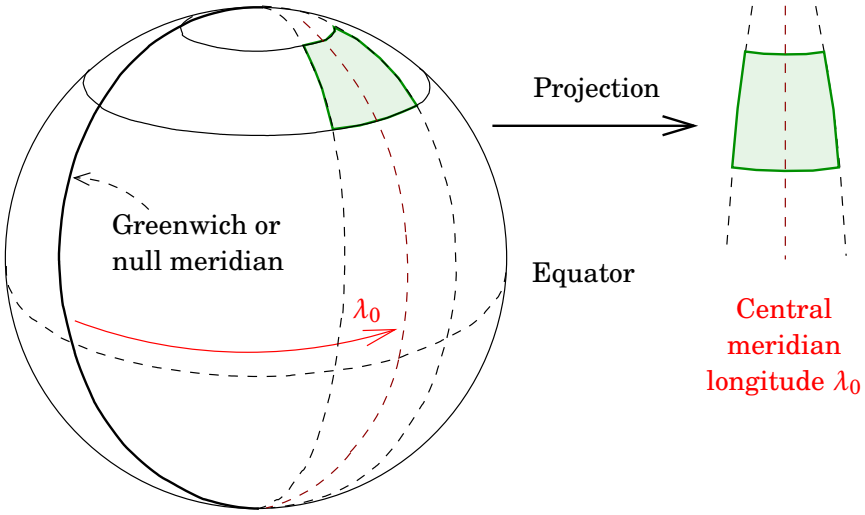


Figure 3.3. Imaging the curved Earth's surface as a narrow zone onto a plane. This is the principle of both the Gauss-Krüger and the UTM projection as used in Finland. The distortions remain acceptably small only in a relatively narrow area.

of Geodesy IAG[7] has created the geocentric reference system ETRS89, the European Terrestrial Reference System 1989. The system has been *realized* by measurements many times: the first Europe-wide realization, or *co-ordinate reference frame*, was EUREF89, which was based on a space geodetic measurement campaign (Overgaauw et al., 1994). Later, Finland created its own national realization under the name EUREF-FIN[8], which has in recent years been taken into use by the national mapping agencies and most other actors.

In subsection 3.3.3 we tell more about the map projections chosen for use with the EUREF-FIN reference frame.

3.3 Map projections used in Finland

In Finland we encounter currently two different plane or projected co-ordinate reference frames: the old KKJ, and the new system based on EUREF-FIN. In the old KKJ, the Gauss-Krüger projection was used, in the new system is used, in addition to Gauss-Krüger, also the UTM (Uni-

[7]In fact this was done by its European reference frame subcommission, the *Reference Frame Subcommission for Europe*, http://www.euref.eu/. The name EUREF was proposed by the Danish geodesist Knud Poder.

[8]The often used name WGS84 refers to the system – which actually has half a dozen realizations – maintained by the U.S. defence authorities, agreeing with the various ETRS89 realizations only roughly (±0.01 – 1 m), see http://www.navipedia.net/index.php/Reference_Frames_in_GNSS. Often the name WGS84 is used – erroneously – as shorthand for these.

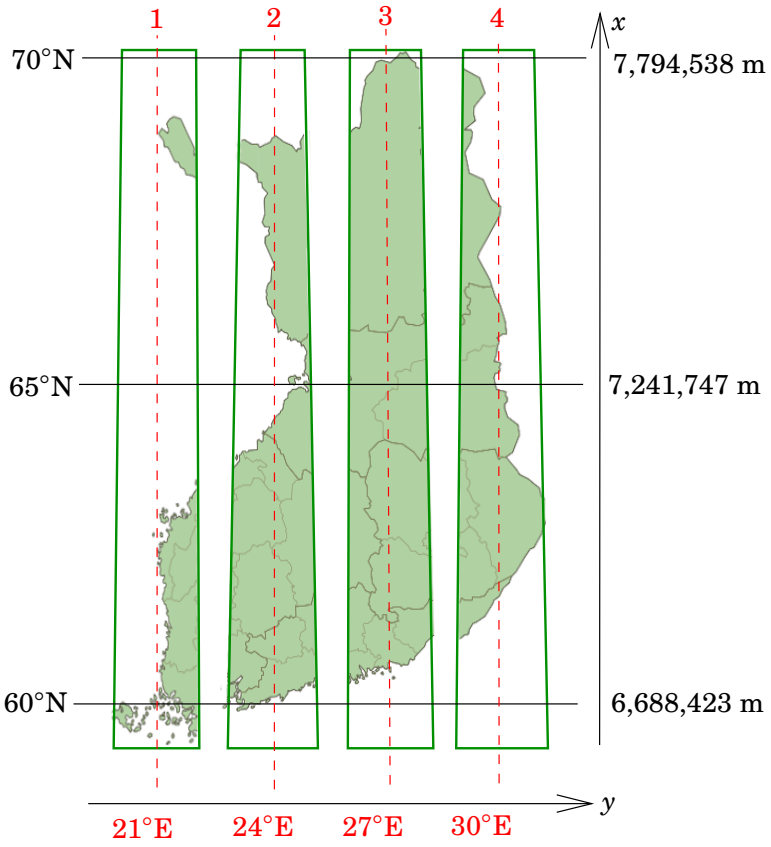


Figure 3.4. The zone division of the (obsolete) Finnish KKJ system's Gauss-Krüger projection. The zones 0 (central meridian 18°) and 5 (33°) have been left out.

versal Transverse Mercator) map projection. See figure 3.3.

3.3.1 The map projection system of the KKJ

In order to map the Earth's surface to the map plane without large distortions, Finland was divided into six *KKJ projection zones*, every one of which has its own co-ordinate frame in the projection plane. The zones were numbered from zero to five: the central meridians of the zones were at longitudes 18°, 21°, 24°, 27°, 30° and 33° East.

The co-ordinates within a projection zone are x (Northing) and y (Easting). The projection used is *conformal* and goes by the name of Gauss-Krüger, one type of transversal Mercator projection. The reference ellipsoid used in the projection calculations was the Hayford or international ellipsoid of 1924.

In the several zones the same co-ordinates appear multiple times. Therefore, to obtain unique values, the zone number was prepended to the y co-ordinate as its first decimal (except when it was 0). See figure 3.5. This multi-zone system is referred to as the *basic co-ordinate system* of the KKJ. peruskoordinaatisto

The x co-ordinate origin is on the equator, and the values grow along

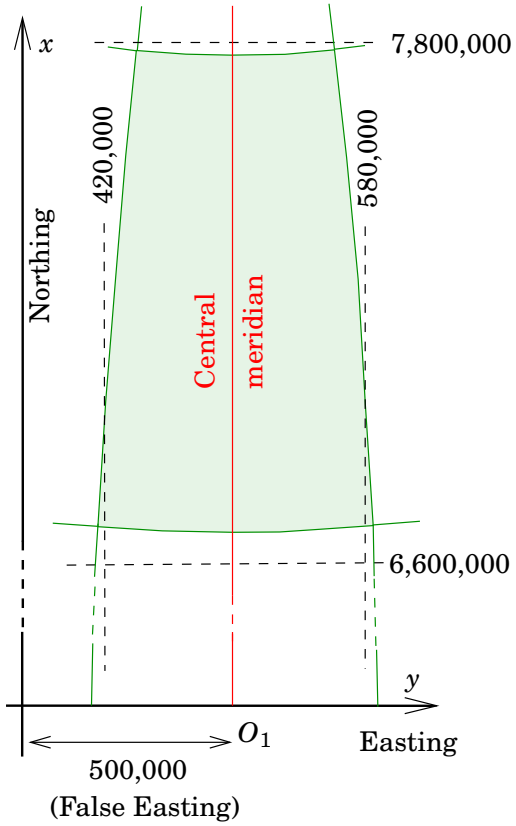


Figure 3.5. The geometry of one zone of the Finnish KKJ co-ordinate frame (stretched in the East-West direction).

the central meridian. Because of this, the x co-ordinates in Finland are numbers in the range 6,600,000 – 7,900,000 m.

The y co-ordinate describes *distance from the central meridian*. In order to avoid negative y co-ordinates, 500 km was added to them, i.e., a point on the central meridian has a y co-ordinate of 500,000 m (*false Easting*). The y axis is *perpendicular* to the x axis. The y co-ordinates could theoretically be in the interval 0 – 1,000,000 m, in practice however the range of values is 420,000 – 580,000 m due to the narrowness of the zones at Finnish latitudes.

E.g., according to figure 3.5, the bottom left corner co-ordinates of the first zone would be

$$x = 6{,}600{,}000.000\,\text{m},$$
$$y = 1{,}420{,}000.000\,\text{m}.$$

Similarly in the third zone:

$$x = 6{,}600{,}000.000\,\text{m},$$
$$y = 3{,}420{,}000.000\,\text{m}.$$

On the central meridian of the projection, the distortion is zero, i.e., the nominal scale of the map applies exactly.

In the case where KKJ co-ordinates were used *within a small area*, often the leftmost digits were left off, as they are all the same. Thus one obtained *truncated co-ordinates*, which one may still come across in municipal calculation documents.

The map projection system employed by the Helsinki system (VVJ) was similar to that of KKJ, except that the notation was $y = 21°420000$, i.e., the y co-ordinate was prefixed with the longitude of the central meridian itself.

3.3.2 The KKJ's Uniform Co-ordinate System

In addition to KKJ, for small-scale maps[9] only, the *Uniform Co-ordinate System* (Finnish acronym: YKJ) of KKJ was taken into use, which mapped the whole territory of Finland using one and the same Gauss-Krüger projection with a central meridian of 27°. As there is only one zone, no zone number is prepended to the y co-ordinate.

yhtenäiskoordinaatisto

3.3.3 The map projections of EUREF-FIN

In connection with the new reference frame we use EUREF-FIN map projection co-ordinates (JUHTA, 2016b), which thus are *plane co-ordinates* (x, y). The new map projections, which use the geocentric GRS80 reference ellipsoid, are, depending on the map scale, either the familiar Gauss-Krüger, or alternatively UTM, Universal Transverse Mercator, a projection type that is also conformal and is in broad international use.

- For *small-scale* maps, i.e., maps which depict all of or a large part of Finland, one chooses the new ETRS-TM35FIN projection co-ordinate frame, which is based on the three-dimensional EUREF-FIN co-ordinate frame and the UTM projection for central meridian 27°E. This replaces the old KKJ Uniform Co-ordinate System. The projection also forms the basis for the whole Finnish map-sheet division.
- Also for topographic maps ETRS-TM35FIN is used. maastokartta
- For *large-scale* maps, intended for local use, one uses the Gauss-Krüger projection like before, but with a zone width of 1°. These projections are named ETRS-GKn, in which n is the degree number of the central meridian, e.g., 23°E, in which case the name of the projection is ETRS-GK23. For practical reasons, however, this rule is bent, e.g., in order to get a whole municipality mapped onto a single zone. Due to the choice of projection type, these maps are useful for applications where no larger scale distortions can be

[9] A small-scale map is a map the *scale number* M of which is *large*, if the scale is $1 : M$. A scale of 1 : 1,000,000 is small: even big objects look small on the map, but the area mapped is large. A scale of 1 : 2000 is large: even small details are well discernible, but the area covered by the map is small.

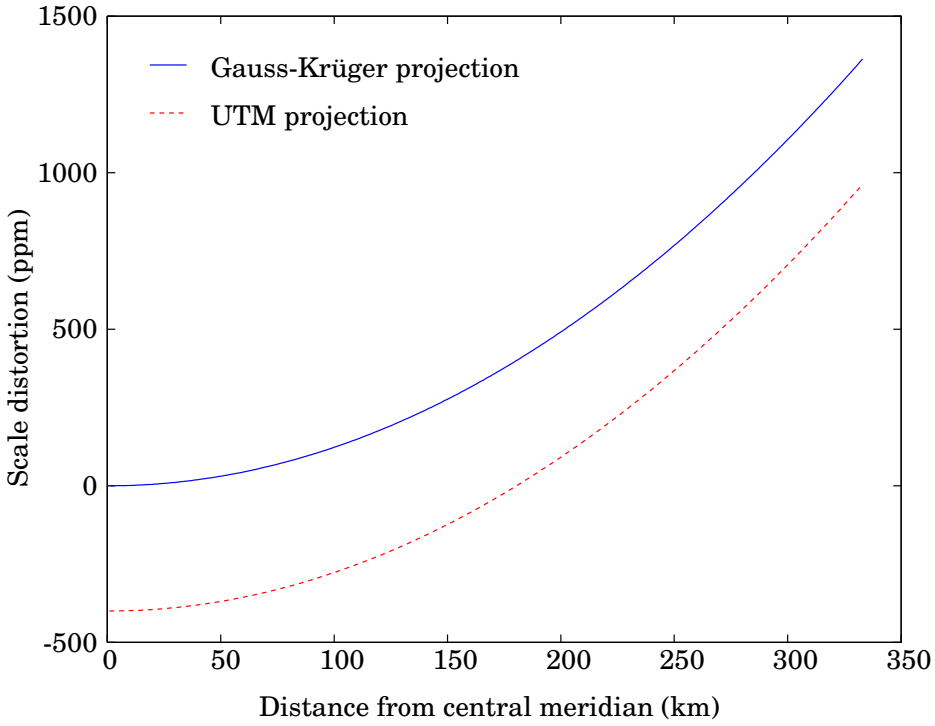


Figure 3.6. Scale distortion of Gauss-Krüger and UTM projections.

accepted, e.g., zoning and infrastructure construction, where map co-ordinates are moved straight to the construction projects as terrain co-ordinates.

Also internationally the UTM system is much used. UTM differs from Gauss-Krüger in two ways:

- The scale on the central meridian is 0.9996 rather than 1.0. This means that on the central meridian the map depicts details about 400 ppm (*parts per million*) smaller than they ought to be based on the map's nominal scale. That amounts to 40 cm for every kilometre.
- The width of the projection zones is 6°, not 3° or 1°. This means that the scale distortion, which on the central meridian amounts to −400 ppm, turns positive going to the edges of the zone, to about +1000 ppm, calculated on the equator. At Finnish latitudes the distortion remains quite a bit smaller than that.

By these two choices one aims to keep the scale distortion over the whole area of the zone within certain bounds, in spite of the zone's large width.

Both Gauss-Krüger and UTM are so-called *conformal* projections: angles and length ratios are preserved locally, small squares map to small squares, small circles to small circles.

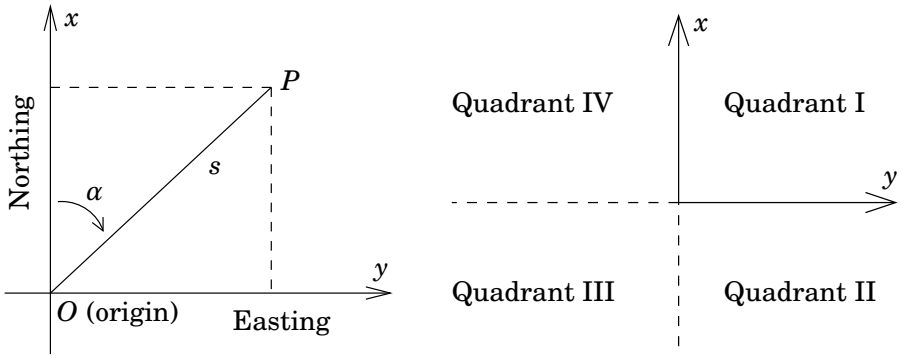


Figure 3.7. Geodetic plane co-ordinates and the quadrants of the plane I – IV.

3.4 More about plane co-ordinates

The plane co-ordinates used in geodesy differ a bit from the familiar mathematical (x, y) system. See figure 3.7. While in mathematics, the x axis points to the right and the y axis to the upper edge of the paper, it is in geodesy the habit that the x axis points North ("Northing") and the y axis East ("Easting").

In addition to rectangular co-ordinates (x, y), polar co-ordinates (α, s) are used. Sometimes the symbols used are (A, s). In geodesy, the *azimuth* or horizontal direction angle α (or A) turns clockwise from the North[10], i.e., through the East, unlike in mathematics. s is the distance from the co-ordinate origin O.

Between rectangular and polar co-ordinates exist the following trigonometric relations:

$$\begin{aligned} y = s\sin\alpha \quad &\Longrightarrow \quad \sin\alpha = \frac{y}{s} \\ x = s\cos\alpha \quad &\Longrightarrow \quad \cos\alpha = \frac{x}{s} \\ \tan\alpha = \frac{\sin\alpha}{\cos\alpha} = \frac{y}{x} \quad &\Longrightarrow \quad \alpha = \arctan\frac{y}{x} + k\cdot 180^\circ. \end{aligned}$$

In the last equation, the integer k is chosen such, that the result α lies in the *correct quadrant*: $\arctan\frac{y}{x}$ is always in the interval $\left(-\frac{\pi}{2}, +\frac{\pi}{2}\right]$, i.e., in quadrants I or IV.

The Pythagoras theorem yields the distance s:

$$s = \sqrt{x^2 + y^2}.$$

3.4.1 Local co-ordinates

Local, often old, co-ordinate frames were long used in many Finnish municipalities, and one comes across them in older documents.

[10] In astronomy, sometimes, especially in older texts, azimuth turns from the South to the West, i.e., also clockwise. Also in geodesy, practices vary: one must always check.
The name "azimuth" originates from Arabic *as-sumût*, "directions".

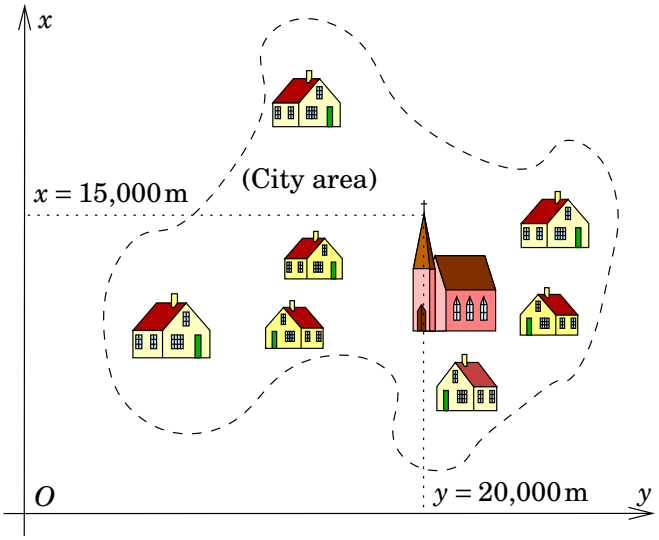


Figure 3.8. A local co-ordinate frame.

The origin is generally located so, that everywhere in the municipality there are only positive x and y values. Often the origin is just a computational point without realization in the terrain. E.g., "truncated" KKJ co-ordinates.

The connection to the national frame may be, that one knows the co-ordinates of, e.g., a church — or some other landmark — also in the KKJ system. In this case one can transform local and national co-ordinates into each other by applying a constant shift, a *translation*, to each co-ordinate x and y.

In more precise work, one point is not enough: a sufficient number of common points is needed, and their co-ordinates must be known in both the local and the national system.

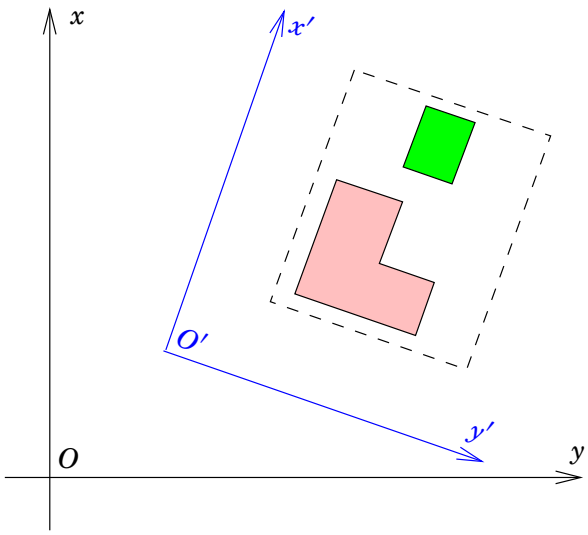


Figure 3.9. Temporary co-ordinates.

3.4.2 Temporary co-ordinates

Sometimes it makes sense to use in measurements a temporary co-ordinate frame that deviates from the general system. Even the directions of the axes may deviate from the customary North and East directions.

A temporary or project specific co-ordinate frame is only used during measurement, or, e.g., during a construction project. Later on, the co-ordinates may be transformed to a more permanent, local or national, correctly oriented system.

The origin and axes orientation can be chosen in accordance with the measurement at hand, e.g., along the walls of a building.

3.5 The geodetic forward and inverse problems

The *direct* or *forward geodetic problem* means the determination of the co-ordinates of an unknown point, when the co-ordinates of a starting point and both the azimuth (horizontal direction) and the distance from the starting point are given.

geodeettinen päätehtävä

In the general case, on an arbitrary curved surface, the forward geodetic problem has no easy solution. Already on a sphere, however, a closed — though not particularly simple — solution exists. On the surface of a reference ellipsoid the solution is obtained numerically. An on-line service is offered, e.g., by the US National Geodetic Survey on their web site.

In a plane co-ordinate system, two-dimensionally, the forward geodetic problem is simpler, as we shall see in the next subsection.

3.5.1 The forward geodetic problem in the plane

Let there be given *in the plane* two points P_1 and P_2 (figure 3.10). The plane co-ordinates (x_1, y_1) of starting point P_1, as well as the azimuth α_{12} and length s_{12} of vector $\overrightarrow{P_1P_2}$, are given. To be calculated are the co-ordinates (x_2, y_2) of the unknown point P_2.

The solution is obtained as follows:

$$\sin\alpha_{12} = \frac{\Delta y}{s_{12}} \quad \Longrightarrow \quad \Delta y = s_{12}\sin\alpha_{12},$$
$$\cos\alpha_{12} = \frac{\Delta x}{s_{12}} \quad \Longrightarrow \quad \Delta x = s_{12}\cos\alpha_{12},$$

using this:

$$x_2 = x_1 + \Delta x = x_1 + s_{12}\cos\alpha_{12},$$
$$y_2 = y_1 + \Delta y = y_1 + s_{12}\sin\alpha_{12}.$$

Example:

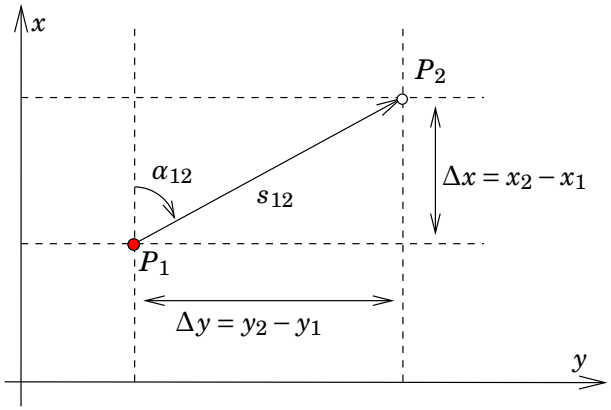


Figure 3.10. The forward geodetic problem in the plane.

Given a point A with co-ordinates

$$x_A = 6{,}800{,}000\,\mathrm{m},$$
$$y_A = 400{,}000\,\mathrm{m}.$$

If the distance from point B is $s = 2828.427\,\mathrm{m}$, and the azimuth (direction angle) $\alpha = 50\,\mathrm{gon}$, solve the forward geodetic problem for the points A, B.

Solution:

$$\Delta x = s\cos\alpha = 2828.427\,\mathrm{m}\cdot\cos(50\,\mathrm{gon}) = 2000\,\mathrm{m},$$
$$\Delta y = s\sin\alpha = 2828.427\,\mathrm{m}\cdot\sin(50\,\mathrm{gon}) = 2000\,\mathrm{m},$$
$$x_B = x_A + \Delta x = 6{,}800{,}000\,\mathrm{m} + 2000\,\mathrm{m} = 6{,}802{,}000\,\mathrm{m},$$
$$y_B = y_A + \Delta y = 400{,}000\,\mathrm{m} + 2000\,\mathrm{m} = 402{,}000\,\mathrm{m}.$$

3.5.2 The inverse geodetic problem in the plane

geodeettinen käänteistehtävä

The *inverse geodetic problem* means the determination of the azimuth (horizontal direction) and distance between two given points.

Let there be given again two points in the plane, P_1 and P_2 (figure 3.10). Let their rectangular co-ordinates be (x_1, y_1) and (x_2, y_2). To be calculated are α_{12} and s_{12}.

Solution:

$$s_{12} = \sqrt{\Delta x^2 + \Delta y^2} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2},$$
$$\tan\alpha_{12} = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}.$$

(We rather don't use $\sin\alpha_{12} = \frac{\Delta y}{s_{12}}$ or $\cos\alpha_{12} = \frac{\Delta x}{s_{12}}$. E.g., the sine formula becomes imprecise when $\alpha_{12} \approx \pm 90°$, and the cosine formula when $\alpha_{12} \approx 90° \pm 90°$. There, the functions $\sin\alpha_{12}$ and $\cos\alpha_{12}$ are *stationary*: a large change in α_{12} causes only a small change

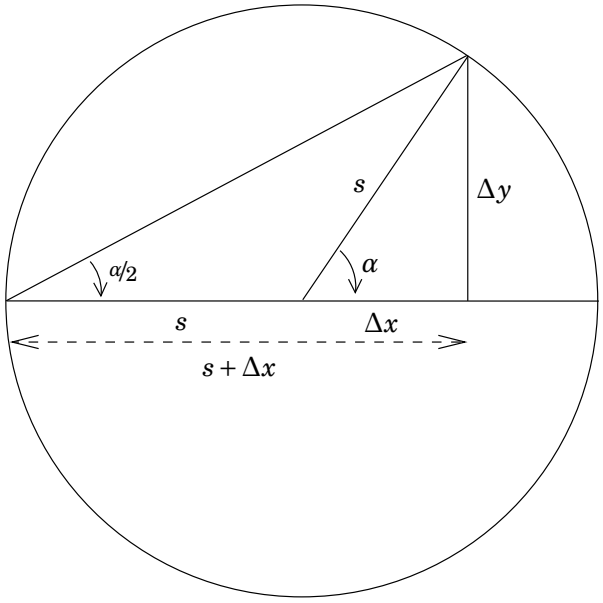


Figure 3.11. The half-angle formula.

in the function value $\sin\alpha_{12} = \frac{\Delta y}{s_{12}}$, and therefore, from the given values Δy and s_{12}, precise as they are, one can calculate α_{12} only imprecisely.)

Then:

$$\alpha_{12} = \arctan\left(\frac{y_2 - y_1}{x_2 - x_1}\right) + k \cdot 180^\circ, k = \begin{cases} 0 & \text{if } (x_2 - x_1) > 0, \\ 1 & \text{if } (x_2 - x_1) < 0, \end{cases}$$

$$\alpha_{12} = \begin{cases} 90^\circ & \text{if } x_1 = x_2 \text{ and } y_2 > y_1, \\ 270^\circ & \text{if } x_1 = x_2 \text{ and } y_2 < y_1. \end{cases}$$

It must be remembered here that the values of the arctan function are, according to the definition, always in the interval $(-\pi/2, +\pi/2)$. The correct α value may however well be outside this interval, like when $(x_2 - x_1) < 0$. This is why there is the conditional term $k \cdot 180^\circ$.

A more elegant solution[11] is to use the *half-angle formula*[12]:

$$\alpha = 2 \cdot \left(\frac{\alpha}{2}\right) = 2\arctan\frac{\Delta y}{\Delta x + s} = 2\arctan\frac{\Delta y}{\Delta x + \sqrt{\Delta x^2 + \Delta y^2}}.$$

See figure 3.11.

Example:

Given point A: $x_A = 6{,}800{,}000\,\text{m}$, $y_A = 400{,}000\,\text{m}$, and point C, co-ordinates $x_C = 6{,}793{,}000\,\text{m}$, $y_C = 407{,}000\,\text{m}$. solve the inverse geodetic problem for the points A, C.

[11] Many programming languages offer the function `atan2(x,y)` with two arguments, which finds also the correct quadrant automatically.

[12] This formula too breaks down in the borderline case $\Delta y = 0$ and $\Delta x < 0$, because then $\alpha = 2\arctan\frac{0}{0}$ is undefined, when the correct solution is $\alpha = 180^\circ$.

Solution:

$$\Delta x = x_C - x_A = 6{,}793{,}000\,\mathrm{m} - 6{,}800{,}000\,\mathrm{m} = -7000\,\mathrm{m},$$
$$\Delta y = y_C - y_A = 407{,}000\,\mathrm{m} - 400{,}000\,\mathrm{m} = +7000\,\mathrm{m}.$$

1. The traditional method:

$$\begin{aligned}\alpha_{AC} &= \arctan\frac{\Delta y}{\Delta x} + k\cdot 200\,\mathrm{gon} = \\ &= \arctan(-1) + k\cdot 200\,\mathrm{gon} = \\ &= -50\,\mathrm{gon} + k\cdot 200\,\mathrm{gon}.\end{aligned}$$

   The correct solution is apparently

$$\alpha_{AC} = -50\,\mathrm{gon} + 200\,\mathrm{gon} = 150\,\mathrm{gon}.$$

2. The half-angle formula:

$$\begin{aligned}\alpha_{AC} &= 2\arctan\frac{\Delta y}{\Delta x + \sqrt{\Delta x^2 + \Delta y^2}} = \\ &= 2\arctan\frac{7000}{-7000 + 7000\sqrt{2}} = \\ &= 2\arctan\left(\frac{-1}{1-\sqrt{2}}\right) = \\ &= 2\cdot 75\,\mathrm{gon} = 150\,\mathrm{gon}.\end{aligned}$$

$$\begin{aligned}s_{AC} &= \sqrt{\Delta x^2 + \Delta y^2} = 7000\,\mathrm{m}\cdot\sqrt{2} = \\ &= 9899.495\,\mathrm{m}.\end{aligned}$$

3.6 The similarity co-ordinate transformation

See Kahmen and Faig (1988, pages 246 – 255).

yhdenmuotoisuus-muunnos

The similarity[13] or Helmert transformation is a transformation between two rectangular co-ordinate frames, usually in the plane, i.e., two-dimensionally. It is encountered very often in practical measurement and computation tasks, when bodies of co-ordinate material in two or more co-ordinate frames need to be combined for joint use. This is a challenging task. In the general case we find a sufficient number of *common points* — often benchmarks — from the area, the co-ordinates of which are known in both systems, and carry out an *adjustment*. Simpler is the special case in which we have only two common points, which is minimally sufficient for determining the transformation.

[13] In Swedish: *likformighetstransformation*, in German: *Ähnlichkeitstransformation*, in French: *similitude*.

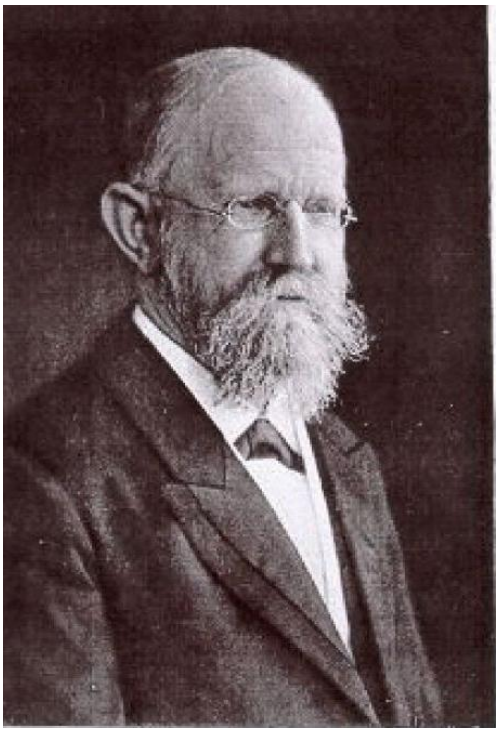


Figure 3.12. Friedrich Robert Helmert (1841 – 1917) was a great German geodesist and a developer of adjustment and probability calculation.

We know the co-ordinates of two benchmarks A and B in *both* coordinate systems:

$$(x_A, y_A), (x_B, y_B), (u_A, v_A), (u_B, v_B).$$

Additionally we are given a set of points with co-ordinates only in the (u, v) system:

$$(u_1, v_1), (u_2, v_2), \ldots, (u_i, v_i), \ldots, (u_n, v_n).$$

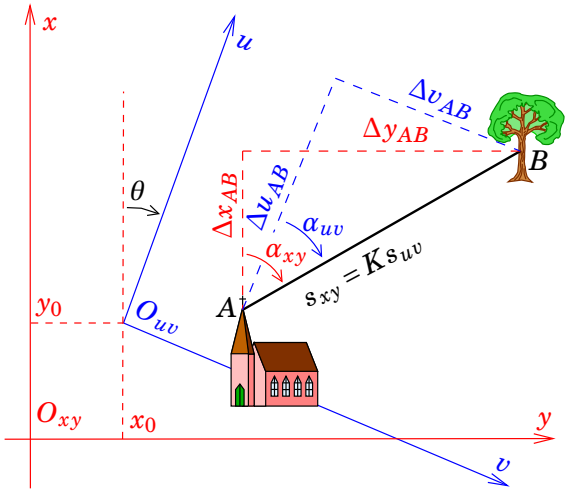


Figure 3.13. A similarity or Helmert co-ordinate transformation in the plane.

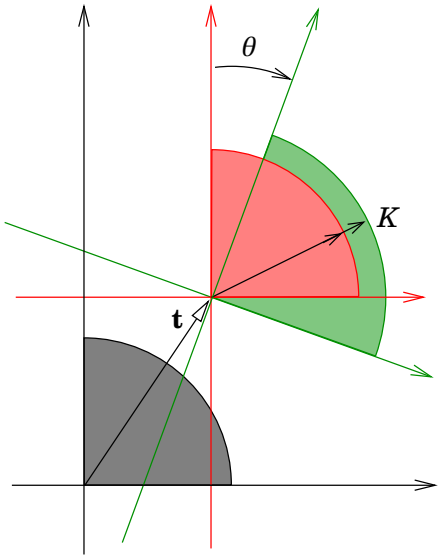


Figure 3.14. The stages of the Helmert transformation in the plane: translation vector $\mathbf{t} \stackrel{\text{def}}{=} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$, rotation angle θ, scaling K.

The problem is now to compute a similarity transformation for this point field

$$(u_i, v_i) \Rightarrow (x_i, y_i), i = 1, \ldots, n.$$

The transformation is carried out in the following steps:

- A *shift*, or translation, of the origin $O_{uv} \Longrightarrow O_{xy}$, translation parameters (x_0, y_0), see figure 3.13.
- A *rotation* of the whole (u, v) co-ordinate system by an angle $-\theta$ — a rotation angle θ, see figure 3.13.
- A s*cale transformation* of the (u, v) co-ordinate system, to become the same as that of the (x, y) co-ordinate system, by applying the scale factor or *scale ratio* K.

The Helmert transformation in the plane is also called *four-parameter transformation*. The parameters are x_0, y_0, θ and K.

The general form of the Helmert transformation is:

$$x = x_0 + K\cos\theta \cdot u - K\sin\theta \cdot v,$$
$$y = y_0 + K\sin\theta \cdot u + K\cos\theta \cdot v,$$

or, in matrix form — see appendix A:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + K \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}. \tag{3.1}$$

3.6.1 Determining the transformation parameters

Determining the transformation parameters unambiguously requires at least *four* "*observations*" — e.g., the altogether four co-ordinates of two

points (x_A, y_A), (x_B, y_B). Then we obtain four equations:

$$\begin{aligned} x_A &= x_0 + K\cos\theta \cdot u_A - K\sin\theta \cdot v_A, \\ y_A &= y_0 + K\sin\theta \cdot u_A + K\cos\theta \cdot v_A, \\ x_B &= x_0 + K\cos\theta \cdot u_B - K\sin\theta \cdot v_B, \\ y_B &= y_0 + K\sin\theta \cdot u_B + K\cos\theta \cdot v_B, \end{aligned}$$

or, again in matrix form,

$$\begin{bmatrix} x_A \\ y_A \\ x_B \\ y_B \end{bmatrix} = \begin{bmatrix} 1 & 0 & -v_A & u_A \\ 0 & 1 & u_A & v_A \\ 1 & 0 & -v_B & u_B \\ 0 & 1 & u_B & v_B \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ K\sin\theta \\ K\cos\theta \end{bmatrix}.$$

If, of the same points A and B, also the old co-ordinates (u_A, v_A),(u_B, v_B) are known, one may from these equations solve four transformation parameters x_0, y_0 and $K\sin\theta, K\cos\theta \Longrightarrow K, \theta$ *uniquely*.

3.6.1.1 *Difference transformation formula*

Subtraction yields

$$\begin{bmatrix} \Delta x_{AB} \\ \Delta y_{AB} \end{bmatrix} = \begin{bmatrix} x_B - x_A \\ y_B - y_A \end{bmatrix} = \begin{bmatrix} -(v_B - v_A) & u_B - u_A \\ u_B - u_A & v_B - v_A \end{bmatrix} \begin{bmatrix} K\sin\theta \\ K\cos\theta \end{bmatrix} = \\ = \begin{bmatrix} -\Delta v_{AB} & \Delta u_{AB} \\ \Delta u_{AB} & \Delta v_{AB} \end{bmatrix} \begin{bmatrix} K\sin\theta \\ K\cos\theta \end{bmatrix}$$

with the logical definitions

$$\begin{aligned} \Delta x_{AB} &= x_B - x_A, \quad \Delta u_{AB} = u_B - u_A, \\ \Delta y_{AB} &= y_B - y_A, \quad \Delta v_{AB} = v_B - v_A. \end{aligned}$$

Arrange the terms cleverly anew:

$$\begin{bmatrix} \Delta x_{AB} \\ \Delta y_{AB} \end{bmatrix} = \begin{bmatrix} -\Delta v_{AB} & \Delta u_{AB} \\ \Delta u_{AB} & \Delta v_{AB} \end{bmatrix} \begin{bmatrix} K\sin\theta \\ K\cos\theta \end{bmatrix} = \\ = K \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \Delta u_{AB} \\ \Delta v_{AB} \end{bmatrix}. \tag{3.2}$$

This equation applies for arbitrary point pairs.

3.6.1.2 *Scale ratio*

the scale ratio or scale factor is obtained with the Pythagoras theorem:

$$K = \frac{s_{xy}}{s_{uv}} = \frac{\sqrt{\Delta x_{AB}^2 + \Delta y_{AB}^2}}{\sqrt{\Delta u_{AB}^2 + \Delta v_{AB}^2}}.$$

3.6.1.3 Rotation angle

The rotation angle[14] is:

$$\theta = \alpha_{xy} - \alpha_{uv} = \arctan\frac{\Delta y_{AB}}{\Delta x_{AB}} - \arctan\frac{\Delta v_{AB}}{\Delta u_{AB}} \tag{3.3}$$

3.6.1.4 Translation vector

For computing the translation vector, we start from the Helmert difference transformation formula, equation 3.2.

Let us look at the point pair A and O_{uv}, the origin of the old (u,v) co-ordinate system. The points have co-ordinates (x_A, y_A) and (x_0, y_0) in the new system, and co-ordinates (u_A, v_A) and $(u_0, v_0) = (0,0)$ in the original system. Then

$$\begin{aligned}\Delta u_{AO} &= u_A - u_0 = u_A,\\ \Delta v_{AO} &= v_A - v_0 = v_A,\end{aligned}$$

and at the same time

$$\begin{aligned}\Delta x_{AO} &\stackrel{\text{def}}{=} x_A - x_0,\\ \Delta y_{AO} &\stackrel{\text{def}}{=} y_A - y_0.\end{aligned}$$

The difference transformation formula gives now

$$\begin{bmatrix} x_A - x_0 \\ y_A - y_0 \end{bmatrix} = K \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} u_A \\ v_A \end{bmatrix},$$

i.e., the translation vector

$$\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} x_A \\ y_A \end{bmatrix} - K \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} u_A \\ v_A \end{bmatrix},$$

just the translation parameters we are after.

For an arbitrary point (x,y) the Helmert transformation formulas are now, after all transformation parameters have been solved for:

$$\begin{aligned}x &= x_0 + K\cos\theta \cdot u - K\sin\theta \cdot v,\\ y &= y_0 + K\sin\theta \cdot u + K\cos\theta \cdot v,\end{aligned}$$

the same equations already given above, 3.1.

[14] Forget for a moment the quadrant problem. In principle, a term $k \cdot 180^\circ$ should be added, with k a small integer.

3.6.1.5 Symbolic matric form

The Helmert transformation formulas in matric form are

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + K \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}.$$

This can be written compactly:

$$\overline{\mathbf{x}} = \overline{\mathbf{x}}_0 + K\overline{R}\overline{\mathbf{u}}, \tag{3.4}$$

in which the definitions of the vectors (column matrices) and matrices are

$$\overline{\mathbf{x}} \overset{\text{def}}{=} \begin{bmatrix} x \\ y \end{bmatrix}, \quad \overline{\mathbf{x}}_0 \overset{\text{def}}{=} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix},$$
$$\overline{\mathbf{u}} \overset{\text{def}}{=} \begin{bmatrix} u \\ v \end{bmatrix}, \quad \overline{R} \overset{\text{def}}{=} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

Often one writes $K = 1+m$, in which m is the *scale distortion*. Generally the number is small and expressed in the unit ppm (parts per million).

The transformation equations 3.1, 3.4 are called the similarity or Helmert transformation in the plane.

Example:

1. Given the co-ordinates of points A, B in the (u,v) co-ordinate system:

$$u_A = 0\,\text{m}, v_A = 0\,\text{m},$$
$$u_B = 1500\,\text{m}, v_B = 1500\,\text{m};$$

and in the (x,y) co-ordinate system:

$$x_A = 2000\,\text{m}, y_A = 3000\,\text{m},$$
$$x_B = 3500.150\,\text{m}, y_B = 4500.150\,\text{m}.$$

Assuming that the transformation between the (u,v) and (x,y) systems is a Helmert transformation:

$$\begin{bmatrix} x \\ y \end{bmatrix} = K \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \cdot \begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} x_0 \\ y_0 \end{bmatrix},$$

calculate its parameters K, θ, x_0 and y_0.

2. Given the co-ordinates of point C in the (u,v) system:

$$u_C = 1000\,\text{m}, v_C = 2000\,\text{m}.$$

Calculate x_C, y_C .

Solution:

1. We see immediately that

$$\begin{aligned}\Delta u_{AB} &= 1500\,\mathrm{m},\\ \Delta v_{AB} &= 1500\,\mathrm{m},\\ \Delta x_{AB} &= 1500.150\,\mathrm{m},\\ \Delta y_{AB} &= 1500.150\,\mathrm{m}.\end{aligned}$$

   From this we infer visually, with the help of the *difference transformation formula* 3.2, that $K = 1.0001$ and $\theta = 0$.
   After this, for point A:

$$\begin{aligned}x_A = x_0 + 1.0001 \cdot u_A \quad &\Longrightarrow \quad x_0 = x_A - 1.0001 \cdot u_A = 2000\,\mathrm{m},\\ y_A = y_0 + 1.0001 \cdot v_A \quad &\Longrightarrow \quad y_0 = y_A - 1.0001 \cdot v_A = 3000\,\mathrm{m}.\end{aligned}$$

2. Calculate

$$\begin{aligned}x_C &= x_0 + 1.0001 \cdot u_C = 2000\,\mathrm{m} + 1000.1\,\mathrm{m} = 3000.1\,\mathrm{m},\\ y_C &= y_0 + 1.0001 \cdot v_C = 3000\,\mathrm{m} + 2000.2\,\mathrm{m} = 5000.2\,\mathrm{m}.\end{aligned}$$

3.7 Datums and datum transformations

Geodetic co-ordinates aren't just mathematical quantities. Points are measured in the terrain, and their co-ordinates are computed, based on given starting or *datum points*. The choice of starting points is always to some extent arbitrary. Every choice made creates what geodesists call a *geodetic datum*. In other words, when geodetic measurements are made on a part of the Earth's surface using a certain set of measurement points, and conventionally assigning starting co-ordinates to starting points chosen from these, we get, in real life, a solution that represents only the *realization* of a certain reference system[15].

A co-ordinate reference frame or *datum* is generally established locally. When it meets another, similarly established (but based on different starting points) frame, the co-ordinates of the same points are generally not the same. E.g., in the place where the Finnish and Swedish precise levelling networks meet at the border in the Torne river valley, we get for the same point two different height values *which are both correct*.

Also in the case of *horizontal networks* we speak of datums, horizontal datums: where they meet at borders, the horizontal co-ordinates (φ, λ)

[15] In English, we call the formal definition a *co-ordinate reference system*, wheras we call the realization in the terrain a *co-ordinate reference frame*. E.g., ETRS = *European Terrestrial Reference System* and ETRF = *European Terrestrial Reference Frame*. Also in Finland, the corresponding terms are gaining traction: *vertausjärjestelmä* against its realization or *vertauskehys*.

are generally not precisely the same. The differences are, in the case of classical triangulation networks, of the order of a few seconds of arc.

For the purpose of transforming the co-ordinates of points in one datum to co-ordinates of another datum, the literature offers *datum transformation* formulas.

3.7.1 Example: height network

Let us look at the example of height measurement, *levelling*. The Finnish official height system until September 25, 2007, N60, is based on the height of a certain point. This starting or datum point is a granite pillar located in the garden of Helsinki astronomical observatory, and on that pillar a certain polished surface. Here a role is played by historical accident, by the fact of Helsinki being the Finnish capital. The height value of the datum point is chosen such, that the heights are rather precisely above the mean sea level in Helsinki harbour at the start of 1960. Scientifically the choice of Helsinki was arbitrary.

The new Finnish height system, N2000, uses as its starting or datum point the fundamental benchmark PP2000 at the Metsähovi research station – 40 km to the East of Helsinki — the height value of which is chosen so, that the heights are relative to the Amsterdam mean sea level N.A.P., one of the oldest in the world. Also the choice of Amsterdam is the product of history, not science.

The national precise levelling has brought official heights to everywhere in the country. It is clear that the precision of a calculated point height in this system will depend on the point's distance from Helsinki. A height in Kevo, Northern Lapland, will be clearly poorer known than a height in Jyväskylä. And the height of Turku is somewhat imprecise, because the levelling from Helsinki to Turku was not absolutely precise. On the other hand, the heights measured for points in the Helsinki area are very precise, because the *datum point*, be it Observatory Hill or Metsähovi, is so nearby.

Imagine for a moment that not Helsinki but Turku were the capital of Finland, and that as the datum point for the Finnish height system were chosen a benchmark in the wall of Turku Cathedral. In that case, all heights of points close to Turku would be very precise, but the points in the Helsinki area would be similarly imprecise as in the present system are the Turku area points.

Precision depends on the viewpoint, on the chosen *datum*.

In figure 3.16 is shown a levelling network of four points. Given the height differences AB, BC, CD and DA. Furthermore are given the heights above mean sea level in coastal points A and B, measured by a mareograph or tide gauge.

First we adjust the levelling loop:

Figure 3.15. Fundamental benchmark PP2000 of the N2000 height datum at Metsähovi research station. In the background the 14 metre astronomical radio telescope of Aalto University — the "Pumpkin" —, with which also geodetic very long baseline interferometric (VLBI) observations are being made. Photo © National Land Survey of Finland, Finnish Geospatial Research Institute.

□

Interval	Observed	Correction	Adjusted
AB	−0.925	−0.004	−0.929
BC	+0.548	−0.004	+0.544
CD	+0.321	−0.004	+0.317
DA	+0.072	−0.004	+0.068
Closing error	+0.016		-

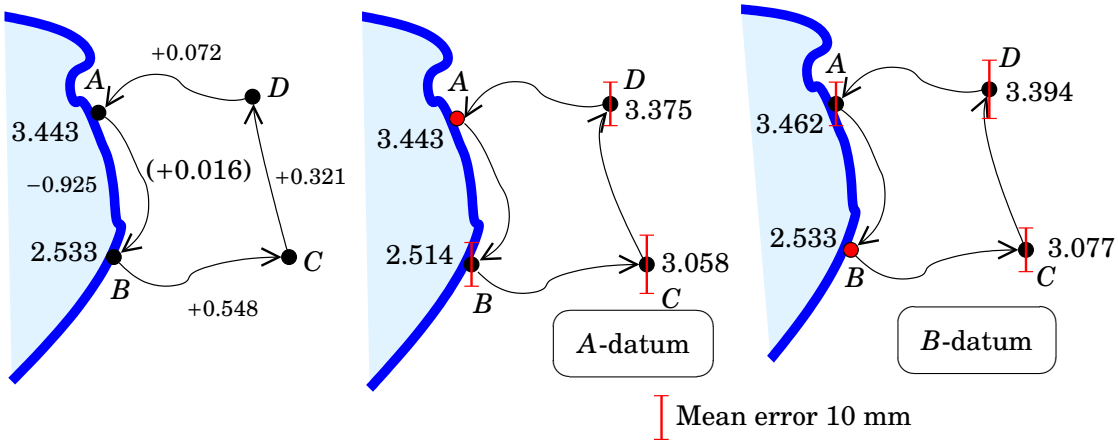


Figure 3.16. Alternative vertical datums A and B.

□

1. Use point A as *datum point* and calculate the heights of the points[16]:

Point	Height	Mean error
A	3.443	±0.000
B	2.514	±0.010
C	3.058	±0.014
D	3.375	±0.010

2. Do the same thing, but now using point B as the datum point:

Point	Height	Mean error
B	2.533	±0.000
C	3.077	±0.010
D	3.394	±0.014
A	3.462	±0.010

It is seen that in the latter case all calculated heights are greater by 0.019 m. The height *differences* are of course the same. The difference 0.019 m is precisely the "difference of height differences" of points A and B between the two methods: (1) levelling plus adjustment, and (2) mareographs. The difference stems from measurement errors and the circumstance that the true sea surface is not a level surface.

The *datum difference* between datum A and datum B is 0.019 m.

The *datum transformation* is

$$H_i^{(B)} = H_i^{(A)} + 0.019\,\mathrm{m},$$

more generally

$$H_i^{(B)} = H_j^{(A)} + \left(H_A^{(B)} - H_A^{(A)}\right).$$

Every adjustment produces precision estimates of the computed values, or *mean errors*. If we assume that, as a given value, the height value of the datum point is *errorless* (mean error zero), the mean errors of the point heights will grow moving away from the datum point. The above table contains (invented) mean errors behaving in just this way. They are also drawn in the figure as error bars. We see that the precision behaviour of the network depends on the datum point choice.

3.7.2 Horizontal co-ordinate datums

tasoverkko

As already said, also horizontal networks can have different datums, so-called *horizontal datums*. In figure 3.17 is depicted a datum AB and a datum PQ.

- The AB datum is created by taking already known approximate co-ordinate values for points A and B as the *formal truth*, and adjusting the whole network without changing these. In this way one

[16] The mean errors in the table are made up though realistic-looking.

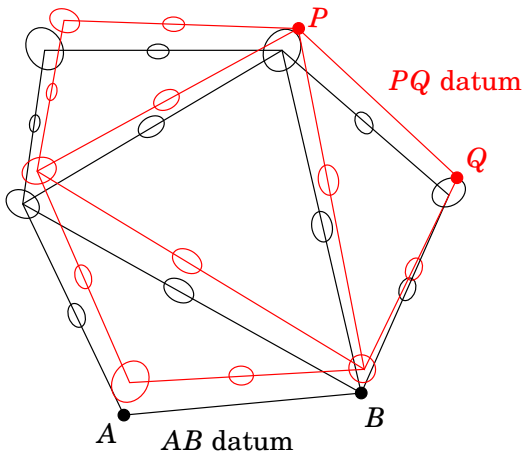


Figure 3.17. Two different datums of a horizontal network, the AB and the PQ horizontal datums with their starting or datum points. The point error ellipses and the inter-point (relative) error ellipses have been drawn. Note how the error vanishes in the datum points, the co-ordinates of which are part of the datum definition and thus conventional.

□

calculates the co-ordinates of the other points, including P and Q, in the same AB datum.

- The PQ datum again is created in the same way, but by fixing the co-ordinates of points P and Q to prior given approximate values, and adjusting the network while keeping these fixed. Thus one also solves for the locations of the other points in the letwork, but now in the PQ datum.

As can be seen from the figure are the co-ordinates of the points different in the AB and PQ datums. The *shape* of the whole network is nevertheless the same, independently of whether the datum AB or the datum PQ is chosen for the solution. In this case, the transformation between the datums is a similarity or *Helmert* transformation.

The prior known co-ordinates for the points A,B,P,Q originate generally from previous network adjustments, astronomical position determinations, or are read from a map: they are *approximate values* of the co-ordinates. Determining a datum is thus the same as choosing the datum points, points the approximate co-ordinates of which are taken as the *formal truth* in the network adjustment.

It would be a coincidence if the computation of the network in the AB datum would produce the same co-ordinates as computation in the PQ datum. The differences between the co-ordinates computed in different ways are comparable in magnitude to the *goodness of approximation* of the approximate co-ordinates used. The differences are often so small, that the transformation parameters are close to zero or unity: in the

Helmert formula

$$x = x_0 + K\cos\theta \cdot u - K\sin\theta \cdot v$$
$$y = y_0 + K\sin\theta \cdot u + K\cos\theta \cdot v$$

the rotation angle θ is so small, that $\sin\theta \approx \theta$ and $\cos\theta \approx 1$; if we also write $K = 1 + m$, with m the scale distortion, then also m is a small number, and we obtain

$$x = x_0 + (1+m)u - (1+m)\theta v \approx x_0 + u + m \cdot u - \theta v,$$
$$y = y_0 + (1+m)\theta u + (1+m)v \approx y_0 + \theta u + v + m \cdot v,$$

or as a matric equation

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix} + \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \begin{bmatrix} m & -\theta \\ \theta & m \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix},$$

an elegant equation in which the second and third term on the right-hand side are *small*, because they contain only the small transformation parameters x_0, y_0, m and θ. Thus also the co-ordinate differences $x - u$ and $y - v$ are small, as we observed already above.

3.8 Map projections and height systems in a three-dimensional world

Although the real Earth and her gravity field are three-dimensional phenomena, which we can describe and handle correctly only in three dimensions, nevertheless means of description are very widely used that are based on "two-plus-one-dimensional" thinking. Like map projections and height systems, which together describe the world by means of 2+1 co-ordinates (x, y, H).

Although we have here three co-ordinates, one cannot speak of three-dimensional coordinates because on the one hand x, y and on the other, H, are not comparable.

Among ordinary people — and even among land surveyors — lives a conceptual model of the "shoebox world": rectangular, the sides oriented in the Northern and Eastern direction, and the height co-ordinate being simply the distance from the bottom of the shoebox, "sea level".

It is easy to be judgmental about this way of thinking. Remember, however, that *in a small area the shoebox model* is — may well be — *an acceptable approximation*. E.g., inside cities, plan maps and rectangular co-ordinates are used, without bad repercussions. The question of acceptability of this approximation demands careful analysis.

If the (x, y, H) representation is acceptable, then it is simpler, as a representation of location and height, than the Earth's true geometry, true

Figure 3.18. The connections between three-dimensional rectangular co-ordinates, geodetic co-ordinates on the reference ellipsoid, and projection co-ordinates in the map plane.

$$(X,Y,Z) \overset{\text{(ref. ellipsoid)}}{\Longleftrightarrow} (\varphi,\lambda,h) \begin{cases} (\varphi,\lambda) \overset{\text{(map projection)}}{\Longleftrightarrow} (x,y) \\ h \overset{\text{(geoid model)}}{\Longleftrightarrow} H \end{cases} \tag{3.5}$$

locations with their rectangular geocentric co-ordinates and true heights with their geopotential numbers.However, *in the general case* it is actually a *more complicated* way of representing location and height. The complexity of map projections and the gravity field pervades all co-ordinates thus defined. Errors caused by misconceptions are very easy to make.

Therefore: *Always in precise scientific work* in geodesy, exclusively *geocentric, three-dimensional co-ordinates* and *geopotential numbers* should be used. "Plane co-ordinates" — i.e., map projection co-ordinates — and metric heights should always be seen as *derived quantities*, on the basis of which no precise computations should be attempted.

See diagram 3.18 in which the signs "$\Longleftrightarrow$" designate operations used:

Reference ellipsoid: a co-ordinate conversion between rectangular and geodetic co-ordinates is a mathematical, exact operation. The choice of reference ellipsoid is however arbitrary. Today GRS80 is the standard, in Finland however also the Hayford or International Ellipsoid of 1924 has historically been used.

Map projection: a mathematical, exact operation. Many alternatives are on offer.

Geoid model: The height type may be orthometric, normal (or variants of those two) or dynamic. Always a *geoid model* (or similar) is needed, see section 4.1.

In diagram 3.5 on the left are the more abstract quantities, whereas on the right are the more concrete quantities, closer to daily life.

3.9 The time co-ordinate

In geodesy we use, in addition to co-ordinates of place, also other co-ordinates. The first of these is *time*. Time describes the changes happening in the Earth, the research into which belongs to the field of *geodynamics*, on which more in section 17.1. With the measurement precision of modern geodesy, the Earth *lives* and *changes* continuously:

- As a result of the solid-Earth tide, the Earth below our feet moves periodically, even in Finland a couple of decimetres, two times a day, up and down. We don't notice this of course, because an even more stable reference point is lacking.
- According to plate tectonics, all continental plates move evenly. The velocity of motion is typically a few centimetres per year, and can be precisely monitored with GNSS technology.
- The rotation of the Earth is irregular. With space geodetic observation techniques it is possible to follow the variations in the direction of the Earth's rotation axis, both relative to the solid Earth — polar motion — and relative to the celestial sphere — precession and nutation —, and variations in rotation rate — LoD, length of day. These phenomena together are called "Earth orientation parameters". vuorokauden pituus
- In Fennoscandia, Canada and elsewhere, the Earth's surface is rising slowly after the last ice age, the so-called glacial isostatic adjustment (GIA).
- There are other, more local motions too, partly caused by human activity.

Self-test questions

1. How are map projections classified based on what they distort and what they preserve?
2. What is the main distinction between "traditional" and "modern" co-ordinate reference frames?
3. What are the differences between the old KKJ datum and the new EUREF-FIN datum for the territory of Finland?
4. Name and describe geodetic plane co-ordinates and geodetic polar co-ordinates. How do you convert between them?
5. Describe the geodetic forward problem in the plane, and its solution.
6. Describe the geodetic inverse problem in the plane, and its solution.
7. Describe the Helmert transformation in the plane. How many free parameters does the transformation contain? Describe them.
8. What is a datum, and how is one established?
9. Describe shortly the various phenomena that geodynamics studies.

Exercise 3 – 1: Distances

We find the co-ordinates of two points in Finland: the GNSS fundamental station Metsähovi (METS), and the GNSS station at the Sodankylä

Geophysical Observatory (SODA). According to the EUREF data centre (http://epncb.oma.be/_networkdata/stationlist.php, accessed March 29, 2017) there are the following approximate ITRF co-ordinates:

Station		X (m)	Y (m)	Z (m)
METS	1	2,892,571.00	1,311,843.28	5,512,634.01
SODA	2	2,200,147.00	1,091,638.20	5,866,870.60

From this, we

1. compute the latitude and longitude φ and λ, again using the online service of the Finnish National Land Survey of Finland. The software asks for 3D Cartesian (rectangular) co-ordinates in the ETRS89 system, which is not quite the same as approximate ITRF; we just pretend it is.
2. From these, we compute the distance, on the surface of the GRS80 reference ellipsoid, between the points, i.e., we solve the *inverse geodetic problem*. Use the NGS web site to obtain the distance s_1.
3. Instead of a reference ellipsoid, we may use a spherical approximation, with the mean Earth radius being $R = 6371.008\,\text{km}$. The angular distance between METS and SODA can be computed from their latitude and longitude, using the equation (cosine rule on the

sphere[17])

$$\cos\psi = \sin\varphi_1 \sin\varphi_2 + \cos\varphi_1 \cos\varphi_2 \cos(\lambda_2 - \lambda_1).$$

And then, the metric distance along the surface of the sphere is

$$s_2 = \psi R.$$

4. Alternatively, the *chord distance* in spherical approximation:

$$s_3 = 2R \sin\frac{\psi}{2}.$$

5. We return to the NLS web site and convert the co-ordinates (φ, λ) to the two-dimensional *map-projection system* ETRS-TM35FIN.
6. Now, using these two-dimensional map projection co-ordinates (x, y), for METS and SODA, we compute the distance between the two stations using the Pythagoras theorem in two dimensions:

$$s_4 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

7. Now, as the icing on the cake, compute the *chord* (three-dimensional) *distance*, also with Pythagoras, from the original rectangular co-ordinates:

$$s_5 = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2 + (Z_1 - Z_2)^2}.$$

[17] Smart readers will note that this equation produces, for short distances ψ, a cosine close to unity, leading to loss of accurate digits when ψ is recovered by the arccos function.
It is possible to convert it to a "half-angle version" in the following way: substitute

$$\begin{aligned} \cos\psi &= 1 - 2\sin^2\frac{\psi}{2}, \\ \cos(\lambda_2 - \lambda_1) &= 1 - 2\sin^2\frac{\lambda_2 - \lambda_1}{2}, \end{aligned}$$

yielding

$$\begin{aligned} 1 - 2\sin^2\frac{\psi}{2} &= \sin\varphi_1 \sin\varphi_2 + \cos\varphi_1 \cos\varphi_2 \left(1 - 2\sin^2\frac{\lambda_2 - \lambda_1}{2}\right) = \\ &= \left(\sin\varphi_1 \sin\varphi_2 + \cos\varphi_1 \cos\varphi_2\right) - 2\cos\varphi_1 \cos\varphi_2 \sin^2\frac{\lambda_2 - \lambda_1}{2} = \\ &= \cos\left(\varphi_2 - \varphi_1\right) - 2\cos\varphi_1 \cos\varphi_2 \sin^2\frac{\lambda_2 - \lambda_1}{2} = \\ &= 1 - 2\sin^2\frac{\varphi_2 - \varphi_1}{2} - 2\cos\varphi_1 \cos\varphi_2 \sin^2\frac{\lambda_2 - \lambda_1}{2}, \end{aligned}$$

from which

$$\sin^2\frac{\psi}{2} = \sin^2\frac{\varphi_2 - \varphi_1}{2} + \cos\varphi_1 \cos\varphi_2 \sin^2\frac{\lambda_2 - \lambda_1}{2},$$

the "half-angle" version of the spherical cosine rule which is well behaved for points that are close together (compared to the size of the Earth).

All these distances are different. Some of the differences are small, some substantial. Complete the exercise by explaining where all these differences come from. Do all calculations in *millimetres* rounding accuracy.

4. Height measurement and the levelling instrument

4.1 Height, geopotential and the geoid

Heights express the locations of points in the vertical direction, the direction of the local *gravity vector*, the vertical or plumb line.

Intuitively this is based on the naive "shoebox model" of the Earth's figure, where height is the third co-ordinate, the straight, metric distance from the bottom of the shoebox, sea level.

The shoebox model is also called the "flat Earth approximation": somewhere below the land surface there is a reference surface, assumed to be a plane, that coincides with mean sea level. Height is the distance in metres from this level surface.

In reality the Earth is not flat and the reference surface is curved, even undulating. The reference surface is called the *geoid*. It is an *equipotential surface* of the Earth's gravity field, a surface on which every point has the same *geopotential*, the potential of the Earth's gravity field. The direction of gravity, the *plumb line*, is everywhere *perpendicular* onto this surface. The distance of a point from this surface, measured along the plumb line, is called its *orthometric height*. Thus, orthometric height has a simple geometric interpretation, and is of course a *metric* quantity.

4.1.1 The geopotential

Physical geodesy is the branch of geodesy concerned with the gravity field and gravity potential of the Earth. The geopotential W can be taken as the fifth "co-ordinate" after the three co-ordinates of place X, Y, Z and time. The geopotential describes the *energy level* of points in relation to sea level. This corresponds to the popular conception of "height". We have a habit of expressing height as a metric quantity, which is mostly appropriate in daily life... but what really interests us is the potential *energy* that comes with height. The effect of gravity on our daily activities is so strong, that "height determination" , i.e., the study, determination, and presentation of the geopotential, forms a large part of practical geodesy and surveying.

Water: ”height”, i.e., potential, represents *energy*. The energy may be recovered or stored — hydro-power. The energy may also be destructive — floods — for which one has to be prepared.

Air: The levels of equal pressure in the air follow fairly precisely the levels of the geopotential. The phenomenon is exploited in barometric height determination. Also an aircraft measures its “height” using an air-pressure sensor.

Traffic: Gravity affects the planning of traffic routes. Slopes may not be too steep, desirable are limited variations in potential, i.e., energy level, along the route. In the case of waterways, this happens automatically in the natural way.

The geopotential is closely related to gravity. Surfaces having the same value for the geopotential, level surfaces or *equipotential surfaces*, are what are ordinarily called “horizontal surfaces”. A freely flowing fluid — sea water, lake water, air — will settle along an equipotential surface. In the sea, the hydrostatic pressure is a constant along the equipotential surfaces of the gravity field, just like, in the atmosphere, barometric pressure is a constant along level surfaces — at least approximately: disturbances are due to salinity and temperature variations in water and temperature variations in air, and the currents these cause.

4.1.2 Metric heights

Heights are being “humanized” in a similar way as is the case with map projections. One invents a way in which heights “above sea level” can be expressed as a *metric* quantity, as a *height* H from some *reference surface*, usually mean sea level. The first step is always to compute the difference between the geopotential values of the point under consideration and of mean sea level, the *geopotential number* C. This number is positive from mean sea level upward.

Unfortunately, just like with map projections, there is no solution that would be in all respects satisfactory. Always something is “distorted”. In the same way as with map projections, there are also different *height types*, like

- orthometric heights H
- normal heights H^*
- dynamic heights H^{dyn}.

They all have their good and bad properties.

Earlier, in the section on geometric co-ordinates (section 2.8), we became acquainted with co-ordinates bound to the reference ellipsoid, of which one was the *height from the reference ellipsoid*, h. This co-ordinate describes the location of a point in the “vertical direction”, i.e., in a way also the point’s height. It is nevertheless reckoned from the *reference*

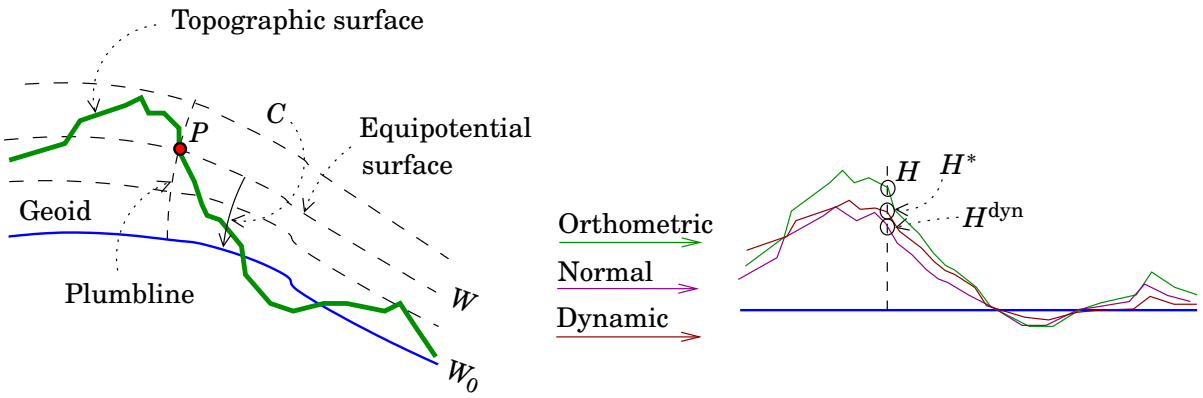


Figure 4.1. Different height types map geopotential numbers C in different ways to metric heights (differences exaggerated).

ellipsoid, a surface that is not physically useable as a reference surface in daily life. Also, it doesn't describe the energy level with respect to sea level, like H, H^* and H^{dyn} (and of course geopotential numbers C) do.

If all this appears at this stage difficult and theoretical, it may be worthwhile to come back to this section later in the course and read it again.

4.2 Orthometric height

Orthometric (Greek "correctly measured") heights H correspond the most precisely to our concept of "height above sea level". They are in principle just metric heights above the *geoid*. The geoid is that equipotential surface of the gravity field which on average is on the same level as mean sea level. In other words, mean sea level continued under the land masses.

If we could excavate under the continents a network of tunnels (figure 4.2) on the level of the sea surface, the water would spread thoughout the network in such a way, that its surface would be a physical realization of the geoid. A point's orthometric height would be its *distance from this fluid surface*. This direct physical interpretation is the reason why many geophysicists, and many countries — among which Finland until 2007 — have chosen to use an orthometric height system.

Building a tunnel network like the one depicted is of course not practical. Inland, the geoid is realized *computationally*, by carrying out the calculation of a height or *levelling network*, starting from a chosen coastal point or set of coastal points. Thus we disseminate orthometric heights throughout the country, to everywhere that the levelling network extends to.

In Finland up until 2007, the official height system or *vertical datum* was N60, the zero point of which was Helsinki mean sea level at the start of 1960. N60 heights are to good approximation orthometric. In

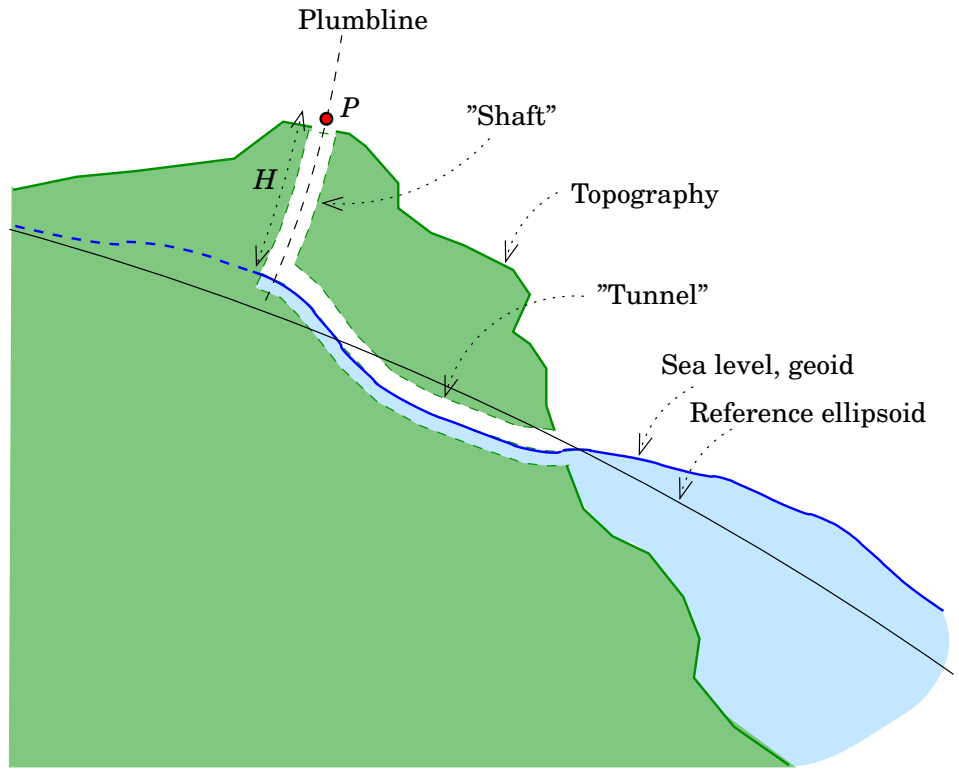


Figure 4.2. *Orthometric heights* are metric distances from the *geoid*, the water surface that would form if sea water could freely move under the topography in an imaginary tunnel network. In that case, orthometric heights could be directly measured along the *plumb line* through a shaft like the one depicted.

2007, the N2000 height system was taken into use, with as the zero level the official Amsterdam or N.A.P. (Normaal Amsterdams Peil) sea-level datum. N2000 heights are *normal heights*, the definition of which differs a little from that of orthometric heights. The difference is practically of little significance.

4.3 Height determination and levelling

4.3.1 Exotic methods for height determination

The most direct way of measuring height differences is to realize an equipotential surface of the gravity field by means of a *fluid surface*. This is how one can transfer geopotential values from one place to another.

- In Denmark and the Netherlands *hydrostatic levelling* has been employed, in which a long tube filled with distilled water is used to transfer the level of the geopotential (the "height") from one island to another or between island and shore. The distances over which measurements have been done have been tens of kilometres.

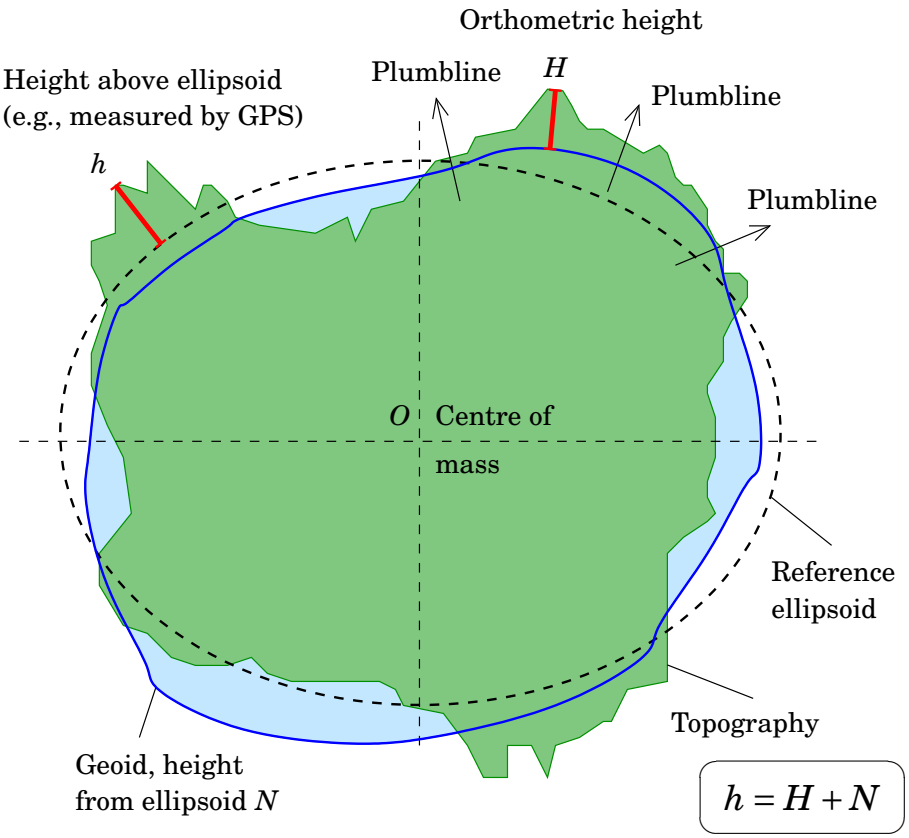


Figure 4.3. Important reference surfaces and height concepts.

- By using *water gauges* in interior waters, heights may also be transferred hydrostatically. Like in the water tube technique, also here the air pressure difference between the terminals must be taken into account, as well as the effects of wind and currents. The method, which works best under an ice cover, has been tested, e.g., in the Netherlands (IJsselmeer, Rijkswaterstaat 1996 – 97 (Reijnoudt, 1996)) and Finland.
- Also *barometers* have been traditionally used for measuring height differences. A careful procedure which takes into account the natural air pressure variations due to the weather will yield a best-case precision of about one metre. See Heiskanen and Härmälä (1963, pages 84 – 87).

The geophysical modelling of sea currents has been attempted, e.g., in the Åland Sea.

Also trigonometric traverse levelling must be mentioned here, e.g., figure 5.37 on page 152 (Takalo, 1995).

A certain *hi-tech* method for measuring potential differences uses precise atomic clocks and the slowing of clocks predicted by general relativity theory. At the time of writing there exist so-called *optical lattice clocks*, atomic clocks operating at optical frequencies, which should have the relative precision of $1 : 10^{18}$ required for a one-centimetre precision.

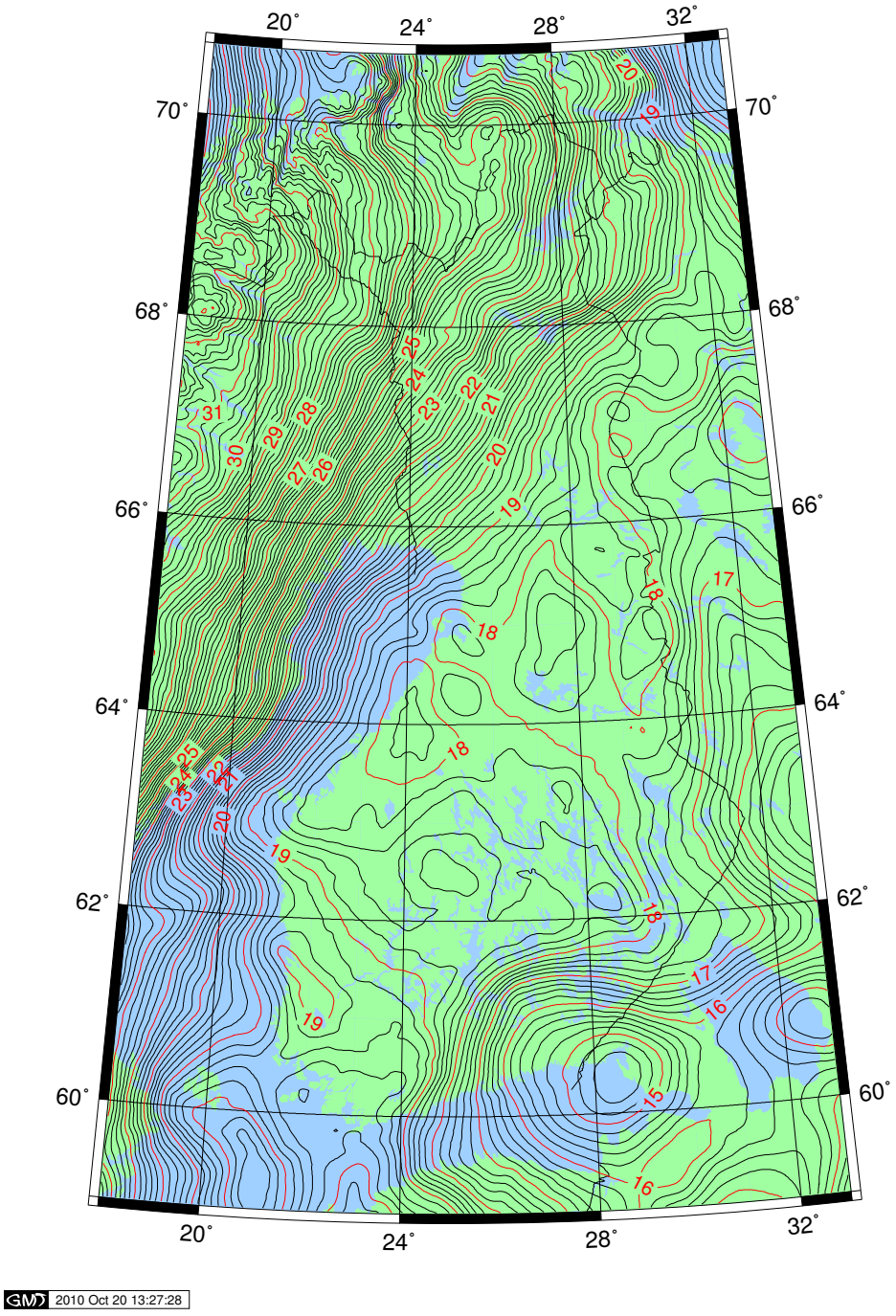


Figure 4.4. The Finnish geoid model FIN2000 (data © National Land Survey of Finland). This map plots the heights of the geoid above the geocentric GRS80 reference ellipsoid. Unit m.

Another *hi-tech* method which is already being realized (Gruber et al., 2014), is the construction of precise, high resolution geopotential models, which may be used to calculate a point's precise geopotential immediately when satellite positioning has determined the point's precise geocentric location. The already completed satellite gravity mission GOCE

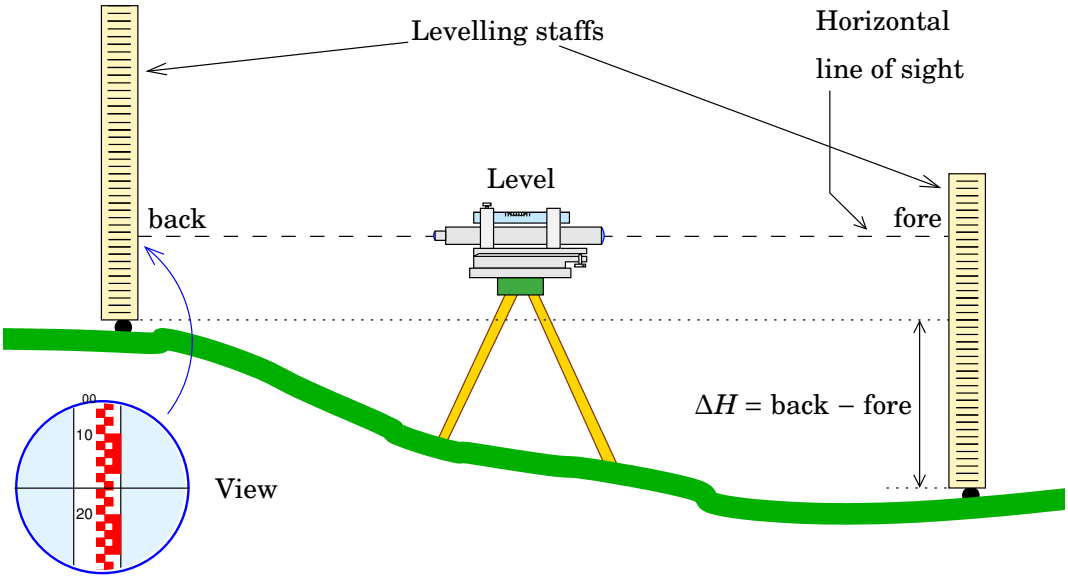


Figure 4.5. The geometry of levelling.

(2009 – 2013) is key to this.

4.3.2 Levelling

The already mentioned *levelling* technique is the standard method for determining heights referred to mean sea level. Levelling measures the height difference between two points, figure 4.5. The distance between points measured in one measurement set-up is short; by chaining point intervals into a traverse levelling, height differences between points at great distances from each other may be determined, and thus heights mapped over wide areas.

In Finland, like elsewhere, the levelling network covers the whole country and provides the opportunity to determine the heights of points in the *reference system* of the network. There is a *hierarchy* in the levelling network: the *precise levelling network*, which was measured and maintained in Finland by the Finnish Geodetic Institute, covers the whole country but is sparse: the loops of the network are hundreds of kilometres long. The lower-order levellings by the National Land Survey densify this network, thus bringing official heights within reach of all users, and many local actors — municipalities, builders — connect their own levelling networks to this system.

The height contours appearing on topographic maps are also in the official system, earlier N60, nowadays N2000.

4.3.3 Creating a height system

The height differences ΔH provided by geometric traverse levelling may be added together only *within a small area*, where local gravity is constant. In larger areas, the height differences ΔH must first be converted

to geopotential differences ΔC:

$$\Delta C = g \cdot \Delta H,$$

in which g is local gravity. After that, it holds for geopotential differences:

$$\sum_{\text{closed loop}} \Delta C = 0,$$

although for raw height differences:

$$\sum_{\text{closed loop}} \Delta H \neq 0!$$

In other words, while the sum of height differences $\sum_{A}^{B} \Delta H$ depends of the path chosen from A to B — and is thus not unambiguous — is the sum of potential differences $\sum_{A}^{B} \Delta C$ independent of the choice of path. Being unambiguous, the geopotential is better suited as the basis for the height *system* of an area.

4.4 The levelling instrument ("level")

Levelling (so-called *geometric levelling*, figure 4.5) depends on a horizontal line of sight: the optical axis of the levelling instrument's measuring telescope, or *sight axis*, is horizontal. To achieve this, the instrument has a *spirit level*. Both the spirit level and the telescope are connected to the body of the instrument.

A traditional levelling instrument (figure 4.6) includes, i.a., a measuring telescope, a small circular or *bull's-eye level* for approximate levelling, and a precise tubular level. To it belongs also a tripod and footscrews. The optical exis of a well adjusted levelling instrument, the *sight axis*, — i.e., the line defined by the crosshairs in the eyepiece — is parallel to the horizontal as defined by the level, i.e., the *horizon*.

rasiatasain

hiusviivaristikko
okulaari

At every instrument station, the levelling instrument must be levelled anew. Many levelling instruments have a separate *lifting screw* for precisely levelling the instrument. This is done before every forward and every backward measurement.

The accurate parallellity of the sight axis and the tubular level's horizon is achieved using an *adjustment screw* when checking up the instrument.

4.4.1 Classification of levelling instruments

Levelling instruments are classified according to accuracy, purpose of use, and construction, in order of increasing accuracy. See table 4.1.

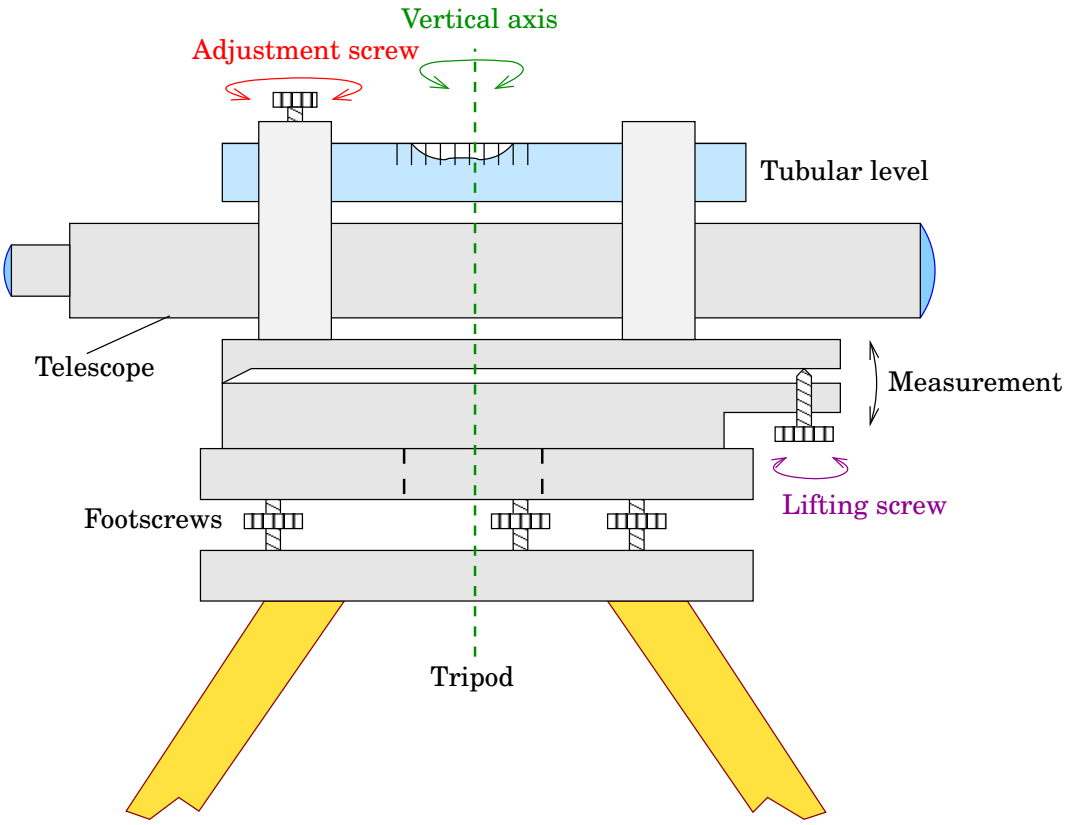


Figure 4.6. Levelling instrument.

We speak of *precise levelling*, if the precision on a one kilometre double run is below 0.5 mm. Devices of this accuracy class have an optical micrometer, see figure 4.16. The size and magnification of the measuring telescope grows with accuracy class, from 20× to 40× magnification. Also the levelling staff is chosen to correspond to the accuracy class, see table 4.2.

4.5 The measuring telescope

The tasks of the measuring telescope are to

Table 4.1. Classification of levelling instruments.

Instrument type	Levelling type
Builder's level	Construction, earthwork levelling
Engineer's level	Engineering, construction levelling
Precise level	Base network levelling
High-precision level	Precise levelling

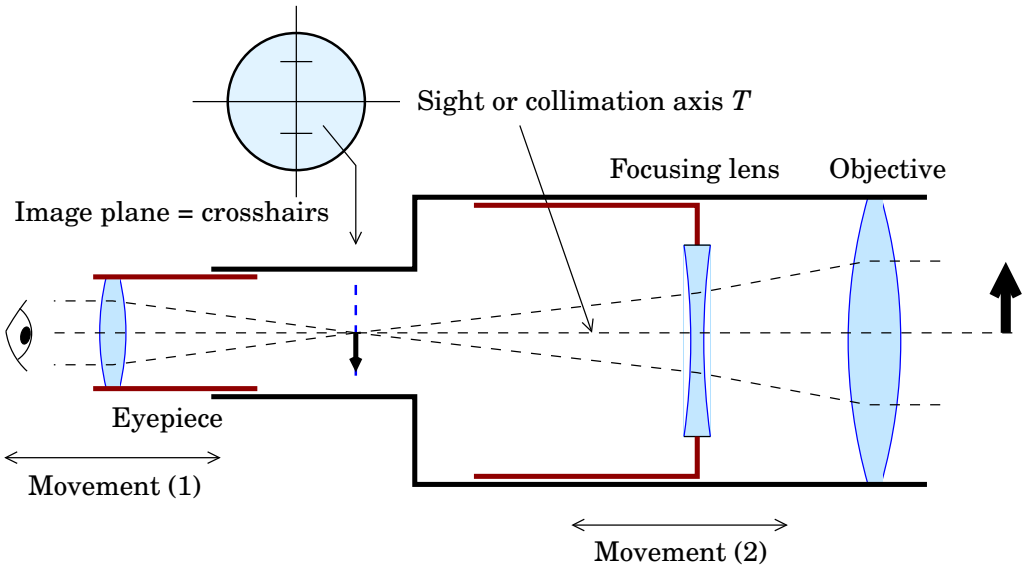


Figure 4.7. Measuring telescope. The object of study is to the right (at a great distance!), the observer's eye to the left.

1. give a sharp image of the aiming target
2. form the sight axis by placing the crosshairs[1] which are in the eyepiece, onto the image of the far away levelling staff.

Both tasks demand precise *focusing*. See figure 4.7.

Focusing is done as follows, usually by turning actuator rings or screws on the telescope:

1. The eyepiece is turned so, that the image of the crosshairs becomes sharp.
2. The focusing element of the instrument is turned so, that also the image of the target appears sharp.

In this case, the focal planes of objective and eyepiece[2] and the plane of the crosshairs all coincide.

In levelling, commonly equal distances to the front and back staffs are chosen. If this is not possible because of the terrain, on should *focus carefully* at every instrument station. If not, one may get *parallax*: the

[1]The crosshairs are nowadays usually engraved onto a glass plate, a *reticule*. Still in the 1930s, threads from a spider's nest web were used! See Bedini (2005).
Crosshairs were apparently invented by the astrometrist William Gascoigne (1612 – 1644), who accidentally noticed the suitability of spider silk for this purpose. Sadly, he was killed in battle in the English civil war.

[2]This only applies if the target is at infinity and the eye of the observer is error free. More precisely the eyepiece + the possible eyeglasses of the observer + the living optics in his own eye project a sharp image of crosshairs and target image onto the retina.

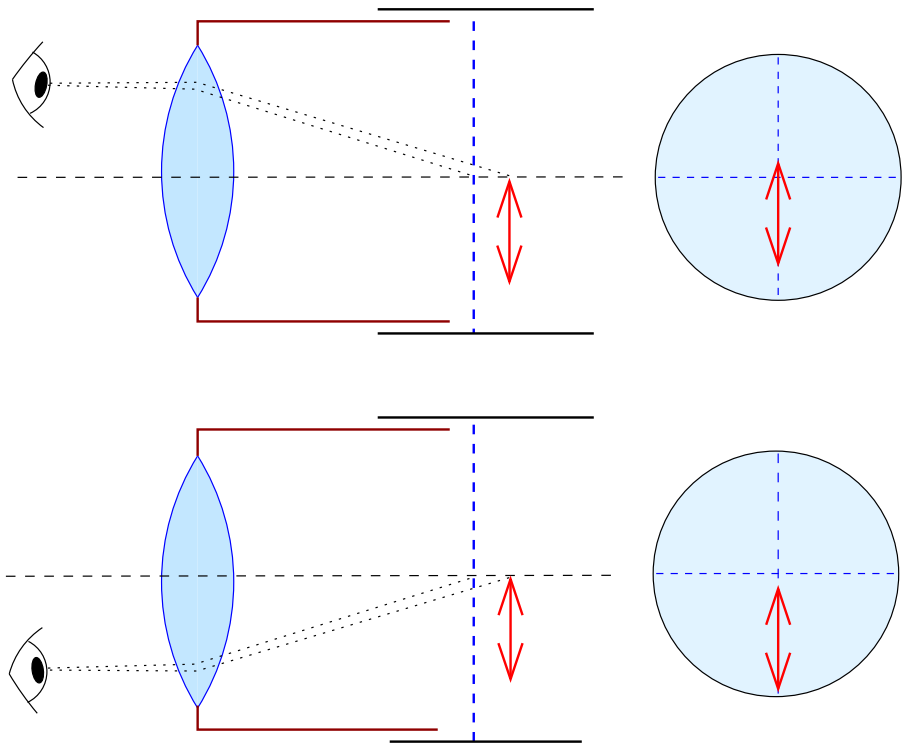


Figure 4.8. Parallax of a measuring telescope. If image and crosshairs are not in the same plane, moving the eye with respect to the eyepiece will cause them to move with respect to each other.

apparent direction of the telescope's optical axis will depend on the position of the observer's eye in relation to the eyepiece. Observing through a poorly focused telescope also tires the eyes.

One must always focus carefully!

Eyeglasses can be taken off if they aren't so-called cylindrical glasses (astigmatism), because near- or farsightedness can be corrected by focusing the eyepiece.

4.6 The tubular level

The construction of a tubular spirit level is explained in figure 4.9. The adjustment screw seen in the picture is only used for adjusting the instrument, rarely in the field. Its purpose is to get the level axis or *horizon* L and the sight axis of the telescope T (figure 4.7) precisely parallel. The level axis is horizontal when the bubble is in the middle, a precondition for measurement.

The task of the tubular level is to help the observer get the sight axis of the levelling instrument horizontal, i.e., perpendicular to local gravity.

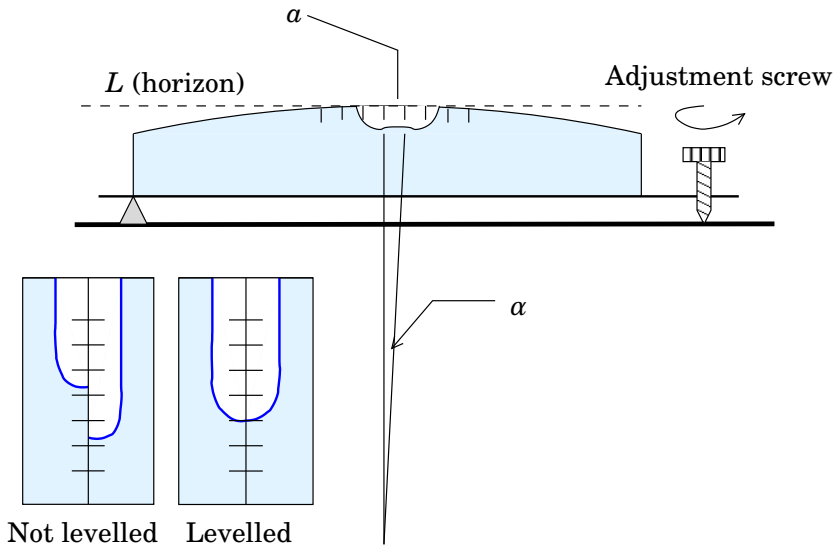


Figure 4.9. Tubular spirit level. Above, the construction, left below, the bubble seen through the prism system in a *coincidence level*.

With it, the instrument is levelled in the measuring direction in connection with every reading.

The distance a is the interval between the level's graduations. Generally $a \sim 2\,\text{mm}$. The *sensitivity* of the level is expressed by the angle α.

In a *coincidence level*, a system of reflective prisms is used to show the opposite heads of the bubble side by side, improving the precision of the levelling achieved.

4.7 Checking and adjusting a levelling instrument

4.7.1 Field check

At certain time intervals one must verify that the sight axis T of the levelling instrument is *parallel* with the axis, or horizon, L of the level. Due to the influence of the environment, every instrument "lives" and changes, e.g., with variations in temperature and air pressure, and due to handling and wear.

A check is carried out using levelling observations (field check): as measurement distance ℓ is chosen $25 \ldots 50\,\text{m}$, depending on weather conditions: the measurements are best done during overcast weather.

The checking method is based on the circumstance that when measured from A, the measurement result, i.e., the height difference, is *correct*, whereas when measured from B, the measurement result contains an error $2v$, in which v is the error caused by the difference in direction between the sight axis and level horizon, at staff distance ℓ. We readily

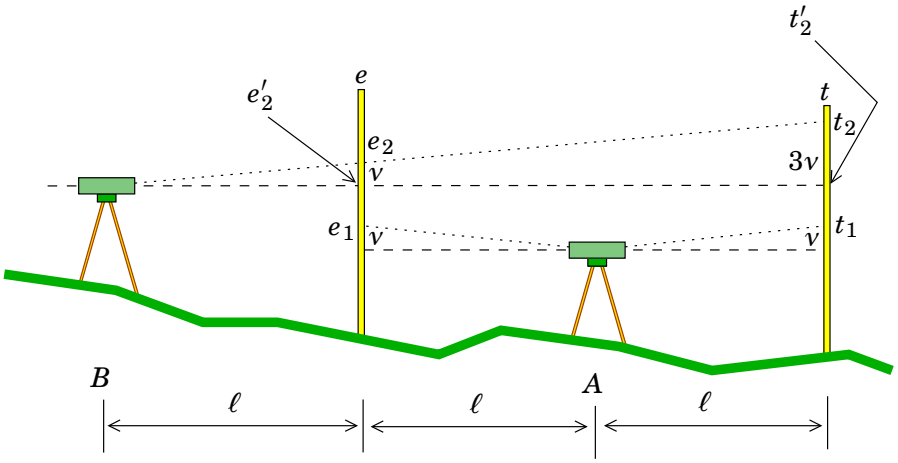


Figure 4.10. The geometry of a field check (Kukkamäki method).

obtain

$$\Delta h_A = (t_1 - v) - (e_1 - v) = t_1 - e_1,$$
$$\Delta h_B = (t_2 - 3v) - (e_2 - v) = t_2 - e_2 - 2v.$$

These height differences are *identical*. From this condition we solve for v:

$$t_1 - e_1 = t_2 - e_2 - 2v \Longrightarrow v = \frac{1}{2}[(t_2 - e_2) - (t_1 - e_1)].$$

The figure shows that

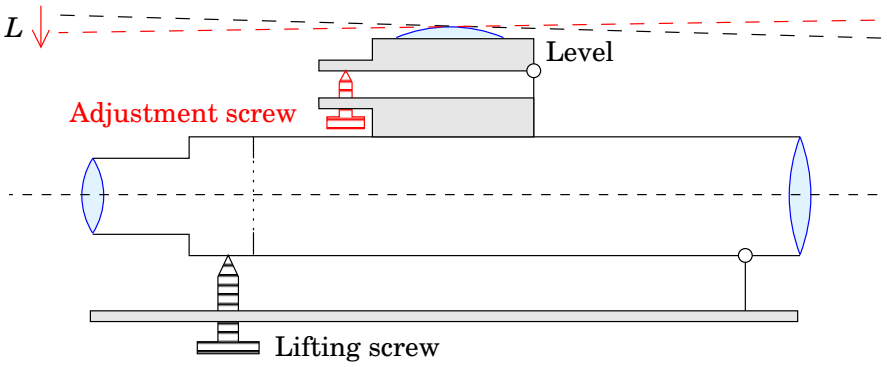
$$e_2' = e_2 - v,$$
$$t_2' = t_2 - 3v,$$

readings that now can be *calculated*. It enables the correction of the difference in direction between sight axis and level horizon in the field, using the *adjustment screw* meant for just that.

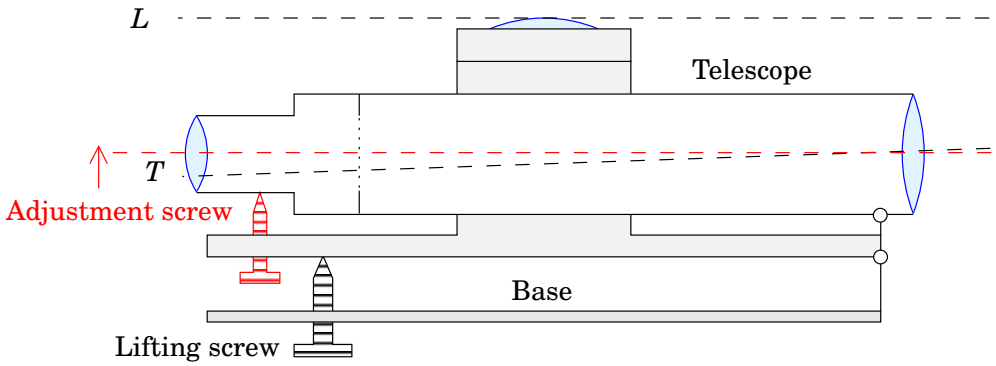
4.7.2 Adjusting a levelling instrument

When, as a result of the field check, it is known that the staff reading is off by an amount v, one proceeds as follows to get the instrument adjusted:

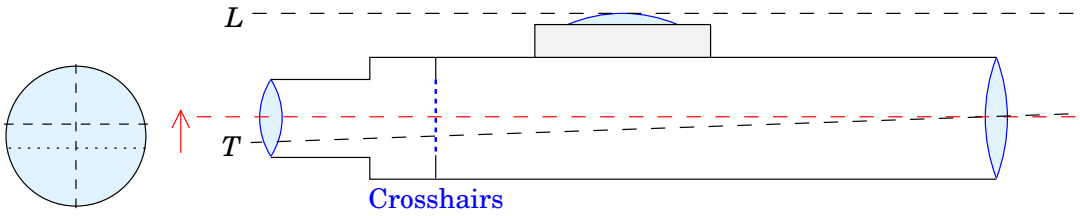
1. The instrument has an adjustment screw for the *tubular level*, which tilts the level with respect to the telescope (figure 4.11a). The telescope – tubular level assembly is attached to the base by a *lifting screw*.
   (a) Level the instrument first approximately using the footscrews, then precisely using the lifting screw, all the time keeping an eye on the tubular level.
   (b) Take the staff reading e.



(a) The tubular level has an adjustment screw



(b) The telescope has an adjustment screw



(c) The crosshairs plate can be adjusted

Figure 4.11. Adjusting the horizon of a levelling instrument.

□

(c) Move, using the *lifting screw*, to the staff reading $e' \stackrel{\text{def}}{=} e - v$, where v comes from the field check.

Note! Don't use the footscrews for this, because then the instrument may also tilt in the transversal direction — the footscrews are not in line with the sight axis of the measuring telescope.

(d) Now the bubble in the tubular level is no longer in the middle. Use the *adjustment screw* for the level to get the bubble in the middle again. After that, $L \parallel T$.

2. The instrument has an adjustment screw for the *measuring telescope*, i.e., the telescope tilts with respect to the level (figure 4.11b). The telescope – tubular level assembly is again attached to the base with a *lifting screw*.

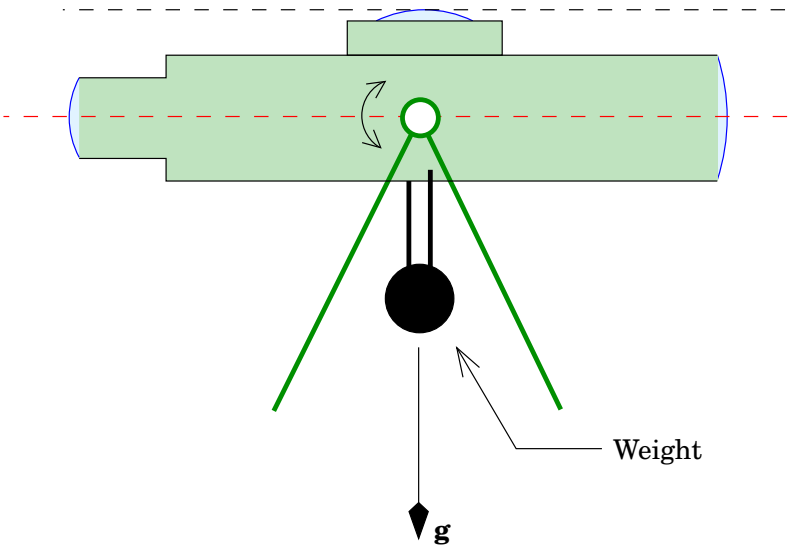


Figure 4.12. Principle of an old-fashioned self-levelling levelling instrument. An instrument built in this way is not very practical.

An equivalent, popular technical solution is an adjustment screw that shifts the crosshairs plate in the vertical direction within the image plane (figure 4.11c).

(a) Level the instrument.

(b) Take staff reading e.

(c) Move, using the telescope's (or crosshair glass plate's) *adjustment screw*, to the new staff reading $e' \stackrel{\text{def}}{=} e - v$.

(d) The tubular level bubble is still in the middle!

4.8 Self-levelling instrument

Self-levelling or automatic levelling instruments use *gravity* to obtain a horizontal sight direction.

Old-fashioned models used gravity to level the whole measuring telescope according to the pendulum principle (figure 4.12). It is clear that such an instrument is difficult to use under field conditions due to disturbances caused by wind and observer proximity.

Nowadays only a prism or mirror guiding the light is used as a pendulum. It is suspended inside the telescope: a *pendulum compensator* is built into the telescope. In order to function, the instrument must already be approximately levelled with the aid of a bull's-eye level.

The principle of the pendulum compensator is shown in figure 4.13.

The principle of operation of the pendulum compensator is explained conceptually in figure 4.14, in which the path of the light beam has been "folded open". The figure shows the situation in a co-ordinate frame

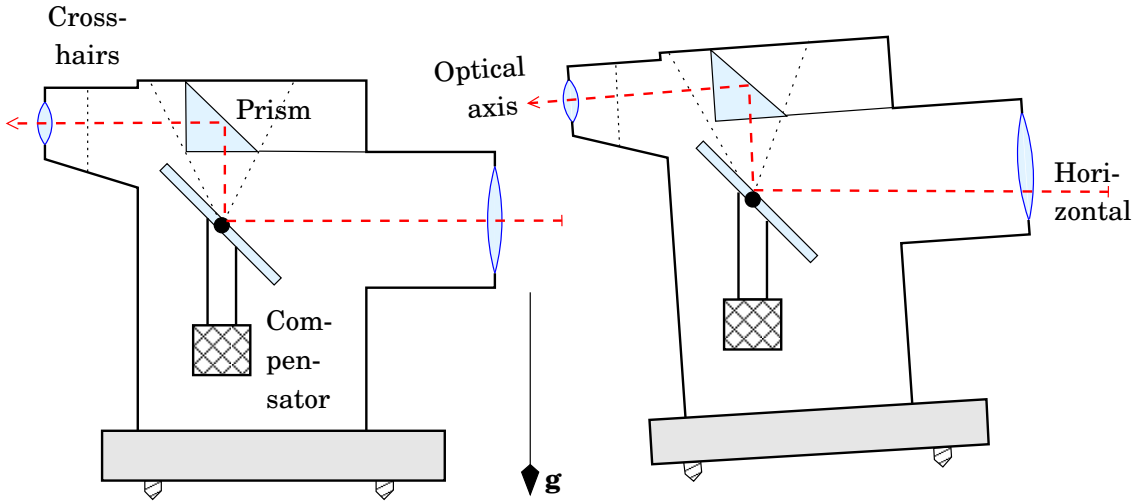


Figure 4.13. Modern self-levelling levelling instrument.

attached to the measuring telescope.

A small tilt of the telescope away from the horizontal causes a tilt of the incoming beam of α. In order for the image to remain in the same place in the image plane of the telescope, the compensator bends the light beam by an amount 2α, assuming that the distance between objective and compensator s is equal to that between compensator and image plane, in other words, the compensator is precisely in the middle between them. A freely suspended mirror turns, relative to the telescope, by an amount α, and the direction of the reflected beam changes by an amount 2α, just as intended. See Kahmen and Faig (1988, pages 334–336).

The strength of compensator instruments is their ease of use. However, in the early days there were technical issues, like magnetism of the mirror suspension (Kukkamäki and Lehmuskoski, 1984). These problems appear to be solved.

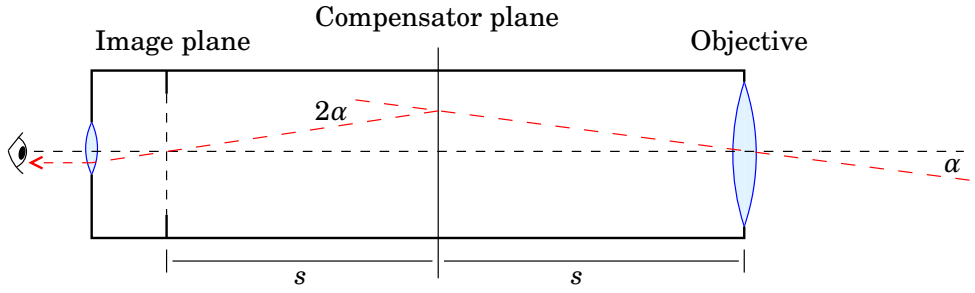


Figure 4.14. Principle of operation of the pendulum compensator.

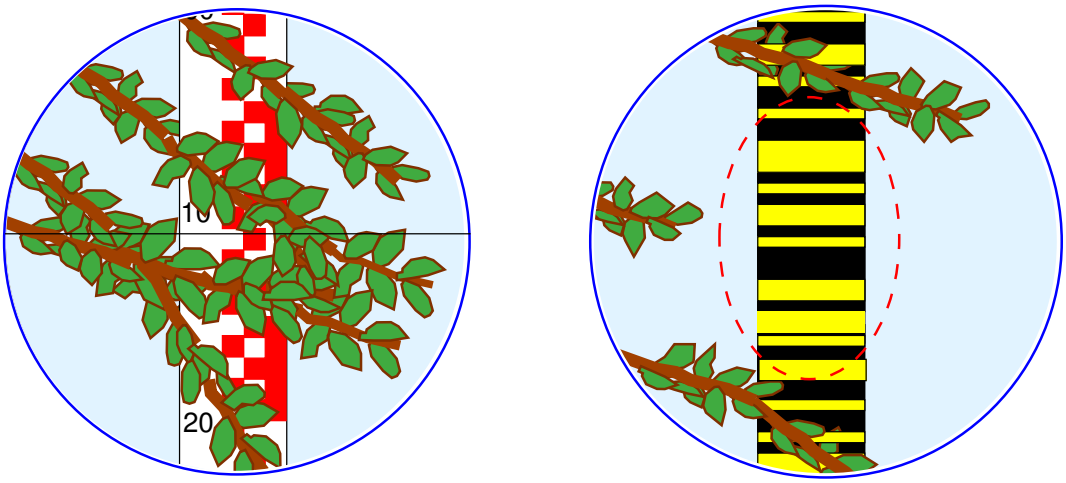


Figure 4.15. With a traditional levelling staff, one may read also through foliage. A bar-code staff causes confusion more easily.

4.9 Digital levelling instrument

Nowadays generally digital levelling instruments are used, as the automation of measurement that they bring with them saves costs. The measurements are stored directly into the instrument's memory, and necessary checks are done immediately.

The staff used together with a digital levelling instrument bears a *bar code* that, at least in principle, is no different than that found on merchandise. Thanks to it, height values can be machine read using the CCD sensor — nowadays that also available off the shelf — and a processor system. As a side result, also a rough staff distance is obtained, as well as a warning signal if fore and back distances are too different.

Unlike a tradional levelling staff, on which the measurement always uses the edges of one or at most two graduation lines, with a digital or bar-code staff always a *whole area* is used, of size 30 cm in the case of the Zeiss (Trimble) DiNi12 instrument. This has both advantages and disadvantages.

Advantage: In the measurements, some sort of average over the edges of many staff graduations is used. Therefore the accuracy of manufacture of the graduations and the accuracy of calibration of the staff are less critical. The staffs last longer in a useable state.

Disadvantages:

- The whole interval on the staff that is being used should be visible. In forested areas, this may cause problems.

- The calibration must always be done as *system calibration*: instrument and staff are calibrated as a "black box", together. On the other hand, by combining the calibration of staff *graduations* and system calibration, one may reconstruct how the instrument weights the graduations it uses, and thus get the black box ajar.

The digital method of levelling is in widespread use, even in precise levelling, and has been the subject of active research (Takalo et al., 2001; Takalo and Rouhiainen, 2004).

4.10 The levelling staff

The *levelling staff* or levelling rod is a scale, in Finland a metric one, with which the height differences between two points are measured using a levelling instrument. There are many alternatives for the graduation of the scale, figure 4.16:

- The "E" graduation. Simplest of all. Its weakness is that a bright white small square appears a little larger than a dark red small square — the Helmholtz[3] brightness illusion. When one interpolates the millimetres visually, small systematic effects easily occur.
- Chessboard graduation. Here, the above mentioned weakness has been corrected.
- Line graduation. Used in precise levelling. The line interval is 10 or 5 mm.

  Instruments for precise levelling have an *optical micrometer*, with which a better accuracy is obtained than by visually interpolating the graduations. The micrometer contains a turnable glass plate in front of the telescope objective, with which one gets the horizontal crosshair placed on one of the graduation lines. The reading from the line gives the crude value, the turning scale of the micrometer plate completes it to a precise reading.
- To be used with digital levelling instruments: bar-code graduation.

The staffs are often manufactured from wood, better ones from aluminium. In precise levelling, an "invar staff" is used, on which the graduations are painted on an invar tape, which is mounted on a wooden or aluminum-alloy frame, and kept under tension by a spring[4]. For precise-

[3] Hermann Ludwig Ferdinand von Helmholtz (1821 – 1894) was a German physician and physicist and a student of vision.

[4] The force of the spring is known — 200 N — and its effect on the length of the invar tape is computable. The coefficient of thermal expansion of invar is close to zero, and metal — unlike wood — is insensitive to moisture.

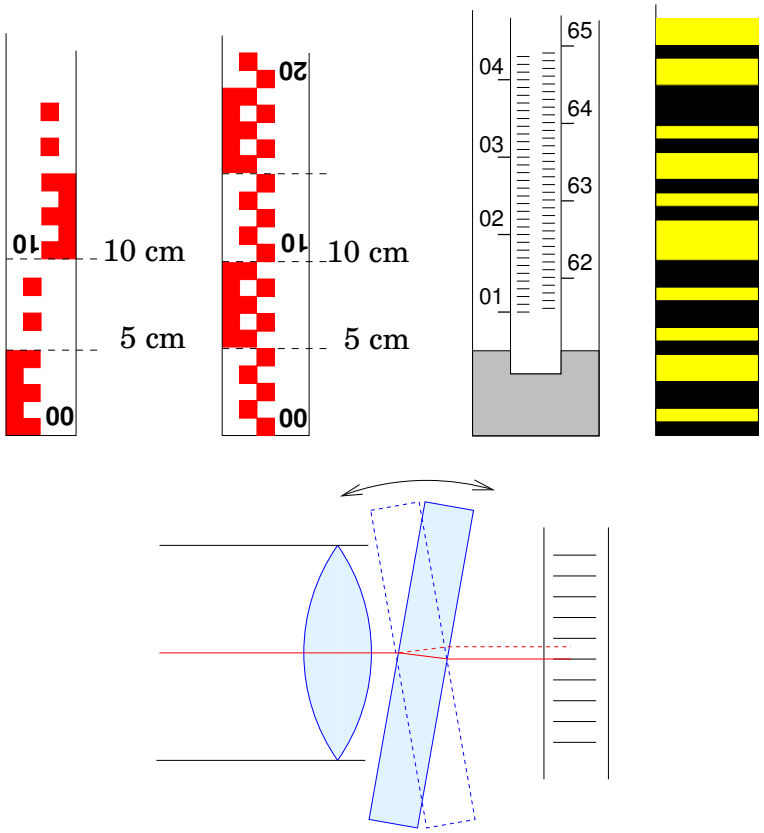


Figure 4.16. Graduation alternatives for the staff scale: "E" graduation, chessboard graduation, precise-levelling staff, bar-code graduation. Below, an optical levelling micrometer to be placed in front of the objective.

□

levelling staffs, a graduation of half a centrimetre is used — the "staff unit" is 5 mm. There are two graduation scales, slightly shifted with respect to each other, as a double check on reading mistakes and to randomize reading errors. Sometimes the scales are on different sides of the staff: "reversion staff"

Upmarket levelling staffs always sport a built-in bull's-eye level. The staff must be precisely vertical at the moment of reading!

Staffs are classified according to purpose of use, see table 4.2, which describes Central European practice.

A *self-calculating* staff or area-levelling staff has the following properties, see figure 4.20:

- The graduation increases from below going upward.
- At the lower end is an adjustable foot, which may be pulled out on a known point, so that the correct fraction of a metre becomes visible. After that, one goes into the terrain to map any number of point heights. See Tikka (1991, page 81).

In high-precision work, the staffs need to be regularly calibrated, at least before and after the field season. Digital levelling instruments and the

Table 4.2. Classification of levelling staffs.

Type of staff	Length (m)	Other
Simple	3 – 5	Foldable or telescoping, wood or aluminium, 2 to 4 parts, "E" graduation
Base-network levelling staff. In German: *Zweiskalenlatte*	3	Stiff, wooden, chessboard or line graduation, double scale or scale on both sides (reversal staff), bull's eye level
Precise-levelling staff	3	Wooden or aluminium frame, invar tape on which double scale
Bar-code staff	3	Aluminium, invar tape. Used with digital levelling instrument
Industrial staff		Double-scale staff of extreme precision
Self-calculating		See text

bar-code staffs they use should be calibrated *as a system*.Suitable staff supports may be benchmarks, temporary wooden spikes hammered into the ground, etc. Standard staff supports are depicted in figure 4.17. A levelling spike is used when the soil is soft, whereas a change plate is used on a hard substrate. A rail shoe is used when levelling along railroads. During the first and second precise levellings of Finland, many levelling lines ran along railroads, but since, more and more, highways have been used.

kiila
kilpikonna

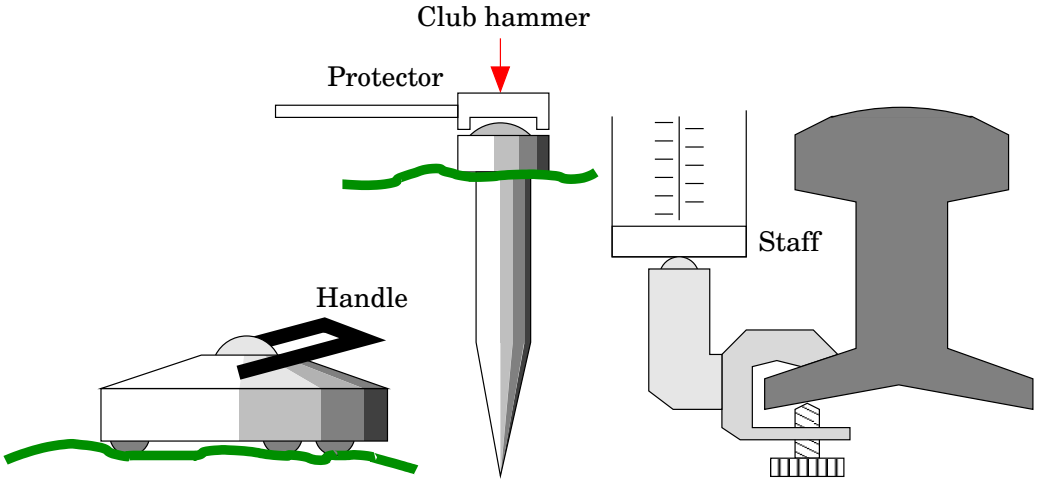


Figure 4.17. Various temporary levelling-staff supports: change plate (German: *Frosch* — "frog" —), levelling spike, rail shoe.

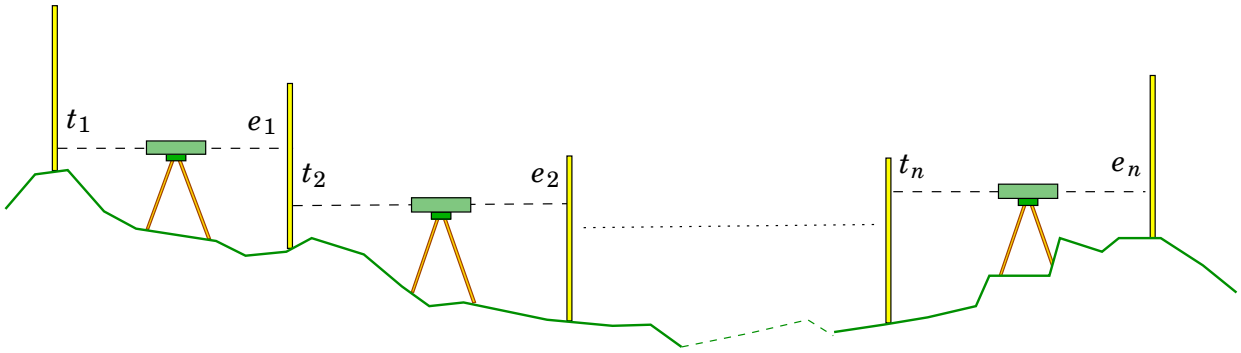


Figure 4.18. Traverse levelling.

4.11 Levelling methods

4.11.1 Traverse levelling

Line or traverse levelling is used in base network measurements. Its objective is to bring the official height system close to all users in the nation, e.g., to be used as the reference level for mapping surveys.

The precise levelling network is the national height base network. It is realized hierarchically by successive network densifications, see section 7.3.

The heights are computed for the user community in the national height system. The heights for the third, i.e., last, Finnish precise levelling were still computed in the N60 system, today the national height system is N2000. All levelling network computations are however first carried out with *geopotential numbers*, from which then either orthometric (N60) or normal (N2000) heights are calculated.

Base levelling is a lower-order traverse levelling serving the creation of height benchmarks for use close to the user. The terminology varies.

Traverse levelling is done by summing together the measurements between successive instrument stations (figure 4.18):

$$\Delta = (t_1 - e_1) + (t_2 - e_2) + \ldots (t_n - e_n) = \sum_{i=1}^{n} (t_i - e_i).$$

- The line runs from known point to known point. Sometimes this is impossible, and we speak of a “spike”. In this case we measure carefully in both directions: *check*.
- A well-planned levelling network incorporates all measurements and points into closed loops: *check*.
- In order to minimize the impact of weather and instrument related errors, one makes the fore and back staff distances as equal

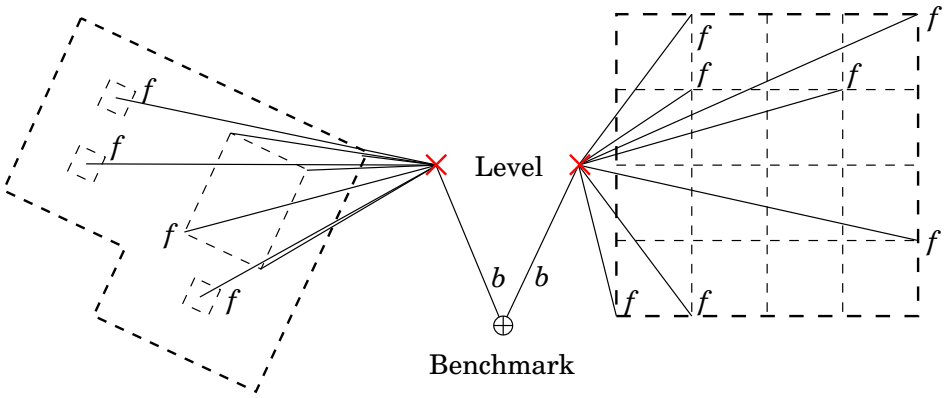


Figure 4.19. Area levelling. One backward (b) observation, many forward (f) observations.

as possible: $\ell_{\text{back}} \approx \ell_{\text{fore}}$. The staff distances also may not be too long, e.g., in precise levelling, 50 m, however depending on weather conditions. In overcast weather one can use longer staff distances; in sunny weather with strong turbulence, staff distances have to be shortened (Kääriäinen, 1966).

- If a levelling line runs along a railway or highway, the *safety arrangements* must be in order.

4.11.2 Area levelling

Area levelling is explained in figure 4.19. With this method, the height situation of a whole area is mapped using one known starting point. Among the points to be measured, there must be at least one other known point: *check*.

Area levelling is an ideal opportunity to use a *self-calculating levelling staff*, see figure 4.20. At the starting point one does the following: the foot of the staff is pulled out until in the levelling instrument one sees the correct decimal fraction of a metre: if the height of the point is known to be 12.75 m, the leg is pulled out until the observer sees < something > .750. The leg is screwed tight, and the staff is transferred to the first observation point. The observer must memorize the whole metres, the fractions show directly.

The results of area levelling are needed and used for

- mapping the height situation on a building site before starting the construction of the foundation
- creating digital terrain models locally and at high resolution
- calculating earthwork volumes to be moved.

Proper *checks* are important: in addition to the starting point, other points with known heights should be included in the measurement. An erroneous starting height would propagate in full to the whole area,

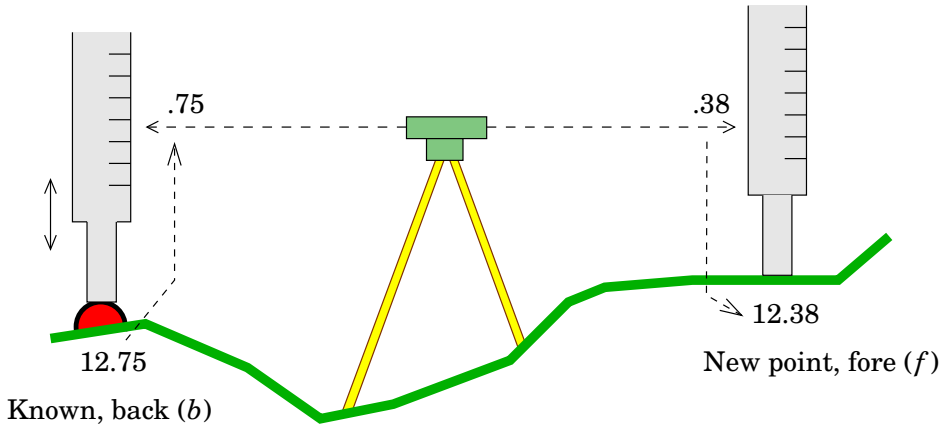


Figure 4.20. Self-calculating levelling staff. How the leg of the staff is set to the right length.

which would be, e.g., for sewer construction, a fatal and expensive mistake.

4.11.3 Technical levelling

4.11.3.1 Installation measurement in industry and construction

This belongs to the field of *engineering geodesy*.

- Extreme case: the CERN Large Hadron Collider (LHC) in Geneva, circumference 27 km, precision on the level of millimetres (Schrock, July 2, 2014)
- paper machines, shipyards
- road construction, bridges, tunnels, railroads
- etcetera.

4.11.3.2 Deformation measurement and monitoring

For this purpose, levelling is one of many methods: generally deformations are three-dimensional. See section 17.1.

- Extreme case: post-glacial land uplift
- deformations caused by the pumping of natural gas, oil, or drinking or irrigation water, antropogenic land subsidence. Venice.
- Deformations of dams, reservoirs, bridges, high steel
- old buildings, the tower of Pisa
- etcetera.

4.11.4 Levelling of profiles and cross sections

Profiles and cross sections are measured in connection with construction projects, especially of roads, railroads and canals.

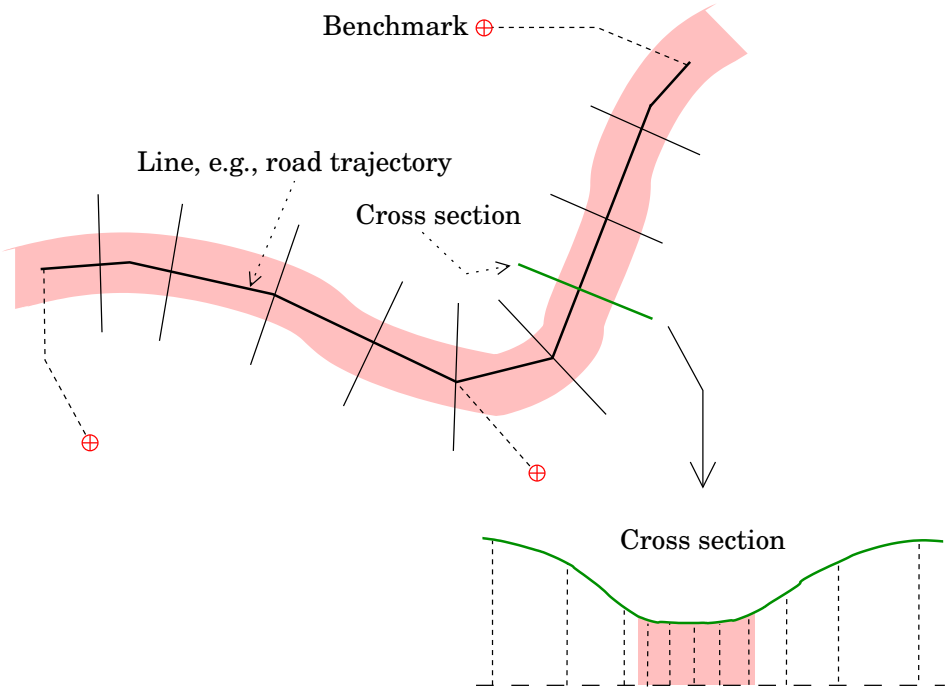


Figure 4.21. Profile and cross sections.

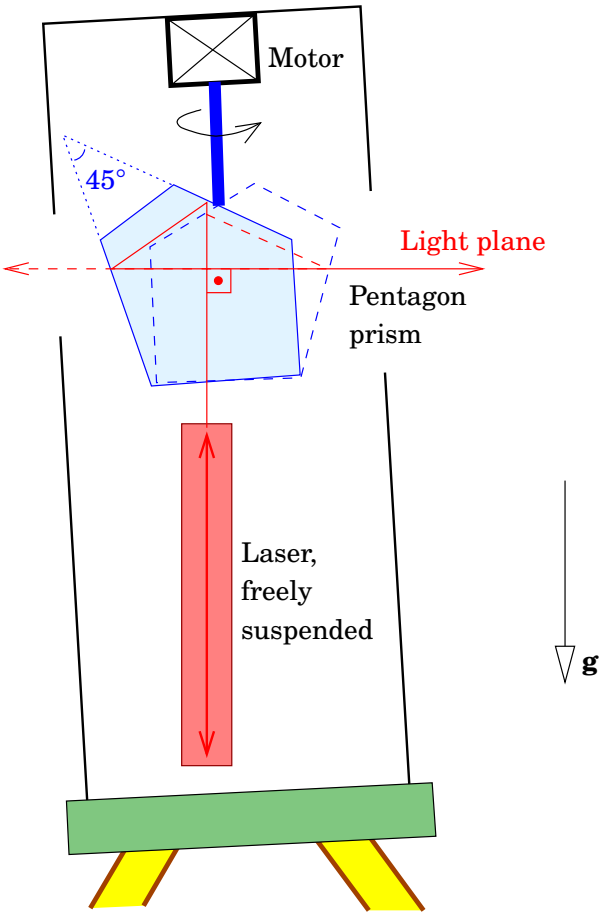


Figure 4.22. Principle of operation of a laser level.

A profile is a longitudinal section of the Earth's surface, usually along a planned trajectory of a road, railroad or waterway[5].

The work starts with setting out the trajectory into the terrain. Markers are placed at 25, 50 or 100 metres intervals, as well as at sharp terrain features, etc. The markers are numbered: the number signs are placed outside the work area. The heights are measured with a builder's level.Tie measurements to benchmarks of the height base network in the area are made *at least two*, at the start and end points of the profile. If use of two benchmarks is not possible, the loop must be closed as a double check by measuring forth and back. Closing errors are corrected in proportion to the levelled distances.

A cross section is a transversal section of the Earth's surface, perpendicular to the trajectory of the profile. At corners, the angle is split equally. Cross sections are typically $20-50\,\mathrm{m}$ long. The purpose of cross sections is to support the planning work and enable the calculation of earthwork masses to be moved. The height measurement of the cross sections is carried out in the same manner as area levelling. The point density is chosen in accordance with the terrain and the intended use.

4.11.5 Laser level

Nowadays in surface levelling often *laser levels* are used, figure 4.22. Laser levels are instruments stabilized by a pendulum compensator, which project laser light though a rotating *pentagon prism*[6] into the environment, in order to form a horizontal plane.

The instruments are handy on building sites, where they realize a horizontal plane, which the user can make visible with a stick. Spreading sand, laying floors, or bricklaying a wall along a straight line becomes easier. A staff equipped with a suitable sliding receiving device will directly give the height of the point on which it stands.

A laser level usually can also be quickly reconfigured to produce a vertical plane of laser light. Also this is useful on building sites.

Self-test questions

1. What is the reference surface for orthometric heights?

[5] In this case, a parallel line is used.

[6] Inside the prism are two reflective surfaces with an angle between them of 45°. The angle between the incoming and outgoing rays is thus always 90° or 100^{g}.

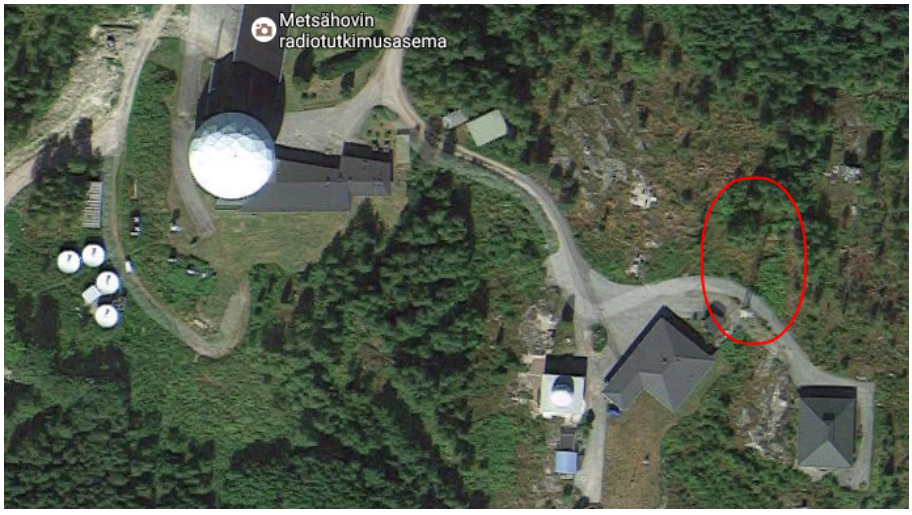
2. What is the relationship between the geoid surface and deviations of the plumb line?
3. How is the *sight axis* of a levelling istrument (or a theodolite) defined?
4. What is parallax, and how does one avoid it?
5. What are the graduation alternatives for levelling staffs?
6. Name three standard levelling-staff supports.
7. Describe the Kukkamäki method of field testing a levelling instrument. What is the purpose of the test?
8. Describe how an automatic or self-levelling levelling instrument works. Drawing!
9. Describe how a laser level works.

Exercise 4 – 1: Heights

Earlier we gave the geographical or geodetic co-ordinates (on the GRS80 reference ellipsoid) of the GPS antenna at Metsähovi research station, equation 2.5:

$$\begin{aligned} \varphi &= 60^\circ 13' 2''.89046, \\ \lambda &= 24^\circ 23' 43''.13336, \\ h &= 94.568\,\mathrm{m}. \end{aligned} \qquad (2.5)$$

1. Use the geoid model given on the U.S. NGA web page for the geopotential model EGM96, which is good to±25 cm on the Finnish territory, to determine the geoid height N at Metsähovi. Alternatively use Charles Karney's web page, which also gives the newer model EGM2008.
2. Compute the height H above "sea level", as represented by the EGM96 model. The relationship you need is given in figure 4.3.
3. From the air, Metsähovi research station looks like this (© Google Maps™). The GNSS antenna METS



Google for “google maps height info” and find the approximate height of the terrain in this area, ±1 m.

Question: how high is this steel grid mast?

4. Go to Google Maps and measure latitude and longitude of the foot of the mast. Compare with the above values. How large are the differences? Discussion.

5. The theodolite and angle measurement

The theodolite was probably invented by Leonard Digges. The invention was published by his son Thomas in his land surveying textbook *Pantometria* in 1571. The origin of the name is unclear. The first theodolites did not yet have a telescope, which was possibly invented — or at least attempted to be patented — not until 1608 by Hans Lippershey in the Netherlands.

The theodolite is an angle measurement device which measures horizontal and vertical angles relative to the local horizon (horizontal plane) and plumb line (direction of gravity). Modern instrument, electronic tacheometers or total stations, measure also distances. Here, we shall call all of those "theodolites".

5.1 Horizontal angles and zenith angles

Theodolite measurements are always made inside the Earth's gravity field. The vertical axis of the theodolite is levelled to be in the direction of the local plumb line. Directions and direction differences may be expressed, in these natural instrument co-ordinates frame, as horizontal and vertical angles.

Let there be (figure 5.2) a difference in direction between the points A and B. The points are projected along the plumb line onto the local horizontal plane, yielding the points A',B'. The difference in direction between points A' and B', the angle α, is the *horizontal angle* between points A and B. Both horizontal directions and horizontal angles are counted positive in the clockwise direction.

The angles ζ_A, ζ_B are the *vertical angles* or *zenith angles* of points A and B.

Horizontal angle: The angle α formed by the projections KA' and KB' of the rays KA and KB in the horizontal plane.

Zenith angle: The angle formed by the plumb line and the ray KA (ζ_A) or KB (ζ_B). The zenith angle is positive going down from the zenith.

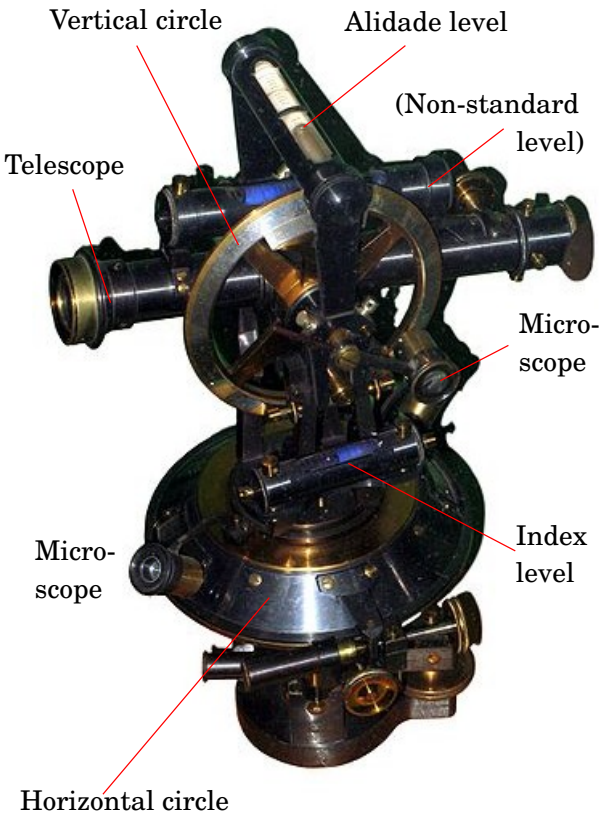


Figure 5.1. An old-fashioned theodolite. Note the external horizontal and vertical circles and reading microscopes. Base image © Wikimedia Commons.

The *plumb line* or vertical, the direction of a plumb wire, is the local direction of the Earth's gravity vector. The plumb line points at the centre of mass of the Earth, but *only approximately*[1].

5.2 The axes of a theodolite

A theodolite has *three axes* (figure 5.3),

- the vertical axis or *standing axis*, which must be directed along the local plumb line or gravity vector
- the horizontal axis or *trunnion axis*, around which the telescope turns

[1] Approximately, mainly because the Earth is an ellipsoid of revolution and not a sphere. The deviation of the plumb line from the direction the the Earth's centre of mass due to the flattening is at its largest as much as 11′ at latitudes ±45°. The plumb line is also ever so slightly curved. Additionally there are local, varying deviations of the plumb line from the surface normal to the ellipsoid of revolution, order of magnitude from a few seconds of arc in even terrain to over a minute of arc in the mountains.

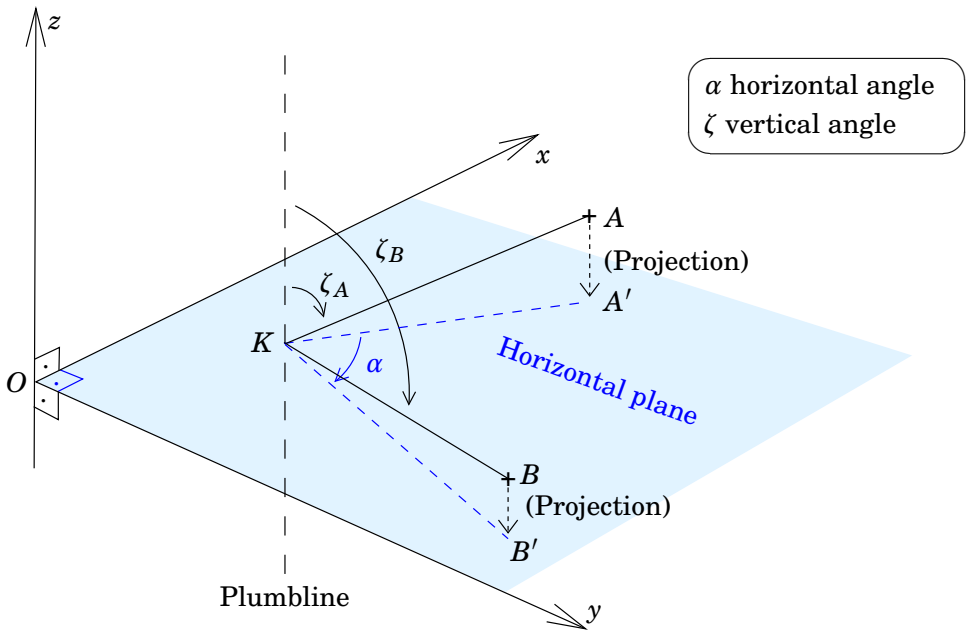


Figure 5.2. Horizontal angle and zenith angle.

- the sight or collimation axis, the axis of the telescope tube passing though the crosshairs of the eyepiece.

kehä A theodolite has two *circles*, the horizontal and vertical circle, figure 5.3.

Theoretical objective for a well adjusted theodolite:

- V,P and T intersect in the same point
- $P \perp V$. If not, the error is called *trunnion axis tilt*
- $T \perp V$ (but not $T \perp P$! Why not?). The error is called *collimation error*.
- V and P pass though the centres of the horizontal and vertical circles. If not, we speak of *circle eccentricity*.

5.3 Construction of a theodolite

See figure 5.4.

5.3.1 Measurement telescope

The measuring telescope of a theodolite is in principle the same as that of a levelling instrument (figure 4.7), a complicated assembly of lenses, prisms and mirrors. The telescope turns around the horizontal axis and it is joined in its movement by either the vertical circle itself, or by the index and reading microscope of the vertical circle, depending on instrument type.

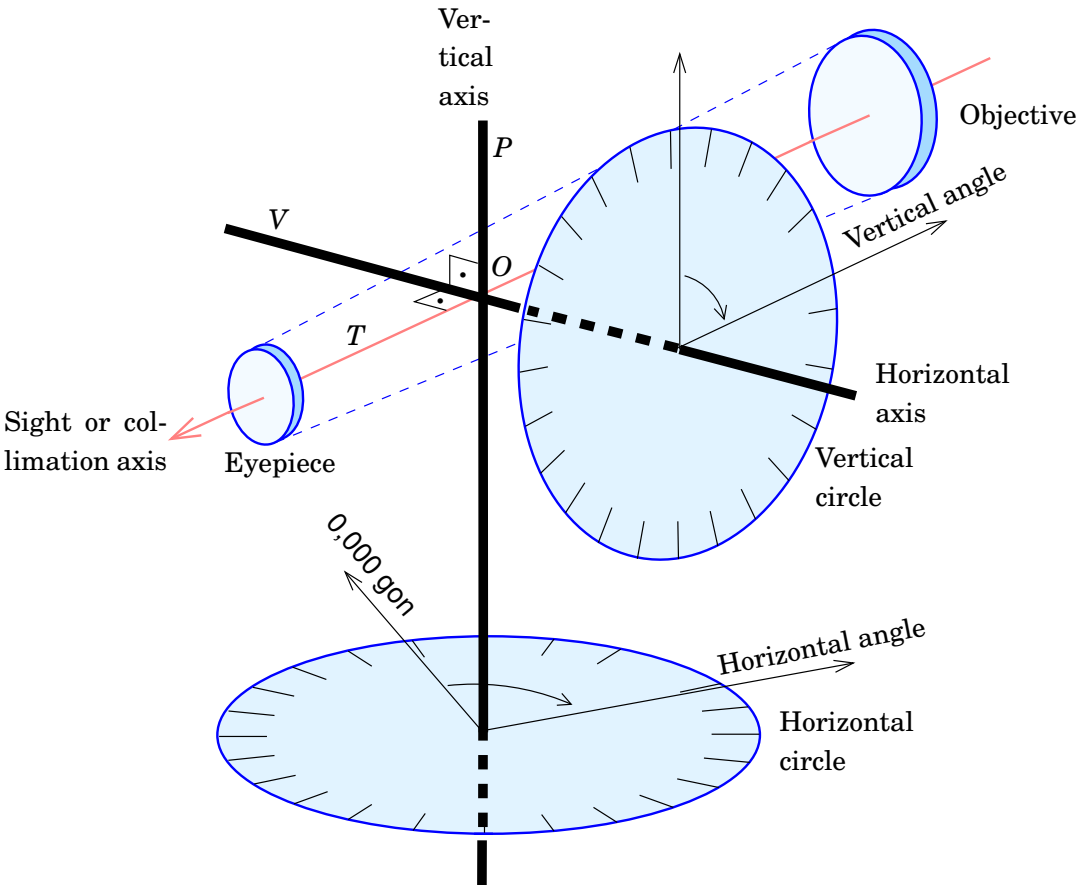


Figure 5.3. The axes and circles of a theodolite.

5.3.2 Alidade

The alidade (Arabic *al-idhâdah*, ruler) is the central part of the theodolite, which turns around the vertical axis, carrying the telescope with it. It contains the following parts:

- a tubular level, the *alidade level*, to level the theodolite
- reading devices for the horizontal and vertical circle:
  - Most often, the images are led by prisms to a microscope eyepiece beside the measurement telescope eyepiece, to facilitate the work of the observer.
  - The scale microscope and the optical micrometer enable a greater reading accuracy, see (Kahmen and Faig, 1988, pages 66 – 67).
  - In a coincidence microscope, readings are taken from two opposite places on the horizontal or vertical circle. In this case, also always an optical micrometer is used (Kahmen and Faig, 1988, pages 72 – 76).

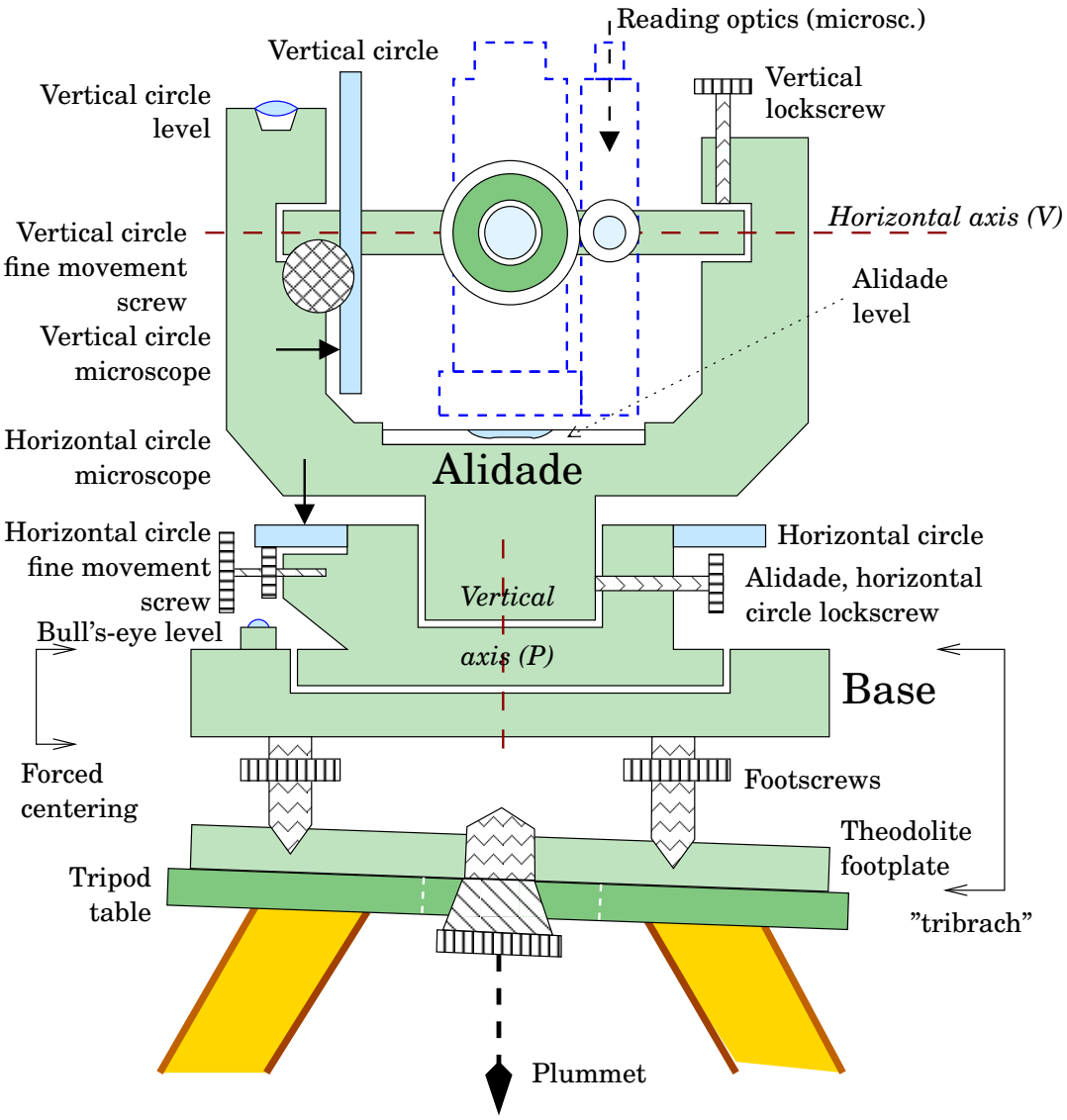


Figure 5.4. Theodolite construction.

5.3.3 Base

The *base* of the theodolite is the fixed part to which the horizontal circle is attached, and on which the alidade rests in its bearing.

- The base contains a forced-centring or footscrew device called *tribrach*.
- The upper part of the base sits *forcibly centred* in this device: it may be detached and reattached precisely to the same place.
- The forced-centring device is attached to the tripod head using a big screw[2]. In the middle of the tripod head there is a large round

[2]The screw conforms to the geodetic standard of 5/8 inch, 11 threads per inch.

Figure 5.5. Forced-centring device or plate.

hole and a mechanism to allow enough horizontal motion for *centring*.

- The footscrews are for *levelling* the theodolite: the vertical axis of the theodolite is oriented along the local plumb line.
- A bull's-eye level is used for rough levelling, whereas the alidade level helps in precise levelling.
- The attachment for a string plummet, or an optical plummet, are for centring.

5.4 Theodolite handling in the terrain

As an expensive fine mechanical and optical device, a theodolite should always be treated with due respect:

1. Transport, especially over longer distances, happens always in the carrying case.
2. The parasol is for the instrument, not the observer. It also offers protection against rain.
3. The instrument is never aimed directly at the Sun: the glass plate with the engraved crosshairs would go pop, and the instrument would need to be fixed and re-calibrated. Further damage occurs when the instrument contains a range finder.
4. *Careful book-keeping* is important: write up anything that might be relevant, e.g., weather conditions (so-called metadata).

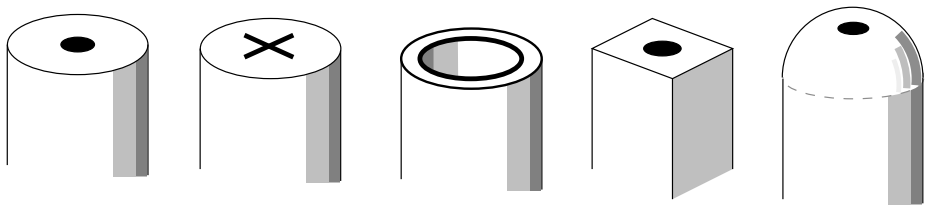


Figure 5.6. Various monument types.

5.4.1 Monuments and point descriptions

Unlike a levelling instrument, a theodolite must be placed precisely over the point to be measured, so that the theodolite measurement refers to the point.

In figure 5.7 the point K has been monumented in the terrain, a measurement point with a central mark. Figure 5.6 shows some examples of monument types in use.

The choice of monument type must be made such that

- The point is clearly and uniquely defined. If one intends to do both a precise positioning and heighting, it is best to use a steel bolt with a round head with a small hole in the middle.
- The monument withstands the impacts of weather and environment — frost heaving! A bedrock point is the best. Iron monuments should be protection painted (rust protection paint, "red lead" is no longer available for toxicity reasons).
- The monument can be easily found — point description!
- The point number is marked on the monument or painted (and chiseled!) next to it.

Always when a point is monumented for future use, a *point description* must be drafted, which helps to find the monument even after decades. The point description may contain the following information:

- the distances of the point to at least three features in the nearby landscape which are believed to be permanent — trees, building corners, etc. —, measured by tape. In the point description a sketch is included of the relative positions of point and reference features
- An approach map on which kilometre counts for car drivers, road signs, description of the landscape, and other useful details
- precise co-ordinates for hand-held GNSS
- Well-chosen photographs.

In order to ensure that angles measured from the point will be correct, two actions need to be performed: *centring* and *levelling*.

centring: in figure 5.7, the points K and O, the monument and the

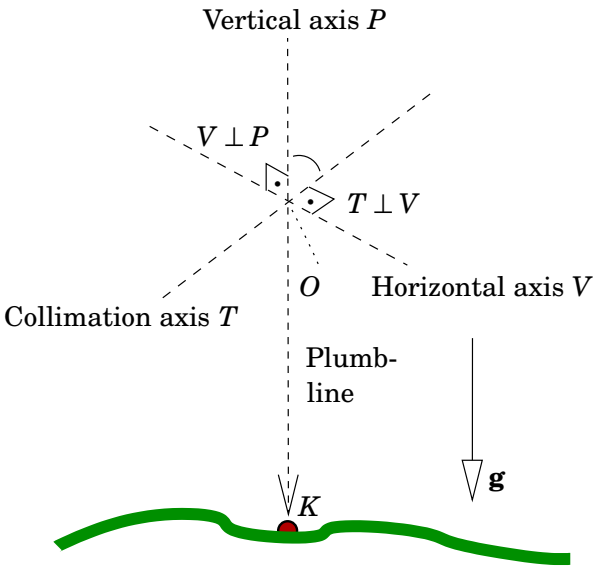


Figure 5.7. Theodolite axes. K is a point monumented in the terrain.

intersection point of the theodolite's axes, must be on the same plumb line.

Levelling: The vertical axis P if the theodolite must be aligned with the plumb line. Then, when the theodolite is turned around the vertical axis P, the horizontal axis V will maintain a constant angle with the plumb line: Ideally, without trunnion axis error, 90°. The horizontal axis will thus always be in the horizontal plane.

centring and levelling are commonly done alternatingly, until the desired end result is achieved.

5.4.2 Rough centring

- The tripod is placed on the right spot, judging by eye and by changing the lengths of the legs.
- It is also judged by eye that the tripod head is horizontal.

5.4.3 Precise levelling

The precise levelling of the theodolite using the alidade level is done in the following steps:

1. Rough levelling is done with the bull's eye level.
2. Level first in the direction of footscrews 1 – 2. Take into account, if necessary, the *zero error* from step 4.
3. Turn the alidade 100 gon, and level also in this direction.
4. Turn the alidade 200 gon. If the alidade level is adjusted correctly, the bubble must again be in the middle. If not, we move the bubble *half way* the difference, i.e., to the equilibrium position, using the

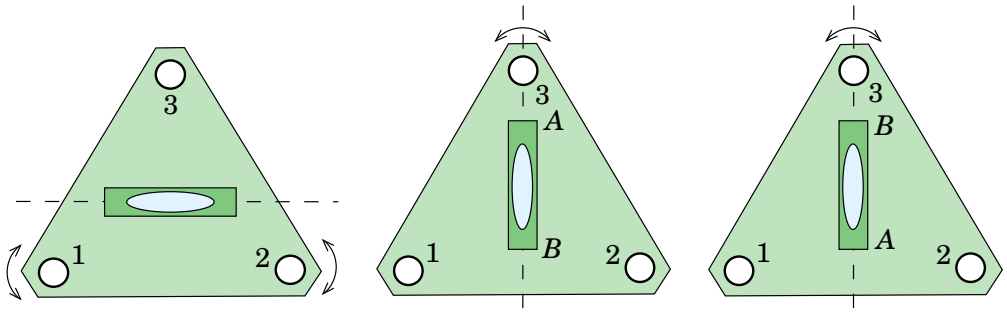


Figure 5.8. Precise levelling of a theodolite using the alidade level.

footscrews. The *zero error* of the level is half the shift of the bubble between steps 3 and 4.

5. Repeat 2–4 until the levelling no longer changes.

Make sure that the bubble can move freely and its end doesn't attach to the edge of the level.

If the difference found in step 4 is very large, then the alidade level is in need of *adjustment*.

5.4.4 Precise centring

The method will depend on the type of plummet used. Nowadays the most common plummet is the *optical plummet*.

No plummet is exact. Also measuring the height of the instrument over the monumented point is subject to error. Because of this, in special measurements, when the precision requirement is sub-millimetre, one uses pillars permanently anchored in the bedrock or on a deep support (frost heaving!) instead of tripods.

1. *String plummet:*
   - traditional
   - attached in such a way, that levelling doesn't change centring
   - centring: *shift* the forced-centring device of the theodolite over the surface of the tripod head in such a way, that the tip of the plummet points at the monumented point K. Tighten the attachment screw of the theodolite. The central hole in the tripod head gives room to play
   - levelling with the theodolite's footscrews.

   Problem: sensitive to the wind.
2. *Rod plummet*:
   - a telescoping tube, the upper end of which is attached to the forced-centring device (through the hole in the tripod head), and the lower end is placed precisely on the central mark
   - a bull's-eye level on the rod

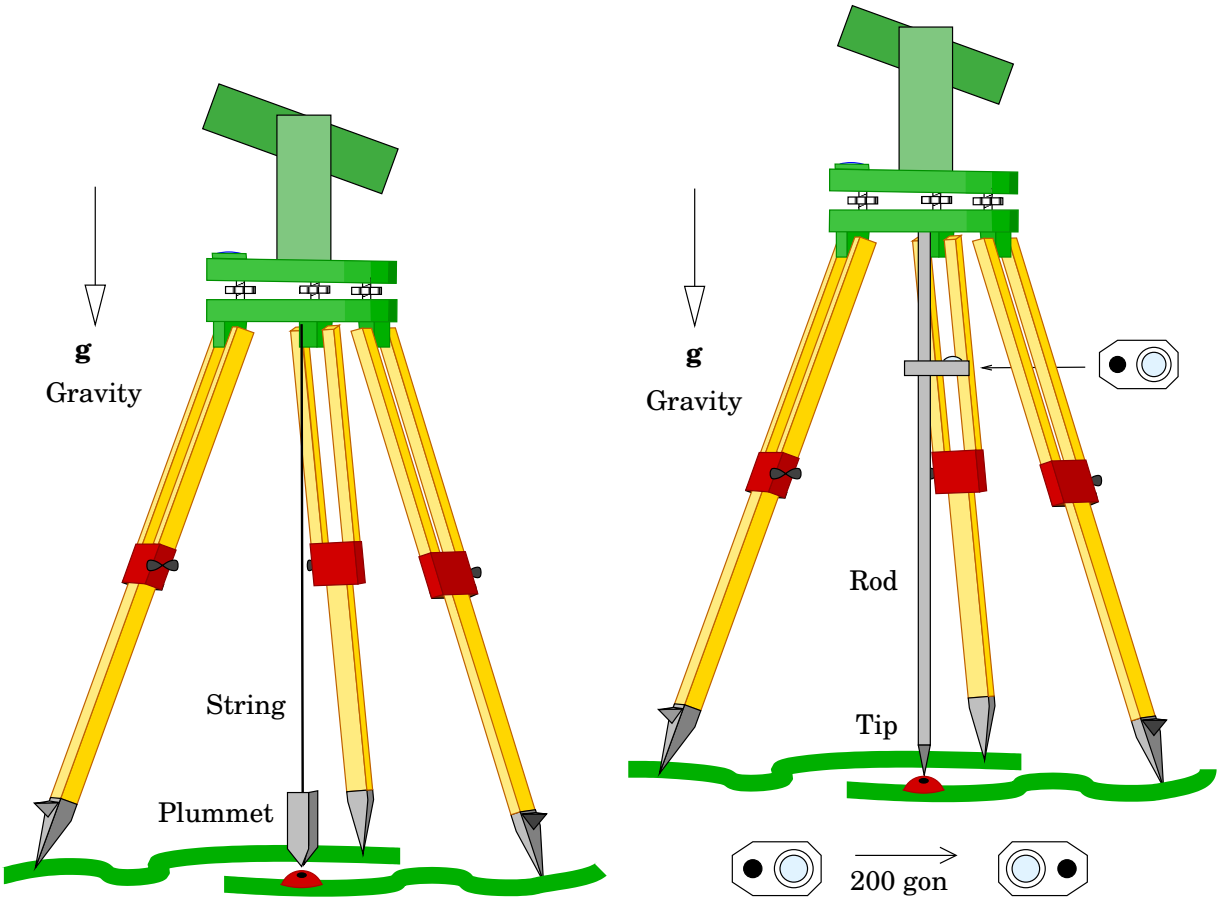


Figure 5.9. String plummet and rod plummet.

- centring:
  - Move the forced-centring device so, that the bull's-eye level on the rod is in the middle.
  - Turn the level 200 gon to the opposite side of the rod: if the bubble stays in the middle, centring has been achieved.
  - If not, level the bubble half way back to the middle, i.e., to the *equilibrium position*.

The rod plummet comes as a part of certain types of tripods (Kern), which are often used in accurate engineering surveying measurements.

3. Optical plummet:
   - In the forced-centring device's side (or in the theodolite's alidade's side) there is a small telescope with prism, which looks straight down. In the focal plane of the telescope's eyepiece there are crosshairs or something similar, i.e., an *index*.
   - The forced-centring device must have been levelled using the bull's-eye level.
   - Centring and levelling procedure using an optical plummet

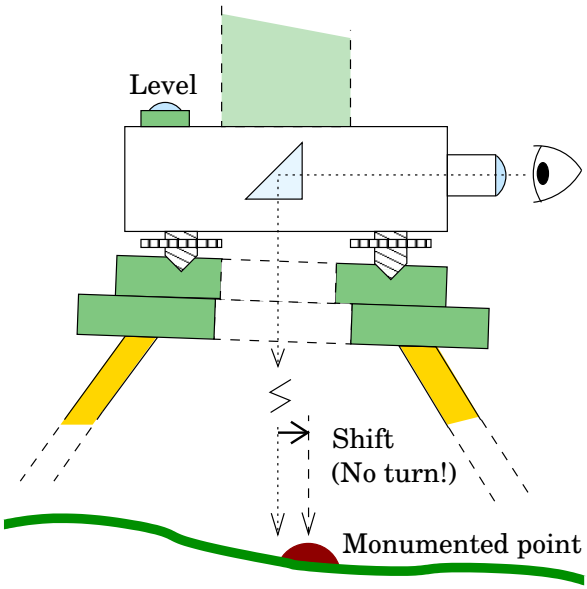


Figure 5.10. Optical plummet.

(figure 5.12):

(a) Using the footscrews of the forced-centring device, place the index initially on the point marker (meaning that levelling is lost!).

(b) By changing the lengths of *two* of the tripod's legs, get the bubble of the bull's-eye level again in the middle (the

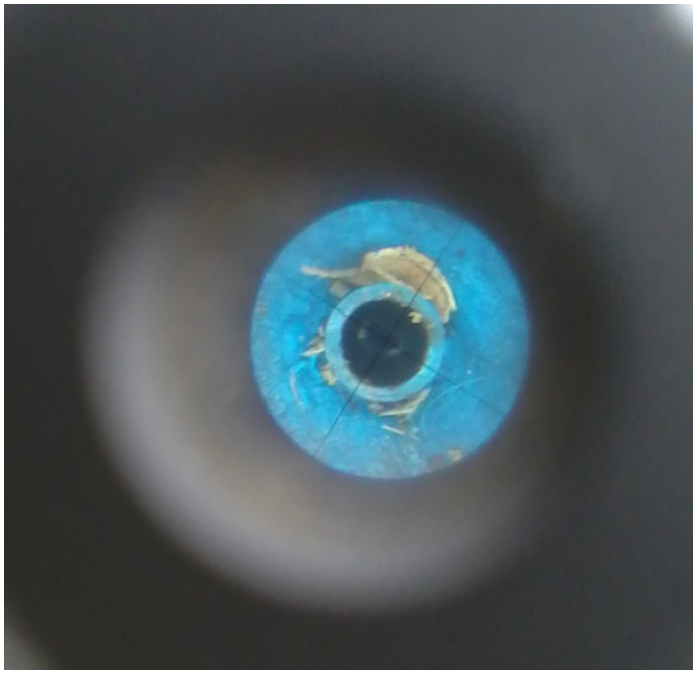


Figure 5.11. A benchmark (tube monument) seen though an optical plummet. After focusing, both the target and the crosshairs are sharp.

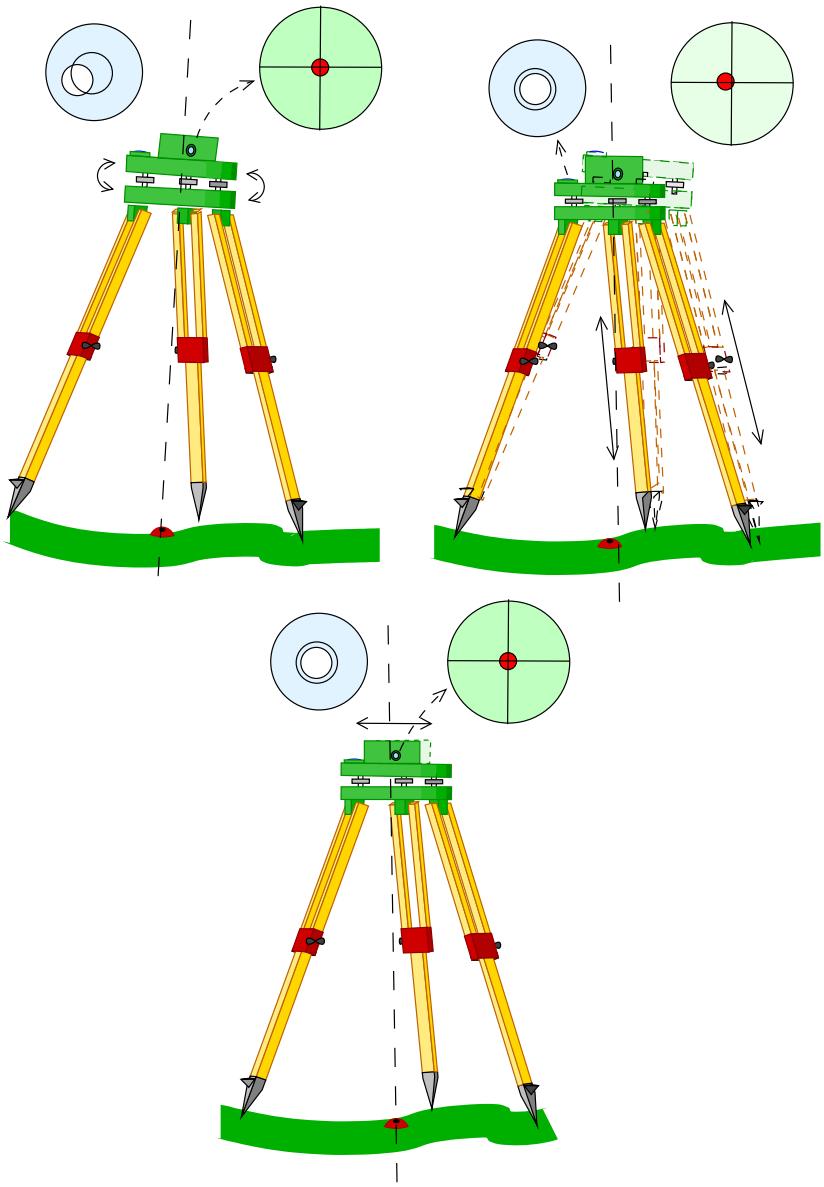


Figure 5.12. An optical plummet and a bull's-eye level are used at the same time to achieve centring and levelling.

□

forced-centring device describes a *circular movement* around the tip of the third leg, and the place of the index on the image of the point changes only little).

(c) By a parallel shift (loosen the big screw of the forced-centring device just a little) we get the index of the optical plummet on the point.

Note! Shift, *don't turn*, because then, the levelling changes!

- Because in practice both the levelling (3c) and the place of the index (3b) will nevertheless always change a little, we *repeat* procedure 3a — 3c until the desired end result is achieved. Fortunately it converges rapidly.

Figure 5.13. Problem situation.

5.4.5 Problem situations

Normally the centring and levelling procedure described above will quickly produce a satisfactory end result. In real-life field work, however, all kinds of pathological situations arise, like in figure 5.13. We leave these as an exercise for the reader.

5.4.6 Forced centring

In precise measurements, a forced-centring device or tribrach is used, i.e., instrument and signal fit onto the same device, figure 5.14.

In this way one eliminates many of the errors coming from the centring of a tripod.

In really precise measurements, like in engineering geodetic measurements, we use a *separate optical plummet*. Over short distances, centring may be the largest error source.

In the measurement, care is taken that every side is measured in both directions: in a network of three points, the procedure may be according to figure 5.15. From every instrument station, one measures to all

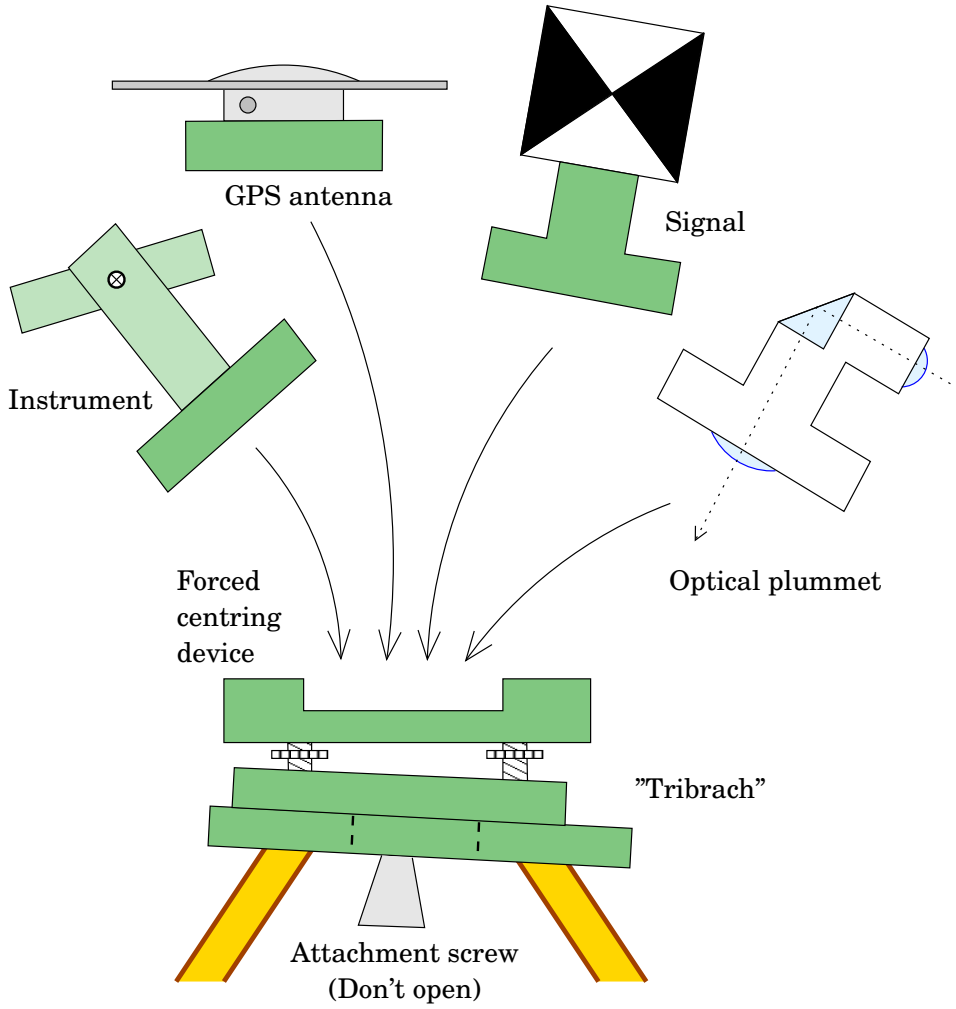


Figure 5.14. Principle of forced centring. The instrument and the signal fit into the same device.

(nearby) points equipped with signals.

5.14.Forced centring is also useful in situations where we measure on the same point with both the satellite technique (GNSS) and a terrestial instrument. The antenna is then attached, though an adapter, to the forced-centring device, figure 5.14.

5.4.7 Checking an optical plummet

The most common plummet today is the optical plummet — though the laser plummet is gaining ground[3].

The weak spot of the optical plummet is, that it must be adjusted so that it looks really straight down when the bubble of the level is in the middle — the same problem as with a levelling instrument’s sight axis and

[3]A laser plummet works practically in the same way as an optical plummet, though the light travels in the opposite direction. http://www.pobonline.com/articles/93629-back-to-basics-quick-setup-with-a-laser-plummet.

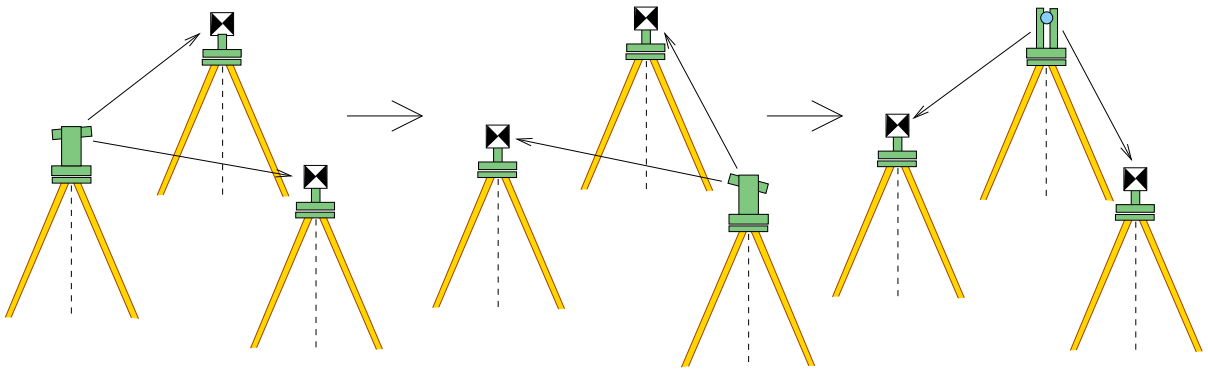


Figure 5.15. Measuring a network using forced centring.

level horizon having to be parallel, figure 4.9. This property, achieved by adjustment, can easily be lost in the handling of the device. For this reason it must be *checked* at regular intervals.

One procedure for checking is described in the book Kahmen and Faig (1988), on pages 95 – 96:

1. Mount the forced-centring device — or the whole theodolite, if the optical plummet is built in — onto a tripod, level it, and mark the point on the floor seen in the plummet's eyepiece on a paper taped to the floor.
2. The outline of the forced-centring plate is marked with pencil or chalk on the tripod head.
3. The forced-centring device is taken off the tripod, rotated 120°, and placed carefully back into the drawn outline. Level, and mark on the floor again the point under the crosshairs of the plummet's eyepiece.
4. Repeat 3.
5. If the points drawn in steps 1, 3 and 4 on the floor are identical, then the plummet is in good adjustment. If not, one shifts the crosshairs of the eyepiece, using its adjustment screws, to the centre of mass of the three points on the floor.

A non-adjusted optical plummet makes the observational material collected with it worthless. A regular check of the state of adjustment, like before every field work, is just as important as measuring the theodolite instrument height over the point!

5.4.8 Sighting and targeting

- The focusing is, like for a levelling instrument, always to be done carefully.
- By adjusting the eyepiece one focuses the crosshairs at the beginning of the observation work. If one wears glasses that are not cylindrical (only in case of near- or farsightedness), one can leave

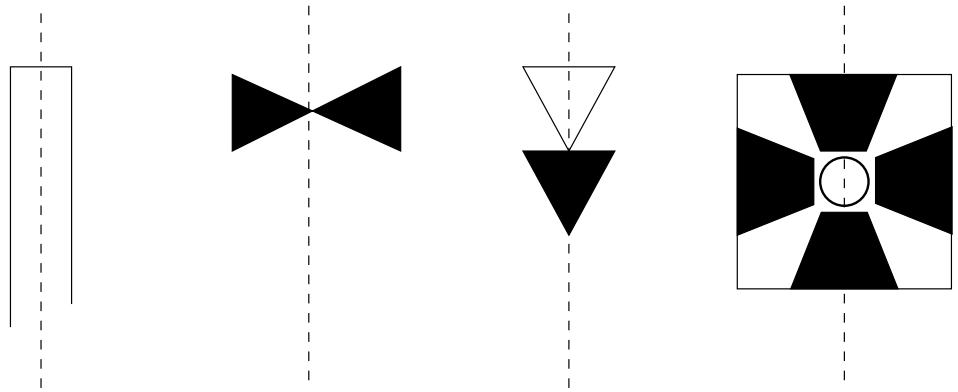


Figure 5.16. Ideal targeting situations for horizontal angles.

these off.

- For every object, one focuses carefully on the signal using the focusing screw.
- During precise measurements using the method of complete sets, one is not allowed to focus during measurement, as then, the collimation will change a little. This presupposes that all targets are sufficiently far away.

Difficult situations:

- The signal is *thin*, e.g., very far way, compared to the crosshair line. This is not a problem if the crosshairs look like figure 5.17.
- Black-and-white signals are always best.
- See figure 5.18.

5.5 Taking readings

In demanding theodolite measurements it is good practice to measure always *in both faces*, face left and face right. The instrument has two

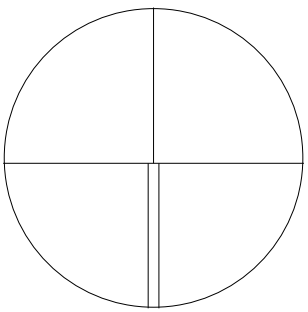


Figure 5.17. Better crosshairs. Here one may choose between a simple and a double line.

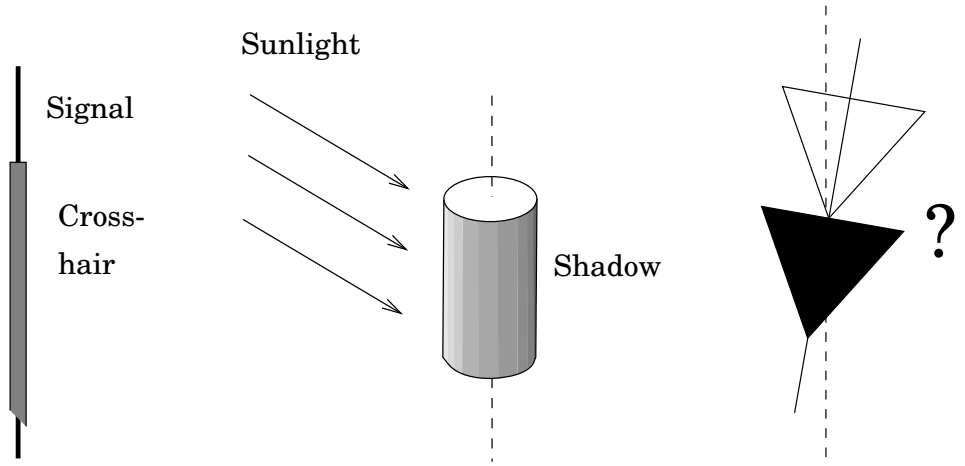


Figure 5.18. Problems caused by the signal.

face positions because it has two axes: turning it around both axes by an amount of 200 gon will make the measurement telescope point again at the same object. This redundancy[4] allows the elimination of a number of systematic errors. The first position, "face left", i.e., the vertical circle to the left of the telescope as seen by the observer, yields height angles near 100 gon if the sighting direction is close to the horizontal. The second position, "face right", yields values close to 300 gon.

The traditional method of complete sets, which further reduces instrument error, is no longer used:

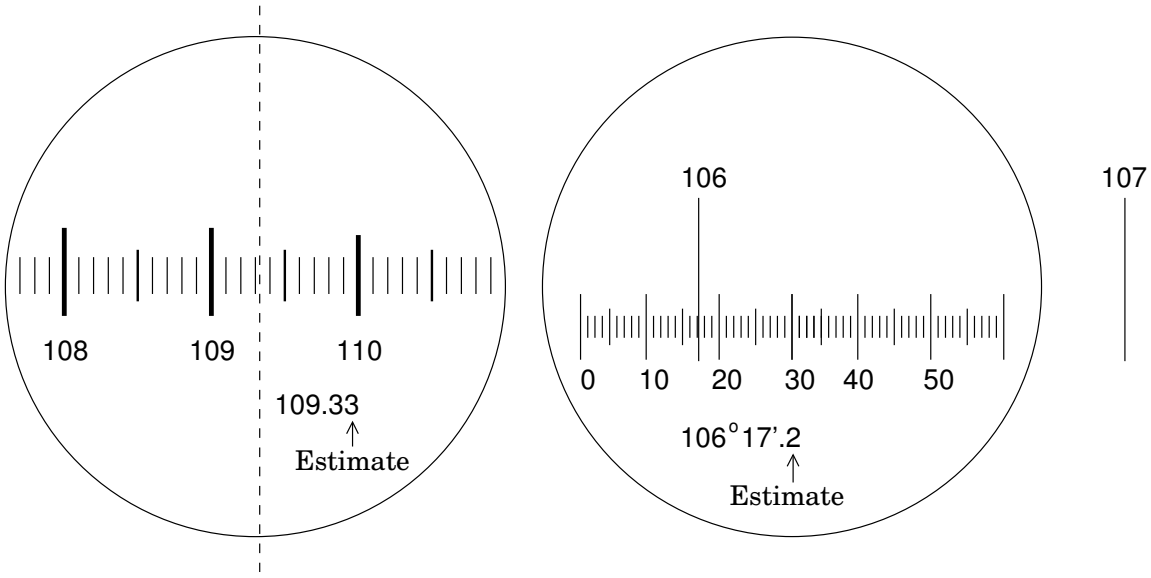
1. The measurement of first-order or national base networks has fully transitioned to using satellite positioning.
2. Digital angle measurement techniques automatically carry out a procedure equivalent to the method of complete sets, without user intervention.

5.5.1 Graduation circles and the classification of theodolites

The graduation circles of a theodolite are generally made of glass. The diameter is 60 – 100 mm, for the most precise instruments 250 mm. The interval for the main scale is 1, 0.5, 0.2 or 0.1 gon, depending on the precision class of the instrument. The traditional method draws a dense line pattern using a "graduation machine" into a layer of wax covering the glass, the lines are etched into the glass using acid, and filled with dye. This was for a long time a carefully protected Swiss business secret (Penry and Ingram, 2013).

Theodolites are classified, traditionally and somewhat inofficially, into "one-minute theodolites", "one-second theodolites", and precision theodolites. The boundaries separating these classes aren't well defined. One

[4] See for theoretical background https://en.wikipedia.org/wiki/Rotation_group_SO(3)\#Topology.



(a) Line microscope **(b)** Scale microscope

Figure 5.19. Various types of reading microscopes.

source (Simonen, 2012) gives > 1 mgon for one-minute theodolites, 0.5 – 1.0 mgon for one-second theodolites, and < 0.5 mgon for precision theodolites. See also Anon. (1971).

5.5.2 Reading device

The reading device *magnifies* the image of the scale on the graduation circle though a microscope. With it, a fine reading is *interpolated* between the numbers on the main scale.

5.5.3 Fine reading methods

Reading microscopes are divided into the following types (figure 5.19):

1. Line microscope:
   - magnification of the main scale and an index line.
   - The fine readings are estimated visually, precision ~ $1/10^{\text{th}}$ part of a graduation interval.
2. Scale microscope:
   - an additional scale, the length of which is the same as the graduation interval of the main scale: generally, main interval 1 gon, interval of the additional scale 0.01 gon, 100 graduation lines.
   - The graduation line of the main scale is used as the index of the additional scale.

The line microscope may be equipped with an *optical micrometer* in order to achieve a better reading precision, see figure 5.20.

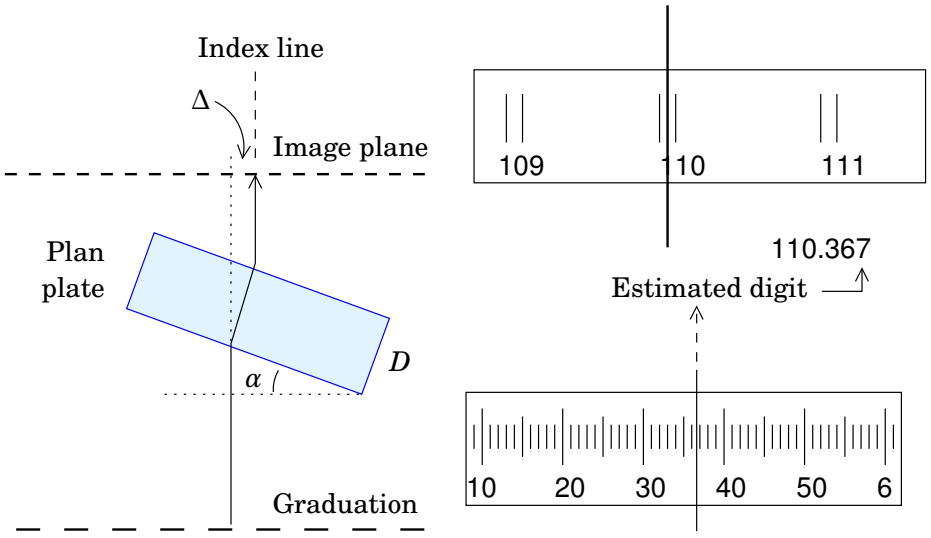


Figure 5.20. Optical micrometer and its reading.

When the glass plate is turned by an angle α, the light ray shifts in a parallel fashion by an amount

$$\Delta \approx \alpha D \left(1 - \frac{1}{n}\right),$$

where n is the index of refraction of the glass. The mechanism for turning the glass plate has a scale showing the value Δ in the angular units of the main scale.

Generally the graduation lines of the main scale are double lines, aiming at them is easy.

The *nonius*[5] or *vernier* is outdated and no longer used. See Kahmen and Faig (1988) page 65.

5.5.4 Reading the circle

1. *One* circle location is read.
2. In *precision theodolites*, simultaneously *two opposite circle locations* are read. In this way the eccentricity error of the circle is canceled out. See figure 5.22. The crude reading is taken from the circle, the fine reading from the micrometer, i.e., $244 + 0.4 + 0.0417$ (in which 7 estimated) or 244.4417.

A single circle reading is used in one-minute theodolites, whereas the method using two opposite readings, i.e., using a coincidence microscope, is found usually in one-second and precision theodolites.

[5] In Heiskanen (1943) it is told that the Portuguese Pedro Nunez (Nonius) invented a precursor to the nonius already in 1542, and the Frenchman Pierre Vernier today's nonius in 1631.

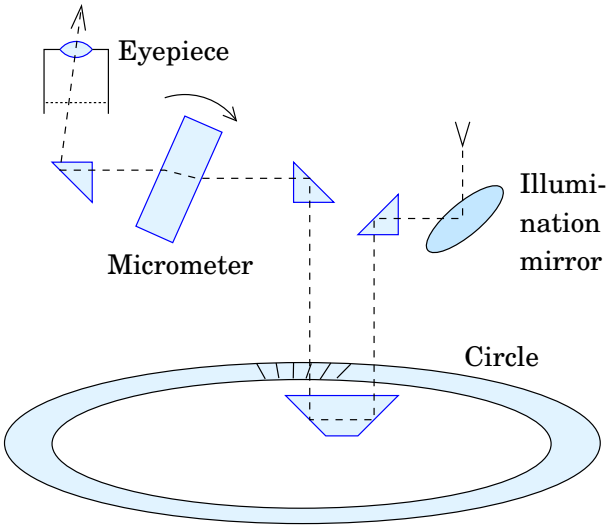


Figure 5.21. Reading the graduation circle. One circle location.

For reading the horizontal and the vertical circles, the same methods are used. Generally the diameter of the vertical circle is less than that of the horizontal circle, so that the reading precision is correspondingly less.

The reading devices include furthermore:

1. a *reading microscope*, to which all the readings of the optics are guided. In the field of view of the microscope, usually both horizontal and vertical readings can be seen. One has to be careful to choose the correct and wanted numbers. Often, the vertical reading is marked with the letter V and the horizontal reading with H or Hz.
2. An illumination system:
   - a mirror that can be turned to guide light to the reading optics
   - a light source with battery to be mounted in the place of the mirror or in its own fitting.

More about optical theodolites in Tikka (1991, pages 21 – 35).

5.6 Electronic theodolites

In *electronic* theodolites all measurements are obtained in *numerical* form, facilitating the automatic storage, correctness check already in connection with the measurements, and forwarding. The monetary savings in the measurement activity can be substantial, not just in measurement time saved, but also in *quality gain*, when the error prone reading and manual writing down of observations goes away.

However, though electronic theodolites record the observations on their own — sometimes “raw data” isn’t even recorded but already prepro-

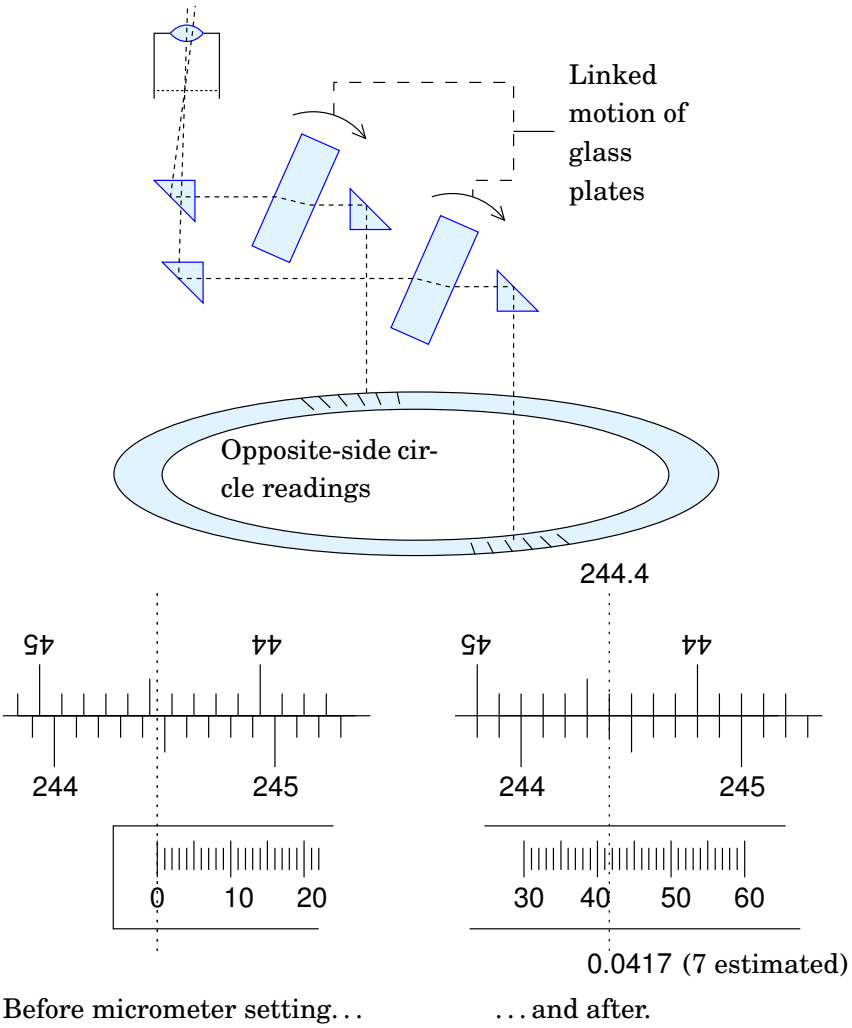


Figure 5.22. Reading the graduation circle. Two opposite circle locations.

cessed data —, the information to go with the data ("metadata") must still be carefully recorded.

In electronic theodolites, the edges of the graduation circles have been imprinted with different line patterns.

5.6.1 Absolute encoding circles

Often used is the so-called Gray code[6], which is made up of bit strings in which, at every step, only one bit changes. In figure 5.23 we see an example of a four-bit Gray code, which has 16 different values.

In realistic applications, more bits are used, e.g., if one wants a horizontal-angle resolution of 10^{-4} gon, there must be $400 \cdot 10^4 = 4 \cdot 10^6$ different values. This requires already 22 bits[7].

[6] Frank Gray (1887 – 1969) was an American physicist and electronicist, a developer of television technology. The Gray code is part of a patent awarded to him in 1953.

[7] ... because $2^{22} = 4194304 \gtrsim 4000000$.

15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0
0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0
0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0

Figure 5.23. The Gray code.

The code is a black-and-white pattern imprinted on the circle, which a row of light sensitive diodes, photodiodes, scans, figure 5.24 left.

The advantage of the Gray code is, that one knows all the time unambiguously, which place on the circle one is reading, i.e., there is a *marked zero direction* on the circle. This is why the encoding is called absolute. In the example case, four photodiode rows side by side are needed, one for each bit field.

The patterns on the circles are observed electro-optically, e.g., using a row of photodiodes, nowadays an integrated microelectronic circuit, like a CCD sensor. The light detector observes the edges between the black and white fields on the circle. Crude, absolute values originate from the edges of the pattern on the circle, more decimals are obtained from the interpolating property of the diode array — or imaging light sensor.

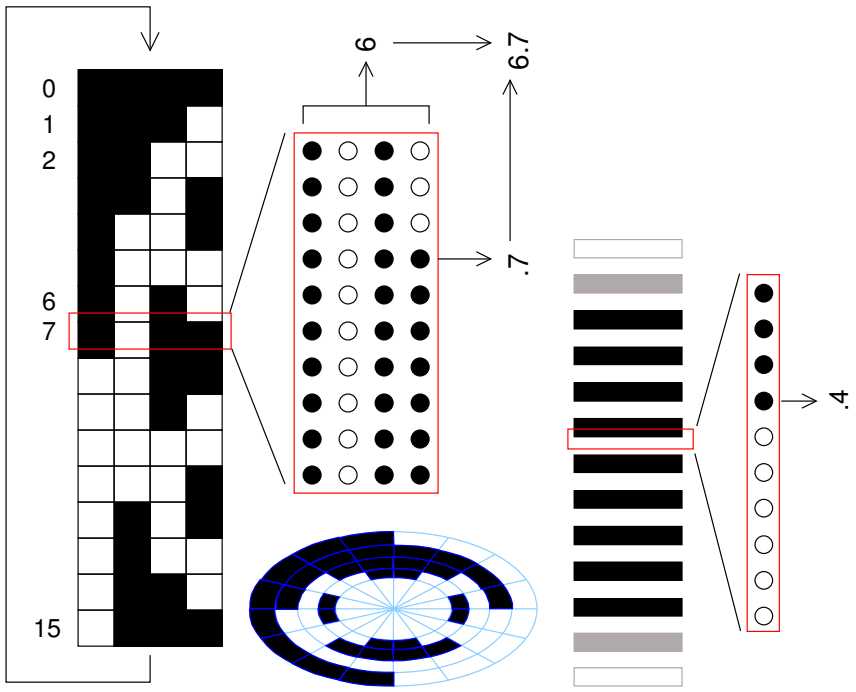


Figure 5.24. On the left, an absolute, on the right, an incremental encoding circle.

5.6.2 Incremental encoding circles

There is *no* marked zero direction on the circle. Here, we may only follow *changes* in instrument orientation, by counting lines, figure 5.24 right. In the incremental solution, at least two diode detectors are needed to establish the turning direction, in the depicted example there are ten. The incremental method of course can only measure direction *differences*, i.e., angles. Crude readings come from counting lines on the circle, more decimals are again obtained by interpolation.

5.6.3 Modern automatic instruments

In electronic theodolites, commonly a so-called ATR, Automatic Target Recognition, system is used. For this purpose, a CCD image sensor has been placed in the image plane of the measuring telescope. The telescope is aimed visually at the object. The uncorrected reading comes, as usual, from the digital encoding circle system. To the reading is added programmatically, by means of image processing, a *correction* from the CCD image.

With electronically readable circles one uses many read-out detectors of which each reads the code from a different place on the circle. All readings are collected in a microprocessor, which averages them. In the average, the periodic and random errors of the circle have been eliminated or essentially reduced. Compare this with the method of complete sets used with optical theodolites, section 5.8 on page 133.

With electronic theodolites, locking and unlocking the horizontal circle, and turning it — and thus applying the method of complete sets by hand — is unnecessary and impossible. Turning over the telescope and measuring in both faces is however still necessary, for the same reasons it is with optical instruments (section 5.5 on page 122).

In some instruments, a *rapidly spinning circle* is used. In this dynamic solution, the light detectors transform the line pattern on the circle into a block signal (figure 5.26). The time shift between two block signals together with the circle's spinning velocity ω gives directly the angle $\alpha(t)$ between the detectors as a function of time — plus an unknown constant — the "ambiguity[8]" — that does not depend on time. For this reason, also the spinning circle method is *incremental*: it can only measure *changes* in the angle.

kokonaisluku-tuntematon

The advantages of the method are:

1. It converts angle measurement into electronic time difference measurement, which can be extremely precise.
2. The precision with which the graduation lines on the circle are produced is not critical, the even spinning speed of the circle is.

[8] Obviously because to the detector, every line on the circle looks like every other.

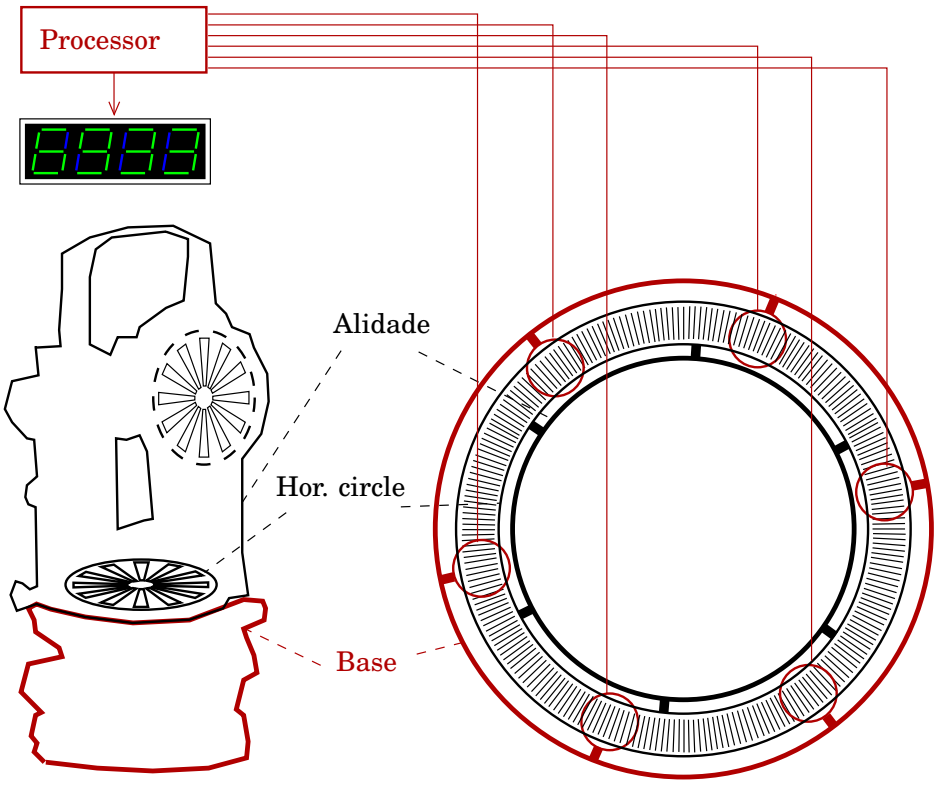


Figure 5.25. Electronic readout of the horizontal circle: readings are taken from all parts of the circle. This replaces the method of complete sets known from optical theodolites.

The range of measurement systems and technical solutions for electronic theodolites is broad and rapidly developing. Therefore we only take a closer look at one case.

5.7 Case: Leica robotic tacheometer TCA2003

TCA2003 (Leica, 1997) is a good example of a fairly modern electronic theodolite or *tacheometer*. The instrument measures horizontal and zenith angles as well as slant ranges. The calculation capabilities of its built-in software are quite versatile.

Measurement precision (as stated by the manufacturer) is $0''.5$, i.e., 0.15 mgon, in horizontal and vertical angle measurement, and 1 mm + 1 ppm in distance measurement. The longest measurement distance is 3.5 km under normal conditions.

The instrument has co-axial ATR (Automatic Target Recognition, subsection 5.6.3). The system also takes part in the angle measurements. The CCD sensor measures the deviations in the horizontal and vertical directions of the laser beam reflected by the prism, and guides the motors so, that the crosshairs are placed almost over the prism. The small remaining deviation is measured in the CCD image and the angle

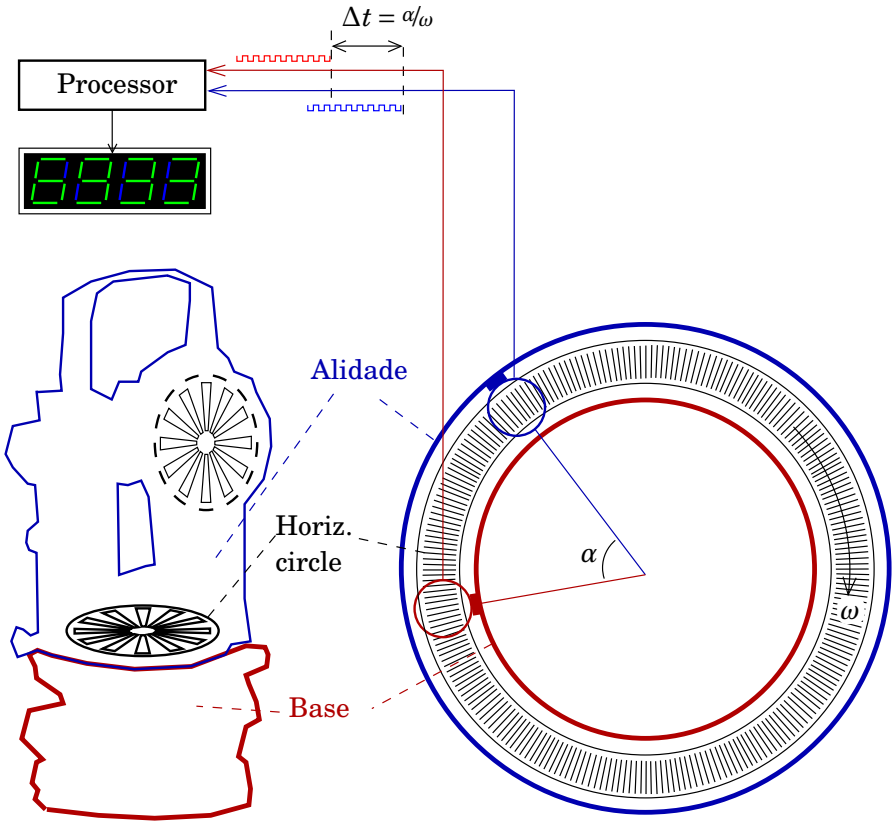


Figure 5.26. A spinning circle converts the measurement of angles into one of time differences.

readings are corrected correspondingly.

ATR can also be programmed to follow a moving target, or to systematically seek (scan) a target if it is not in the expected place.

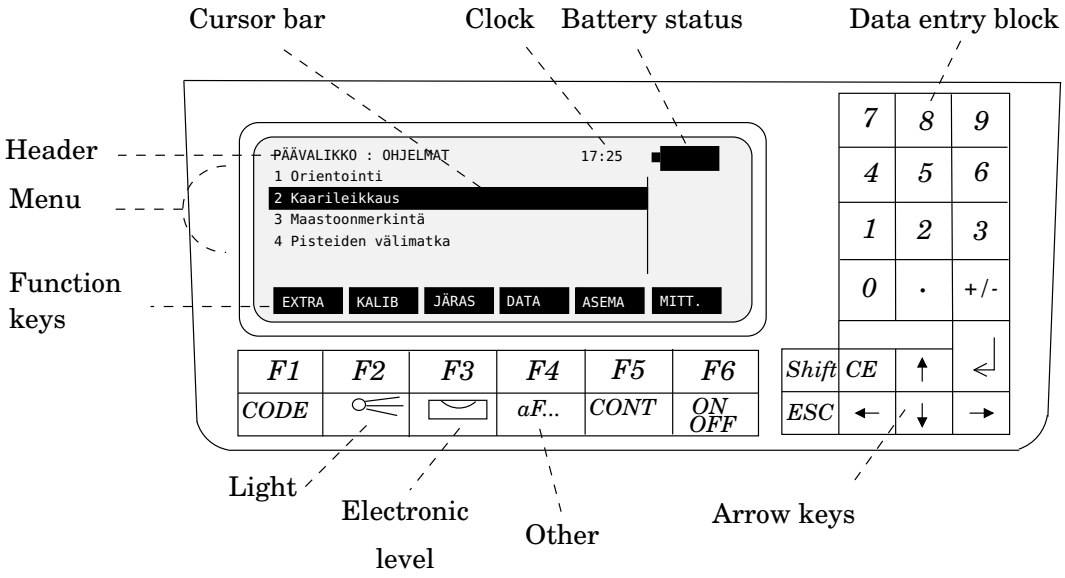


Figure 5.27. Leica TCA2003 control panel (in Finnish).

Figure 5.28. Leica TCA2003 theodolite.

The instrument contains *self-calibration facilities*, a measurement programme for determining the vertical circle index error (subsection 5.9.3), the collimation error (sight axis not perpendicular to horizontal axis, section 5.9), the trunnion axis tilt (horizontal and vertical axes not perpendicular to each other) as well as the so-called *zero point* or *constant error* of the distance measurement device or range finder (subsection 6.4.4).

Many "corrections", rather, *reductions*, of distance measurement, like weather corrections, reduction to the horizontal plane, and even the map projection reduction (e.g., Gauss-Krüger or UTM) can be done already in the instrument.

The instrument has a laser plummet, which works in the same way as an optical plummet — also the precision is similar, a bit better than ±1 mm —, but the light moves in the other direction, from the theodolite down to the ground. Levelling and centring are otherwise the same as with an optical plummet.

The instrument has, in addition to a bull's-eye level, also an accurate electronic level. The display is LCD and looks the same as a real bull's-eye level. It is self calibrating, and turning the instrument by 100 gon or 200 gon every time it is levelled (figure 5.8) is not needed.

The instrument is highly automated: on offer is a pre-programmed monitoring measurement of 50 points. This is useful especially in deforma-

tion measurements in industry and construction. On offer are even two programming languages: GSI for simple use, and GeoCOM for advanced use. Additionally there is the GeoBasic environment, which allows the development of more applications in the PC environment, and their upload to the instrument.

The exchange of data between the instrument and a computer can be done in two ways:

1. through a serial interface (RS232). This technique is obsolete, modern devices use a USB port or Bluetooth
2. using a non-volatile memory card, a so-called PCMCIA card. Storage capacity may be 512 kB – 4 MB. The format of the card is the MS-DOS file system FAT. Also this solution has been obsolesced by USB memory sticks.

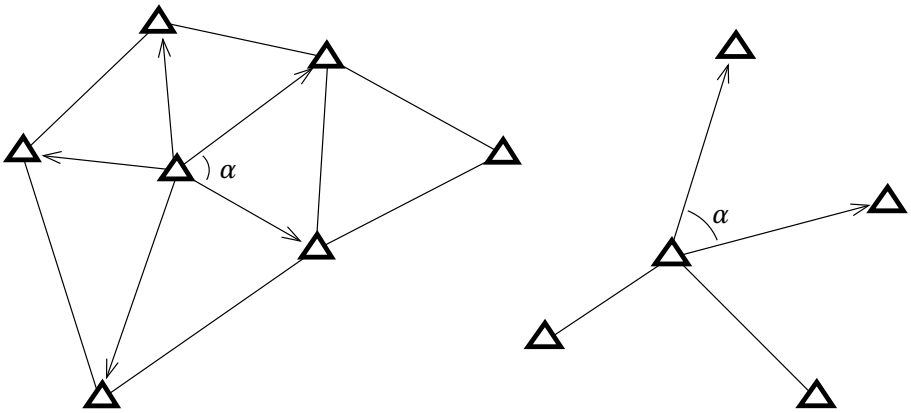
The format of the observations themselves is the Leica designed GSI (Geo Serial Interface), documented in the manual.

5.8 Horizontal angle measurement

In traditional topographic surveying, horizontal angles are observed for three reasons:

1. As a part of *triangulation*, where horizontal angles and distances, or side lengths, are measured between points of the triangle network. The triangle sides may be as long as tens of kilometres. From the observations, the geodetic co-ordinates of the triangle points are computed, hierarchically in densification stages of the network. In this way, the first-order triangulation network measured by the Finnish Geodetic Institute has been densified by the Finnish National Land Survey.
2. As a further densification of the triangle network, *traverse measurements* are carried out. In this method, side lengths and bending angles are measured between successive points of a traverse (figure 5.29c) in order to determine traverse point co-ordinates. To make calculation possible, the starting and closing points of the traverse need to be known, e.g., from a higher-order triangulation.
3. From the points of a traverse, e.g., *radial surveys* may be conducted (figure 7.13).

Many of these observation techniques have nowadays been replaced by GNSS measurement. Triangulation is no longer done anywhere, while traverse measurements are done only in situations where GNSS is not usable, like underground (tunnels) or high-rise urban landscapes.



(a) Triangulation **(b)** Radial survey

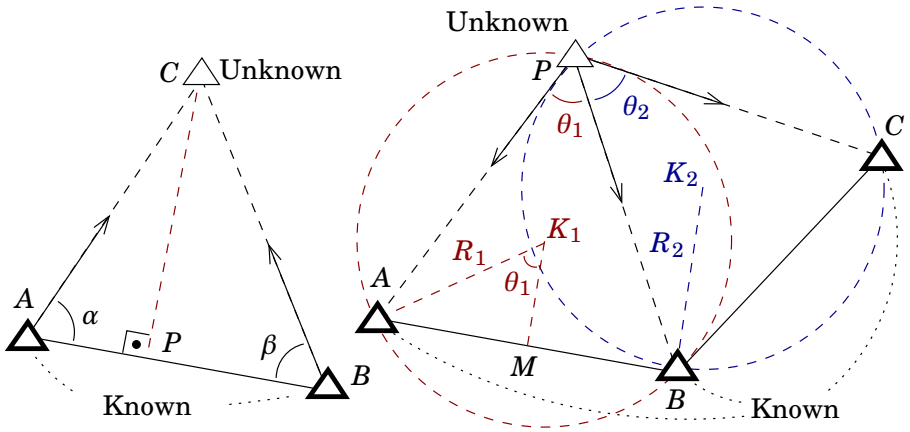
(c) Traversing

Figure 5.29. Use situations for horizontal angle measurement.

5.8.1 Intersection and resection

Often used geometries are *intersection* and *resection*, figures 5.30. In Finnish *eteen- ja taaksepäin leikkaus*, in Swedish *avskärning och inskärning*, in Dutch *voorwaartse en achterwaartse insnijding*, in German *Vorwärtsschnitt und Rückwärtsschnitt*.

Intersection works as follows: let the points A and B have co-ordinates (x_A, y_A) and (x_B, y_B). Let the projection point between these two points of the unknown point C be P, and its co-ordinates (x_P, y_P).



(a) Intersection **(b)** Resection

Figure 5.30. Intersection and resection.

Then

$$PC = AP\tan\alpha = PB\tan\beta,$$

i.e.,

$$AB = AP + PB = PC\cot\alpha + PC\cot\beta = PC\left(\cot\alpha + \cot\beta\right).$$

From this

$$PC = \frac{AB}{\cot\alpha + \cot\beta}$$

and

$$AP = \frac{AB\cot\alpha}{\cot\alpha + \cot\beta}, \quad PB = \frac{AB\cot\beta}{\cot\alpha + \cot\beta}.$$

Now we use these distances, i.e., the coefficients $\cot\alpha$, $\cot\beta$, as *weights* in calculating the co-ordinates of point P as weighted averages of the co-ordinates of points A and B:

$$x_P = \frac{x_A\cot\beta + x_B\cot\alpha}{\cot\alpha + \cot\beta}, \quad y_P = \frac{y_A\cot\beta + y_B\cot\alpha}{\cot\alpha + \cot\beta}.$$

After this, the direct calculation formulas for the co-ordinates of intersection point C are:

$$x_C = x_P + \tan\alpha\,(y_P - y_A) = x_P + \frac{y_B - y_A}{\cot\alpha + \cot\beta},$$
$$y_C = y_P - \tan\alpha\,(x_P - x_A) = y_P - \frac{x_B - x_A}{\cot\alpha + \cot\beta}.$$

Resection is inherently more difficult, because it is an *inverse problem*. The problem was studied already by Snellius and Laurent Pothenot[9] and is often named after them.

Let (figure 5.30b) M be the midpoint of points A and B. If from the unknown point P has been measured the angle θ_1 between points A and B, then this point must be on the circle that passes through the points A and B, and seen from the centre point of which, K_1, the angle between points A and B is $2\theta_1$.

The line segment MK_1 stands perpendicularly on AB: the co-ordinates (x_M, y_M) of the midpoint M are calculated using the formulas derived above for the special case $\alpha = \beta$. The co-ordinates of point K_1 are calculated in the following way — remember that in the role of α we now have $90° - \theta_1$:

$$x_{K_1} = x_M + \frac{1}{2}\cot\theta_1\,(y_B - y_A) = \frac{1}{2}(x_A + x_B) + \frac{1}{2}\cot\theta_1\,(y_B - y_A),$$
$$y_{K_1} = y_M - \frac{1}{2}\cot\theta_1\,(x_B - x_A) = \frac{1}{2}(y_A + y_B) - \frac{1}{2}\cot\theta_1\,(x_B - x_A).$$

[9] Laurent Pothenot (1650 – 1732) was a French mathematician and geodesist.

The radius of the circle is obtained by the Pythagoras theorem:

$$R_1 = \frac{1}{2\sin\theta_1}\sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}.$$

Now, because there is also another point pair B, C, through which runs a circle of its own, with centre point K_2 and radius R_2, we have a pair of quadratic equations describing two circles:

$$\begin{aligned}(x - x_{K_1})^2 + (y - y_{K_1})^2 &= R_1^2,\\ (x - x_{K_2})^2 + (y - y_{K_2})^2 &= R_2^2.\end{aligned}$$

From this, the co-ordinates (x, y) of the unknown resection point can be solved — at least in principle. There are many ways of solving this, e.g., linearization with respect to a pair of approximate values (x_0, y_0) with an iterative solution.

Note that two circles clearly have *two intersection points*, from which the right solution must be chosen, and it is not point B...

Singularity: The precision of resection depends on the geometry of the points. If the points A, B, C and the unknown point lie on the same circle, the solution is even impossible: we speak of a *singularity*. Then, the blue and the red circle are identical, and any point on that circle serves as a solution. We also speak of the *dangerous circle*: already close to this geometry, precision deteriorates ominously.

Note also, that the situation would be conceptually the same if there had been measured *two distances* $R_1 = K_1P$ and $R_2 = K_2P$ from points K_1 and K_2. Here, P is the unknown point, co-ordinates (x, y). When using distance measurement, we do not speak of intersection and resection, but they are in a way the same thing.

5.8.2 Observation method of complete sets

The graduation of the horizontal circle of an optical theodolite has been manufactured with great care, but is always imprecise. Every gon of the graduation should be the same size, precisely one gon, but in reality is not. However, the sum of all the gons of the graduation should always be exactly 400 gon. This is why the effect of graduation errors can be diminished by measuring the same angle *using different sectors of the horizontal circle*, and taking the average of those. This principle is implemented in the so-called *observation method of complete sets*.

sarjahavaintomenetelmä

In this method, traditionally used in the measurement of base networks, the same fan of directions is measured multiple times — complete sets — in this way, that between the sets, the horizontal circle is *unlocked*[10]

[10] The locking screw of the horizontal circle is protected by a lid. That way the circle isn't turned by accident during measurement.

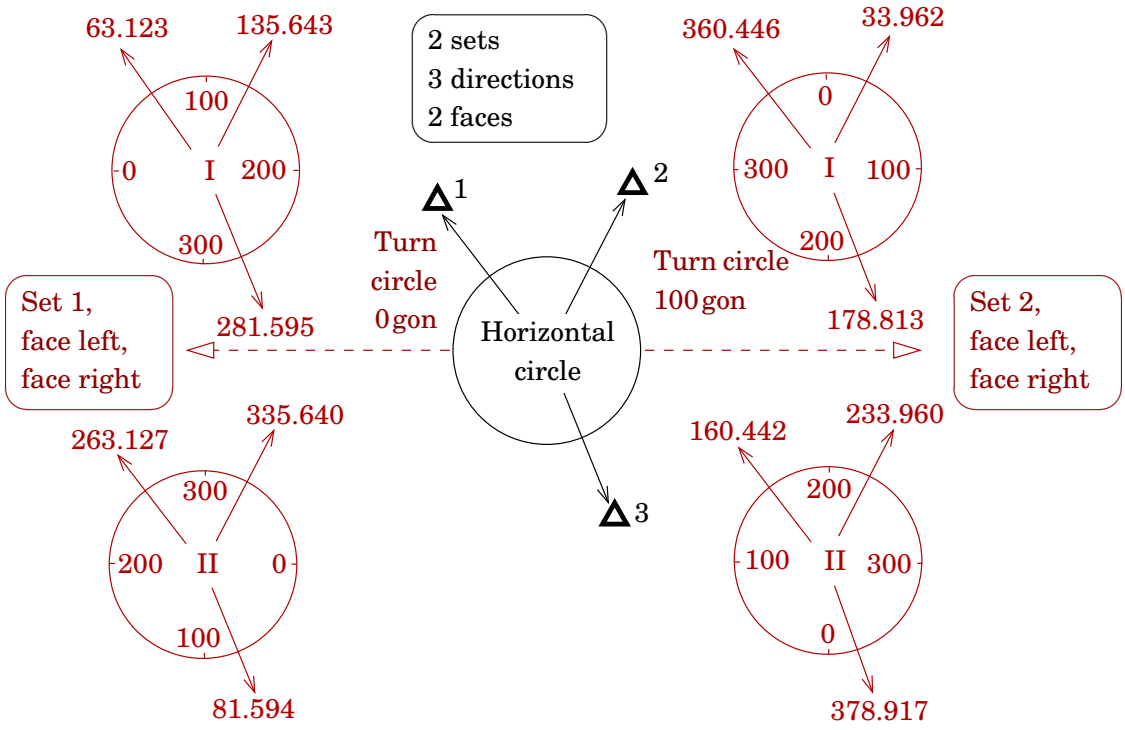


Figure 5.31. Observation method of complete sets. In this case a measurement of two complete sets of three directions is made in both instrument faces.

□

and *turned* (by an amount $\frac{200\,\text{gon}}{n}$, in which n is the number of complete sets). In this way one makes use of the different sectors of the horizontal circle, and systematic errors, e.g., due to graduation errors of the horizontal circle, are minimized as one takes the average of the sets.

In this observation method, also always the instrument's telescope is turned over — both axes by 200 gon — i.e., one measures in both faces, face left and face right.

The following errors are eliminated completely by the method of complete sets: collimation error and trunnion axis tilt, axes intersection error, eccentricity of horizontal circle and measuring telescope. The impact of horizontal circle graduation errors is essentially diminished. Levelling or centring errors are *not* eliminated! See Tikka (1991), pages 38 – 40.

More technical remarks:

- If, in the method of complete sets, the sets do not agree with each other, they must *all* be measured again. One is not allowed to reject only a single set, because that would distort the statistical properties of the observational material.
- An individual sighting direction may be rejected from all sets.

One does not apply the method of complete sets with an electronic theodo-

lite — or more precisely, the instrument does so itself internally, see section 5.6.

5.8.3 Station adjustment of horizontal angles

Station adjustment, the merging of the several complete sets of horizontal direction measurements into one optimal solution or measurement set, is an *adjustment problem*, though a simple one. Typically it was solved already in the observation notebook.

Modern electronic tacheometers or total stations don't need station adjustment as they don't collect multiple observation sets as described in section 5.8.2. We describe it here nevertheless shortly for historical and methodological interest. There will be a more thorough discussion of adjustment calculus in chapters 13 and 14.

Here we describe a calculation example in which four complete sets to three sighting directions have been measured. The two instrument faces for each series, face left and face right, have already been merged.

In this example, the number of complete sets is $s = 4$, the number of directions $r = 3$, so the number of observations is $n = rs = 12$. The number of unknown directions in the complete-set average is $r - 1 = 2$. In each complete set there is one orientation unknown[11]. So, the *number of excess observations* (redundancy, number of degrees of freedom) is

$$rs - (r - 1 + s) = (r - 1)(s - 1).$$

The observation equations are[12]

$$\underline{t}_{ij} + \underline{v}_{ij} = \widehat{\alpha}_j - \widehat{\Omega}_i,$$

in which $i = 1,\ldots,4$ — generally $i = 1,\ldots,s$ — is the number of the complete set, $j = 1\ldots3$ — generally $j = 1,\ldots,r$ — is the number of the sighting direction, Ω_i the orientation unknown of the horizontal circle for the complete set, α_j the direction unknown of the target (the azimuth of the measured direction), and $\underline{t}_{ij}$ the (raw) direction reading. Additionally one must fix one direction, or linear combination of directions, so that the whole problem can be solved (from relative angle measurements one

[11]The situation is somewhat similar to that in GPS measurement. Also there, all observations from every receiver contain a clock error unknown ΔT, and that is why the observable is called pseudo-range and not range. Similarly we could here talk about "pseudo-direction observations". These uninteresting additional unknowns are called *nuisance parameters*.

[12]So the design matrix is, for the case of four complete sets and three sighting

cannot solve absolute directions). I.e., additionally is required, e.g.[13]:

$$\alpha_1 = 0. \tag{5.1}$$

In the following we present, without proof, how one may compute the result of a station adjustment, and how this can be arranged as a simple calculation template. The procedure is statistically optimal.

5.8.4 Angle transformation

First, we carry out the *angle transformation*: we subtract from every observation $\underline{t}_{ij}$ except the first of every complete set, the first observation of that set $\underline{t}_{i1}$:

$$\underline{t}'_{ij} \stackrel{\text{def}}{=} \underline{t}_{ij} - \underline{t}_{i1}, j = 2, \ldots, r.$$

Thus we also implicitly redefine the unknowns:

$$\widehat{T}_j \stackrel{\text{def}}{=} \widehat{\alpha}_j - \widehat{\alpha}_1, j = 2, \ldots, r.$$

These new unknowns are now in relation to the first direction, they are thus *angle transformed* direction unknowns.

---

directions:

$$A = \left[\begin{array}{ccc|ccc} 1 & & & -1 & & \\ & 1 & & -1 & & \\ & & 1 & -1 & & \\ \hline 1 & & & & -1 & \\ & 1 & & & -1 & \\ & & 1 & & -1 & \\ \hline 1 & & & & & -1 \\ & 1 & & & & -1 \\ & & 1 & & & -1 \end{array}\right],$$

if the vector of observations is

$$\underline{\ell} = \left[\begin{array}{ccc|ccc|ccc|ccc} \underline{t}_{11} & \underline{t}_{21} & \underline{t}_{31} & \underline{t}_{12} & \underline{t}_{22} & \underline{t}_{32} & \underline{t}_{13} & \underline{t}_{23} & \underline{t}_{33} & \underline{t}_{14} & \underline{t}_{24} & \underline{t}_{34} \end{array}\right]^{\mathsf{T}}$$

and the vector of unknowns $\widehat{\mathrm{x}} = \left[\begin{array}{ccc|ccc} \widehat{\alpha}_i & \widehat{\alpha}_2 & \widehat{\alpha}_3 & \widehat{\Omega}_1 & \widehat{\Omega}_2 & \widehat{\Omega}_3 & \widehat{\Omega}_4 \end{array}\right]^{\mathsf{T}}$, bringing the observation equations into the form

$$\underline{\ell} + \underline{\mathrm{v}} = A\widehat{\mathrm{x}},$$

with $\underline{\mathrm{v}}$ the vector of residuals.

We use here the standard "hat" notation from statistics, as well as underscoring stochastic quantities as customary.

[13] Completely arbitrary. Just as well $\alpha_2 = 0$, $\alpha_3 = 0$, or, e.g., $\alpha_1 + \alpha_2 + \alpha_3 = 0$ which would at least be "democratic".

The solution is

$$\widehat{T_j} = \frac{1}{s}\sum_{i=1}^{s} \underline{t}'_{ij},$$

The adjusted angle transformation is the average over complete sets of the angle transformations.

Furthermore we may write, arbitrarily, due to the arbitrariness of choosing the first direction,

$$\widehat{T_1} = \frac{1}{s}\sum_{i=1}^{s} \underline{t}'_{i1} = \frac{1}{s}\sum_{i=1}^{s} \left(\underline{t}_{i1} - \underline{t}_{i1}\right) = 0.$$

The T values are adjusted, but still relative directions, not absolute azimuths.

Now we obtain as the solution for the orientation unknowns:

$$\widehat{\Omega_i} = -\frac{1}{r}\sum_{j=1}^{r} \left(\underline{t}_{ij} - \widehat{T_j}\right).$$

5.8.5 Residuals and degrees of freedom

Now that the unknowns $\widehat{T_j}$ and $\widehat{\Omega_i}$ have been calculated, or more precisely, *estimated*, with the least-squares method, we can also calculate the residuals:

$$\underline{v}_{ij} = \widehat{T_j} - \widehat{\Omega_i} - \underline{t}_{ij}.$$

Let the mean error σ of the direction observations $\underline{t}_{ij}$ be the same for all, and let the direction observations be uncorrelated. In that case the *variance matrix* of the observations is

$$\Sigma_{\ell\ell} = \mathrm{Var}\{\ell\} = \sigma^2 I_{s\times r},$$

in which $I_{s\times r}$ is the unit matrix of size $(s\times r)\times(s\times r)$.

In this case, the quantity ("shifting variate", Baarda)

$$\underline{\mathscr{E}} = \underline{\mathrm{v}}^{\mathsf{T}}\Sigma_{\ell\ell}^{-1}\underline{\mathrm{v}} = \frac{1}{\sigma^2}\sum_{i=1}^{s}\sum_{j=1}^{r} \underline{v}_{ij}^2 = \frac{1}{\sigma^2}\sum_{i,j} \underline{v}_{ij}^2.$$

is distributed according to the so-called χ_b^2 distribution, the expectancy of which is b, the number of *degrees of freedom*, or "excess" observations. See section 14.4.

Earlier we saw that $b = (r-1)(s-1)$, the number of angle transformations (independent direction measurements) multiplied by the number of *excess* complete sets. Thus, $\underline{\mathscr{E}}$ is distributed according to $\chi^2_{(r-1)(s-1)}$ and its expectancy is $(r-1)(s-1)$. So:

$$E\left\{\frac{1}{\sigma^2}\sum_{i,j}\underline{v}_{ij}^2\right\} = (r-1)(s-1) \Longrightarrow \sigma^2 = E\left\{\frac{\sum_{i,j}\underline{v}_{ij}^2}{(r-1)(s-1)}\right\},$$

and we see that

$$\widehat{\sigma^2} \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{s}\sum_{j=1}^{r} \underline{v}_{ij}^2}{(r-1)(s-1)}$$

is the unbiased *estimator* of the variance σ^2 of a single observation:

$$E\left\{\widehat{\sigma^2}\right\} = \sigma^2.$$

For the quantity σ we use the name "*mean error of unit weight*". It represents the uncertainty, the mean error, of a "typical observation", in this case, a single, raw direction reading.

We shall return to this subject in chapter 13.

5.8.6 The calculation table for station adjustment

A station adjustment consists of the following steps:

1. We fix the arbitrariness in the definition of the directions, e.g., by fixing the first direction to zero, $T_1 = 0$; this is called the *angle transformation*. We do it by subtracting the first observed direction reading from the other directions in the same complete set.
2. We compute the *average over all complete sets* ("complete-set average") $\frac{1}{s}\sum_{i=1}^{s} \underline{t}_{ij}$ for every sighting direction j;
3. We compute the residuals of the observations relative to the *complete-set averages*

$$\underline{v}'_{ij} = \underline{t}_{ij} - \frac{1}{s}\sum_{j=1}^{s} \underline{t}_{ij};$$

4. We compute the sums of the "first residuals" $\underline{v}'_{ij}$ over all directions within every complete set:

$$\frac{1}{r}\sum_{i=1}^{r} \underline{v}'_{ij} = \frac{1}{r}\sum_{i=1}^{r}\left(\underline{t}_{ij} - \frac{1}{s}\sum_{j=1}^{s} \underline{t}_{ij}\right);$$

5. We compute the final residuals $\underline{v}_{ij} = \underline{v}'_{ij} - \frac{1}{r}\sum_{i=1}^{r} \underline{v}'_{ij}$, which corresponds precisely to those residuals we would obtain *if* we would give each complete set its own orientation unknown $\widehat{\Omega}_j$ in addition to the direction unknowns $\widehat{\alpha}_i$, like in the original system of observation equations.

The end result is the *calculation template* presented in table 5.1, which contains these operations and ought to be self evident.

After filling out the template (in the order: average over complete sets, v', sum per series, v), the mean error of a single direction (i.e., the unit

Table 5.1. Computing table for station adjustment.

	Angle transformation (gon)	v' (cc)	v (cc)
1. Complete set	0.0000	0	-3,7
	68.8430	+4	+0,3
	209.1880	+7	+3,3
Complete set sum, average		+11	+3,7
2. Complete set	0.0000	0	-1,3
	68.8425	-1	-2,3
	209.1878	+5	+3,7
Complete set sum, average		+4	+1,3
3. Complete set	0.0000	0	+2,7
	68.8424	-2	+0,7
	209.1868	-5	-2,3
Complete set sum, average		-7	-2,7
4. Complete set	0.0000	0	+3,7
	68.8422	-4	-0,3
	209.1866	-7	-3,3
Complete set sum, average		-11	-3.7
Average over complete sets	0.0000		
("complete-set average")	68.8426		
	209.1873		

weight) is calculated by first calculating the quadratic sum of the residuals over all series and directions:

$$\underline{v}^{\mathsf{T}}\underline{v} = \sum_{i=1}^{s}\sum_{j=1}^{r}\underline{v}_{ij}^2 = [vv]$$

(the latter notation was invented by Gauss and is sometimes still used). Then, we calculate (i.e., *estimate*[14])

$$\widehat{\sigma} = \sqrt{\widehat{\sigma^2}} = \sqrt{\frac{\underline{v}^{\mathsf{T}}\underline{v}}{rs-(r+s)+1}} = \sqrt{\frac{\underline{v}^{\mathsf{T}}\underline{v}}{(r-1)(s-1)}} = 3,7^{\mathrm{cc}}.$$

Here:

[14]Note that $E\left\{\widehat{\sigma^2}\right\} = \sigma^2$, but however $E\left\{\widehat{\sigma}\right\} \neq \sigma$! The reason for this is, that the E operator and the square root may not be interchanged: in general, $E\left\{\sqrt{\underline{x}}\right\} \neq \sqrt{E\left\{\underline{x}\right\}}$.

Counter example: if the probability distribution of $\underline{x}$ is $p(1) = 0.5, p(4) = 0.5$, and elsewhere $p(x) = 0$, we obtain $E\left\{\sqrt{\underline{x}}\right\} = 1.5$ but $\sqrt{E\left\{\underline{x}\right\}} = \sqrt{2.5} \approx 1.58$. So $\widehat{\sigma}$ is not unbiased although $\widehat{\sigma^2}$ is.

- rs the number of observations (r directions in s complete sets)
- $r+s$ the number of unknowns (r direction unknowns α_i, s orientation unknowns Ω_j)
- The number $+1$ represents the *defect* of the problem, the dimension of the solution space: the circumstance mentioned above, that adding an amount Δ to both the α_i and the Ω_j does not affect the observations in any way that can be observed, and that thus, the solution is not unique without *one* extra condition[15], e.g., fixing α_1 to some value, like, zero.

The number $rs-(r+s)+1=(r-1)(s-1)$ is called the *number of degrees of freedom*.

After this, the estimate of the mean error of the average taken over all complete sets is $\widehat{\sigma}_s = \frac{\widehat{\sigma}}{\sqrt{s}} = 1.8^{\text{cc}}$.

5.9 Instrumental errors of a theodolite

The instrumental errors, of a theodolite are divided into the following groups:

1. axis errors:
   - the axes are not perpendicular to each other: *collimation error* and *trunnion axis tilt*, subsections 5.9.1 and 5.9.2.
   - The axes don't intersect in the same point.
2. eccentricity errors:
   - eccentricity of the graduation circles
   - eccentricity of the measuring telescope.
3. graduation errors of the circles.

In modern theodolites the eccentricity and graduation errors are small. The errors cannot generally be corrected, but they can be determined, in *calibration*.

5.9.1 Collimation error

The most consequential error is the *collimation error*, which is easy to determine and correct. Collimation error means that the angle between the sight axis T and the horizontal axis V is not a right angle: $T \not\perp V$.

The sight axis T is realized by the crosshairs in the measuring telescope (more precisely, T is the straight line going through the optical centre of the objective and the crosshairs, figure 5.32). This is why by shifting the crosshairs one adjusts to $T \perp V$.

[15] This extra condition can also be understood as a "pseudo observation": then, the number of observations is $rs+1$.

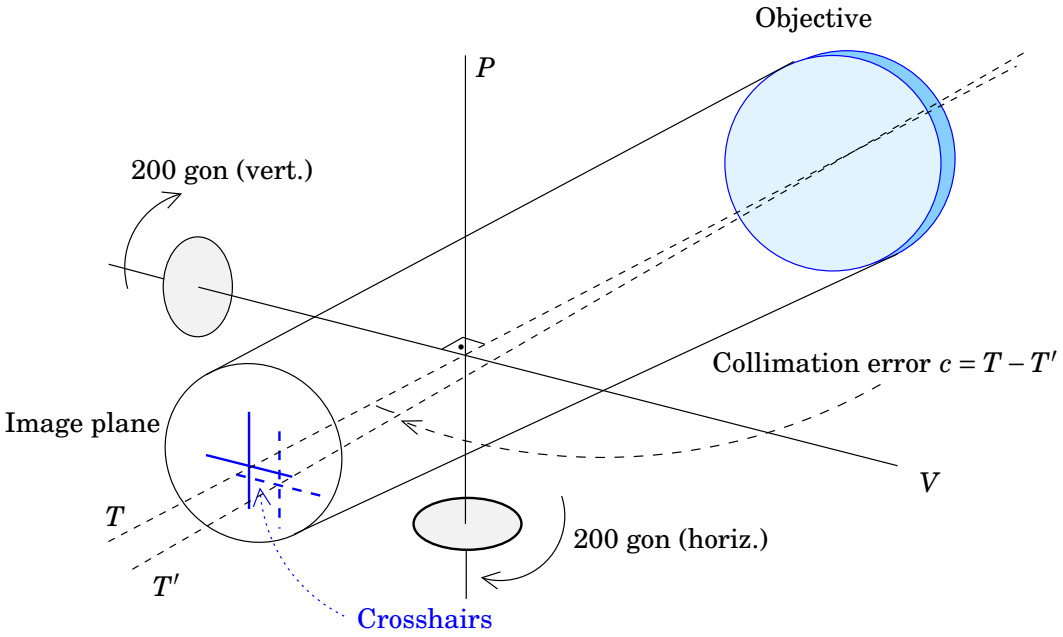


Figure 5.32. Turning the sight axis by shifting the crosshairs.

- If there is no collimation error ($c = 0$), the readings of the same target A in face right and face left are a_1 and a_2, where $a_1 = a_2 \pm 200\,\text{gon}$.
- If there is collimation error ($c \neq 0$) and the telescope is turned through to face right (200 gon around both the horizontal and the vertical axis), then A won't show under the crosshairs until the telescope is turned an amount $2c$ more.

Aim approximately horizontally in face left at object A, take reading a_1, and, in face right, take reading a_2.

The correct readings, without collimation error, *would* be A_1 and A_2, and we would have exactly $A_1 = A_2 \pm 200\,\text{gon}$.

In reality we obtain $a_1 = A_1 + c$ and $a_2 = A_2 - c$. Their difference is

$$a_1 - a_2 = 2c \pm 200\,\text{gon},$$

from which one obtains

$$c = \frac{1}{2}(a_1 - a_2 \pm 200\,\text{gon}). \tag{5.2}$$

In this way one may determine c, usually a small number.

Because c is so small, we are only interested in fractions of a gon, not the integer number of gons. Therefore we use the following notation:

> $[a]$ signifies the *rounding residue* of a, the difference between the precise value and the value rounded to an integer. So $[127.4531] = 0.4531$, $[16.9850] = -0.0150$, and so on. The outcome of the operation is always between the values $-0.5\,\text{gon}$ and $+0.5\,\text{gon}$.

Then

$$c = \frac{1}{2}[a_1 - a_2].$$

This can also be calculated from the observation notebook, if the same objects have been measured in both faces.

$$[c] = \frac{1}{2}\frac{\left[\sum[a_1] - \sum[a_2]\right]}{n} = \frac{\left[\sum[a_1] - \sum[a_2]\right]}{2n}.$$

$\sum[a_1]$ the sum of the readings taken in face left, without gons

$\sum[a_2]$ the corresponding sum in face right (correspondence both in fully measured set, and in direction)

n number of readings (sets × directions).

□ *5.9.1.1 Correcting collimation error:*

Carry out, in face left, a turning of the instrument so that the reading is the precomputed true value" $A_1 = a_1 - c$. *Shift*[16] the crosshairs (focal plate) on top of the image of the aiming target A. Verify, in face right, that the second "true value" $A_2 = a_2 + c$ is obtained when the crosshairs are on the target.

A *calculation example* from an observation notebook (all values in gon):

Face		$\sum[a_i]$	$\left[\sum[a_i]\right]$
Left	I	4.77315	77315
Right	II	5.77328	77328
			−13

The number of observations per face $n = 12$, i.e.,

$$[c] = \frac{-0.13\,\text{mgon}}{24} = -0.00542\,\text{mgon} = -0^{\text{cc}}.0542.$$

This is a good *field check*. In this case there is hardly any collimation error.

□ **5.9.2 Trunnion axis error**

The second axis error is the *trunnion axis tilt* t. It means that the horizontal axis, or trunnion[17] axis, V, is not actually horizontal even after levelling the instrument: it and the vertical axis P are not perpendicular, $P \not\perp V$.

[16] This is a task for maintenance, the screws are small and hidden.

[17] https://en.wikipedia.org/wiki/Trunnion.

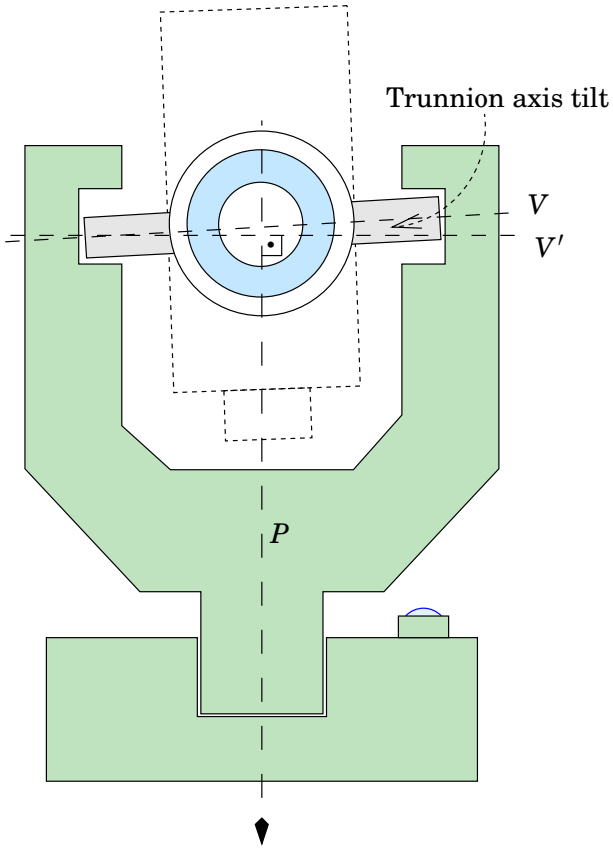


Figure 5.33. Trunnion axis tilt.

The trunnion axis tilt may be determined by using a target far away from the horizontal ($\zeta \neq 100\,\text{gon}$). Then it holds that

$$a_1 = A_1 + c + t\cos\zeta,$$
$$a_2 = A_2 - c - t\cos\zeta,$$

from which the combined term $c + t\cos\zeta$ can be determined in the same way as explained above for c. Separating the errors c and t requires measurements at two different height angles ζ. If one of these is $\zeta = 100\,\text{gon}$ ($\cos\zeta = 0$), we get back the original formula 5.2 for determining the collimation error.

5.9.3 Zenith-angle measurement and index error

Measuring a zenith angle with a theodolite requires that the index of the vertical circle — the place where the values for the zenith angle are being read — is in the horizontal plane. For this purpose a theodolite has its own fine motion screw (see figure 5.34) with which the frame is moved on which the index or indices of the vertical circle and the vertical index (or collimation) level are mounted together. The frame can however be turned slightly around the horizontal axis. Also the measuring telescope and the vertical circle are similarly connected to each other.

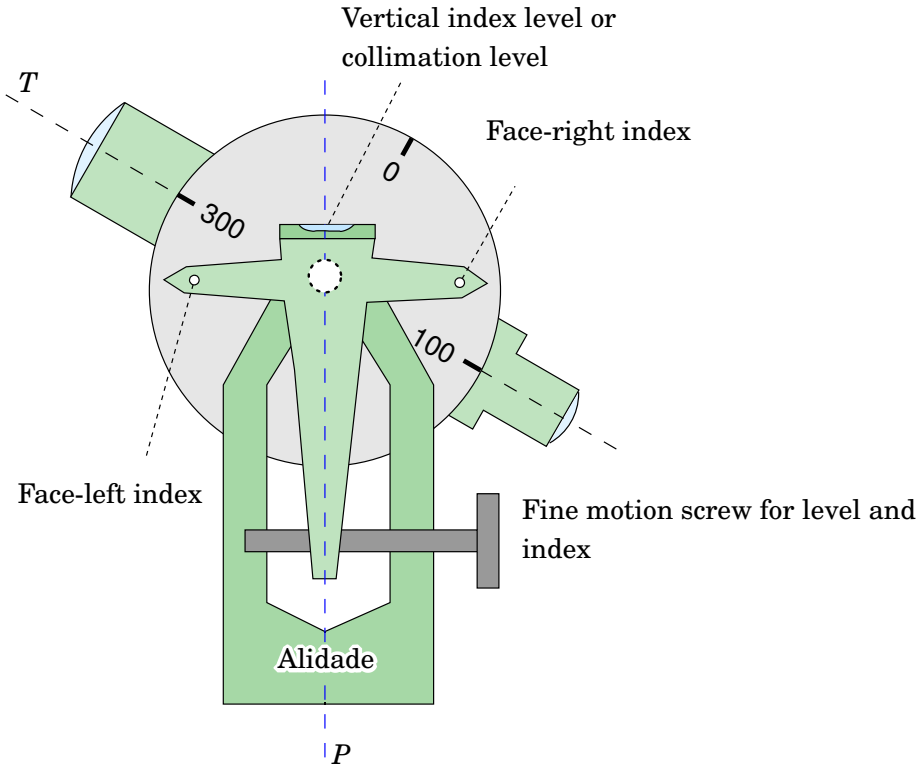


Figure 5.34. Observing a zenith angle.

Before every zenith-angle measurement it must be ascertained, using the fine motion screw of the level and index, that the vertical-circle indices are really in the horizontal plane. Of course one may not assume that the vertical index level has been adjusted so, that the index really gives exactly 100 and 300 gon precisely when the optical axis of the measuring telescope is horizontal; this error is called the *index error* (i). It can be eliminated by measuring *in both faces*, left and right. One moves between faces by "turning the instrument through", i.e., turn the telescope 200 gon around the vertical axis, and 200 gon around the horizontal axis. The measurement values obtained in both faces are added together:

$$\zeta_1 + \zeta_2 = 400 + 2i,$$

from which the index error i follows.

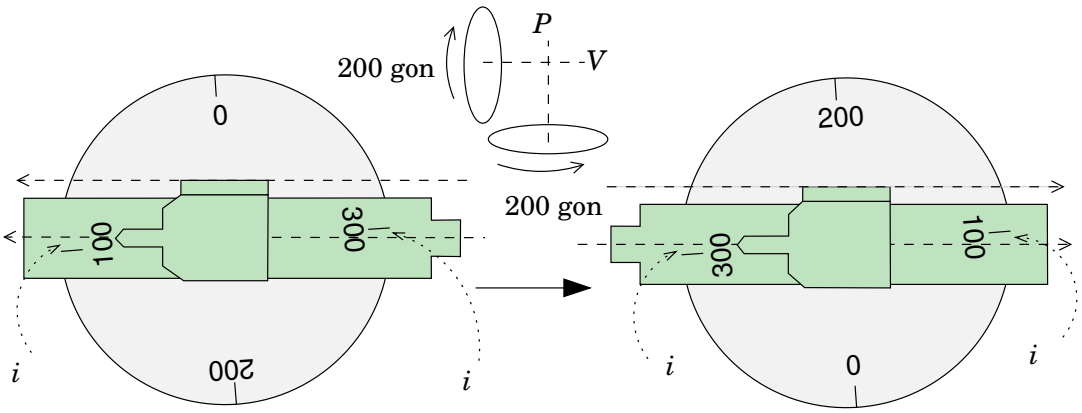
Both observations ζ_1 and ζ_2 are corrected by an amount $-i$, i.e.,

$$\zeta_1' = \zeta_1 - i,$$
$$\zeta_2' = \zeta_2 - i.$$

After this, the condition

$$\zeta_1' + \zeta_2' = 400$$

holds exactly.



True zenith angle: ζ	100
Reading: ζ_1	98
Index error: $i = \zeta_1 - \zeta$	-2

True zenith angle: ζ	100
Reading: ζ_2	298
Index error: $i = \zeta_2 + \zeta - 400$	-2

"True value":	100
$\zeta = \frac{\zeta_1 - \zeta_2 - 400}{2}$	

"Inferred index error":	-2
$i = \frac{\zeta_1 + \zeta_2 - 400}{2}$	

Figure 5.35. Index error.

See figure 5.35, which depicts (unlike figure 5.34) a theodolite of which the vertical circle is read only in one place. In the figure, the sight axis is in the horizontal plane, i.e., $\zeta = 100\,\text{gon}$. The formulas given below, however, apply generally.

In the left image, the angle ζ is measured and the reading $\zeta_1 = \zeta + i$ is obtained, in which i is the index error. In the right image, the same angle ζ is measured, but in face right, and the reading obtained is $\zeta_2 = (400 - \zeta) + i$. We obtain:

$$\zeta_1 + 400 - \zeta_2 = \zeta + i + \zeta - i$$

The angle ζ:

$$\zeta = \frac{1}{2}(\zeta_1 - \zeta_2 + 400)$$

The angle i:

$$\zeta_1 + \zeta_2 - 400 = \zeta + i - \zeta + i \Longrightarrow i = \frac{1}{2}(\zeta_1 + \zeta_2 - 400).$$

The corrected reading is

$$\zeta = \zeta_1 - i = 400 - \zeta_2 + i.$$

Removing the index error: Assume the construction in figure 5.34.

- Aim at a target of which the "true angle" ζ has been calculated.
- Turn the index and collimation level together, using the fine motion screw, to reading ζ.

- The bubble of the level will move, adjust the level with its adjustment screw(s)[18] until the bubble is again in the middle.

Check:

- Observe the target in face left (ζ_1) and immediately after, the same target in face right (ζ_2)
- Calculate the sum $\zeta_1 + \zeta_2$. If the index error $i = 0$, $\zeta_1 + \zeta_2 = 400$. This is also a good field check.

Generally one measures the same object in two full sets, quickly in succession. Set averages are calculated and entered into the computations.

In figure 5.34 there is a manual vertical-circle index (height index). Nowadays most often an automatic height index is used. The principle is somewhat similar to that used in a self-levelling levelling instrument, and like it, the index must already be approximately level:

- pendulum compensator
- liquid compensator, silicone oil. The light to the reading microscope either is refracted through the liquid or reflected from its surface. See Kahmen and Faig (1988) pages 394, 395.
- In electronic theodolites, also this compensation mechanism has been implemented digitally, using a digital tilt meter. So, the index is not adjusted but rather the reading corrected computationally based on the reading from the tilt meter.

Instrumental errors are largely the same as for horizontal angles. Part of them are eliminated by measuring in both faces. Measuring complete sets will certainly not diminish the impact of vertical circle graduation errors (because the vertical circle, unlike the horizontal circle, can not be loosened and turned, see subsection 5.8.2), it does allow however for a check on the correctness of the measurements.

The observations in face left and face right should be taken *as quickly as possible in succession*. For this reason, vertical angles should always be observed *separately*, never together with horizontal angles.

5.10 Zenith angles and refraction

In the measurement of zenith angles, because of the stratification of the atmosphere, the overwhelmingly greatest source of error is *refraction*. On must always be careful concerning weather and the landscape when choosing time and place, and when processing observations. Hot *asphalt* in summer is especially treacherous.

[18] The adjustment screws are small and may be somewhat hidden. Adjustment is a maintenance job.

The effect of refraction doesn't show in the check $\zeta_1 + \zeta_2 = 400$, see subsection 5.9.3. So, measuring in both faces doesn't help.

5.10.1 Refraction coefficient

In geodesy, the habit has formed to use, for describing refraction, a quantity named the refraction coefficient, symbol k. This quantity characterises the curvature of the measurement ray or path in the atmosphere compared to the curvature of the Earth:

$$k = \frac{\text{ray curvature}}{\text{Earth curvature}} = \frac{R}{\rho}, \tag{5.3}$$

in which

R radius of curvature of the Earth; $\frac{1}{R}$ is the curvature of the Earth's surface[19]

ρ the radius of curvature of the measurement ray; $\frac{1}{\rho}$ is the curvature of the measurement ray.

Typical k values in the atmosphere are $k_L = 0.13$ for visible light, and $k_M = 0.25$ for microwaves, see Kahmen and Faig (1988) page 167. The curvature of the measurement ray is thus $4-8$ times weaker, and the radius of curvature $4-8$ larger, than the curvature, respectively radius of curvature, of the Earth's surface. However, during an atmospheric *inversion*, exceptionally large k values may occur, even $0.3 \ldots 0.4$, see Grafarend et al. (1987).

5.10.2 The refraction equation

In figure 5.36 one sees on the left-hand side, how one may determine the height difference between two points A and B using zenith-angle measurement. The applicable trigonometric equation is

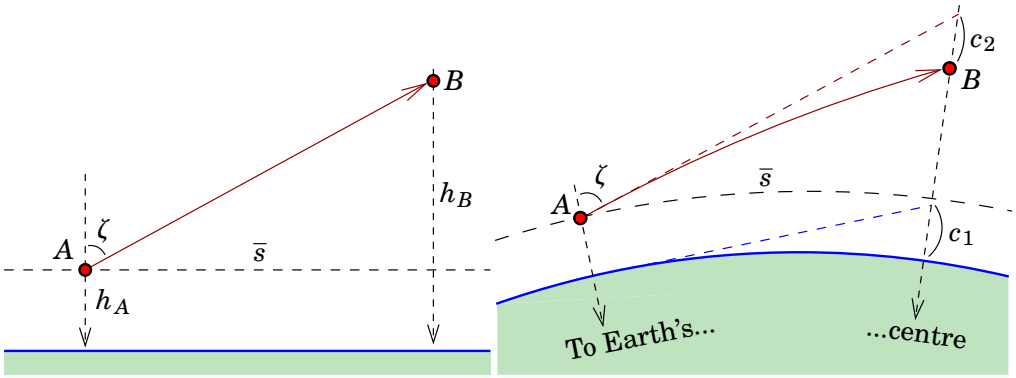
$$h_B = h_A + \bar{s} \cot \zeta,$$

in which $\bar{s}$ is the *horizontal distance* (i.e., the slant range projected onto the horizontal plane!) and ζ the measured zenith angle.

When, in reality, both the surface of the Earth and the measurement ray's path in the atmosphere are curved, in practice the right-hand side figure applies, in which however all angles have been exaggerated. Based on the figure, one should add to the above equation two *correction terms*:

$$c_1 \approx \frac{\bar{s}^2}{2R},$$

[19] Remember that the curvature is the inverse of the *radius* of curvature!



(a) Straight geometry **(b)** Taking curvatures along

Figure 5.36. The effects of refraction and Earth curvature on zenith-angle measurement.

the correction due to the Earth's curvature, and

$$c_2 \approx -\frac{\bar{s}^2}{2\rho} = -k\frac{\bar{s}^2}{2R},$$

the correction due to the curvature of the measurement ray — assuming the angle ζ not differing very much from a right angle.

Everything together:

$$h_B = h_A + \bar{s}\cot\zeta + (1-k)\frac{\bar{s}^2}{2R}.$$

If we still take along the height of the theodolite, or instrument, i over point A and the height of the signal or target t over marker B underneath it, we obtain

$$h_B = h_A + \bar{s}\cot\zeta + (1-k)\frac{\bar{s}^2}{2R} + i - t, \tag{5.4}$$

the *fundamental equation of trigonometric heighting*.

5.10.3 Simultaneous measurement in opposite directions

If we carry out zenith-angle measurements simultaneously in point A and point B, we obtain

$$h_B = h_A + \bar{s}\cot\zeta_A + (1-k)\frac{\bar{s}^2}{2R} + i_A - i_B,$$
$$h_A = h_B + \bar{s}\cot\zeta_B + (1-k)\frac{\bar{s}^2}{2R} + i_B - i_A,$$

in which we now called, for the measurement done in A, $i \overset{\text{def}}{=} i_A, t \overset{\text{def}}{=} i_B$, and for the measurement done in B, $i \overset{\text{def}}{=} i_B, t \overset{\text{def}}{=} i_A$. So, we assume the heights of instrument and signal to be the same on the same point

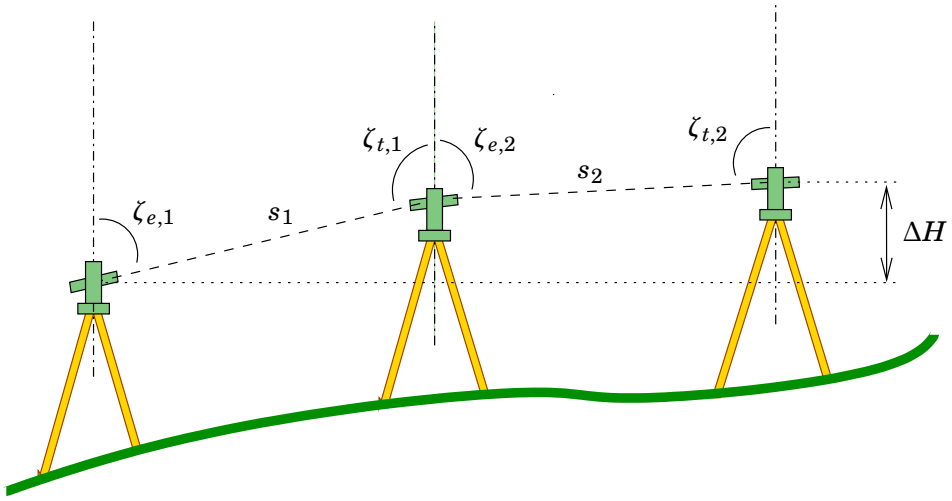


Figure 5.37. Trigonometric levelling traverse.

(easy to do with *forced centring*, see subsection 5.4.6). Then, rearranging terms and subtraction yields

$$h_B = h_A + \frac{1}{2}\bar{s}(\cot\zeta_A - \cot\zeta_B) + i_A - i_B, \tag{5.5}$$

from which the *term describing atmospheric refraction has vanished*.

This method for determining height differences between points has been found good and has been much used also over long distances. It requires the measurement or determination of the distance $\bar{s}$ with sufficient precision.

This method is used in trigonometric levelling (figure 5.37), which can replace traditional levelling, e.g., in terrain with great height variations, where the staff distance of levelling would become very short and the work laborious. In the method, two total stations and two signal-reflector assemblies are used, and the observation data is transmitted by radio modems from one instrument to the other for processing, error control and storage. Movement to the next point is by car where the terrain allows it.

V. R. Ölander[20] used already in the 1930s a method for *refraction modelling* which was based, not on simultaneous measurements in opposite directions on all triangle sides, but on (approximately) simultaneous measurements *from each triangulation point to all neighbouring points*.

Ölander assigned to every triangulation point in the network its own refraction coefficient as an unknown, which were all solved for by means of network adjustment (Ölander, 1932, see also Grafarend et al., 1987).

[20]Victor Rafael Ölander (1897 – 1973) was a Finnish geodesist who played a central role in the Finnish primary triangulation.

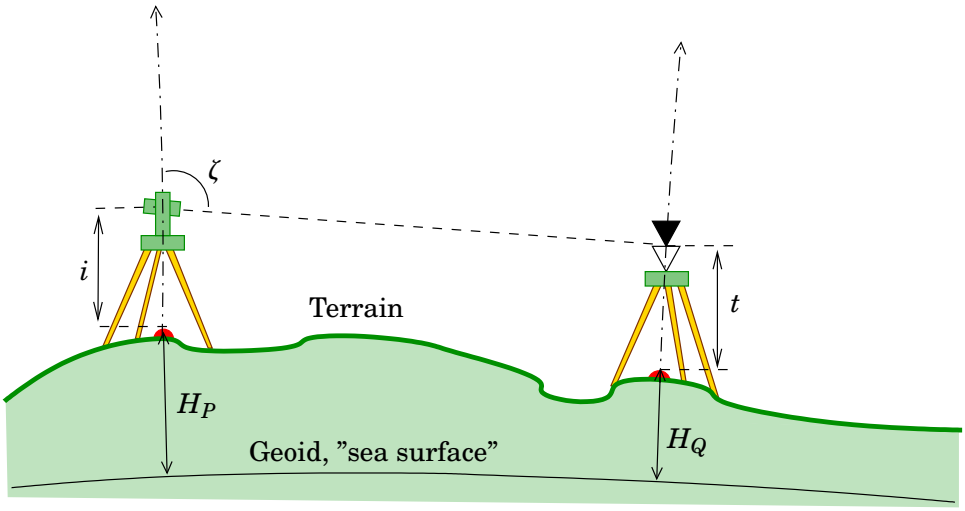


Figure 5.38. Heights of instrument and signal.

5.11 Heights of instrument and signal

Even though the instrument is levelled and centered on the point, it is nevertheless always *eccentrically set up in the height direction*.

The zenith angle to be measured is the angle between the plumb line and the sight axis of the telescope, figure 5.38. Thus, one must measure the height H_K of the instrument. Generally the aiming is at a signal on a tripod, not the point marker (monumented point) itself, so also the signal or target height H_T must also be measured. So, one must always measure

- At the instrument, the height difference between marker or monumented point, and horizontal instrument axis
- At the signal, the height difference between marker (monument) and signal target point.

In the depicted situation (figure 5.38), the aiming point was the upper edge of the white triangle, and it is the height H_T of that point that must be measured.

In the height direction, both the instrument and the signal are *eccentrically set up*.

Also drawn in the picture are the *heights* of the monumented points from a computational reference surface ("sea level"), which one tries to determine from measurements:

- height of the instrument from the reference level: $H_P + i$
- height of the signal or target from the reference level: $H_Q + t$.

Self-test questions

1. Decribe the three axes of a theodolite. Which of them are mutually perpendicular?
2. In order to measure from a known point, a theodolite must be *centred* and *levelled*. Describe the stages in which this is done, for a theodolite equipped with an optical or laser plummet.
3. Explain the idea of forced centring. Why is it useful?
4. Reading the circles. What is a scale microscope, what a coincidence microscope? Drawing!
5. Explain Gray codes.
6. What is ATR, Automatic Target Recognition?
7. Describe collimation error, trunnion-axis error and index error. How are they determined?

6. Distance measurement

6.1 Mechanical distance measurement

The SI unit of distance, more precisely of *length*, is the metre (section 2.1). In precise distance measurements *traceability* to the metre standard is important.

Although today, even short distances are measured electronically or electro-optically, it is good to understand the oldest[1] and technically simplest length measurement method, tape measurement. It continues to be used in local measurements, when distances are short and the precision afforded by tape measurement suffices[2]. And of course the equipment required is inexpensive!

In tape measurement, *four corrections* must be taken into account:

1. *Tape correction*

   This is a tape specific reduction. It is determined by *calibration* in a *comparator*, the true length of which is known precisely —

[1] Back in the days, the scale of the Finnish primary triangulation was transferred from the Nummela baseline to the principal sides of the network using invar wires. Invar is an iron-nickel alloy (64% iron, 36% nickel) having a very small coefficient of thermal expansion. Also in the Lapland grade measurement of Maupertuis, the scale of the triangle network was obtained mechanically from a baseline — during winter 1736 on the ice of the Torne river!

[2] *Careful* tape measurement is surprisingly precise!

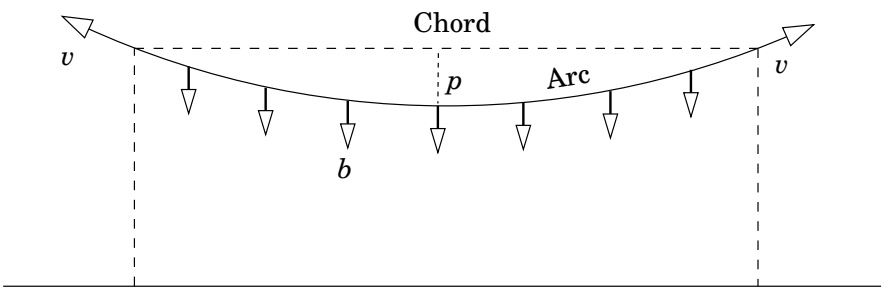


Figure 6.1. Sag correction of measuring tape.

thanks to comparisons, though the traceability chain, with the standard metre. The tape correction $\Delta\ell_0$ is now the difference between the comparator's true length $\ell_0 + \Delta\ell_\mathrm{k}$ and its length measured by the tape to be calibrated, $\ell_0 + \Delta\ell_\mathrm{m}$, i.e.[3]:

$$\Delta\ell_0 = \Delta\ell_\mathrm{k} - \Delta\ell_\mathrm{m}.$$

Here, ℓ_0 is the nominal length of the tape, e.g., $\ell_0 = 30\,\mathrm{m}$.

2. *Temperature correction*
   The temperature correction is caused by the thermal expansion of steel, and thus requires measuring the temperature of the tape. The temperature correction is stated for a standard temperature of $t_0 = 20^\circ\,\mathrm{C}$. If the coefficient of thermal expansion of steel is α, the temperature correction is

$$\Delta\ell_\mathrm{t} = \alpha\ell_0(t - t_0).$$

   Here, α is expressed in micrometres per metre and degree. E.g., the coefficient of thermal expansion of a certain steel alloy is $11.34\,\mu\mathrm{m/m}^\circ\mathrm{C}$ i.e., $\alpha = 11.34\cdot10^{-6}(^\circ\mathrm{C})^{-1}$, because $\mu\mathrm{m/m} = 10^{-6}$. If a measuring tape is 30 metres long and the temperature is 28 °C, it follows that

$$\Delta\ell_\mathrm{t} = 2.7\,\mathrm{mm}.$$

3. *Sag correction*
   During measurement, the tape is tensioned with a known force v. No matter how large this force is, the tape will always settle into a so-called catenary, in fact a cosinus hyperbolicus (cosh function) figure. We thus need measurement of the *tensioning force* of the tape, see figure 6.1. The theory of the phenomenon is surprisingly complicated. The end result, the sag correction, i.e., the difference in length between chord and arc, is

$$\Delta\ell_\mathrm{p} = -\frac{b^2}{24v^2}\ell^3,$$

   proportional to the length of the tape to the third power. In this formula, b is the weight of the tape per metre, and v the tensioning force of the tape. Alternatively one can measure the sag p in the middle and use the equation

$$\Delta\ell_\mathrm{p} = -\frac{8p^2}{3\ell}.$$

[3] Question: why the algebraic sign like this? Why not

$$\Delta\ell_0 = \Delta\ell_\mathrm{m} - \Delta\ell_\mathrm{k}?$$

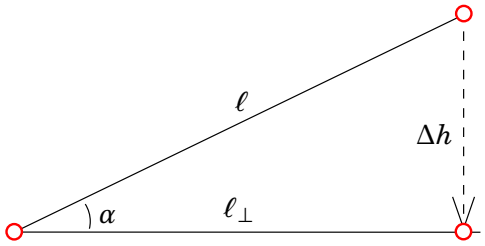


Figure 6.2. slope correction of slant ranges.

4. *Slope correction*

   This isn's actually a correction but rather a *reduction*, which is necessary if, instead of the slope distance or *slant range*, one wishes to obtain the horizontal distance between the points, the projection of the slant range onto the horizontal plane.

In routine tape measurements, the tape correction, temperature correction and sag correction can most often be ignored. Their detailed description with formulas is found in the literature, e.g., Kahmen and Faig (1988, pages 122 – 130). The *slope correction* can however be significant: if it is given that the height difference of the end points of a distance ℓ is Δh, then the horizontal distance is, according to the Pythagoras theorem,

$$\ell_\perp = \sqrt{\ell^2 - \Delta h^2}. \tag{6.1}$$

If we only know the slope angle α, $\sin\alpha = {}^{\Delta h}/_{\ell}$, between end points, we obtain the horizontal distance as follows (figure 6.2):

$$\ell_\perp = \ell \cos\alpha.$$

Equation 6.1 may be written with sufficient precision for small slopes

$$\ell_\perp = \ell\sqrt{1-\left(\frac{\Delta h}{\ell}\right)^2} = \ell\sqrt{1-\left(\frac{\kappa}{100}\right)^2} \approx \ell - \frac{1}{2}\frac{\kappa^2}{10{,}000}\ell \overset{\text{def}}{=} \ell + \Delta\ell_k,$$

in which the slope correction

$$\Delta\ell_k = -\frac{\ell\kappa^2}{20{,}000}$$

is presented as a function of the *slope percentage* κ.

6.2 Electromagnetic radiation

Light, radio waves and many other forms of radiation are examples of *electromagnetic radiation*.

In physics, it has been long thought about whether visible light is a wave motion (Huygens) or a stream of particles (Newton). The finding of interference phenomena resolved the controversy in favour of the wave-motion theory. This was decisively helped also by the development, by

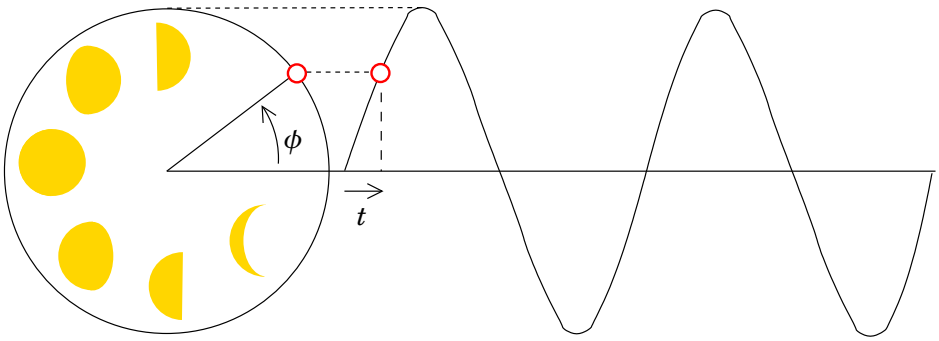


Figure 6.3. The phase of a wave motion. The name "phase" probably originates from the Moon's phases.

James Clerk Maxwell[4], of the field theory of electromagnetism, which, using partial differential equations, describes *electromagnetic waves* as a natural wave phenomenon occurring in this field. Maxwell even succeeded in calculating theoretically the propagation speed c, which was close to the already observed speed of light...

Being a wave motion, electromagnetic radiation has a *phase* ϕ. When we describe a wave motion as the projection of a uniform circular motion onto one dimension[5] (figure 6.3), then ϕ is the angle at the centre of that circle that measures this uniform motion. The relationship between phase ϕ and frequency f is

$$\phi(t) = \phi(t_0) + 2\pi f\,(t - t_0),$$

in which t is time and t_0 reference time. Clearly, the phase is periodic and repeats after one cycle, or 2π. Therefore we may always reduce the phase angle to the interval $[0, 2\pi)$.

Nowadays one can measure the wavelengths and frequencies of the various forms of electromagnetic radiation very precisely; between them there is the relationship

$$\lambda f = c$$

in which f is the frequency and λ the wavelength. The quantity c is the *speed of light* (in vacuum), which according to Einstein (or actually already Maxwell) is a constant of nature. See figure 6.4.

---

[4]James Clerk Maxwell FRS FRSE (1831 – 1879) was a Scottish mathematical physicist who set the theory of the electromagnetic field on a mathematical footing, and also contributed essentially to theoretical thermodynamics.

[5]Equivalently: as the real part of the complex wave function $\exp(i\phi) = \cos\phi + i\sin\phi$.

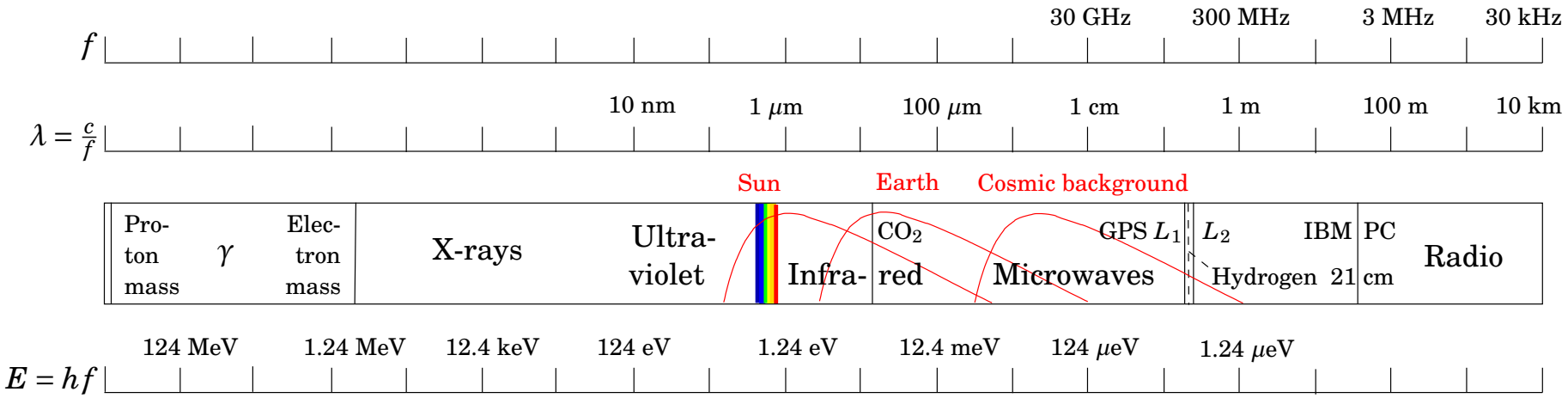


Figure 6.4. The electromagnetic radiation spectrum.

However, *quantum theory* has made the particle model relevant again. Light can be described as a stream of particles, *photons*, the energy of which is

$$E = hf,$$

in which h is *Planck*[6]*'s* constant. The particle model is fertile especially for high energy levels, the left side of figure 6.4.

The electromagnetic field is a *vector field*. Therefore electromagnetic radiation is a *transversal wave motion*, and may be *polarized*[7]. In figure 6.5 is shown both linearly and circularly polarized radiation. One can choose two independent directions of polarization, e.g., up-down and left-right, of which all others can be composed by combination. E.g., circularly polarized radiation is obtained by combining two mutually orthogonal, linearly polarized rays with a phase difference of $\pi/2$. This also works the other way around: by combining clockwise and counterclockwise circularly polarized rays, one obtains again a linearly polarized ray.

An intermediate form between linear and circular polarization is represented by elliptically polarized radiation, the field vector of which turns along an ellipse-shaped path.

In particle language one might say, that the electromagnetic field is the quantum theoretical wave function of the photon, which thus is a

---

[6]Max Karl Ernst Ludwig Planck (1858 – 1947) was a German physicist and organizer of German physics. He is remembered for his discovery that the thermal radiation spectrum of a so-called black body is caused in a natural way by the quantization of electromagnetic radiation.

[7]The correct understanding of polarization was the achievement of Thomas Young (1773 – 1829) and Augustin-Jean Fresnel (1788 – 1827). In studying polarization, a certain mineral, clear calcite ($CaCO_3$), *Iceland spar*, was in a central role. The crystal is *birefringent* ("doubly refracting") and splits light into two parts according to their direction of polarization. Apparently the Vikings used it in navigating by the Sun in overcast weather. The Dutch Christiaan Huygens (1629 – 1695), the father of the wave theory of light, spent ample time on the experimental study of Island spar.

Another background story links polarization to the chemistry of life. Chiral (non-mirror-symmetric, "handed") molecules, like sugar, rotate the plane of polarization of light passing through them, so-called *optical activity*. Already Louis Pasteur (1822 – 1895) studied the background of the optical activity of organic tartaric acid, and today we know that the phenomenon is linked to the chirality of life itself, its "handedness", like the winding direction of the helix of the DNA molecule. The polarimeter is a vital tool in the medical and food industries.

Iceland spar was also used in military equipment, for which reason it was long classified as a strategic material.
http://www.lensrentals.com/blog/2013/12/iceland-spar-the-rock-that-discovered-optics/.

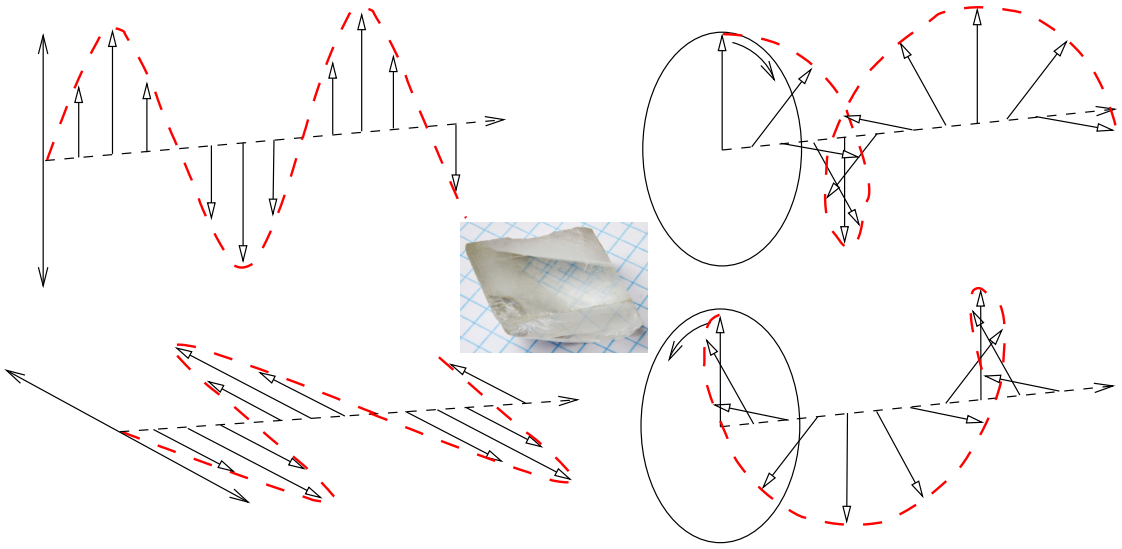


Figure 6.5. Polarization of electromagnetic radiation. The arrows show the field vector **E**. On the left, a linearly polarized wave motion; on the right, a circularly polarized wave motion. The photograph of Iceland spar is © Wikimedia Commons.

vector-valued function. The photon is a *vector particle*, having an intrinsic angular momentum, or *spin*, $\hbar/2\pi$. This spin — which may be visualized as an angular-momentum vector — may be oriented along the flight direction, or opposite to it, which corresponds to either clockwise or counterclockwise circular polarization. Linearly polarized radiation again may be described as an equal mix of both spin directions. See http://en.wikipedia.org/wiki/Photon_polarization.

6.3 Väisälä interferometry

One classical distance measurement technique which continues to be in use, is Yrjö Väisälä's white-light interferometry technique, invented already in the 1920s, for the precise measurement of long baselines.

The method works as follows. White light travels from a source to the observation device along two paths:

1. directly, by being reflected from the far mirror
2. by being reflected multiple times — in the figure, four times — back and forth between the near mirrors.

The light used is *white* and contains all the different wavelengths that make up white light. Therefore the *coherence length* of the light is very short, only $1.3\,\mu$m.

This is why the interference fringes will only show up if both paths are,

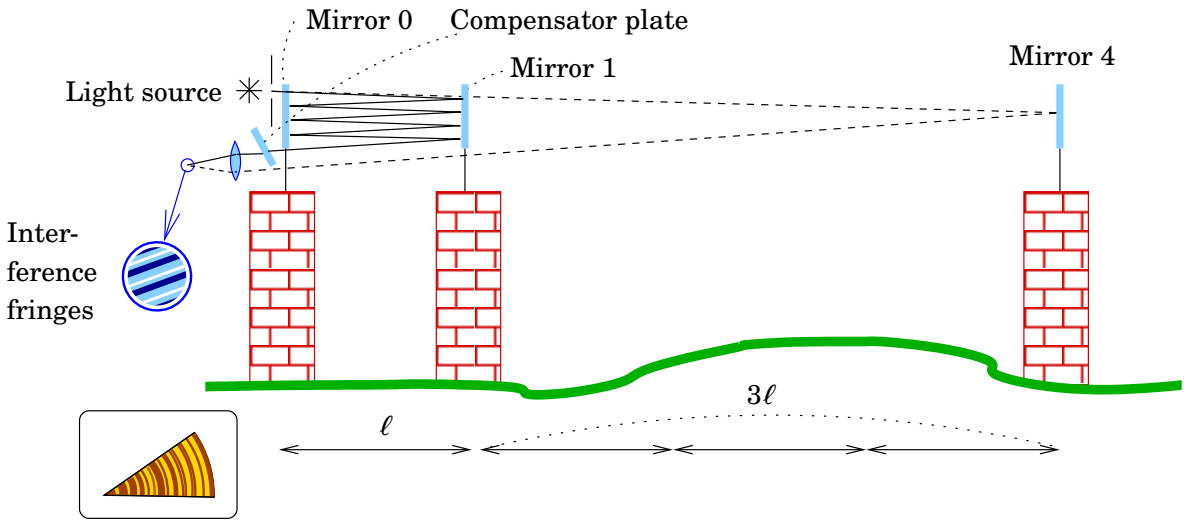


Figure 6.6. Väisälä's interference method.

at this accuracy, equally long[8]. So, if the distance of the far mirror is a *multiple* of the distance between the near mirrors.

This enables the *multiplication* of a given distance. Say that the distance between mirrors 0 and 1 is precisely 1 m. Then we can, using interference, place the far mirror at distance 6 m — precisely. After that, we take mirror 1 away, and using the same method, using now the mirror pair 1 and 6 as near mirrors, place the far mirror at 24 m. And so on...

In practice, this doesn't quite work this way. There will always remain a small difference between the path lengths, which is eliminated, i.e., measured, using a turnable glass plate of constant thickness, a *compensator glass*.

The device described is actually an analogue optical *correlator*. See figure 6.7. We will go deeper into correlation determination in subsection 11.4.1, in connection with GPS.

Realizing the original distance of one metre between the 0 and 1 mirrors isn't quite simple either. For this is used a one metre long *quartz gauge*[9], which is allowed to touch the surface of one mirror. In the air gap between the other mirror and the end of the gauge, now Newton rings will show, an interference pattern. By counting the rings in sodium light the width of the air gap can be determined.

The Väisälä interferometry method is extremely time consuming. Measurement conditions are suitable only very rarely for measuring the longest distance, 864 m. Also setting up the mirrors, their orientation, and the transfer of their measured places to the underground, perma-

[8]This is why laser light won't do! It would produce interference fringes even, when the distances are unequally long.

[9]Calibrating those precisely is again a story of its own...

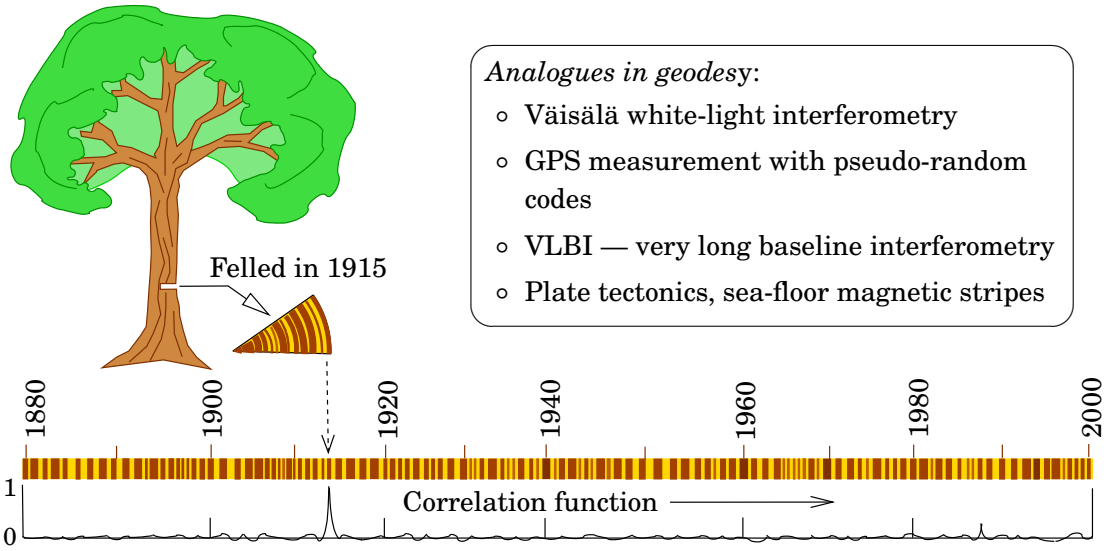


Figure 6.7. Dendrochronology: dating of wood using tree rings.

nent markers by *projection measurements*, is a complex operation demanding its own time. During measurements, the air temperature is continuously read from precision thermometers suspended along the whole line, providing gainful employment and physical exercise to two measurement assistants.

The method has been used for measuring baselines as long as 864[10] metres in Nummela, Finland, and in many places around the world. The precision achievable is at its best ±0.02 mm. The interference measurement itself is more precise still, the bottleneck is the projection measurement, the transfer of the measurement values from the mirrors to the underground markers.

6.4 Electronic distance measurement

6.4.1 The speed of light

The speed of light in vacuum is a constant of nature. Apparently already Galileo (1564 – 1642) tried to measure it in 1638 using two lamps and an assistant: a cover was taken off one lamp, and the assistant on another hill responded in the same way. Of course the result was useless: the speed of light would, according to the experiment, be infinite, or at least very large.

The first terrestrial distance measurement devices or range finders were developed to determine the speed of light. The prototype of the method is Fizeau[11]'s instrument, consisting of a light source, a reflector, and a

[10] $864 = 2 \times 2 \times 3 \times 3 \times 4 \times 6$.

[11] Armand Hippolyte Louis Fizeau MIF FRS FRSE (1819 – 1896) was a French

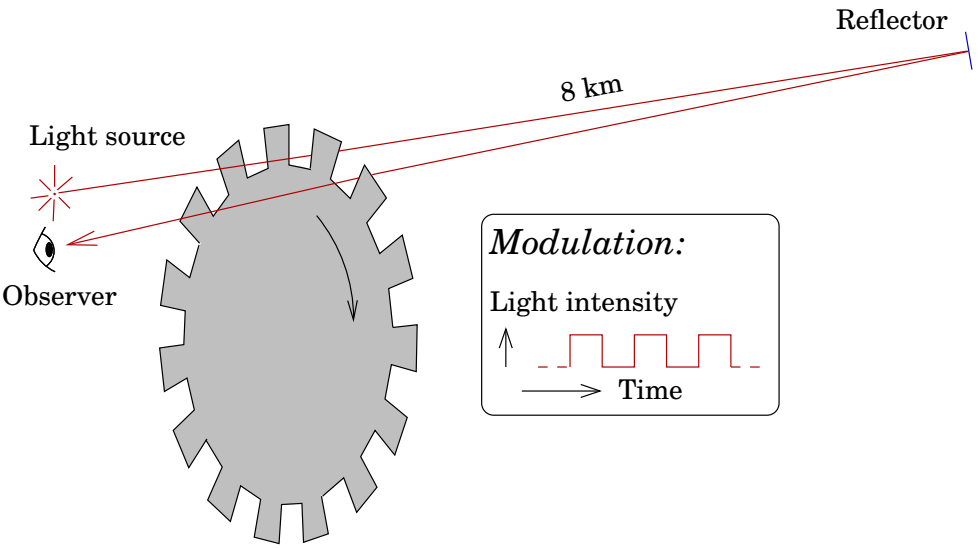


Figure 6.8. Fizeau's method for measuring the speed of light. The distance used for the 1849 measurement was from Montmartre to Mont Valérien, in the municipality of Suresnes, 8,633 m.

rapidly spinning camwheel. If the wheel spins at the right speed, light leaving through one opening between the teeth will return through the next opening. With a slightly greater or smaller rotation speed, however, the light will hit a tooth. In Fizeau's tests, the distance measured was 8 km.

Fizeau's camwheel was a primitive *modulator*. Nowadays electronic or electro-optical modulators are used, which have as their task to vary or *modulate* the intensity of outgoing light periodically at a certain frequency.

6.4.2 Electronic distance measurement instruments

When the superior accuracy of the devices developed to determine the speed of light became clear, the picture changed. Today, the speed of light is no longer measured; it is a quantity derived from the definitions of the metre and the second (see section 2.1) which has the conventionally agreed value of *exactly* 299,792,458 m/s in vacuum.

Electronic distance measurement devices or range finders can be of three types of construction:

1. separate: the device is placed into the forced-centring device. This solution has become rare due to the devices becoming ever smaller.

---

physicist. He also measured the speed of light in flowing water and found in it an anomaly which only special relativity could explain. He is one of the 72 French scientists and engineers who got their names inscribed on the Eiffel Tower, outside the balcony of the first floor. He was also a fellow of the Royal Societies of London and Edinburgh.

2. a separate part that is locked onto the theodolite's measuring telescope. Also this solution has become impractical:
   (a) An instrumental tilt correction must be made.
   (b) E.g., the classical Distomat range finder was placed on top of the theodolite's telescope, preventing it from being plunged through to the other face.
3. integrated with the theodolite. One speaks of a *coaxial* solution: the light moves in both directions through the theodolite's measuring telescope and uses the same optics.

Electronic distance measurement is, based on the frequency area used, divided into two main types:

1. electromagnetic, using microwaves. Obsolete as a terrestrial method — but satellite positioning uses also microwaves!
2. electro-optical:
   (a) visible light, white: Mekometer
   (b) laser light or light-emitting diode (LED), visible or near infrared. Monochromatic.

Independent of device type, the measurement takes place by *modulating* either light (or infrared) or radio waves (microwaves) with a certain frequency, and measuring the *phase difference* between the outgoing radiation and the radiation reflected from the target. Electronic phase measurement can be very precise, but does not tell how many *whole* wavelengths fit in the distance, the so-called *ambiguity problem*.

The travel time of the signal is

$$\Delta t = \left(\frac{\Delta\phi}{2\pi} + n \right) \frac{1}{f},$$

in which f is the frequency, $\Delta\phi$ the measured phase difference in radians, and the integer n the unknown so-called ambiguity. Because the measured phase difference lies always in the interval $[0, 2\pi)$, the measurement is not yet enough to fix Δt. We can certainly calculate a *possible* travel time

$$\overline{\Delta t} = \frac{\Delta\phi}{2\pi} \frac{1}{f},$$

but it is not unambiguous.

Phase measurement takes place through measurement of *time differences*: when the reference signal passes through zero in the positive direction, it starts a counter, and when the returning measurement signal does the same, the counter stops and the value is read out. See figure 6.9.

From the phase measurement $\Delta\phi$ the *distance* is computed:

$$s = \frac{1}{2} c \Delta t = \frac{1}{2} \left(c \overline{\Delta t} + n\lambda \right),$$

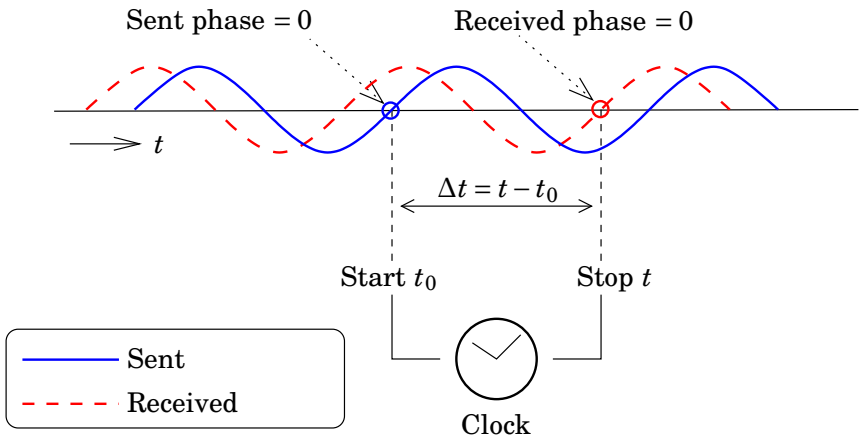


Figure 6.9. The principle of electronic phase measurement: zero phase starts / stops an electronic counter or "clock".

in which $\lambda = c/f$ is the wavelength, and n an unknown number of whole wavelengths. Determining the integer unknowns or *ambiguities* n is a problem we shall also see in connection with GPS carrier-phase measurements. In a range finder many different modulation frequencies f_i (or, equivalently, wavelengths $\lambda_i = c/f_i$), are used, which have been chosen so, that only one distance s, and corresponding travel time Δt, is compatible with integer-valued unknowns n_i for all wavelengths. See figure 6.10.

Nowadays there exist small and inexpensive hand-held range finders, which work either based on an infrared beam or an acoustic (ultrasound) beam. They are handy in construction projects, and even real-estate brokers use them...

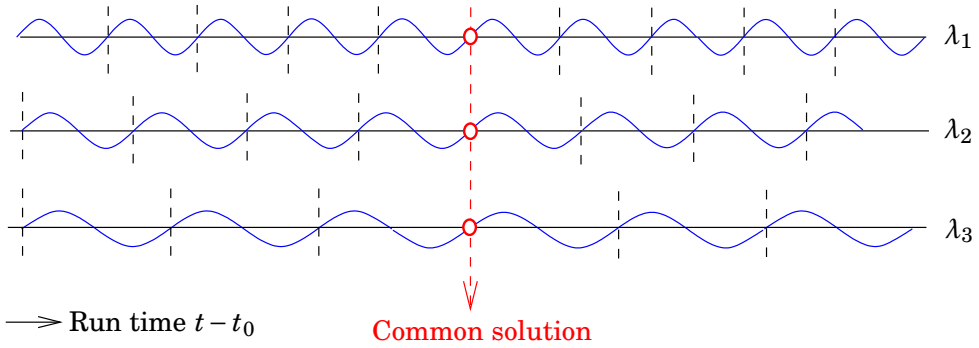


Figure 6.10. Ambiguities, or integer unknowns, are resolved by using several wavelengths. The figure shows signals received on three different wavelengths. It is assumed that the phase angles of the transmitted signals are all zero on the moment of transmission t_0. The only possible travel time is the one for which also the phase angles of all three received signals are zero. It has been marked with an arrow.

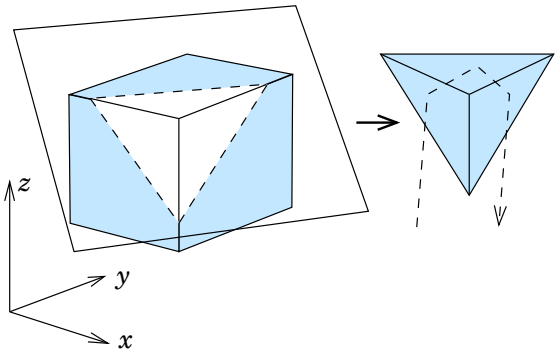


Figure 6.11. Corner-cube prism.

6.4.3 Reflectors

Electro-optical devices need, in order to work, a *reflector* placed at the target. Typically, a so called *corner-cube prism* is used, see figure 6.11. The principle of operation of a corner-cube prism is based on using three reflective surfaces that are perpendicular to each other, of which the first inverts the x co-ordinate of the light ray to the $-x$ direction, the second, the y co-ordinate to $-y$, and the third, the z into the $-z$ direction. The end result is a complete inversion of the light ray:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} \Longrightarrow \begin{bmatrix} -x \\ -y \\ -z \end{bmatrix}.$$

The light ray hitting the prism will be reflected back in precisely the opposite direction, independently of which direction it came from — as long as it is within the prism's opening angle.

Over short distances, also reflective stickers can be used, or the light reflected by the target itself, without aids. Then, the accuracy of measurement isn't necessarily the best possible!

Over longer distances one may use, instead of a single prism, an assembly of three prisms. Over very long distances (tens of kilometres) one can combine many prisms into one "pack", figure 6.12. Nowadays such distances (vectors) are however measured with GNSS.

The prism assembly fits into a forced-centring device.

Warning: When using a signal equipped with a prism, one may not place the theodolite's crosshairs on the prism "crosshairs"! It may be (and presumably is) not aiming straight, and was never meant to be used for this. Use the triangular markings on the signal instead! See figure 6.13.

6.4.4 Instrumental errors of electronic rangefinders

The systematic instrumental error of an electronic range finder consists of two parts:

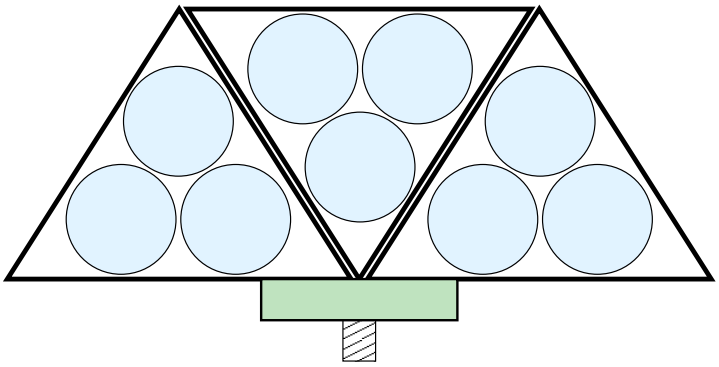


Figure 6.12. A prism pack for measurement over long distances.

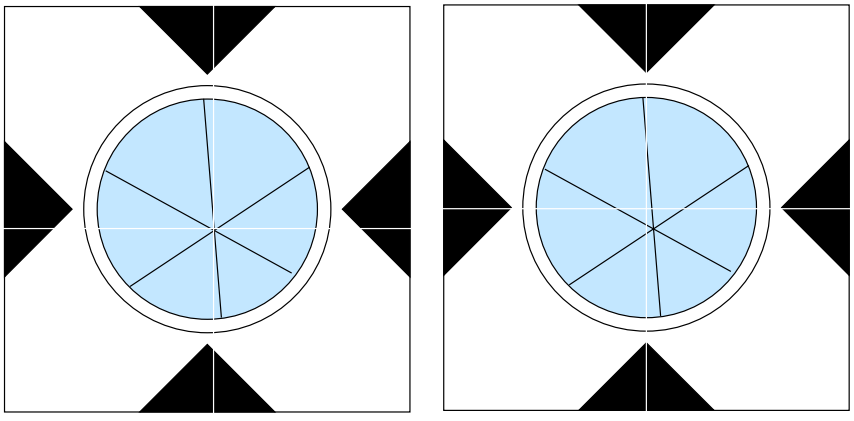
1. the zero-point or constant error
2. the scale or frequency error.

The constant error is an instrumental constant to be determined by *calibration*. The frequency error is a scale error, which is also determined by calibration: frequency calibration.

The constant or zero-point error is caused by the circumstance that the electric centre of the instrument is at a different place than its nominal one. Inside the device, the signal path may contain unknown delays. The constant error may depend on temperature, may change slowly over time ("drift" or "creep"), and may change in connection with repairs. For this reason *regular calibration* is to be recommended.

Also the reflector has a zero-point error, and often, the sum of the zero-point errors of instrument and reflector together are stated.

1. The frequency error is often determined in the laboratory using a



(a) Not like this... **(b)** ...but like this

Figure 6.13. Incorrect and correct targeting at a signal equipped with a corner-cube prism.

Tableau 6.1. Calculating the constant and frequency error by linear regression.

In order to calculate — *estimate* — the constant error a and the frequency error V from measurements on a calibration baseline, we use the following standard pair of formulas for linear regression:

$$\begin{aligned}\widehat{V} &= \frac{n\sum_{i=1}^{n} s_i \Delta s_i - \sum_{i=1}^{n} s_i \sum_{i=1}^{n} \Delta s_i}{n\sum_{i=1}^{n}\left(s_i^2\right) - \left(\sum_{i=1}^{n} s_i\right)^2},\\ \widehat{a} &= \frac{1}{n}\left(\sum_{i=1}^{n} \Delta s_i - \widehat{V}\sum_{i=1}^{n} s_i\right).\end{aligned}$$

In this, n is the number of calibration distances s_i used, at least 2. Δs_i is the difference: measured distance minus nominal distance, for each distance measured.

precise frequency standard:

$$V = \frac{f_{\text{measured}} - f_{\text{nominal}}}{f_{\text{measured}}}. \tag{6.2}$$

In this, f_{measured} is the frequency value obtained in laboratory, i.e., calibration, measurements, f_{nominal} is the frequency value stated by the manufacturer (and acting as the basis for the correction formulas in the instrument's firmware).

2. The constant and frequency errors are determined by calibrating the instrument on a precise *baseline*:

$$s'_i = s_i - a - V s_i,$$

in which

s' is the "correct" distance given by the baseline

s is the distance as measured by the instrument to be calibrated

i is the number of the measurement point on the baseline, e.g., the number of the pillar.

We write

$$\Delta s_i \stackrel{\text{def}}{=} s_i - s'_i = a + V s_i$$

and solve for a and V by means of *linear regression*, see tableau 6.1.

Of course obtaining a good calibration result requires the use of a sufficiently long calibration baseline. On a short baseline one can only determine a with sufficient accuracy. On the other hand, for many distance measuring devices, the electronic signal can be

taken out and compared with a frequency standard directly, without using a baseline.

Thus, the instrumental correction for distance measurement is obtained:

$$s' = s - a - Vs,$$

in which

s' is the corrected distance

s is the measured distance

$-a$ is the instrumental constant or zero-point correction

$-V$ is the instrumental frequency correction.

The *random* total error of distance measurement, i.e., the *mean error* of the measurements, depends generally also on the length of the distance measured. An often useful formula is

$$\sigma = \alpha + \beta s,$$

in which α is the random error for zero distance, β the distance dependent random error, and σ the mean error of the observations computed from these. Here it is assumed that *systematic* errors — like constant error and frequency error — and the necessary observation reductions have already been taken along as corrections.

6.5 Ray propagation in the atmosphere

In air, like in other media, light — and other electromagnetic radiation like infrared or radio waves — travels slower than in vacuum. The slowing-down effect of the medium is expressed by the *index of refraction* n, which is defined as

$$n = \frac{c_0}{c},$$

in which c is the speed of light in air, and c_0 the speed of light in vacuum, a constant of nature. Because air is a gas, i.e., a low-density medium, the values of n are always very close to unity. Therefore also a definition is used that gives the deviation of the index of refraction from unity, in units of ppm, parts per million:

$$N = 10^6 (n - 1).$$

The index of refraction of air for the wavelengths of visible light is, according to the following approximate formula accepted by the International Association of Geodesy IAG in its general assembly of 1999 in Birmingham UK (Rüeger, 1990, page 55), Anon. (1999):

$$N_{\mathrm{L}} = N_0(\lambda) \left[\frac{273.15\,\mathrm{K}}{T} \frac{p}{1013.25\,\mathrm{hPa}} - \frac{11.27\,\mathrm{K/hPa}}{T} e \right], \qquad (6.3)$$

in which

$N_0(\lambda)$ index of refraction of the light used (wavelength λ) in dry air under standard temperature and pressure (STP), i.e., $T = 273.15\,\mathrm{K} = 0°\,\mathrm{C}$, $P = 1013.25\,\mathrm{hPa}$, $e = 0.0\,\mathrm{hPa}$

T air temperature, unit kelvin (K), i.e., absolute temperature

p air pressure, unit hectopascal (hPa) i.e., millibar (mbar)

e the partial pressure of atmospheric water vapour ("absolute humidity"), also in units of hectopascal.

N_0 depends only on the wavelength λ of the light used. An approximate formula for its calculation is

$$N_0 = 287.6155 + \frac{4.8866\,\mu\mathrm{m}^2}{\lambda^2} + \frac{0.0680\,\mu\mathrm{m}^4}{\lambda^4}. \tag{6.4}$$

This is the so-called *group index of refraction*, which differs from the phase index of refraction[12]. In connection with electronic and electro-optic distance measurement equipment one must use the group index of refraction, because *information* travels in the *modulations* on the carrier wave, which propagate at group speed.

Example: for a helium-neon laser (wavelength $\lambda = 632.8\,\mathrm{nm}$) equation 6.4 yields $N_0 = 300.231$.

The index of refraction for *microwaves* again is

$$N_\mathrm{M} = \frac{77.624\,\mathrm{K/hPa}}{T}(p - e) + \frac{64.70\,\mathrm{K/hPa}}{T}\left(1 + \frac{5748\,\mathrm{K}}{T}\right)e. \tag{6.5}$$

Unlike the index of refraction for light and infrared, the microwave index of refraction is *in the troposphere* not dependent upon wavelength. However, the propagation of microwaves in the *ionosphere* is an entirely different matter relevant to satellite positioning, see sub-section 11.6.3.

As an interesting detail we may still note that, in the index-of-refraction formula for microwaves 6.5 at a temperature of $T = 273.15\,\mathrm{K}$, the ratio of the effect of e to the effect of p is over a hundred times larger than in the visible-light refraction formula 6.3!

Index	p	e	Ratio
Optical	0.296,30	0.041,26	0.139,25
Microwave	0.284,18	4.9372	17.374

[12] In fact, the phase index of refraction is similarly (Anon., 1999):

$$N_0 = 287.6155 + \frac{1.62887\mu\mathrm{m}^2}{\lambda^2} + \frac{0.01360\mu\mathrm{m}^4}{\lambda^4}.$$

Microwaves are thus sensitive to water vapour[13], which is one of the drawbacks of this measurement technique. The problem is also manifest with GNSS measurements.

The index of refraction affects the measured distance in the following way:

$$s - s' = (n-1)s, \tag{6.6}$$

in which s is the measured distance, and s' the true distance, which would have been measured in vacuum. The *refraction correction* is now applied as follows:

$$s' = s + K_1,$$

in which

$$K_1 = -(n-1)s = -10^{-6}Ns$$

is the traditional designation for the refraction correction. Because the correction is so small, one is allowed to use an approximate value for the distance s.

6.6 "Curvature corrections"

In distance measurement, the "curvature corrections" are second-order corrections: they are significant only when the distance of measurement exceeds many kilometres. Therefore they are inconsequential in mapping or lower-order base network surveys.

Because nowadays base network surveys are done almost exclusively using GNSS, these corrections are mostly of historical interest, and we present them only shortly. More detailed expositions are found in the literature, e.g., Rüeger (1990, 2002).

In the formulas for the curvature corrections, again the *refraction coefficient* k, which we already encountered in connection with zenith-angle measurement (subsection 5.10.1), appears. There are four different corrections:

1. The measurement-ray curvature correction
   The curving of the measurement ray causes a lengthening of the path. The measured distance is, for this geometrical reason, longer

[13] This is caused by the non-symmetry and large dipole moment, or *chemical polarity*, of water molecules: in the H_2O molecule, the angle between the two $O-H$ bonds is 104°.5. This is also the reason why water is a liquid at room temperature, and why such a good solvent, and why a microwave oven is such a useful device for preparing food. All other molecules in the atmosphere, $N_2, O_2, CO_2, Ar, O_3, CH_4, \cdots$ are non-polar, and gases. https://en.wikipedia.org/wiki/Chemical_polarity.

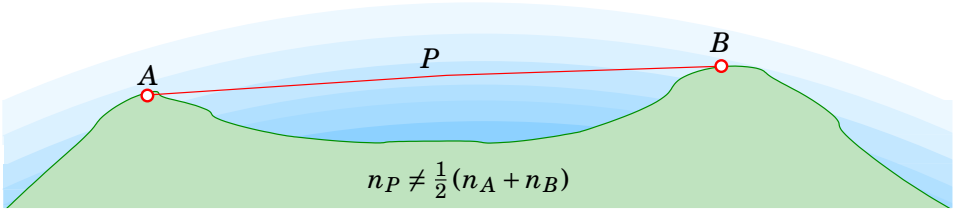


Figure 6.14. The second velocity correction: on a curved Earth, the indices of refraction in the end points of the measurement path are not representative.

□

than the true distance. The formula for the correction is

$$K_3 = -k^2 \frac{s^3}{24R^2},$$

in which k is the refraction coefficient, R the radius of the Earth, and s the distance.

2. The Earth's surface curvature correction
   After various reduction stages, usually the straight-line distance between the two projection points on a reference surface is obtained. Wanted is, however, the distance over the curved Earth's surface, which is longer. The formula for this correction is

$$K_5 = \frac{s^3}{24R^2}.$$

3. The "second velocity correction"
   Usually, the effect of refraction on the propagation of the measurement ray is evaluated based on weather observations — measurements of air pressure, temperature, and humidity — made on both ends of the path. For very long paths, these measurements are no longer *representative* for the whole ray path.
   The second velocity correction is a *systematic* effect caused by the ray curvature differing from the curvature of the Earth's surface, together with the strong vertical gradient of the air pressure — and thus the index of refraction. When the curvature of the ray path is smaller than that of the Earth's surface, the path will "dive" on longer distances deeper into the Earth's atmosphere than what the end points of the path are telling us. See Rüeger (1990, page 81). The traditional symbol and the computational formula — the derivation of which is laborious — are

$$K_2 = -k(1-k)\frac{s^3}{12R^2}.$$

   All three corrections can be combined into one equation:

$$K_{235} = K_2 + K_3 + K_5 = (1-k)^2 \frac{s^3}{24R^2}.$$

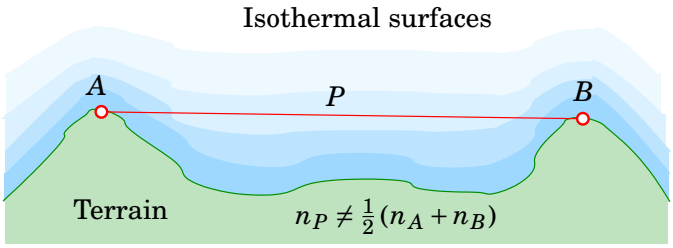


Figure 6.15. The terrain correction of distance measurement. Due to the shape of the terrain, the indices of refraction at the end points of the measurement path are not representative.

Let us calculate some values, assuming $k = 0.2$:

s	1 km	3 km	10 km	30 km	100 km
K_{235}	0.65 μm	18 μm	0.65 mm	18 mm	0.65 m

The effect is thus normally really small.

4. The "terrain correction"
   The terrain correction as studied by Juhani Kakkuri (figure 6.15) is caused by the surfaces of equal temperature within the atmosphere, the *isothermal surfaces*, generally following the forms of the terrain. Therefore, in the same way as in the case of the second velocity correction, are the weather observations in the points A and B non-representative for the mean value of the measurement path. There is no simple formula for this phenomenon.

6.7 Geometric reductions

6.7.1 Reduction to a reference surface

Thanks to the above described "curvature corrections", the original measurement was reduced for the curvature both of the measurement ray and of the Earth's surface. Thanks to corrections K_1 and K_2 also the delays by the atmosphere were taken into account. The measurement path is however still sloping and above the Earth's surface.

The correction, or *reduction*, to a chosen *reference level* corrects both the *slope* of the measured distance in space and its *height* above the chosen reference surface. As reference surfaces may be chosen

- sea level
- the surface of a reference ellipsoid
- the height of a locally defined reference surface ("zero level").

The reduction is carried out as follows:

$$s^2 = \frac{s_0^2 - \Delta h^2}{(1 + h_A/R)(1 + h_B/R)}, \tag{6.7}$$

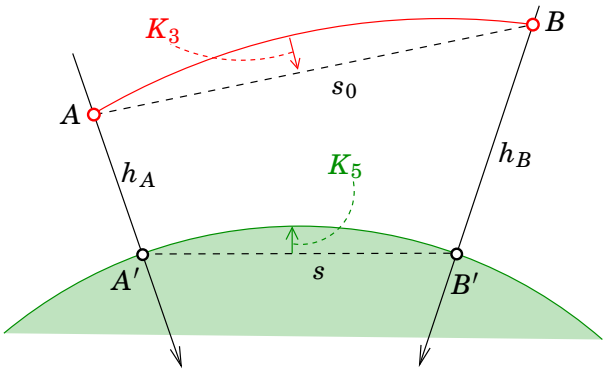


Figure 6.16. Reduction of distance measurement to a reference level, equation 6.7.

in which s_0 is the measured slant range in space, s the reduced distance (i.e., the distance between the points A', B' projected onto the reference surface, see figure 6.16), $h_A = AA'$ and $h_B = BB'$ the heights of the points from the reference surface, $\Delta h = h_A - h_B$, and R the (approximate) radius of curvature of the Earth. In the heights h_A, h_B, the heights of theodolite and signal are included.

Table 6.2 gives some examples of corrections[14] $s - s_0$. As can be seen, can this correction be already substantial also for short distances.

Modern theodolites and total stations are able to compute, in addition to device specific corrections, at least the slope correction to the measured range.

6.7.2 Map projection reduction

A map projection scale reduction is needed if we *wish* to have, instead of the length of an arc reduced to the reference ellipsoid s_{ell}, the *length*

[14] An even more approximate equation for when $\Delta h \ll s_0$ and $h_A, h_B \ll R$:

$$s - s_0 \approx \frac{\Delta h^2}{2 s_0} - \frac{h_A + h_B}{2R}.$$

Table 6.2. Examples of distance reductions.

$s_0 =$	100 m	1 km	10 km	100 km
$h_A = 0, h_B = 10$ m	−0.50 m	−51 mm	−13 mm	−79 mm
$h_A = 0, h_B = 100$ m	-	−5.0 m	−0.58 m	−0.83 m
$h_A = 0$, $h_B = 1000$ m	-	-	−51 m	−13 m
$h_A = 100$ m, $h_B = 100$ m	−1.6 mm	−16 mm	−0.16 m	−1,6 m
$h_A = 1000$ m, $h_B = 1000$ m	−16 mm	−0.16 m	−1.6 m	−16 m

in the map plane, the projected length s_{proj}. E.g., in case of the Gauss-Krüger projection as used in Finland, the approximate reduction is done as follows:

$$s_{\text{GK}} = s_{\text{ell}} \left(1.0 + \frac{y_A^2 + y_A y_B + y_B^2}{6R^2} \right), \tag{6.8}$$

in which the distance s is between the points A and B, map co-ordinates[15] (x_A, y_A) and (x_B, y_B).

This reduction *depends on the map projection chosen*, and is thus different for different map projections. E.g., the equation for the Universal Transverse Mercator (UTM) projection is otherwise the same as equation 6.8, but the constant — the scale on the central meridian — is 0.9996 instead of 1.0.

Self-test questions

1. What is the relationship between wavelength and frequency for electromagnetic radiation?
2. What is the relationship between frequency and the energy of a photon for electromagnetic radiation?
3. Describe the correlation process between two identical but random signals. Why must the signals be random?

[15] Here, y is the raw distance from the central meridian, without the false Easting 500,000 m!

7. Base network and detail survey measurement

7.1 Objective and planning of base network measurement

The *task* of base network measurement is to create, through a *network hierarchy*, the geometric foundation for mapping the country. For this purpose a permanent, sufficiently dense and precise benchmark set is created to which the local measurements of the various user groups will be tied. The co-ordinates of the benchmarks are known in the national co-ordinate reference frame, and by using them, also the locally measured points and drafted maps will be in the same frame.

Benchmarks are used both in detail surveys for mapping, and in the setting out of plans into the terrain — "the inverse problem of mapping".

The *planning* of base network measurements starts from taking inventory of the existing situation and an analysis of needs. The goal is to build a sufficiently precise and dense set of benchmarks, at a minimum cost. It pays off, however, to plan it for the future, especially in choosing the *benchmark substrate* and in *monumentation*. Also, possible future building activity is taken into account in the choice of location, as this may destroy points or render them useless by destroying the visibility conditions for measurement: in case of theodolite measurement intervisibility between points, in case of satellite measurement, a sufficient visibility of the sky from the point.

Reconnaissance is part of the detailed planning of a network: an on-the-spot check that the measurements can be carried out as planned. Before reconnaissance, one does a "map reconnaissance", in which, already in the office, the situation is judged. If new points are to be created, a clear and useable *point description* must be drafted, with the aid of which also others can find the point.

In addition to *precision*, also attention must be paid to *reliability*: reliability is, that possible gross errors are noticed with the greatest possible ease, and that the effect on the end result — the co-ordinate solution — of the greatest possible gross error that remains undetected is as small as possible. To this end, there has to be sufficient *redundancy* in the

network: the measurement plan should always contain enough extra measurements above and beyond the required minimum.

Nowadays, instead of the traditional solution — a triangulation network densified by traverses —, commonly GNSS networks are used. Those too need to be designed right, i.e., hierarchically, and the measurements must be planned so, that the accuracy and point density objectives are achieved in an economic way.

The following alternative methods are on offer for base network measurement:

- satellite positioning (GNSS)
- traditional terrestrial measurement using total stations
- photogrammetric aerotriangulation.

The choice is dictated by the purpose of use (size of area and accuracy requirements) and visibility conditions at the points.

Measurement technology for base network measurement has in recent years undergone a revolutionary change. Traditionally, base network measurements were done in the plane using triangulation and traversing, and precise levelling and lower-order levelling for measuring heights. Nowadays, always satellite positioning is used if at all possible. There are however situations where traditional techniques hold their own, like tunnel and mine surveying where the sky cannot be seen.

A good introduction to the subject is Salmenperä (1998), on which this presentation is partly based.

7.2 Guidance and standards

In order to ensure the quality and effectiveness of geodetic measurements, there exist many different standards and guidance documents. We present here only those with a clear official status.

Kaavoitusmittausohjeet

Important guidance was provided by the Zoning Survey Guide ("*Kaavoitusmittausohjeet*") of the Finnish National Land Survey (Maanmittauslaitos, 2003). The guidance concerns base network measurement, detail survey, aerial mapping, and the drafting of a zoning base map, as well as documenting the work.

Base network measurement, *co-ordinate reference systems*, *map projections*, and *zoning surveying* has of late been the subject of guidance by JUHTA, the Advisory Board on Data Management in Public Administration, who publish the series JHS, Recommendations for Public Administration (Ollikainen, 2013). In 2013 it was decided that from then on, the guidance concerning zoning surveying would be published in the JHS series.

Of the guidance documents relating to these subjects, we may mention

the following on-line publications, unfortunately only in Finnish:

- JHS 196: *Co-ordinates according to the ETRS89 system in Finland* (JUHTA, 2016a).
- JHS 197: *Co-ordinate systems, related transformations and map-sheet division for ETRS89* (JUHTA, 2016b).
- JHS 163: *The Finnish height system N2000* (JUHTA, 2010).
- JHS 178: *The municipal geographic information interface.* This document defines an interface called kuntaGML[1] (a variant of Geographic Mark-up Language) (JUHTA, 2012a).
- JHS 184: *Benchmark measurement in the EUREF-FIN co-ordinate system*. (JUHTA, 2012b).
- JHS 185: *Drafting a base map of a local detailed plan*. This replaces in part the earlier *Zoning Survey Guide* and the *Zoning Base-map Guide* ("*Kaavan pohjakartta*") by the Finnish National Land Survey (JUHTA, 2014).

Kaavoitusmittausohjeet, Kaavan pohjakartta

Important standardization work in the Finnish language area is also *terminology work*. Let us mention the Geoinformatics Word List (*Geoinformatiikan sanasto*) drafted in collaboration by the National Land Survey and the Finnish Terminology Centre TSK.

Guidance and standardization is always a work in progress.

7.3 Network hierarchy and classification

Already during the past couple of decades, almost all base-network measurements are done using the satellite positioning technique — with the exception of precise levelling, for technical reasons, i.e., the geoid problem. Table 7.1 on the next page contains a catalogue of the technologies used back then and now in connection with base network measurement.

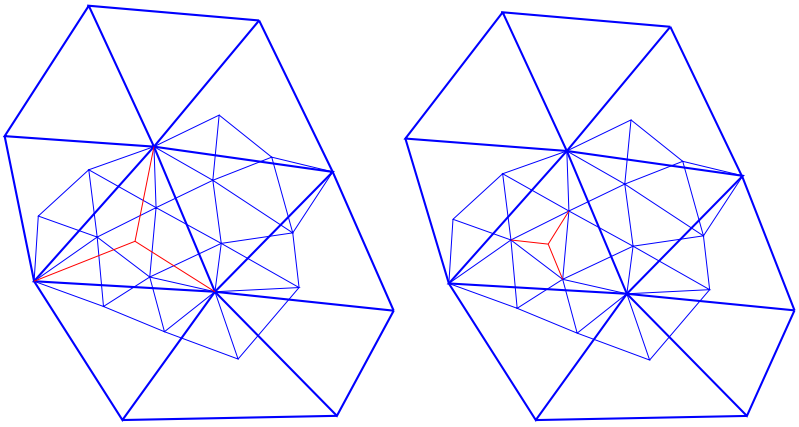
The table also illustrates well the concept of *network hierarchy*: more local networks are always tied to more extended ones, which serve as the former's "formal truth". The order of operations is always *from the large to the small*, first the most extensive networks are measured, which are then *densified* with measurements in a smaller area. In this way, a benchmark set is obtained that

- covers the whole country
- is dense enough: building activity requires starting points sufficiently nearby, at most a few hundred metres from the project area and from each other, and
- is of homogeneous quality.

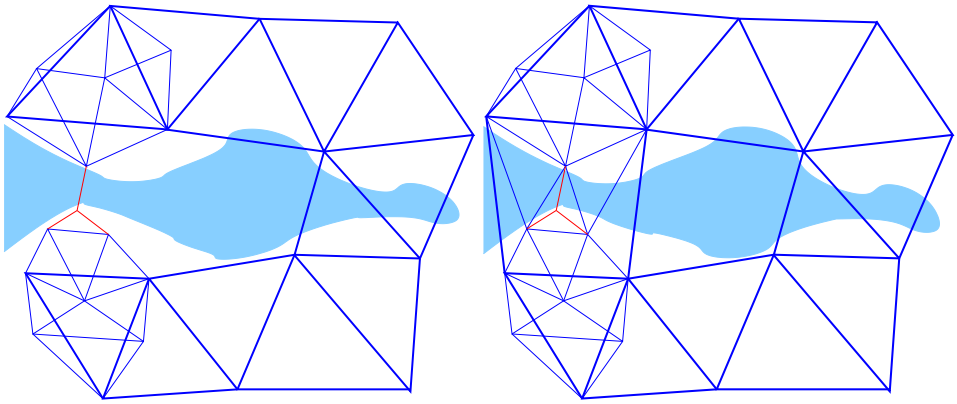
[1] The name translates as "municipalityGML".

Table 7.1. Methods for base-network measurement.

Areal extent	Order old	Order new	Traditional methods	Modern methods
Global	-		-	GNSS, VLBI, satellite laser, DORIS
1000 km	-	E1	-	GNSS, continuously operating (FinnRef™)
100 km	I.	E1b,E2	First-order triangulation	GNSS, EUREF-FIN densification
10 km	II, III	E3	Lower-order triangulation	GNSS, static
1 km	IV, V	E4	Traverses, aero-triangulation	GNSS, aerotriangulation; RTK is not recommended



(a) A hierarchic level is forgotten inbetween



(b) Use of two different hierarchy paths in the same measurement

Figure 7.1. The significance of network hierarchy, and mistakes often made. Not like this... but like this.

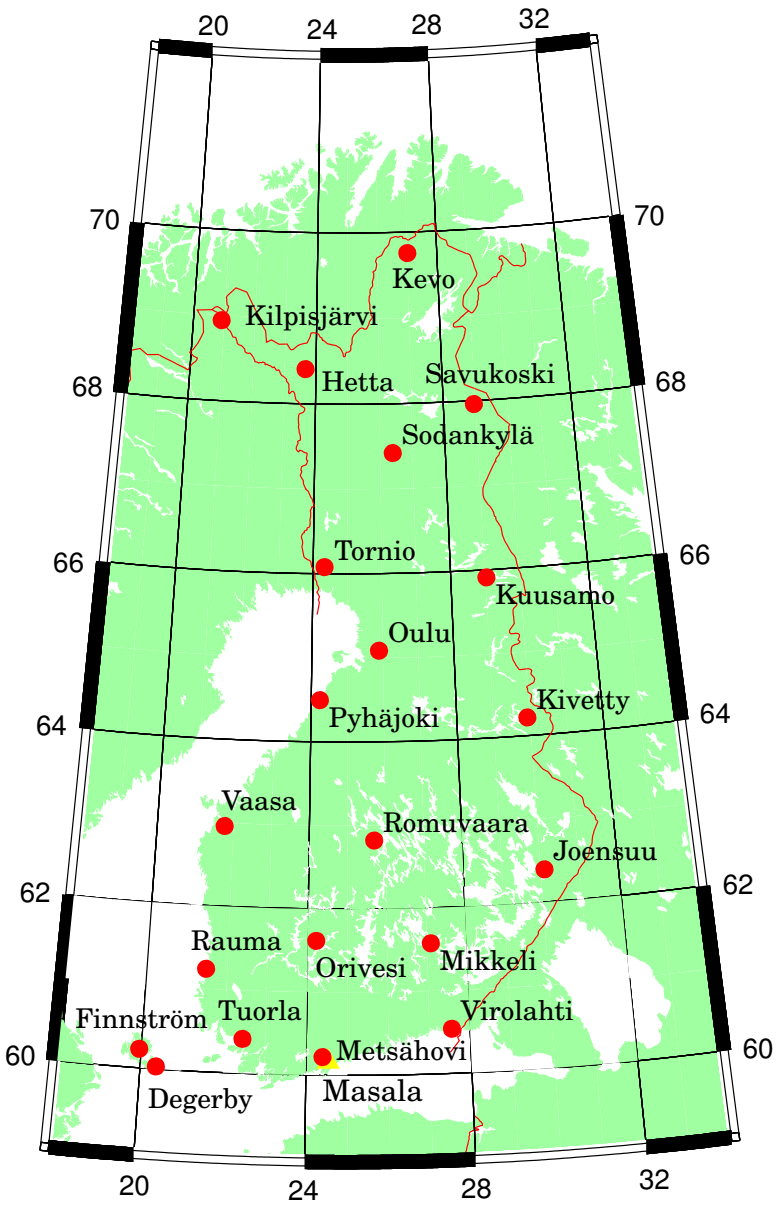


Figure 7.2. The Finnish continuously operating GNSS network FinnRef™. The stations collect GNSS measurements continuously at a rate of one measurement event per second. The computing centre is at the Finnish National Land Survey's Geospatial Research Institute FGI in Masala, Kirkkonummi, 20 km west of Helsinki.

The hierarchical method is meant to prevent the unpleasant situation from occurring in which neighbouring points have co-ordinates determined for them along two different paths, so that the *relative* location precision between them will be weak.

The new measurement technologies mentioned in the table will be discussed in later chapters.

The network of the first-order triangulation mentioned in the table comprises 364 points and covers the whole territory of Finland. The network was measured by the Finnish Geodetic Institute during 1919–

1987. Lower-order triangulations and traverses were measured by the National Board of Survey / National Land Survey. Local measurements were carried out by many players, e.g., municipalities. In a similar way (JUHTA, 2012b) the points of orders E1 and E1b were measured by the Finnish Geodetic Institute, whereas E2 and E3 were measured by the National Land Survey. E4 and the use-point orders E5 and E6 are measured by municipalities.

Nowadays the highest level in the Finnish national GNSS network hierarchy is formed by the continuously operating GNSS network FinnRef™. Earlier on it consisted of 13 stations, in the years 2012 – 1013 a renovation was carried out, bringing it up to 20 stations. The observations are collected by the Finnish National Land Survey's Geospatial Research Institute FGI, the former Finnish Geodetic Institute FGI. In the years 1996 – 1999 the FGI carried out a two-stage EUREF-FIN densification using the static GPS positioning technique, comprising in total some 450 points. The first stage consisted of 100 points, see figure 7.3 and JUHTA (2016a). It was measured in 1996 – 1997. Together with the permanently operating GNSS network FinnRef, it forms the modern order I, or E1. Together they define the EUREF-FIN co-ordinate reference frame. The other EUREF-FIN densification phase, which was measured in 1998 – 1999, comprises 350 points, and was designed to offer easier-to-reach points for practical measurements. Its order is E1b.

The National Land Survey has already for many years carried out base-network measurements using the static GPS positioning technique: there are some 2000 points of order E2, the measurements of which are tied directly to points of orders E1 and E1b. Recently, kinematic GNSS measurement (RTK, *real-time kinematic*) has become common in lower-order base-network measurements, although its suitability for this has been credibly questioned.

A good description of the co-ordinate solutions used in Finland and the relationships between them give Häkli et al. (2009).

7.4 About the terrain, the ellipsoid and the map plane

Geodetic measurement networks, like a triangulation network or a traverse, are in fact three-dimensional networks, figure 7.4. A logical idea is then to also carry out the computation of the network, i.e., the *adjustment*, three-dimensionally: the point co-ordinates are written three-dimensionally in the form of rectangular co-ordinates, and every observable is described as a function of the co-ordinates of those points between which the measurement takes place. This is how one obtains the *observation equations* upon which the adjustment of the network is based.

Three-dimensional network adjustment is a tempting thought, mostly

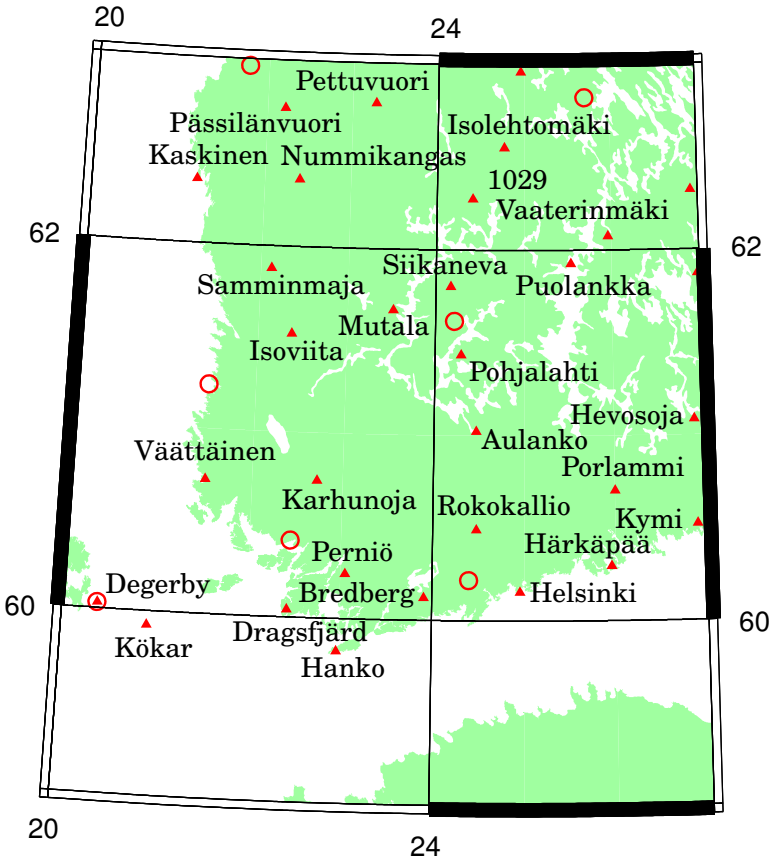


Figure 7.3. The Finnish EUREF-FIN first-stage densification network, detail. These points, together with the FinnRef™ points, form the E1 order.

because of the simplicity of the underlying idea. The formation of observation equations is nevertheless complicated, as the measurements are done, in every point, in *instrument co-ordinates*, i.e., such co-ordinates in which the z axis points upward along the local *plumb line*. The direction of the plumb line, which can be measured by astronomical means, is different in each point, as seen in figure 7.4.

This means that at least in the observation equations for the horizontal angles (azimuths) and zenith angles, the direction of the plumb line must be along. This makes them seriously complicated. For slant ranges, on the other hand, the observation equation is simple.

Terrestrial geodetic measurements are always done close to the physical surface of the Earth, generally between points located on the surface. Thus we may call the network geometry “quasi two-dimensional”, and it would seem to make sense to try and carry out also the computations in two dimensions, on a suitably chosen, mathematically simple *computation surface* close to the Earth’s surface. The Earth’s physical surface with its mountains and depths is however too craggy to serve as a computation surface.

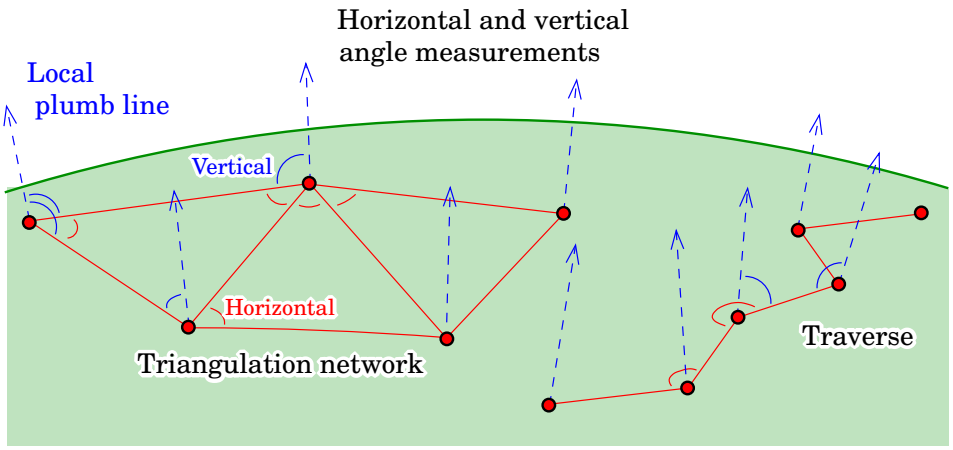


Figure 7.4. A triangulation network and a traverse in space.

More suitable computation surfaces are the *reference ellipsoid*, or — in a small area — the *map projection plane*, figure 7.5. In preparation for computation, the observations are *reduced* to this computation surface.

7.4.1 Adjustment on the reference ellipsoid

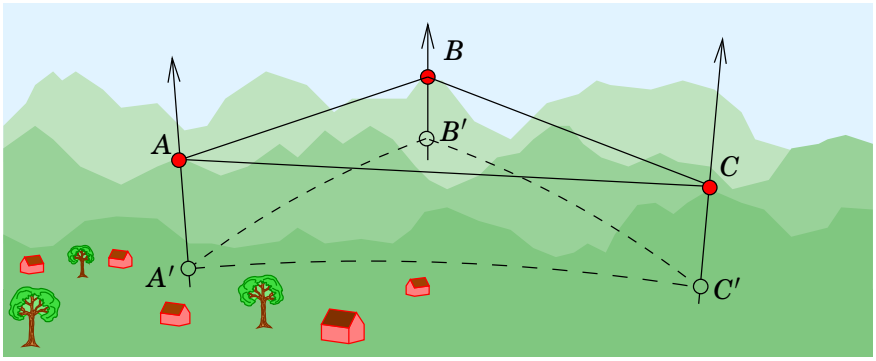
The reference ellipsoid coincides rather well with the Earth's surface, and as a simple mathematical surface it is suitable as a computation surface. Table 7.2 shows the magnitude of the differences between the physical surface of the Earth and the reference ellipsoid[2]. And let us remark still, that mankind lives close to the surface of the solid Earth on land, but close to the sea *surface* on the seas: the impression given by the table exaggerates the thickness of the human living space.

For comparison, the difference between the equatorial and polar radii of the GRS80 reference ellipsoid is already 21.4 km or 0.336 %. A "reference sphere" would be a clearly poorer approximation.

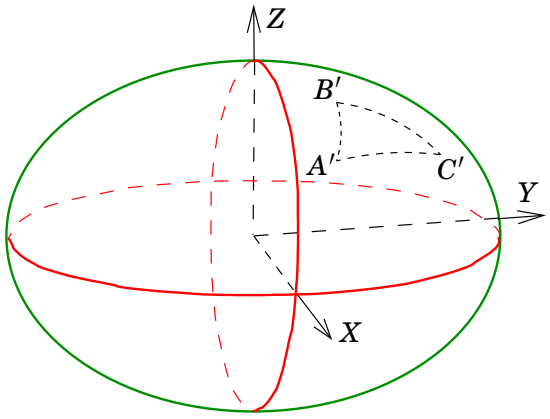
[2]For comparison: the mountain Olympus Mons on the planet Mars is 27 km high above its surroundings, 0.65 % of the radius of Mars. Gravity on Mars is only one third of that on Earth.

Table 7.2. Goodness of approximation by the reference ellipsoid: separation between the physical surface of the Earth and the reference ellipsoid, both in kilometres and in proportion to the Earth's radius.

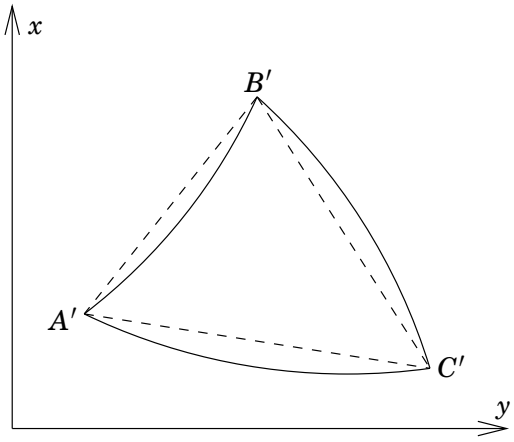
Unit	Highest (Mt. Everest)	Deepest (Mariana Trench)	Land mean height	Sea mean depth
km	+8.8	−11	+0.84	−3.8
%	+0.138	−0.17	+0.013	−0.06



(a) Measurement in the terrain



(b) Projected on the reference ellipsoid



(c) In the map projection plane, "straight lines" are curved

Figure 7.5. Use of a reference ellipsoid and a map projection plane when mapping the Earth.

The reference ellipsoid was widely used as a computation surface already in the 19th century, before the existence of computing machinery and satellite positioning. The mathematics needed is complicated, but the method is more intuitive: terrestial geodetic networks are on the Earth's surface, close to the reference ellipsoid, and local plumb lines, along which the vertical axes of measurement instruments are aligned, are close to the *normal* to the reference ellipsoid surface.

Nowadays base networks are measured with satellite technology, and the traditional method has gone into history. GNSS networks are always adjusted truly three-dimensionally.

7.4.2 Adjustment in the map plane

The adjustment of small, local networks, e.g., traverses, can be done without significant error in the *map projection plane*.

A *map projection* is applied always in such a way that, first, for a terrain point, *geodetic co-ordinates* φ and λ are calculated on the surface

of the reference ellipsoid. This is how the *projection* onto the surface of the ellipsoid is done. Then, the points on the ellipsoid are projected onto the map plane. Of course the curved surface of the ellipsoid cannot be mapped onto the plane without error. The objects projected are *distorted*: directions and distances, and thus also surface areas and volumes, are wrong in the map plane. The map projection is chosen so, that some things that are considered important, are *not* distorted. Some other things then *are* distorted, sometimes badly. E.g., a *conformal* projection maps angles and distance ratios correctly, but, as the classical Mercator projection demonstrates, it can depict surface areas very, very wrongly.

In conformal projections, small objects nevertheless are mapped with their correct shapes: their scale and absolute orientation may be wrong, but their *shape is correct*. For larger objects, the "straight" lines projected from the ellipsoid are curved in the map plane. In the map plane, *directions* may be different from on the surface of the ellipsoid[3] — although the *angles* are identical in conformal projections.

Expressing and solving the adjustment problem in the map projection plane is relatively simple, however it presupposes that

1. the distance measurements are reduced, first to the reference level, then to the map projection plane (so the *scale reduction* of the map projection has been done).
2. The map projection is *conformal*, so the measured horizontal angles are directly useable without a reduction. Projections used for general maps are conformal, like the Gauss-Krüger and UTM projections used in Finland. In Finland, map co-ordinates can thus be used directly in network adjustment.
3. The zenith angles between two measurement points A and B have been measured in both directions, and the angle to be used is the mean of the measurements $\overline{\zeta} = (\zeta_A + \zeta_B)/2$. Then one may calculate heights from the reference level according to rectangular geometry.
4. Points of which the co-ordinates are known, like the starting and closing points as well as the auxiliary points of a traverse, have been projected to the map projection plane using the exact projection formulas.

A visual explanation of this approach is found in figure 7.6.

[3] In the classical Mercator projection they are however identical, a valued property in navigation at sea.

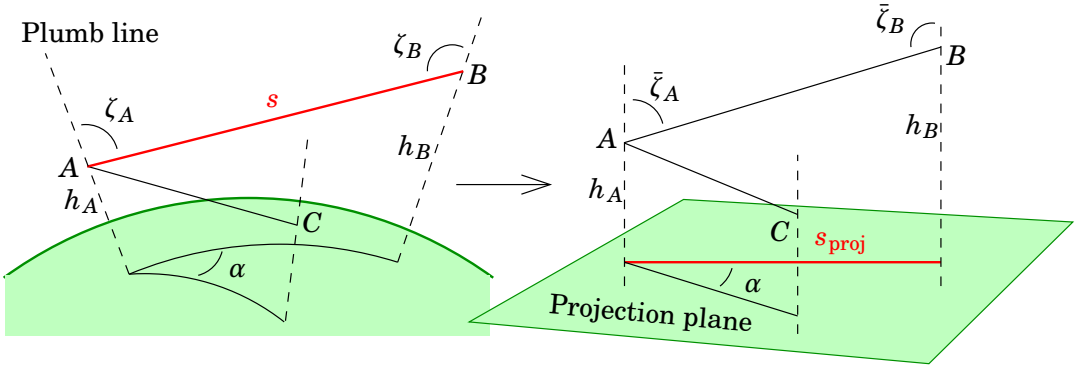


Figure 7.6. Transferring the geometry for adjustment of a small network to the map projection plane.

7.5 Traverse measurement and computation

monikulmiojono-mittaus

A traditional method for densifying lower-order base networks is traverse measurement or *traversing*. In spite of the existence of satellite positioning, and especially real-time kinematic positioning (RTK), there continue to be situations where traversing is the best, or even the only, method of base network measurement. Such situations include underground — mine or tunnel — measurement, and measurements in high-rise urban landscapes — "urban canyons" — where GNSS measurement has problems due to the blocking of signal by buildings, and spurious signal reflections — "multipath" — off buildings.

monitie

Going back in history then, in 1807, the mathematician Nathaniel Bowditch[4] won a competition for the most appropriate way to compute a traverse. The prize was $10 (Cooper, 1982, pages 147 – 150).

The Bowditch method, a traditional separate adjustment of angles and co-ordinates, is suboptimal. Yet is it worth looking at as an example of geodetic computation.

A *traverse* (polygon) is a line of measurement stations, where at every station one measures directions and distances to the previous and the next station. The instrument used is generally a *total station* or tacheometer. The measurement is usually the last stage in bringing co-ordinates (x, y) to the site reference points in the immediate vicinity of objects to be measured.

There are two types of traverses: *open* and *closed*. See figure 7.7. We always know the co-ordinates (x, y) of the starting and closing points.

[4]Nathaniel Bowditch (1773 – 1838) was an American mathematician, student of navigation at sea, and scientific translator and textbook author. He wrote, i.a., "*The American Practical Navigator*", which was published in 1802. The modern version of the book, "*The Bowditch*", can be found on the Internet: http://msi.nga.mil/NGAPortal/MSI.portal?_nfpb=true&_pageLabel=msi_portal_page_62&pubCode=0002.

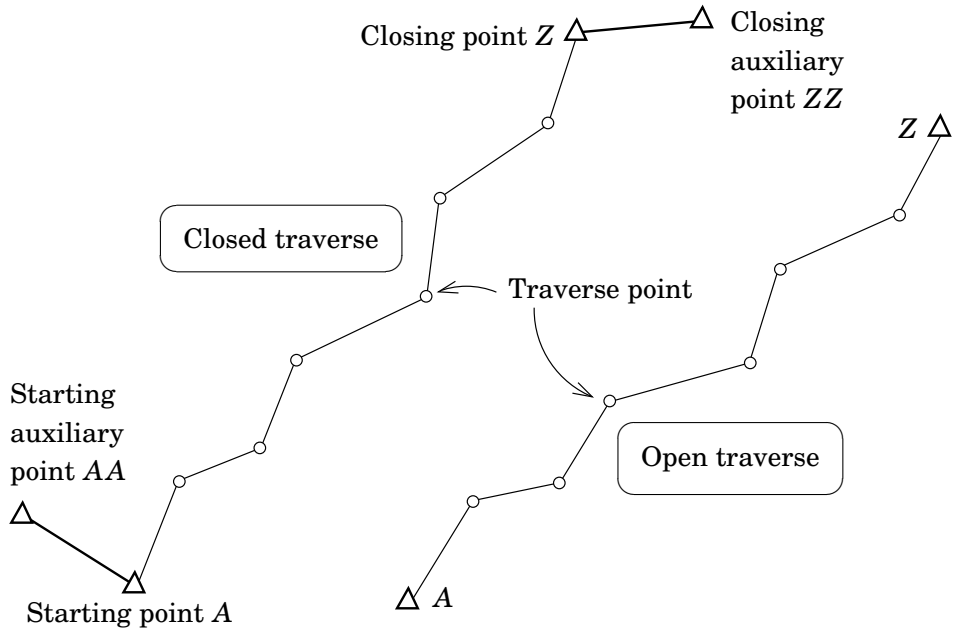


Figure 7.7. An open and a closed traverse.

The difference between an open and a closed traverse is in the use of *auxiliary points* in the latter. These points, the co-ordinates of which are known, help to orient the traverse correctly.

7.6 Open traverse

Generally we try to measure a closed traverse, because it gives the possibility to check and adjust both the measured angles and distances. There are however situations where this is impossible or difficult and where one has to use an *incompletely closed* traverse.

The common case is where we know neither starting nor closing direction. See Figure 7.8.

We know

- (x_A, y_A), starting point
- (x_Z, y_Z), closing point

We need to compute, or *estimate*[5], co-ordinates for the traverse points $\left(\widehat{x}_i, \widehat{y}_i\right), i = 2, \ldots, n-1$.

7.6.1 Starting direction

Because we do not know the starting and closing directions, we cannot adjust directions. Computing the direction α_{12} is not possible as we lack a starting direction. We may obtain one in two different ways:

[5]Using again the "hat" notation for estimators, as well as underscoring stochastic quantities.

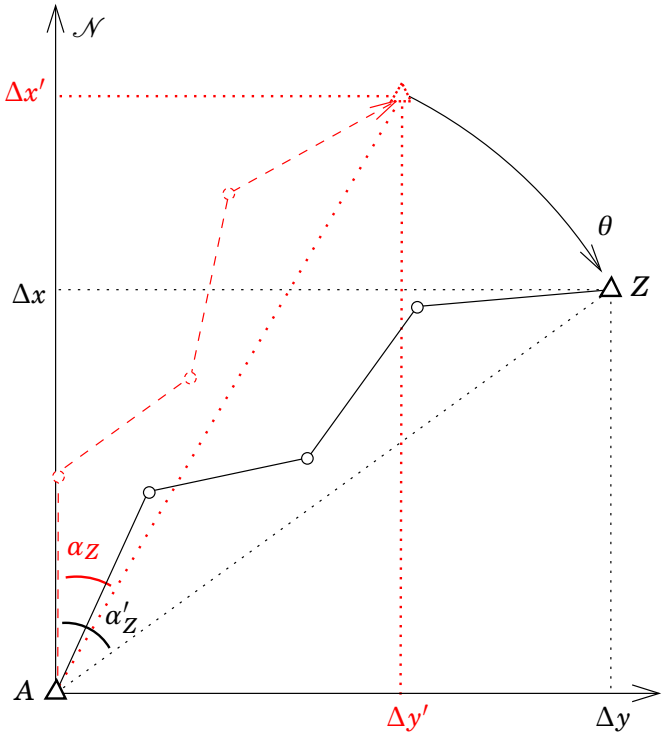


Figure 7.8. An open traverse.

1. *from the observations* by setting the direction angle of the first side $\alpha_{12} = 0$, figure 7.8, or
2. (preferred) by measuring in a map an *approximate value* α'_{12} for α_{12}.

7.6.2 Computing the traverse

We carry out the calculation of directions in the familiar way:

$$\begin{aligned}
\alpha_{12} &= \alpha'_{12}, \\
\underline{\alpha}_{23} &= \alpha_{12} - 200^{\mathrm{g}} + \angle\underline{\alpha}_2, \\
\dots \quad &\dots \quad \dots \\
\underline{\alpha}_{i,i+1} &= \underline{\alpha}_{i-1,i} - 200^{\mathrm{g}} + \angle\underline{\alpha}_i, \\
\dots \quad &\dots \quad \dots \\
\underline{\alpha}_{n-1,n} &= \underline{\alpha}_{n-2,n-1} - 200^{\mathrm{g}} + \angle\underline{\alpha}_{n-1}
\end{aligned}$$

These directions are “final”. We calculate co-ordinates for them using the forward geodetic problem recursively:

$$\begin{aligned}
\underline{x}'_2 &= x_A + \underline{s}_{12}\cos\alpha_{12}, \\
\underline{y}'_2 &= y_A + \underline{s}_{12}\sin\alpha_{12},
\end{aligned}$$

and so forth:

$$\begin{aligned}
\underline{x}'_i &= \underline{x}'_{i-1} + \underline{s}_{i-1,i}\cos\underline{\alpha}_{i-1,i}, \\
\underline{y}'_i &= \underline{y}'_{i-1} + \underline{s}_{i-1,i}\sin\underline{\alpha}_{i-1,i}.
\end{aligned}$$

Finally:

$$\begin{aligned} \underline{x}'_n &= \underline{x}'_{n-1} + \underline{s}_{n-1,n} \cos \underline{\alpha}_{n-1,n}, \\ \underline{y}'_n &= \underline{y}'_{n-1} + \underline{s}_{n-1,n} \sin \underline{\alpha}_{n-1,n}. \end{aligned}$$

This calculation has been carried out in the correct way, but *in a wrongly oriented co-ordinate frame*. The closing errors

$$\begin{aligned} \Delta\underline{x}' &= \underline{x}'_n - x_Z, \\ \Delta\underline{y}' &= \underline{y}'_n - y_Z, \end{aligned}$$

don't tell us anything about measurement errors, but rather about the approximativeness of the assumed starting direction α'_{12}. For this reason, it is not allowed to eliminate the closing errors by adjustment, as is done in a completely closed traverse.

The whole traverse is *rotated* by an angle amount θ, and it should be rotated back by a Helmert or similarity transformation.

Because starting point A is the common turning point of both co-ordinate systems, we may simply calculate the scale ratio $\underline{K}$ and rotation angle $\underline{\theta}$ between them, i.e.

$$\underline{K} = \frac{\sqrt{\Delta x^2 + \Delta y^2}}{\sqrt{\left(\Delta\underline{x}'\right)^2 + \left(\Delta\underline{y}'\right)^2}},$$

$$\underline{\theta} = \arctan \frac{\Delta y}{\Delta x} - \arctan \frac{\Delta\underline{y}'}{\Delta\underline{x}'}$$

– or with the half-angle formula:

$$\theta = 2 \left[\arctan \frac{\Delta y}{\Delta x + \sqrt{\Delta x^2 + \Delta y^2}} - \arctan \frac{\Delta\underline{y}'}{\Delta\underline{x}' + \sqrt{\left(\Delta\underline{x}'\right)^2 + \left(\Delta\underline{y}'\right)^2}} \right]$$

in which

$$\begin{aligned} \Delta x &= x_Z - x_A, \\ \Delta y &= y_Z - y_A, \\ \Delta\underline{x}' &= \underline{x}'_n - x_A, \\ \Delta\underline{y}' &= \underline{y}'_n - y_A. \end{aligned}$$

Let us again construct a transformation:

$$\begin{aligned} \underline{c} &= \underline{K} \cos \underline{\theta}, \\ \underline{s} &= \underline{K} \sin \underline{\theta}, \end{aligned}$$

with the help of which we obtain as the co-ordinates of the traverse points $i = 1, \ldots, n$:

$$\begin{aligned} \widehat{x}_i &= x_A + \underline{c}\Delta\underline{x}'_i - \underline{s}\Delta\underline{y}'_i, \\ \widehat{y}_i &= y_A + \underline{s}\Delta\underline{x}'_i + \underline{c}\Delta\underline{y}'_i. \end{aligned}$$

In matrix form:

$$\begin{bmatrix} \widehat{x}_i \\ \widehat{y}_i \end{bmatrix} = \begin{bmatrix} x_A \\ y_A \end{bmatrix} + \begin{bmatrix} \underline{c} & -\underline{s} \\ \underline{s} & \underline{c} \end{bmatrix} \begin{bmatrix} \Delta\underline{x}'_i \\ \Delta\underline{y}'_i \end{bmatrix},$$

in which

$$\begin{aligned} \Delta\underline{x}'_i &= \underline{x}'_i - x_A, \\ \Delta\underline{y}'_i &= \underline{y}'_i - y_A. \end{aligned}$$

Note that in the matrix

$$\underline{M} = \begin{bmatrix} \underline{c} & -\underline{s} \\ \underline{s} & \underline{c} \end{bmatrix} = \underline{K} \cdot \begin{bmatrix} \cos\underline{\theta} & -\sin\underline{\theta} \\ \sin\underline{\theta} & \cos\underline{\theta} \end{bmatrix}$$

the direction correction $\underline{\theta}$ is dominated by the assumed starting direction, when again the scale correction $\underline{K}$ contains only a correction for the closing error caused by the imprecision of the measurements[6]: even if the starting direction α_{12} were guessed exactly correctly, nevertheless $\underline{x}'_n, \underline{y}'_n$ wouldn't necessarily coincide with the precise point x_Z, y_Z. In the sideways direction however (perpendicular to the traverse) there will be remaining observation error, which is not adjusted away, but more precisely, is "absorbed" into $\underline{\theta}$.

A useful final correctness check is:

$$\begin{aligned} x_Z &\stackrel{?}{=} \widehat{x}_n, \\ y_Z &\stackrel{?}{=} \widehat{y}_n. \end{aligned}$$

In addition to this one should check that $\underline{K}$ is realistically valued, i.e., close to 1. Alternatively one could compute the co-ordinate closing error by putting $K = 1$, and test it against the known precision of distance measurement.

7.7 Closed traverse

We know:

- (x_A, y_A) and (x_{AA}, y_{AA}): starting point A and auxiliary starting point AA;
- (x_Z, y_Z) and (x_{ZZ}, y_{ZZ}): closing point Z and auxiliary closing point ZZ.

See Figure 7.9.

We observe:

- n bending angles $\angle\underline{\alpha}_1, \ldots, \angle\underline{\alpha}_n$

[6] . . . i.e., it is a primitive *adjustment*.

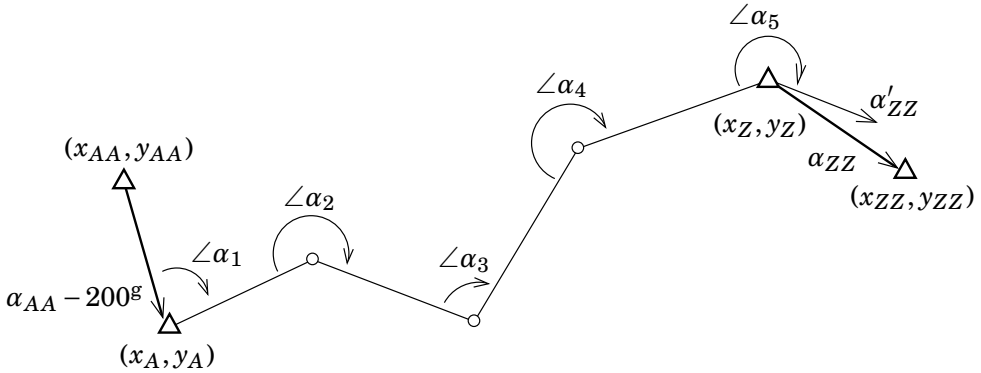


Figure 7.9. The geometry of a closed traverse. Only angles and starting and closing directions marked.

- $n-1$ distances, or side lengths $\underline{s}_{12}, \underline{s}_{23}, \ldots, \underline{s}_{n-1,n}$

We must compute, or *estimate*:

$\left(\widehat{x}_i, \widehat{y}_i\right)$ co-ordinates for traverse points $i = 2, \ldots, n-1$. There are $n-2$ new points.

7.7.1 Computing the auxiliary starting and closing directions

The *auxiliary starting and closing directions* are obtained by solving the inverse geodetic problem:

$$\begin{aligned} \alpha_{AA} &= \arctan \frac{y_{AA} - y_A}{x_{AA} - x_A} + k\pi, \\ \alpha_{ZZ} &= \arctan \frac{y_{ZZ} - y_Z}{x_{ZZ} - x_Z} + k\pi, \end{aligned} \tag{7.1}$$

where $k \in \{0, 1\}$ need to be chosen so the result is in the correct quadrant.

The approximate Bowditch method that we shall use has two stages. First we process the *directions*, then the *co-ordinates*.

7.7.2 Adjustment of directions

Compute the directions or *azimuths* $\alpha_{AA}, \underline{\alpha}_{12}, \ldots, \underline{\alpha}_{n-2,n-1}, \underline{\alpha}_{n-1,n}, \underline{\alpha}_{n,n+1}$, in which we know α_{AA}, and $\underline{\alpha}_{n,n+1}$, which is computed from the observations, corresponds to the above computed auxiliary closing direction

α_{ZZ}. As follows[7]:

$$\begin{aligned}
\underline{\alpha}_{12} &= \alpha_{AA} + \angle\underline{\alpha}_1, \\
\underline{\alpha}_{23} &= \underline{\alpha}_{12} - 200^{\mathrm{g}} + \angle\underline{\alpha}_2, \\
&\cdots\cdots \\
\underline{\alpha}_{i,i+1} &= \underline{\alpha}_{i-1,i} - 200^{\mathrm{g}} + \angle\underline{\alpha}_i, \\
&\cdots\cdots \\
\underline{\alpha}_{n-1,n} &= \underline{\alpha}_{n-2,n-1} - 200^{\mathrm{g}} + \angle\underline{\alpha}_{n-1}, \\
\underline{\alpha}'_{ZZ} \overset{\text{def}}{=} \underline{\alpha}_{n,n+1} &= \underline{\alpha}_{n-1,n} - 200^{\mathrm{g}} + \angle\underline{\alpha}_n.
\end{aligned} \tag{7.2}$$

At the end we thus obtain the value $\underline{\alpha}'_{ZZ}$, which, if all the angle observations $\underline{\alpha}_i$ would be errorless, would be equal to α_{ZZ} computed from co-ordinates by equation 7.1. In reality this is not so, but rather the *direction closing error* amounts to

$$\underline{w}_\alpha \overset{\text{def}}{=} \underline{\alpha}'_{ZZ} - \alpha_{ZZ}.$$

This closing error is removed, or *adjusted*, by dividing it evenly over all bending angles, i.e., to every measured bending angle we apply a *correction*:

$$\delta\underline{\alpha} = -\frac{\underline{w}_\alpha}{n},$$

after which we again carry out the calculation of angles, producing the *adjusted traverse angles*:

$$\begin{aligned}
\widehat{\alpha}_{12} &= \alpha_{AA} + \left(\angle\underline{\alpha}_1 + \delta\underline{\alpha}\right), \\
\widehat{\alpha}_{23} &= \underline{\alpha}_{12} - 200^{\mathrm{g}} + \left(\angle\underline{\alpha}_2 + \delta\underline{\alpha}\right), \\
\cdots \quad &\cdots \quad \cdots \\
\widehat{\alpha}_{i,i+1} &= \underline{\alpha}_{i-1,i} - 200^{\mathrm{g}} + \left(\angle\underline{\alpha}_i + \delta\underline{\alpha}\right), \\
\cdots \quad &\cdots \quad \cdots \\
\widehat{\alpha}_{n-1,n} &= \underline{\alpha}_{n-2,n-1} - 200^{\mathrm{g}} + \left(\angle\underline{\alpha}_{n-1} + \delta\underline{\alpha}\right), \\
\widehat{\alpha}'_{ZZ} &= \underline{\alpha}_{n-1,n} - 200^{\mathrm{g}} + \left(\angle\underline{\alpha}_n + \delta\underline{\alpha}\right).
\end{aligned}$$

7.7.3 Co-ordinate adjustment

Using the thusly adjusted directions we compute, for the points $2, \ldots, n-1$, co-ordinates by applying the forward geodetic problem:

- Point 2:

$$\begin{aligned}
\Delta\underline{x}_{12} &= \underline{s}_{12} \cos\widehat{\alpha}_{12} \\
\Delta\underline{y}_{12} &= \underline{s}_{12} \sin\widehat{\alpha}_{12}
\end{aligned}$$

[7] *For the first bending angle* $\underline{\alpha}_1$ we do not subtract 200 gon! And to keep the directions $\underline{\alpha}_i$ in the interval $[0, 400\,\mathrm{gon})$, you may sometimes have to add or subtract 400 gon.

with the aid of which

$$\begin{aligned} \underline{x}_2 &= x_A + \Delta\underline{x}_{12} \\ \underline{y}_2 &= y_A + \Delta\underline{y}_{12} \end{aligned}$$

- General point i: (and also closing point $i \to n$)

$$\begin{aligned} \Delta\underline{x}_{i-1,i} &= \underline{s}_{i-1,i} \cos\widehat{\alpha}_{i-1,i} \\ \Delta\underline{y}_{i-1,i} &= \underline{s}_{i-1,i} \sin\widehat{\alpha}_{i-1,i} \end{aligned}$$

with the aid of which

$$\begin{aligned} \underline{x}_i &= \underline{x}_{i-1} + \Delta\underline{x}_{i-1,i} \\ \underline{y}_i &= \underline{y}_{i-1} + \Delta\underline{y}_{i-1,i} \end{aligned}$$

- Generally (remember that $x_A = x_1$ and $y_A = y_1$):

$$\begin{aligned} \underline{x}_k &= x_A + \sum_{i=2}^{k} \underline{s}_{i-1,i} \cos\widehat{\alpha}_{i-1,i}, \\ \underline{y}_k &= y_A + \sum_{i=2}^{k} \underline{s}_{i-1,i} \sin\widehat{\alpha}_{i-1,i}. \end{aligned}$$

By substitution of $k \to n$ we obtain the equations and co-ordinates $\underline{x}_n, \underline{y}_n$ of the *closing point*.

If the observations were errorless, we would have $\underline{x}_n = x_Z$ and $\underline{y}_n = y_Z$, but they are not. The *co-ordinate closing errors* are:

$$\begin{aligned} \underline{w}_x &\overset{\text{def}}{=} \underline{x}_n - x_Z, \\ \underline{w}_y &\overset{\text{def}}{=} \underline{y}_n - y_Z. \end{aligned}$$

Closing errors are adjusted by giving a *weight coefficient* $q_{i-1,i}$ for each point interval, and the closing errors $\underline{w}_x$ and $\underline{w}_y$ are distributed over the point intervals in proportion to these weight coefficients.

> *Small weight coefficient* $\Longleftrightarrow$ *large weight,*
> *large weight coefficient* $\Longleftrightarrow$ *small weight!*

We compute the *standard correction* corresponding to the sum of weight coefficients, separately for the x and y co-ordinates:

$$\begin{aligned} \delta\underline{x} &= -\frac{\underline{w}_x}{\sum_{i=2}^{n} q_{i-1,i}}, \\ \delta\underline{y} &= -\frac{\underline{w}_y}{\sum_{i=2}^{n} q_{i-1,i}}. \end{aligned}$$

We obtain for the adjusted co-ordinates

$$\begin{aligned} \widehat{x}_i &= \widehat{x}_{i-1} + \underline{s}_{i-1,i} \cos\widehat{\alpha}_{i-1,i} + q_{i-1,i}\delta\underline{x}, \\ \widehat{y}_i &= \widehat{y}_{i-1} + \underline{s}_{i-1,i} \sin\widehat{\alpha}_{i-1,i} + q_{i-1,i}\delta\underline{y}, \end{aligned}$$

Table 7.3. Traverse computation template according to the Bowditch method.

i	$\angle\underline{\alpha}_i$	$\underline{\alpha}_{i,i+1}$	$\delta\underline{\alpha}_{i,i+1}$	$\widehat{\alpha}_{i,i+1}$
A		345,3750		
1	212,2345	157,6095	+7	157,6102
2	151,4565	109,0660	+15	109,0675
3	221,9823	131,0483	+22	131,0505
4	175,9831	107,0314	+29	107,0343
5	165,3467	72,3781	+37	72,3818
Z		72,3818		
Closing error $\underline{w}_\alpha$		-0,0037	↑	

i	$\widehat{\alpha}_{i,i+1}$	$\underline{s}_{i,i+1}$	$\underline{s}\cos\widehat{\alpha}$	$\underline{s}\sin\widehat{\alpha}$	$\underline{x}_i$	$\underline{y}_i$	$\delta\underline{x}_i$	$\delta\underline{y}_i$	$\widehat{x}_i$	$\widehat{y}_i$
1	157,6102	502,345	-395,038	+310,315	1000,235	256,256				
2	109,0675	487,241	-69,164	+482,307	605,197	566,571	+9	-36	,206	,535
3	131,0505	445,981	-209,001	+393,977	536,033	1048,878	+18	-73	,051	,805
4	107,0343	512,125	-56,472	+509,002	327,032	1442,855	+26	-109	,058	,746
5	72,3818				270,560	1951,857	+35	-146	,595	,711
Z					270,595	1951,711				
			Closing errors $\underline{w}_x, \underline{w}_y$		-0,035	+0,146	↑	↑		

or, counting from the starting point A:

$$\widehat{x}_k = \widehat{x}_A + \sum_{i=2}^{k} \left(\underline{s}_{i-1,i}\cos\widehat{\alpha}_{i-1,i} + q_{i-1,i}\delta\underline{x}\right),$$

$$\widehat{y}_k = \widehat{y}_A + \sum_{i=2}^{k} \left(\underline{s}_{i-1,i}\sin\widehat{\alpha}_{i-1,i} + q_{i-1,i}\delta\underline{y}\right).$$

7.7.4 Computation template

The whole computation may be carried out using the template of table 7.3, which is readily automated. Let it be given that $\alpha_{AA} = 145.3750\,\text{gon}$ and $\alpha_{ZZ} = 72.3818\,\text{gon}$, as well as the co-ordinates of the starting point A and the closing point Z. $n = 5$. Weighting used: uniform. The values for the starting and closing points are underlined, observations are in black, computed values in red, closing errors and adjustment corrections in blue. This diagram is good enough for practical work in local base network measurement.

7.7.4.1 Remarks

- The adjustment method described above is *approximate,* i.e., sub-optimal. In a proper least-squares adjustment the direction and co-ordinate corrections are computed simultaneously, as they *correlate* with each other. The results will then be slightly different.

  Nevertheless, if the traverse is relatively straight — i.e., the sides are in the same direction —, the suboptimality will be small and

a proper adjustment would only lead to small changes in the final co-ordinates. On the other hand it is not desirable that the sides are in *precisely* the same direction: that would make finding gross errors more difficult.

- Here, the computation of the traverse was done *in the plane*. The measurements however have been obtained three-dimensionally, *in space*. This requires that instrument and signal are precisely and correctly *centred* and *levelled*, and that all relevant *reductions* (like slope, reference level, etc., see earlier chapters 5 and 6) to the observations have been done. Only then can it be said that
  - the horizontal angles are plane angles, and
  - the reduced distances are horizontal distances.
- In computing plane co-ordinates, directions are always referred to the *map North*, and also distances are reduced to the map plane, i.e., the scale distortion caused by the map projection used has been accounted for.

7.8 Detail survey

Detail survey (Kahmen and Faig, 1988, pages 285 – 303) is the stage of the measurement process, based on base network measurements, aimed at mapping details in the terrain. It is the most laborious stage of the whole mapping project. A detail survey consists of collecting the data and of its processing into the desired end product: a map or a digital geospatial data set. In the processing stage of detail surveying, the results of the base network measurement are along, assuring the geometric correctness of the result.

In the following we describe in more detail four classical methods, *right-angle survey*, *tie-in survey*, *radial survey*, and *free-stationing survey*.

In local detail surveys, a popular method is also *real-time kinematic* (RTK) satellite positioning, which is often competitive. Its use is however difficult, e.g., in a high-rise urban landscape, “urban canyons”, and impossible underground in tunnel and mine surveying.

7.8.1 Right-angle survey (prism surveying)

For this are needed a measuring tape, a double pentagon prism, ranging rods[8] to mark out mapping lines, and drawing paper, see figure 7.10.

The measurement is carried out according to figure 7.11. A and B are known points, often traverse points from a lowest-order survey. The right angles are created using a *double pentagon prism*: when one stands

[8] German *fluchtstab*, French, Dutch *jalon*.

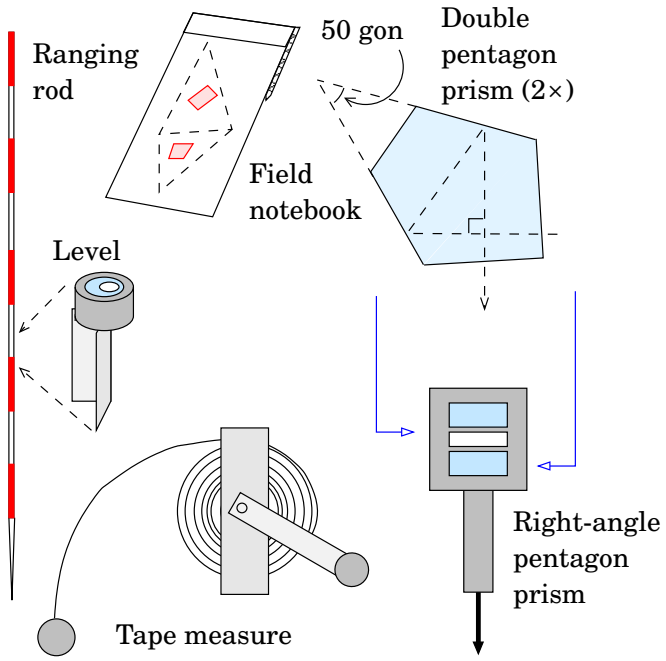


Figure 7.10. Tools of the right-angle survey method.

on the line AB, both end points (or rather, the ranging rods set up on them) show on top of each other in a device containing two pentagonal prisms. One of them looks by an angle 100 gon to the right, the other by the same angular amount to the left, i.e., in the opposite direction.

The distances b may not be longer than one tape length (50 m).

On the right-hand side of the figure is shown how a building is measured in the right-angle method.

One should always take pains that there is sufficient *redundancy* or control, in order to identify mistakes. In this example, the wall measures of the house could be measured.

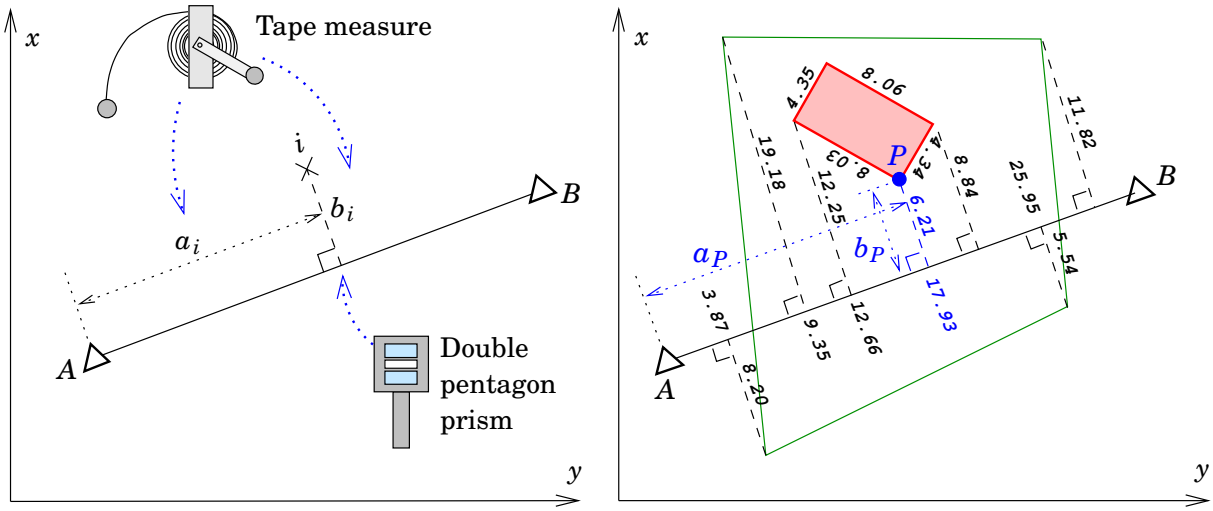


Figure 7.11. Right-angle survey.

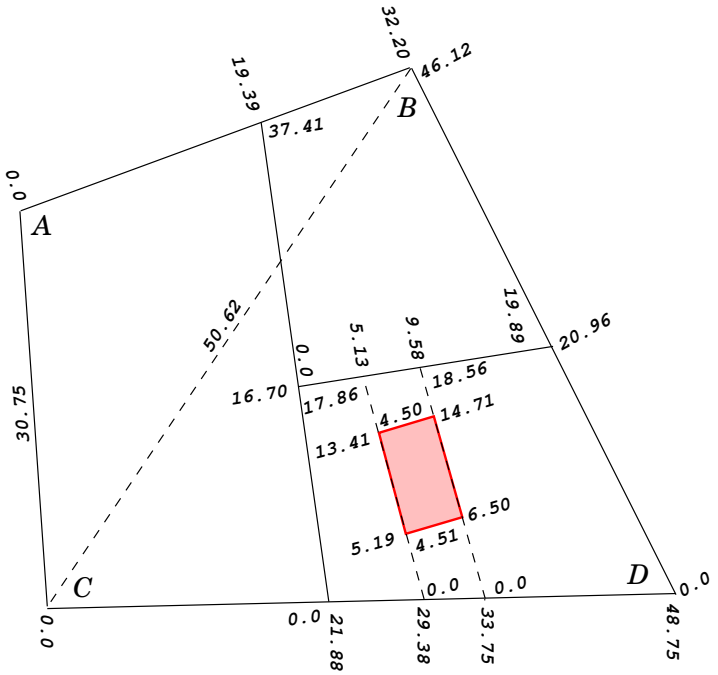


Figure 7.12. Tie-in survey. A cross-measure and the wall measures of the house serve as checks.

The measurements with their number values are written on a field sketch, preferably neatly and systematically, in a way that will be intelligible also to others besides the draftsman at the moment of drawing.

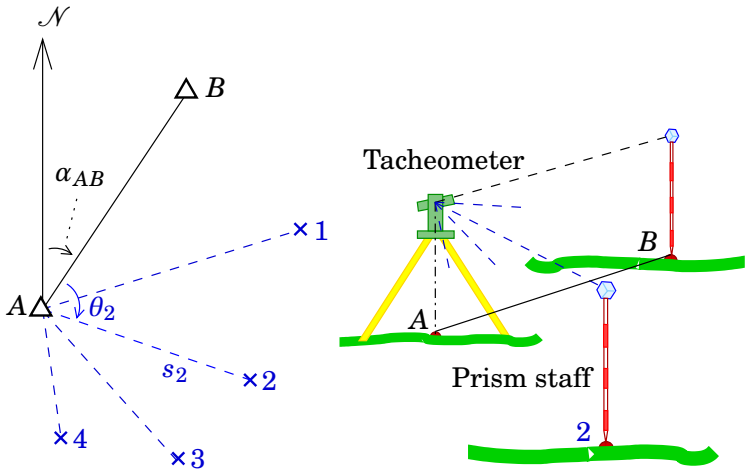
7.8.2 Tie-in survey

Sometimes a method is used in which the measurement is carried out only with distance measurements in several densification steps ("Tie-in survey", Kahmen and Faig, 1988), figure 7.12. In this example were measured a cross-measure (dashed line) of the parcel and the wall measures of the house as a check. In this way one can already in the field find out if there is an error in the measurements. A crude check can be done graphically.

More often, a mixed form is used where the tie-in method *complements* the right-angle survey method.

7.8.3 Radial survey

Radial survey is explained in figure 7.13. Determination of the location in the plane of the points $i = 1,2,\ldots,n$ is done by means of measuring the angles θ_i and the (slant) ranges s_i In the figure, the example point is $i = 2$. From a point the co-ordinates of which are known — generally a traverse point – A the horizontal angle between another known *auxiliary point* B and the point to be determined, P. The orientation direction α_{AB} is calculated by the inverse geodetic problem from the given location co-ordinates of points A and B.



(a) Measurement geometry

(b) Execution

Figure 7.13. The radial survey method.

After this

$$x_i = x_A + \bar{s}_i \cos(\alpha_{AB} + \theta_i) = s_i \sin\zeta_i \cos(\alpha_{AB} + \theta_i),$$
$$y_i = y_A + \bar{s}_i \sin(\alpha_{AB} + \theta_i) = s_i \sin\zeta_i \cos(\alpha_{AB} + \theta_i),$$

Here, the measured slant range has been reduced to the *horizontal distance* $\bar{s}_i = s_i \sin\zeta_i$, in which ζ_i is the vertical or zenith angle, which also should be measured.

The equipment to be used is an electronic tacheometer or total station equipped with suitable software. The instrument is chosen based on the accuracy objective of the mapping to be undertaken.

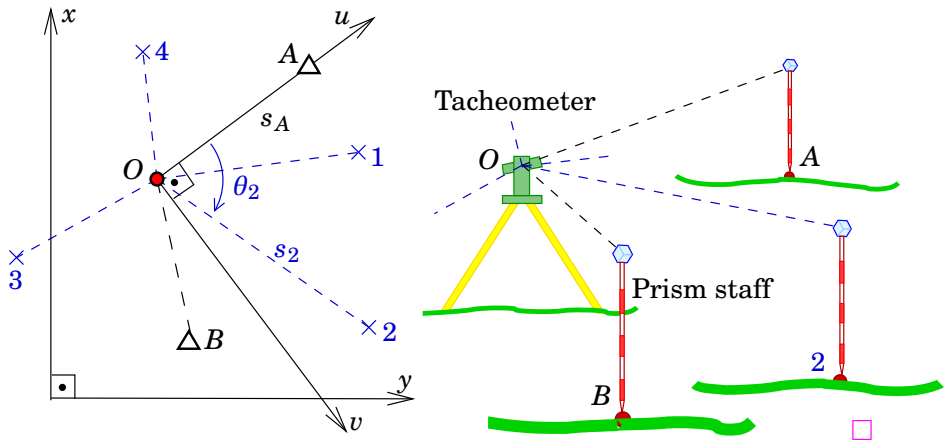
An advantage of radial survey is, that one gets trigonometric height determination thrown in on the deal: if the device measures both the zenith angle ζ_i and the slant range s_i, one obtains, in addition to the horizontal distance $\bar{s}_i$, also the third co-ordinate:

$$z_i = z_A + s_i \cos\zeta_i.$$

The height of either point A or point B must be known as the starting value for the height computation. If the height of point B is given, it is not even necessary to measure the instrument height of the tacheometer over marker A, because the height of the prism staff will not change during measurement.

Radial surveying is a *numerical mapping method*, in which the map product is generated computationally based on the numerical mapping measures θ_i and s_i using their metadata. As the measurements are collected electronically, this is inexpensive and readily automated when the number of points to be measured is large and accuracy requirements high.

Figure 7.14. The free-stationing survey method.



(a) Measurement geometry **(b)** Work mode in perspective

Especially in surveying urban areas, the radial surveying method is practicable, because the setting out of mapping lines required in the right-angle method can be cumbersome due to traffic. Also objects of which both plane and height information are required (special technical measurements, building and utility-line surveys) and busy work sites are suitable for radial survey.

7.8.4 Free-stationing survey

In free-stationing survey — German: *freie Standpunktwahl* — the tacheometer is placed on a *freely chosen* point in the terrain, subject only to good visibility to the points to be measured as well as at least two, preferably three to four, points of which the co-ordinates are known, normally base-network points. The advantage of the method is, that the instrument does not have to be precisely centred over a known point marker or monument: the need for *centring*, and measuring instrument height, goes away. The work proceeds faster.

The method has become widespread with the availability of electronic tacheometers and increased computing power. In principle, however, the method could be used with theodolite and measuring tape. See figure 7.14. *Note* that there is no point marker (monument) under the tacheometer!

Let the co-ordinates of the points $A\,(x_A, y_A)$ and $B\,(x_B, y_B)$ be known. The instrument is set up on the unmarked point O. Observations are made to the points A, B and the unknown points $i = 1, 2, 3, \ldots, n$ (the example point in the figure is $i = 2$):

- horizontal angles $\theta_i = t_i - t_A$
- distances s_i.

The measurement yields, in local or instrument co-ordinates (u,v):

$$\begin{aligned} u_A &= s_A, & v_A &= 0, \\ u_i &= s_i \cos\theta_i, & v_i &= s_i \sin\theta_i. \end{aligned}$$

Now we can use the *forward geodetic problem* to compute co-ordinates for all points $A,B,1,2,\ldots,i,\ldots,n$. The co-ordinates are $(u_A,v_A),(u_B,v_B)$, (u_i,v_i), in a co-ordinate frame the u axis of which is OA.

Using the Helmert transformation (section 3.6) we now *transform* the instrumental co-ordinates (u_i,v_i) to terrain co-ordinates (x_i,y_i). As we know, we can solve for the unknown parameters of the Helmert transformation if minimally, the co-ordinates of two points, A and B, are known in *both* frames.

7.9 Carrying out a detail survey

Detail surveys can be carried out as topographic surveys, in which case they often cover limited areas. The instrument to use then is the *electronic tacheometer*. As an alternative there is GNSS-RTK, — the real-time kinematic method — or aerial mapping, which however may not be always suitable on their own due to lack of visibility in the terrain. In local measurements, traditional prism and tape measurement may be considered, but is used nowadays less due to its low productivity.

7.9.1 Data to be collected, mode of operation

From every measurement station, measure three-dimensionally, for every point to be measured, the horizontal direction (θ), zenith angle (ζ), and slant range (s). The instrument calculates itself the topocentric rectangular co-ordinates (x,y,z) and performs simple checks. When the whole object has been mapped, a draft printout is taken in the site office, and an overall quality check is carried out.

The following data is collected:

General information: site, date, time, weather, observer. Identifying codes as needed (subsection 7.9.2).

From each station point:

- station point (number, type), instrument height
- auxiliary points (number, type), horizontal direction, zenith angle, distance, prism height
- survey points (number, type), horizontal direction, zenith angle, distance, prism height.

7.9.1.1 Work phases

- Choose terrain points based on terrain and purpose of use of the measurement. E.g., if one wants to form a precise terrain model of an uneven area, points need to be collected at sufficient density.

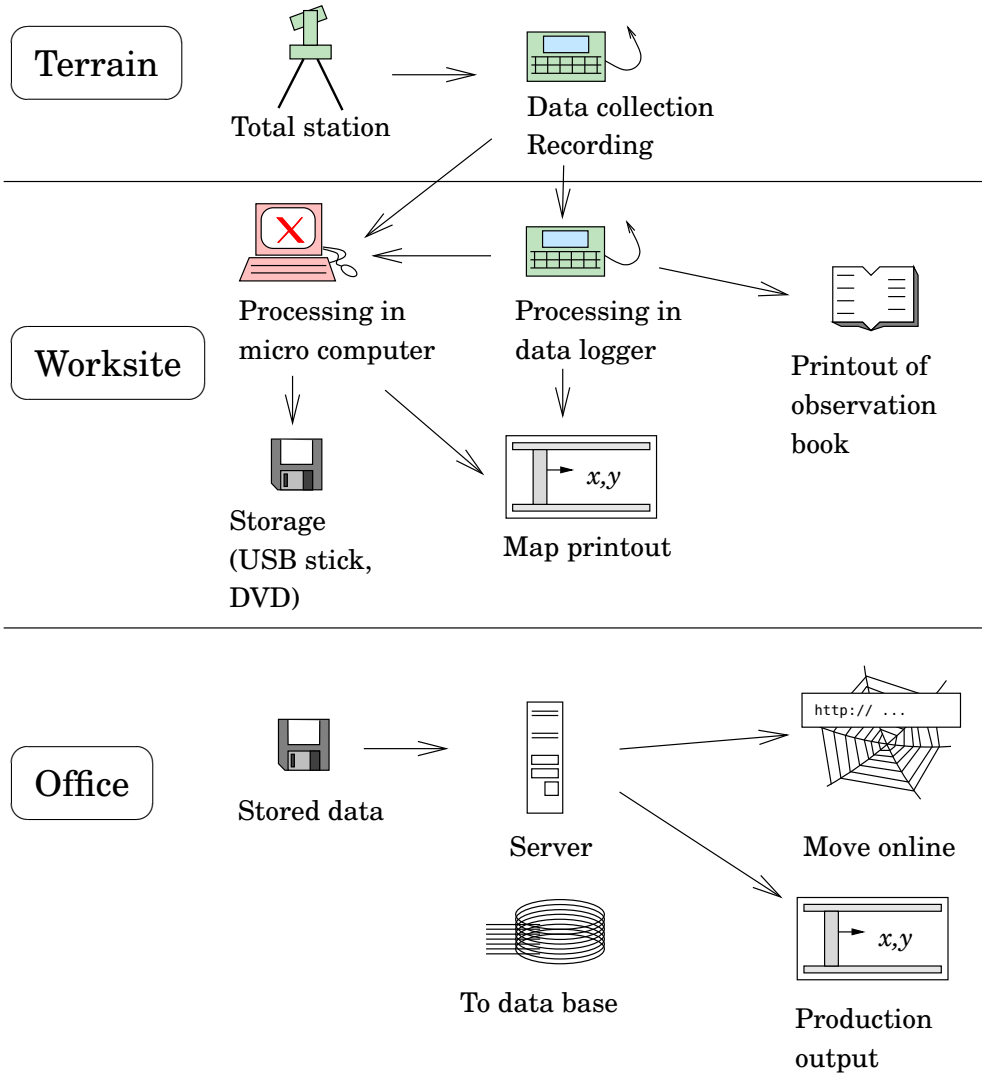


Figure 7.15. Workflow diagram of detail survey. The figure shows some archaic technologies like pen plotters and diskettes, that time has left behind although work stages persist.

- Carry out the measurement: collection and preprocessing.
- Process the material.
- Present and archive the result. The result is a report on the measurement work, containing, e.g., a description of methods used, measurement conditions, point co-ordinates and their estimated accuracy, a draft map, and possibly calculations of areas or volumes or other relevant measurement results.

The working mode is fully digital.

7.9.1.2 Equipment and software

- The processing capacity of total stations is sufficient for many uses. However,
- more and more, one sees a standard tablet or similar, loaded with versatile software, control the total station wirelessly.

- The tablet should preferably be ruggedized for terrain use.
- Software guides the whole observation workchain in the field.
- Software enables *collection*, *testing*, *processing* and *reporting* in the field.

Also in the computations for topographic surveying, the separate phases of *base network measurement* and *mapping* should be distinguished. Both have their own routines.

7.9.2 Encoding of field data

A *topographic information system* is a special case of a geographic information system. It serves the efficient collection of topographic information by geodetic means for further processing, and thus differs from general geographic information systems.

As an extreme example of a topographic information system may be mentioned the proposal by the Finnish Ministry of Agriculture and Forestry for a national topographic information system Karlsson (2015). It is defined as consisting of the following components:

- the national co-ordinate and height system [intended are probably horizontal and vertical reference frames]
- the *topographic database* and information management systems related to it
- all information corpora and information products collected or produced for the maintenance of the topographic database, and services maintained in order to obtain suggestions for its maintenance
- all those information and service products that the administrator of the topographic information system produces in order to bring the topographic database on-line and improve its useability.

Of course most topographic information systems will not be on the national level.

Topographic survey data is collected in numerical form, and to it belongs, besides point measurement data, also so-called *metadata*: data describing data.

The concept of metadata may be described with the aid of a topographic map: much more has been depicted on the map than just the points measured. Points form *objects*, linear (roads, streets, waterways, ...), area shaped objects (parcel boundaries, buildings, forests and fields, ...) or three-dimensional (hills and valleys, terrain forms). Everything is depicted in different ways onto the map, and the way of depiction is documented in the map's *legend*[9]. The legend is thus in a way the metadata of an ordinary paper map.

[9] Legenda, Latin: *what can be read.*

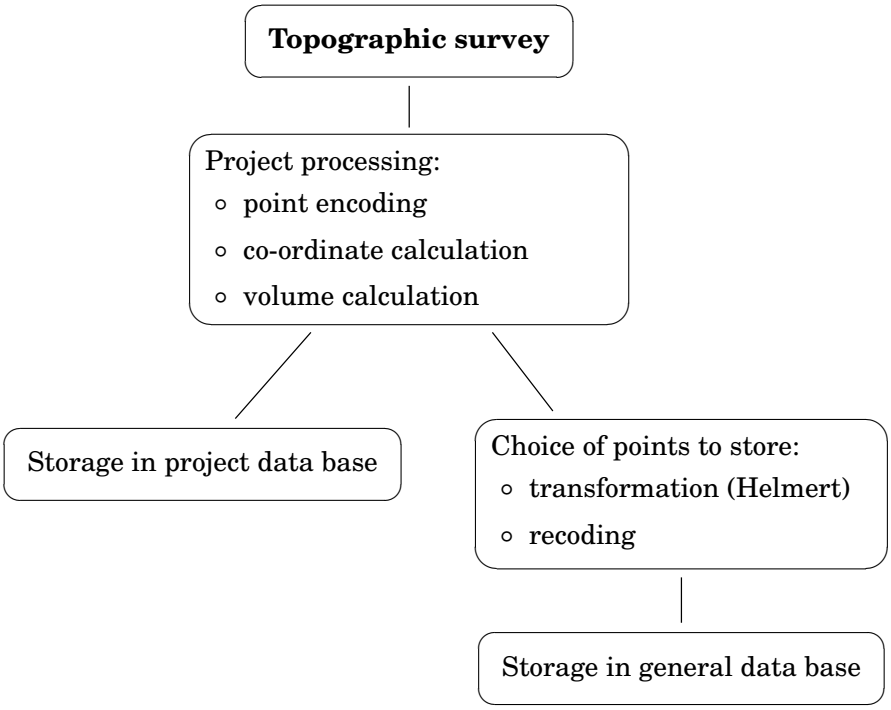


Figure 7.16. The encoding process for topographic data. The information is stored into a general data base, on a case-by-case basis, for later use.

Documenting the measurements already in the measurement phase requires that, at the same time, *also the metadata is recorded*: does this point belong to a parcel boundary, is it the edge of a road, is it a tree (and which species), or is it just a height point in the terrain from which height curves or earthwork volumes will be calculated. To this end, *encoding methods* or *catalogues* have been developed, making the transfer of data easier and as automatic as possible[10].

Some things must be encoded *in the terrain* while doing the topographic survey:

- the point numbers of points measured in the terrain. These can also be generated (semi-)automatically
- point classification codes
- classification of other objects consisting of points, like lines and areas
- certain identifying attributes.

In connection with computation and data base entry, the encoding may be *supplemented* in some respects, e.g., with topology data. The encoding carried out in the terrain is also not final for the following reasons:

1. not all terrain points will be entered into general data bases.

[10] See, e.g., http://www.opengeospatial.org/standards/cat.

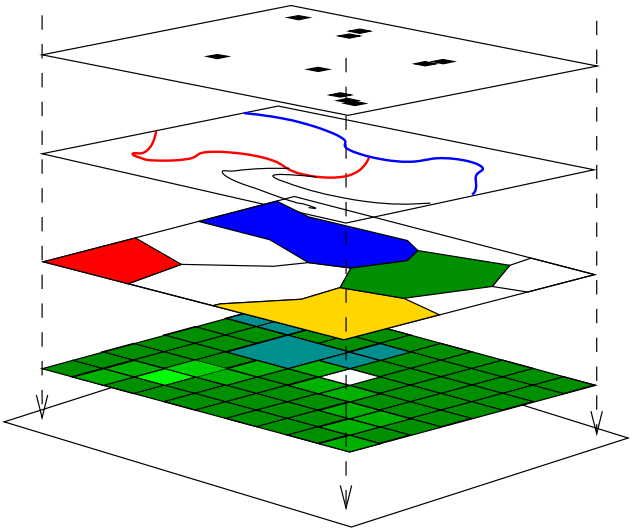


Figure 7.17. Attribute data of objects in multiple layers. With a geographic information system, spatial data can be efficiently combined, analyzed, refined.

□

2. A project specific encoding is not suitable for general use.

Often, project specific data isn't entered into a general data base at all, but the area is mapped again when the work is done — "as-built" -mapping. The idea is, that one maps the finished situation, so no confusiun can arise between what was realized, and what was only planned, but, after a differing realization, was never re-measured.

See figure 7.16.

Kaavoitusmittausohjeet

Old encodings in widespread use are based in the classification of the Zoning Survey Guide, aimed at producing a map. The main starting point is *shared use* of geospatial information by the various players in the field.

Geographic information technology creates possibilities for this and many other efficient uses, e.g.,

- the locations of objects are given in the same common co-ordinate frame. Earlier this system was KKJ and its map projections, see subsections 3.2.1 and 3.3.1. Today, it is always EUREF-FIN and its various map projection co-ordinate frames, see subsections 3.2.3 and 3.3.3. The transformations between the co-ordinate frames are known. This facilitates the combination of different objects into the same system, producing added value.
- Many methods and tools are on offer for combining, analyzing and refining data from many sources.
  - Often, various attribute data are presented on different *data layers* of a digital map, which can be processed together using various operators: *map algebra*. See figure 7.17.

Table 7.4. Classification of topographic data.

Natural data	Cultural data
Topsoil type	Properties
Earth's surface forms	Buildings, structures
Soil, bedrock	Street and utility networks
Vegetation	Zoning
Waters	Street names, local names

Location data	Attribute data
Co-ordinate data (where)	Identifying a.
Geometry data (what shape)	Locating a.
Topology data (*relations with neighbours*)	Timestamp a.
	Descriptive a.

  - The information carried by objects can be sorted and classified according to different attributes, e.g.:
    * all GNSS points on a map sheet
    * the drilling points in square $(x_a, y_a) - (x_b, y_b)$
    * the manhole covers of the municipal sewer network.
- Spatial data may be visualized and, in this way, also made available to people who are not mapping professionals.

Topographic data can be classified *by content* into two main information types: natural data and cultural data (Salmenperä, 1998, sivut 83 – 84). Another way of classifying topographic data is as either location data, or attribute data. See table 7.4.

Part of cultural attributes are invisible in the terrain, like place names, ownership, parcel boundaries, zoning, historical details, etc.

Self-test questions

1. What is the task of base-network measurement?
2. How does network hierarchy work? Why is it important? What could go wrong if not properly done?
3. What are commonly used computation surfaces for geodetic network computation?
4. What is the difference between an open and a closed traverse? How many closing errors (redundancy, degrees of freedom) are there in each, which can be used for checking for mistakes?
5. Explain why aerial photogrammetry cannot be the *only* method for executing detail surveys.

6. Explain why GPS — e.g., real-time kinematic GPS, RTK — cannot be the *only* method for executing detail surveys.
7. What is metadata, and why is it important? Give an example of metadata.

8. Construction surveying

8.1 Zoning plans and setting out

The types of zoning plan in use are regulated by law, in Finland the *Maankäyttö- ja rakennuslaki* ("Land Use and Building Act") 1999/132[1]. They are *local detailed plan* and *the local master plan*[2]. Both are approved by the municipality. Additionally there are still *regional plans*, which are plans at a higher level.

A local master plan gives the outlines of a land use plan for a municipal area. The plans comprise zoning base maps, the scale of which varies between 1 : 20,000 – 1 : 10,000 and 1 : 5000 – 1 : 4000.

Kaavoitusmittaus-ohjeet

According to the new Zoning Survey Guide (JUHTA, 2014), *three measurement classes* are defined. Every measurement class has a corresponding recommended scale for the zoning base map. Digital map products have no actual scale, but the precision at which map material is collected need to be in accordance with the recommended scale. The co-ordinate and height reference frames are the new EUREF-FIN and N2000, and as the map projection, Gauss-Krüger is used: ETRS-GKn, where n is the longitude of the municipality as an integer number.

1. To the *first measurement class* belong local detailed plan areas that are built-up areas where land is extremely valuable, and where there is a local detailed plan with a binding parcel division, or a building ban aimed at drafting such a plan. The base map scale of the local detailed plan is 1 : 500 or 1 : 1000. In maps that are intended to be used as part of a municipal geographic information system and made use of in technical planning requiring great precision, the still more precise *measurement class 1e* may be used.
2. To the *second measurement class* belong local detailed plan areas that are built-up areas in which the local detailed plan to be

[1] http://www.ym.fi/en-US/Land_use_and_building/Legislation_and_instructions.

[2] Internationally the nomenclature is highly variable: *general* or *comprehensive plan* is also used.

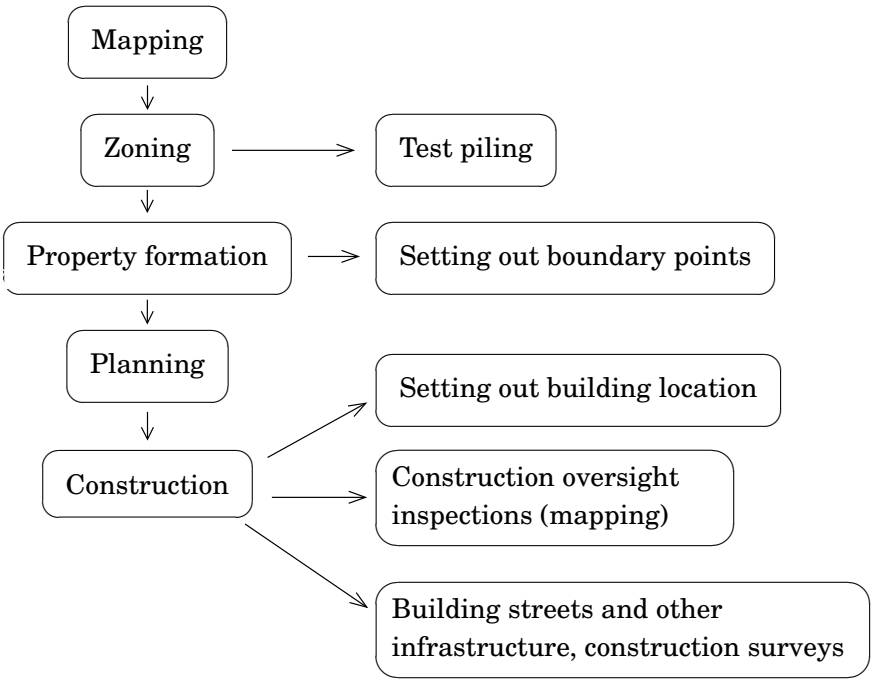


Figure 8.1. Setting out into the terrain, process description.

drafted does not require a binding parcel division. The scale of the base maps is 1 : 1000 or 1 : 2000.

3. To the *third measurement class* belong shore detailed plan areas and lake and sea shore areas, as well as other such areas where land is clearly more valuable than land in agricultural and forestry use, e.g., scattered settlement areas. The base maps are drawn at a scale of 1 : 2000, in special cases 1 : 4000 or 1 : 5000.

The setting out into the terrain of, i.a., the boundary markers and location of buildings, must be done according to need before actual construction starts. The nature of the markers to be used — their quality, the accuracy of benchmarks and the way of carrying out the works — is strictly regulated, in Finland in the Zoning Survey Guide. The base network of the zoned area may have to be brought up to standard.

Kaavoitusmittausohjeet

In setting out, we use an automatic total station, real-time kinematic (RTK) GNSS positioning, or any other sufficiently accurate measurement technology. The co-ordinates of both the known points and the points to be set out in the terrain are entered into the instrument's memory. The method used is radial survey or free stationing. The instrument also pre-calculates *setting-out measures*. Where to place the instrument may be flexibly decided in the terrain: point intervisibility isn't always clear from the map.

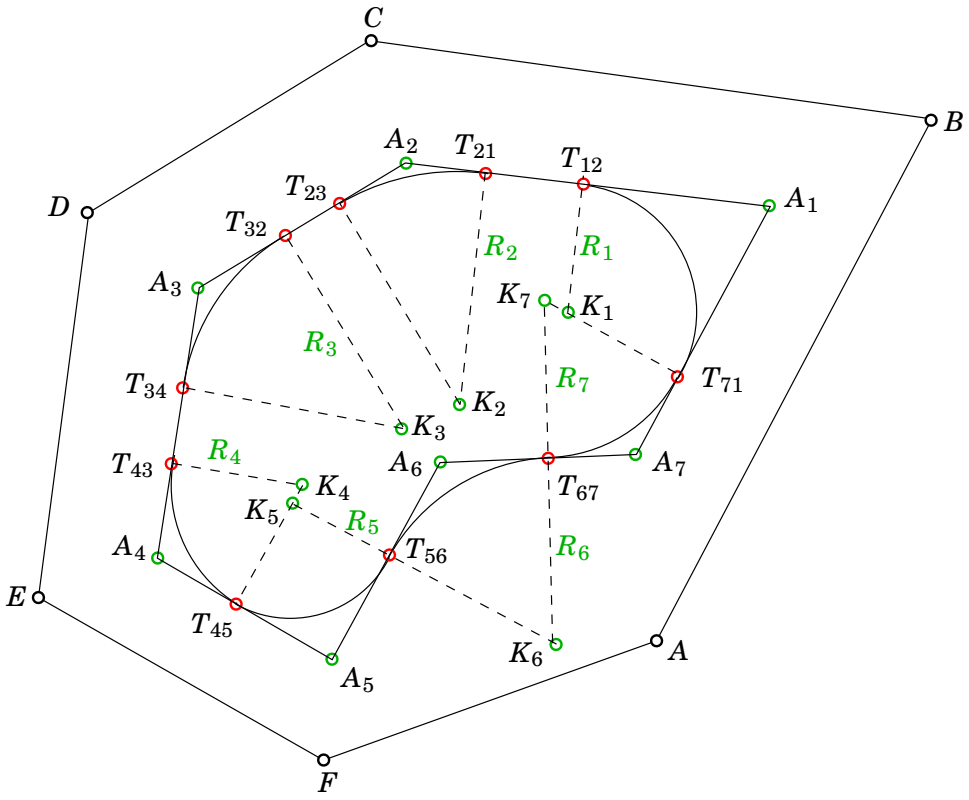


Figure 8.2. Zoning-plan interpretation — an example.

8.2 Setting out and infrastructure

In connection with zoning, technical planning, i.e., infrastructure planning, of the zoned area is carried out:

- In the zoning plan, an area is assigned to a certain use.
- The formation of property organizes the land ownership situation and boundaries as well as easements, etc.
- Planning and construction implement the intended use as stated in the zoning plan, and the area is taken into use.

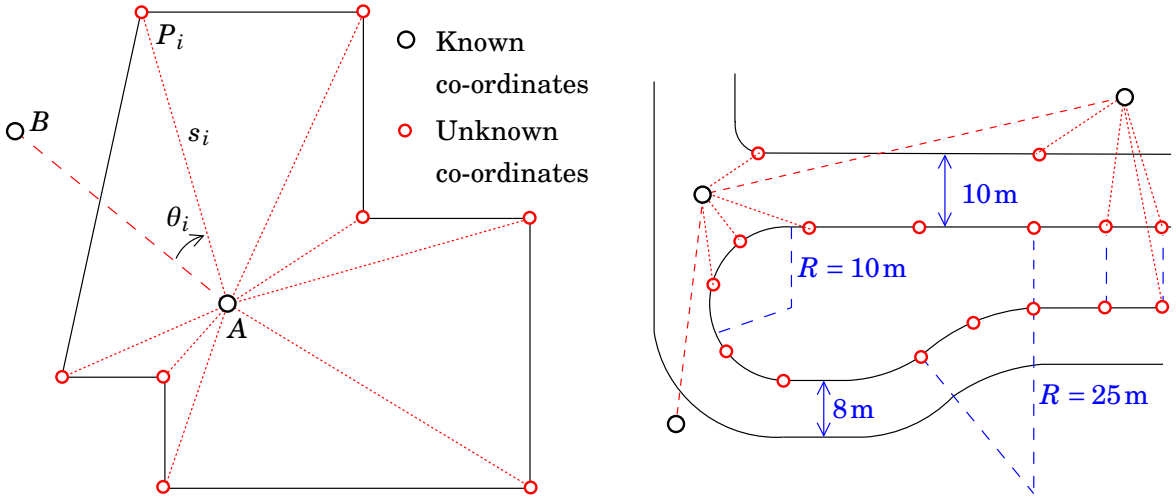
The measurements related to the building of municipal technology or infrastructure (streets, roads, road furniture, utility lines, cables etc.) form their own sub-field of measurement.

Zoning-plan calculation:

- The sketch of the plan, a graphical presentation, is presented in numerical form.
- The plan is interpreted as circular arcs and straight line pieces starting from known elements, figure 8.2.

  First, we compute a *polyline* $A_1,\ldots,A_7$, and fit to this the circular arcs $K_1,\ldots,K_7$.

  To be set out into the terrain are the *zoning-plan boundaries*: the boundaries of blocks of houses, parcels, spaces, general traffic ar-



(a) Radial measurement **(b)** Application to a road area

Figure 8.3. Setting out into the terrain using the radial survey method.

eas, recreational areas as well as areas to be built on.

- For the objects to be set out into the terrain, co-ordinates are calculated.

See figure 8.3. Using the forward geodetic problem one computes directions and distances to the points to be set out, reckoned from the place of the instrument. When using the radial survey method, the place of the instrument is chosen to be a *known point*.

Radial survey (subsection 7.8.3) in setting out works as follows:

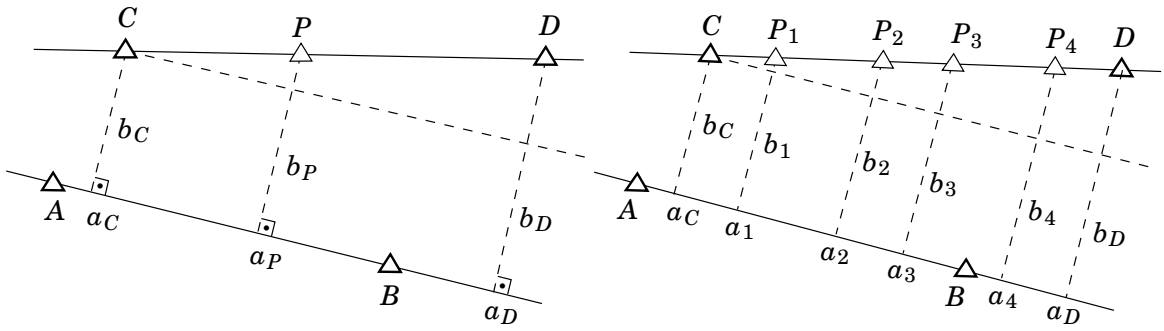
- The angle θ_i and distance s_i are calculated by formula from the co-ordinates of point P_i.
- The instrument is set up on point A and is aimed at signal B, both known. The reading on the horizontal circle is set to zero.
- The telescope is turned until the reading is θ_i.
- The reflective prism is, within the field of view of the telescope, moved so, that its distance reading becomes s_i.
- A *pole* is driven, a precise marker, on the location of which, e.g., a boundary marker is built.

Also the free-stationing survey method and the right-angle survey method are being used.

8.3 Straight lines, circular arcs, rounding of corners

8.3.1 Setting out a straight line into the terrain

The *setting-out measures* a_P, b_P of unknown point P from the given straight line AB must be determined. If we know that point P is on the



(a) Principle **(b)** Several points

Figure 8.4. Straight setting-out method.

straight line CD, we may derive the setting-out measures of P directly from the setting-out measures of points C,D, i.e., a_C,b_C and a_D,b_D, and the distance of P along the straight line CD (, i.e., the distance CP):

$$\begin{aligned} a_P &= a_C + \frac{CP}{CD}(a_D - a_C), \\ b_P &= b_C + \frac{CP}{CD}(b_D - b_C). \end{aligned}$$

This is the so-called *straight method of setting out*, e.g., with a measuring tape and a right-angle or double pentagon prism.

Alternatively one uses *co-ordinates*. To this end, the co-ordinates of points A,B need to be known. Calculate the co-ordinates of C,D and from those, the co-ordinates of P. After this, the setting-out measures of P are calculated. Standard software knows how to do this.

The result is easily generalized to the case in which on the straight line CD there are several points P_i for all of which setting-out measures a_i,b_i are calculated.

8.3.2 Circular arc

Circular arcs are uses a lot in planning, due to their simplicity. A circular arc is defined by four parameters, figure 8.5:

- angle α between the tangents
- half the arc angle θ
- arc radius r
- tangent length t.

There are four dependencies between parameters:

$$\theta = 100\,\text{gon} - \frac{\alpha}{2} \Longleftrightarrow \alpha = 200\,\text{gon} - 2\theta,$$
$$t = r\tan\theta \Longleftrightarrow r = t\cot\theta.$$

Setting out the arc into the terrain is done as follows:

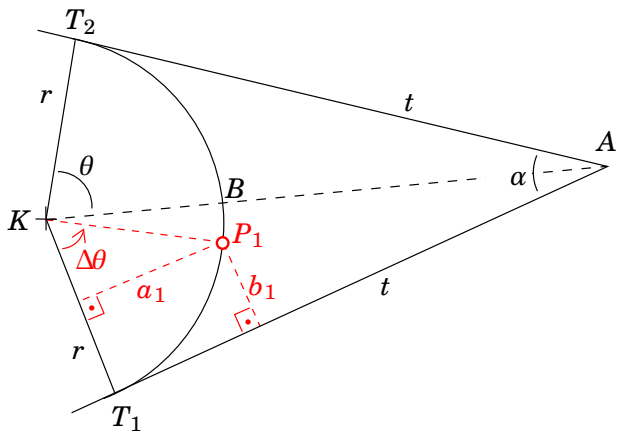


Figure 8.5. Rounding of corners with a circular arc.

1. normally the calculation of the straight lines has already determined the intersection point A of the two tangents, and the angle α.
2. Determine one more parameter, e.g., the radius r, and calculate the others with the formulas given above.
3. Measure, from A along the tangent, the distance t, yielding the tangent points T_1 and T_2.
4. From both of these we determine, with double pentagon prism and measuring tape, the centre point of the arc K (redundance!)
5. On line KA we now mark the middle tangent point B.
6. From centre point K one can mark as many arc points as needed, using the distance r. One example point P_1 is marked in the figure.
7. Also the rectangular setting-out measures of point P_1, i.e., a_1, b_1, are readily obtained.

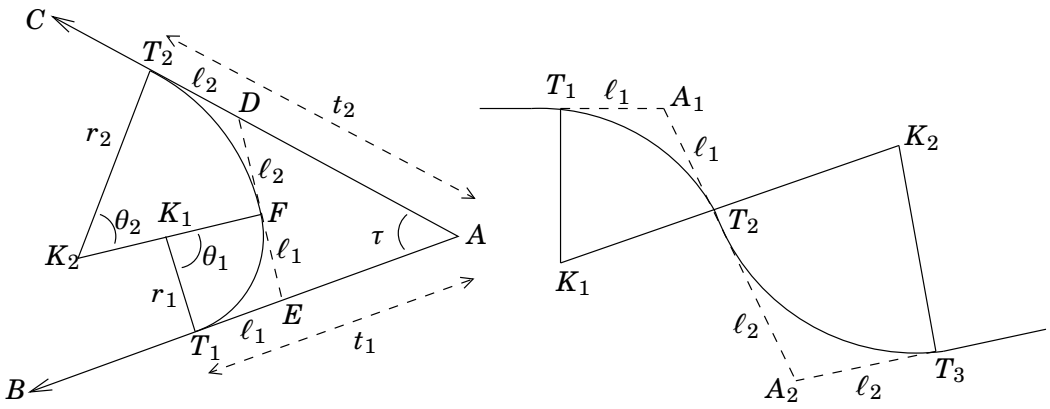
The rounding of corners is commonly done using circular arcs. Combining straight lines and circular arcs is very common in detail zoning maps. The lines are connected according to the situation using various conditions, which produce continuity and the appearance of smooth curvature. Here we present a few examples.

8.3.3 Compound curve

An example case is the connection of two straight lines with two (or more) circular arcs bending in the same direction, a so-called *compound curve*, figure 8.6a, WebTechTix (2014). So, at its simplest we have two straight lines and two circular arcs, that in their junctions — there are three of them — are parallel. korikäyrä

In the situation of figure 8.6a there are the following parameters:

- the lengths of the tangents $t_1 = AT_1$, $t_2 = AT_2$
- the radii of the arcs r_1, r_2



(a) Compound curve **(b)** Reversed curve

Figure 8.6. Rounding of corners with a compound curve. The reversed curve is a special case of the compound curve.

□

- the intersection angle of the tangents τ
- the central angles of the arcs θ_1, θ_2.

There are multiple dependencies between parameters.

1. In triangle $\triangle ADE$ one sees immediately, that $\angle ADE = \theta_2$, because $K_2F \perp DE$ and $K_2T_2 \perp AC$; and similarly $\angle DEA = \theta_1$. Therefore

$$\tau + \theta_1 + \theta_2 = 200\,\text{gon}.$$

   Furthermore $T_1E = EF = \ell_1$ and $T_2D = DF = \ell_2$.
2. And $\ell_1 = r_1 \tan\frac{1}{2}\theta_1$, $\ell_2 = r_2 \tan\frac{1}{2}\theta_2$.
3. By the sine rule

$$\frac{AD}{\sin\theta_1} = \frac{AE}{\sin\theta_2} = \frac{DE}{\sin\tau} \quad\Longrightarrow\quad \frac{t_2-\ell_2}{\sin\theta_1} = \frac{t_1-\ell_1}{\sin\theta_2} = \frac{\ell_1+\ell_2}{\sin\tau}$$

   and by substitution

$$\frac{t_2 - r_2\tan\frac{1}{2}\theta_2}{\sin\theta_1} = \frac{t_1 - r_1\tan\frac{1}{2}\theta_1}{\sin\theta_2} = \frac{r_1\tan\frac{1}{2}\theta_1 + r_2\tan\frac{1}{2}\theta_2}{\sin\tau}.$$

   Thus we can calculate *all* seven parameters catalogued above, if given are
4. two of the three angles τ, θ_1, θ_2, and
5. two of the four lengths r_1, r_2, t_1, t_2.

In the case depicted in the figure, the setting out into the terrain is done as follows:

1. Measure from A along the tangents the distances $t_1 - \ell_1$ and $t_2 - \ell_2$, yielding the points E and D.
2. The tangent intersection angles of the *individual circular arcs* in points D, E are $T_1EF = 200\,\text{gon} - \theta_1$ and $T_2DF = 200\,\text{gon} - \theta_2$.

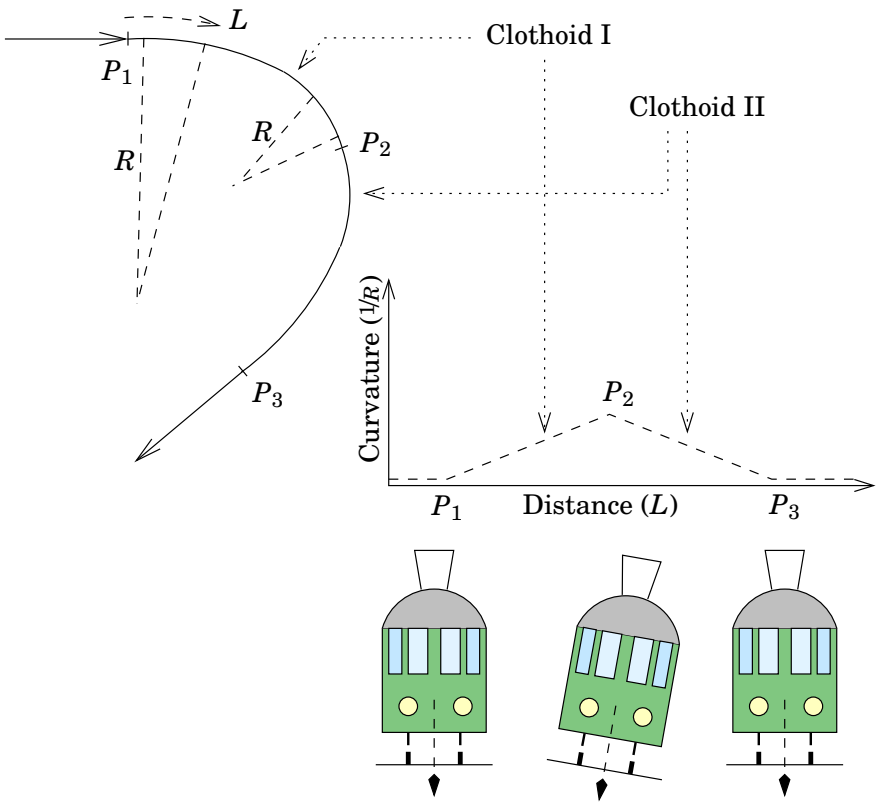


Figure 8.7. Principle of the clothoid.

3. After this, the setting out is done *separately* for circular arcs 1 and 2 in the way already explained above.

8.3.4 Reversed curve

The alternative case where the circular arcs are bending in opposite directions (but otherwise the situation is similar to the compound curve) is the *S-curve*, also reversed curve, figure 8.6b, WebTechTix (2014). What makes the situation a bit cumbersome is the possibility that the intersection point A of the straight lines[3] T_1A_1 and T_3A_2 may not exist if the lines are parallel. S-käyrä

8.4 Transfer curve

A transfer curve, Euler spiral[4] or *clothoid* is used in the planning of railroads and fast motorways. In fast traffic, not only the centre line of the road, but also its *curvature*[5] must be continuous, for the following reasons:

[3]The notation used in the S-curve figure doesn't directly match that in the compound curve figure.

[4]https://en.wikipedia.org/wiki/Euler_spiral.

[5]Curvature is the inverse of the *radius* of curvature!

- e.g., the control of the movement of an articulated lorry through the steering wheel is slow.
- The surface of the road or railroad is tilted in the sideways direction against the centrifugal force. This *transversal tilt*, which is proportional to the curvature of the road, may change only slowly in the longitudinal direction of the road.

For these reasons, a combination of straight lines and circular arcs is unsuitable: a better solution is a spiral or *clothoid*.

The equation of the clothoid is

$$RL = A^2$$

in which A is the *parameter* of the clothoid, L is the distance along the clothoid, i.e., along the road, and R is the local radius of curvature. As can be seen, the radius of curvature changes as a continuous function of the distance

$$R = \frac{A^2}{L}.$$

At constant speed, the centrifugal force[6] F is inversely proportional to the radius of curvature:

$$F = \frac{v^2}{R} = \frac{v^2}{A^2}L, \tag{8.1}$$

i.e., assuming that the *design speed* of the road, v, is a constant, the centrifugal force, and also the transversal tilt of the road surface needed, is a *linear* function of distance travelled. This explains the suitability of the clothoid curve as the shape of motorway and railroad curves.

Of course care must be taken that the clothoid satisfies the continuity condition for the radius of curvature at the start and end points, both with another clothoid, or with a straight line ($A = R = \infty$). See figure 8.7, in which a straight line links to clothoid I (P_1P_2), which links to clothoid II (P_2P_3), which again continues as a straight line from point P_3 onward. The train arrives at point P_1 upright; on the interval P_1P_2 it tilts sideways at a linear rate, and arrives at its maximum tilt angle in point P_2. On the interval P_2P_3 the tilt diminishes linearly, and in point P_3 the train is again upright, and continues its journey straight ahead.

If the speed of the train agrees with the design speed v according to equation 8.1, the resultant of gravity and centrifugal force is always *perpendicular* to the floor of the train, and the passengers don't notice anything.

Also in planning motorways, the clothoid is used, though there, the true speeds of vehicles will vary.

---

[6] . . . more precisely, the *pseudo*-force experienced by the travellers in the vehicle, which acts however from their viewpoint in precisely the same way as gravity.

8.5 Road and street surveying

Roadbuilding comes with the following measurement and calculation tasks:

- drafting the base map for the road plan, road planning, using aerial mapping (photogrammetry) and information technology
- calculating the routes in co-ordinates and setting-out measures
- calculating the co-ordinates, setting-out measures, if necessary earthwork masses etc. for the road structures and special structures (bridges, tunnels, underpasses, etc.)
- on-site measurements for building the road.

Also in other technical measurements related to large-scale construction projects, the same work phases occur.

8.6 Construction surveying

Among construction surveying measurements are the installation measurements of buildings, parts of buildings, bridges, tunnels, reservoir dams, industrial machines and similar structures.

- The measurement starts from the base network. First, a project *measurement base* is created, i.e., a sufficient set of horizontal and vertical benchmarks in the project area. As benchmarks are used base network points. Densification points or "use points" are created as needed, in a hierarchical fashion.
- The actual measurements are carried out from the use points. The measurements are done separately as horizontal and height measurements, also the point sets are partly separate.
- There exist international standards regulating the measurements.

8.6.1 Setting out a building location into the terrain

When the building permit has been obtained, the builder can apply to the authorities for a decision to *set out the building location into the terrain*. The procedure has *three objectives*:

paalutus, maastoon merkintä

1. setting out the building location into the terrain. It is verified that no part of the building is too close to the parcel boundary.
2. Checking the correctness of the wall measures of the building, important information for the builder.
3. Verifying the correctness of the building's height location.

The land survey authorities only measure with a view to assigning where to build, unless agreed otherwise, and the precision and number of points is not enough to actually start building. The builder can continue the work in connection with the actual building job.

As a site reference point, either a traverse point or a boundary marker is chosen. If there are not enough of those nearby, or not of the required precision, the first thing that must be done is measuring new *base points*. For these, map co-ordinates (x, y) are calculated and they are marked on the site plan.

As *control measures* may be used: wall measures, cross-measures, distances from boundaries, and distances from other buildings.

The height location of a building is obtained by traverse levelling, which runs from one general height benchmark to another. Close to the building site, *at least two* height base points are created for the later work, unless nearby there are already enough general points. The correct height of a building is critical for functioning sewers and flood safety.

8.6.2 Location review of a building

sijaintikatselmus

In *location review* it is verified that the building is in the correct place and at the correct height. The review is performed when the foundation of the building is completed. After approval, construction may continue.

8.7 Other measurements

8.7.1 Technical measurements, deformation measurement

Technical measurements, or *engineering geodesy*, form their own specialized discipline. To this belong also precise deformation measurements:

deformaatio-seuranta

- deformation monitoring during construction and/or afterward. Objects: reservoir dams, tunnels, bridges, large structures, etc.
- nowadays often *monitoring measurement* using automated equipment
- industrial measurements indoors or outdoors, installation measurements of large machines, paper machines, shipyards
- engineering surveying
- tunnel and mine surveying. For these measurements, characteristic are the tight constraints on the measurement geometry, as well as the non-functioning of GNSS.

In these measurements, *traceability* of the quantities measured is central. Careful (system) calibration and good metrological practice are of great importance.

8.7.2 Machine guidance of working machines

GNSS technology is used widely, in addition to more traditional positioning technologies, for guiding *working machines* in real time. The reliability requirements are obviously tough if an expensive working machine

Figure 8.8. The AGA/Minilir infrared tracking device — French military technology! — in use on the building site of the Dutch Easter Scheldt storm-surge barrier, 1980. © 2016 Nicolàs de Hilster, http://www.dehilster.info/geodetic_instruments/1980_satsagem_Fennel-Minilir.php, figure 4, with permission.

is guided to lay, e.g., asphalt on a motorway. Operation stops are costly, errors even more so.

In the construction of reservoir dams, bridges, tunnels, and other infrastructure, usually terrestrial real-time guidance is used. In the construction phases of the Dutch Easter Scheldt storm-surge barrier (figure 8.8) myrskyvuoksisuoja and of the Danish Great Belt and Sound bridge-tunnel solutions, this kind of technology was used, as in many similar projects.

Also shipping containers and cranes in harbours are being positioned in real time using GNSS technology in order to improve efficiency (Pitkä, 2009).

Agricultural and forestry machines may be guided by real-time GNSS ("precision farming"), with the aid of which seedstock, fertilizer, and pesticides can be administered with precision according also to very localized needs.

8.7.3 Underground utility lines mapping

Of the underground utility lines, like telephone, data and electric power cables as well as sewage, water and city heating pipes, only a part has been mapped satisfactorily. Most of these are on maps prepared by various municipal institutions, of which there may be many per municipality. When utility-line surveys are this decentralized, the geodetic quality of the maps will vary.

johtokartta-standardi

Cities generally use presentational and preparation practice of the 1974 Utility Lines Mapping Standard SFS 3161. The standard was renewed in 1996.

Utility lines maps are used for many needs: in zoning for the planning of technical maintenance and the network, for construction work, for maintenance of the lines by the owner institution, and in connection with damages and managing crisis situations. The scale is most commonly 1 : 500.

Mapping methods:

- The measurement should be based on general benchmarks available in the area, so that the result is obtained in the same system. If necessary, a densification is carried out, i.e., new benchmarks are created following the Zoning Survey Guide.

Kaavoitusmittaus-ohjeet

- The same measurement methods are used as more generally in detail survey, see section 7.8.
- The mapping of new lines is done during construction, when the lines are still visible.

kaivonkansi

- The visible parts of old lines (manhole covers, distribution cabinets) are mapped. The underground parts can sometimes be located with a metal detector.
- The mapping work chain is fully digital.

johtojen näyttöpalvelu

Underground utility lines marking service: The lines are marked in the terrain for the builder, in order to avoid costly damage caused by excavation work. The service is usually offered by the owner of the line, e.g., a power company.

Self-test questions

1. What is zoning and why is it necessary?
2. How many measurement classes are used in Finland? How are they defined? What scales are the zoning maps on for each class?
3. Describe the various methods of setting out objects into the terrain.
4. What is a compound curve? A reversed curve?
5. How many independent parameters uniquely define a compound curve?
6. Explain why clothoids are used for fast roads and railroids.
7. Describe the Finnish practice of “location review”. In what stage of construction does it take place?

9. Digital terrain models and volume calculation

Construction and its planning, as well as the technical measurements made in connection with them, make extensive use of *digital height and terrain models*.

The term *digital height model*, DHM, or *digital elevation model*, DEM, or *digital terrain model*, DTM, refers to a file made up of points on the Earth's surface, which describes — more or less well — the forms of the Earth's surface. High-resolution terrain models are expensive to produce over large areas, but are nevertheless available for many countries. For Finland, terrain models are produced by the National Land Survey.

Information about the Earth's surface and its forms can be obtained by topographic surveying measurements, photogrammetrically from aerial imagery, and by *scanning* from aircraft or satellites. The scanner may be laser based or can be a microwave radar, so-called SAR, synthetic-aperture radar, in which interferometrically a very high resolution is achieved.

1. Global terrain models:
   (a) The older model GTOPO30 (US Geological Survey, original 1996)[1]. The resolution is 30″ or about one kilometre. Was updated with data from the SRTM, see below. The model contains no sea depth data.
   (b) The GLOBE model (Global One-km Base Elevation project (NOAA and many others, 1999)[2]. Also 30″ = 1 km. The model contains no sea depth data, but this could change in the future.
   (c) The ETOPO1 model, supersedes the older ETOPO5 and ETOPO2 models (NOAA)[3]. The resolution is 1′ (about 1.8 km), the model contains depth data in addition to elevation data.

[1] https://lta.cr.usgs.gov/GTOPO30.

[2] http://www.ngdc.noaa.gov/mgg/topo/globe.html.

[3] https://www.ngdc.noaa.gov/mgg/global/global.html.

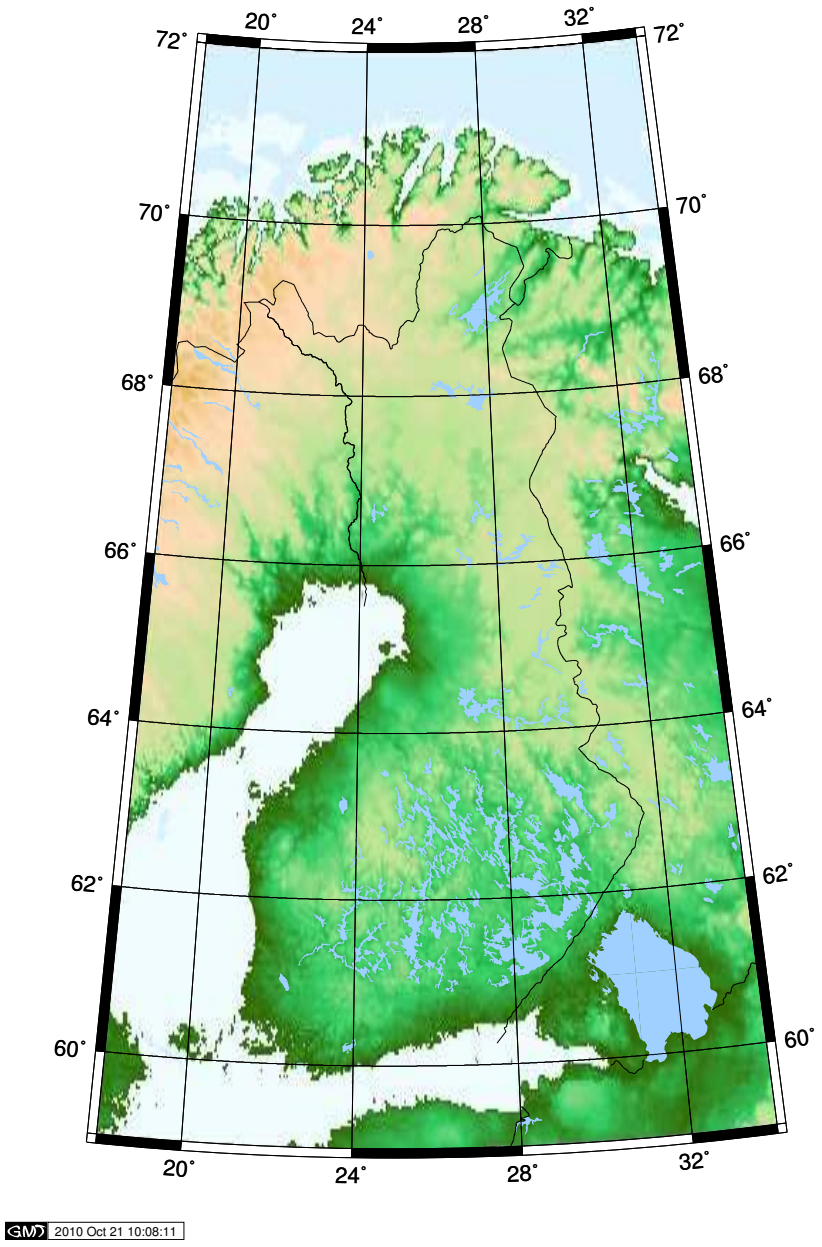


Figure 9.1. The global terrain model ETOPO2 version 2 on the Finnish territory.

(d) The Shuttle Radar Topography Mission (SRTM) imaged the Earth's topography between latitudes 60°N and 56°S. The flight took place in February 2000. The resolution is one second of arc on the Earth's surface, about 30 m. During 2014 *all* SRTM data was declared public[4]. The data does not contain sea depths.

1. The national terrain model:

   New technologies, like airborne laser scanning, have been opera-

[4] http://www2.jpl.nasa.gov/srtm/.

tional for many years already and are in widespread use. Also in Finland, the National Land Survey has for a long time been scanning various Finnish areas in order to build a new, precise national terrain model. There are two models: the model KM10 is currently complete at spatial resolution 10 m[5], and the model KM2 is partially complete at resolution 2 m[6]. The KM2 model is planned to be completed by 2020. Based on the Inspire directive[7], the data is free of charge.

9.1 Terrain models: measurement, construction, presentation

9.1.1 Measuring terrain models

Measurement geometry:

- the points to be measured do not generally form a regular pattern, and may be freely chosen within the constraints of the measurement technique used.
- The random-point method is a statistical sampling method. The sampling density may be higher where terrain forms are more variable.
- The point density is chosen higher near break lines, i.e., edge-like features in the terrain where the surface changes orientation.

Measurement technologies:

- geodetic topographic surveying. This is a low-productivity technique that often complements other methods.
- photogrammetry. Points are measured in a stereo model formed from two aerial photographs, nowadays often automatically by *correlation* of the digital images. However, ground control points (GCPs) and occluded areas are surveyed geodetically.
- airborne laser scanning. This technology collects huge numbers of three-dimensional terrain points in the form of a *point cloud*, from which the terrain surface can be extracted by suitable processing.

From the point data measured, either a triangulated network or a regular point grid is generated.

[5] http://www.maanmittauslaitos.fi/kartat-ja-paikkatieto/asiantuntevalle-kayttajalle/tuotekuvaukset/korkeusmalli-10-m.

[6] http://www.maanmittauslaitos.fi/kartat-ja-paikkatieto/asiantuntevalle-kayttajalle/tuotekuvaukset/korkeusmalli-2-m.

[7] http://inspire.ec.europa.eu/.

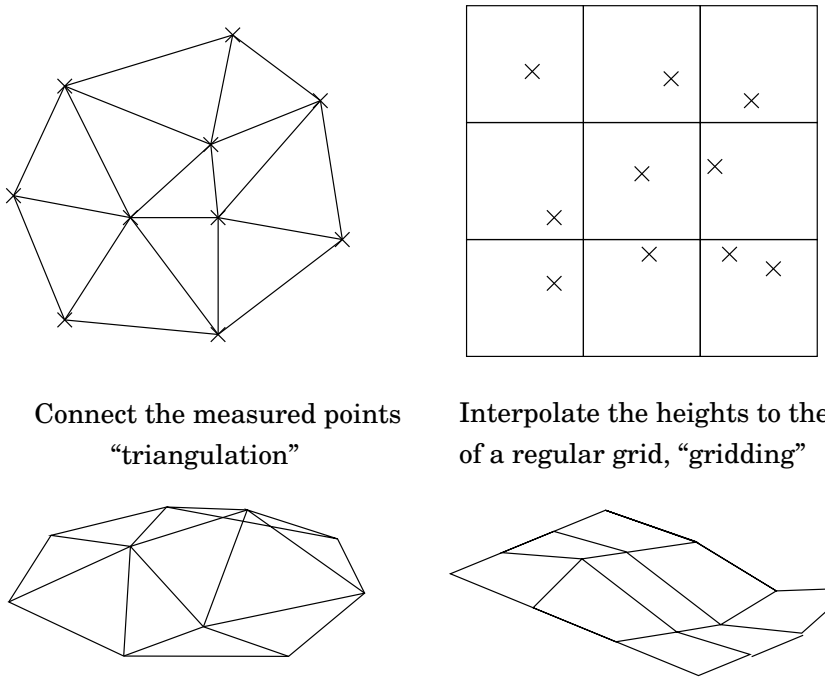


Figure 9.2. Presentation of terrain models: triangulated network or point grid.

With terrain models, different *forms of presentation* are used[8]:

Point grid presentation. Agrees well with the way computers operate: also the handling of large amounts of data is straightforward and easy.

A regular grid may be square, rectangular, or more complicated, like a hexa pattern (“beehive”) or a triangular pattern.

Triangulated-network presentation. Here, points representative of the terrain forms are chosen and connected by lines into a cover of triangles. One known mathematical triangulation technique is *Delaunay*[9] triangulation, which gives beautiful triangles of which the sides are as equal as possible.

A triangulated network type of terrain presentation is more difficult to manipulate, but it is able to present also difficult terrain forms, like sharp edges, better than a grid presentation, using a smaller number of points. Also if the resolution of the terrain model varies by area, the triangulation presentation is better, because the sizes of the triangles vary with resolution.

In the literature, the term TIN method is used: *Triangulated Ir-*

[8]In image processing one speaks in completely analoguous fashion of pixel and vector graphics.

[9]Boris Nikolayevich Delaunay (1890 – 1980) was a Russian mathematician and mountaineer. Not to be confused with French celestial mechanicist Charles-Eugène Delaunay (1816 – 1872).

regular Network.

CAD (computer-aided design) software, in use in planning offices, know how to use digital terrain models and how to display them in many different ways, e.g., as a perspective image. Also the building plans are in digital form and can be combined with this.

Let us mention in this connection also the multi-resolution "tiling" methods which are based on the discrete wavelet transform[10] (DWT) and are meant for the interactive presentation of materials[11]. Also the image format JPEG 2000, as well as Google Earth™are based on this technique. This form of presentation enables extremely fast changes in viewing location and zooming. It is even more suited than Delaunay triangulation for presenting materials of varying resolution interactively.

9.2 Use of terrain models

Terrain models are used, e.g., for the following applications:

- the ortho-rectification of aerial photographs: the removal of the projection errors of aerial imaging in the production of orthophoto maps
- calculating the *height contours* shown in a map
- planning a traffic route (road, street, waterway, electric power line, ...). Minimize (under other constraints, like maximum slope or minimum radius of curvature):

maamassa

  1. earthwork volumes to be moved, section 9.3 on page 228
  2. the difference between soil to be removed and soil to be added
  3. consumption of fuel and/or time by a typical vehicle using the route.

- creating and visualising three-dimensional landscape models, in support of planning and the public debate over the plans
- resolving visibility issues, e.g., in connection with placement of cell-phone or radio masts
- planning ski pistes
- A military application is the automatic navigation in low flight of cruise missiles, but also jet fighters
- calculating the gravity effect of terrain masses (the terrain correction) in gravity field and geoid computation
- realistic landscapes for video games, flight simulators
- many others.

[10] http://en.wikipedia.org/wiki/Discrete_wavelet_transform.

[11] Esimerkiksi http://meru.rnet.missouri.edu/mvl/kolam/.

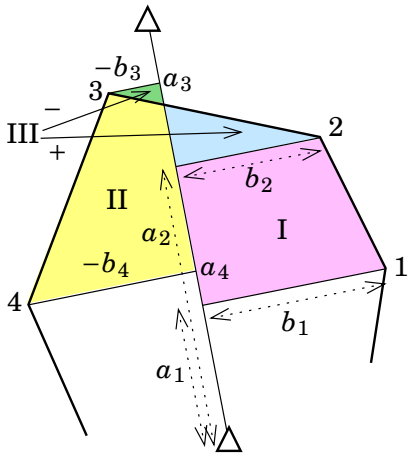


Figure 9.3. The use of setting-out measures in calculating surface areas.

Strictly speaking a digital elevation model (DEM) describes the heights, not only of the terrain, but also of buildings etc., whereas a digital terrain model (DTM) describes only the heights of the terrain itself. Nevertheless the words are often used as synonyms.

9.3 Computing surface areas

Calculating surface areas is discussed in the book Kahmen and Faig (1988, chapter 8.6).

A handy way of calculating is to use setting-out measures with respect to a baseline. In figure 9.3, the setting-out measures form trapezoids.

Area I is calculated as follows:

$$A_{\mathrm{I}} = \frac{1}{2}(a_2 - a_1)(b_1 + b_2),$$

and area II as follows:

$$A_{\mathrm{II}} = \frac{1}{2}(a_4 - a_3)(b_3 + b_4),$$

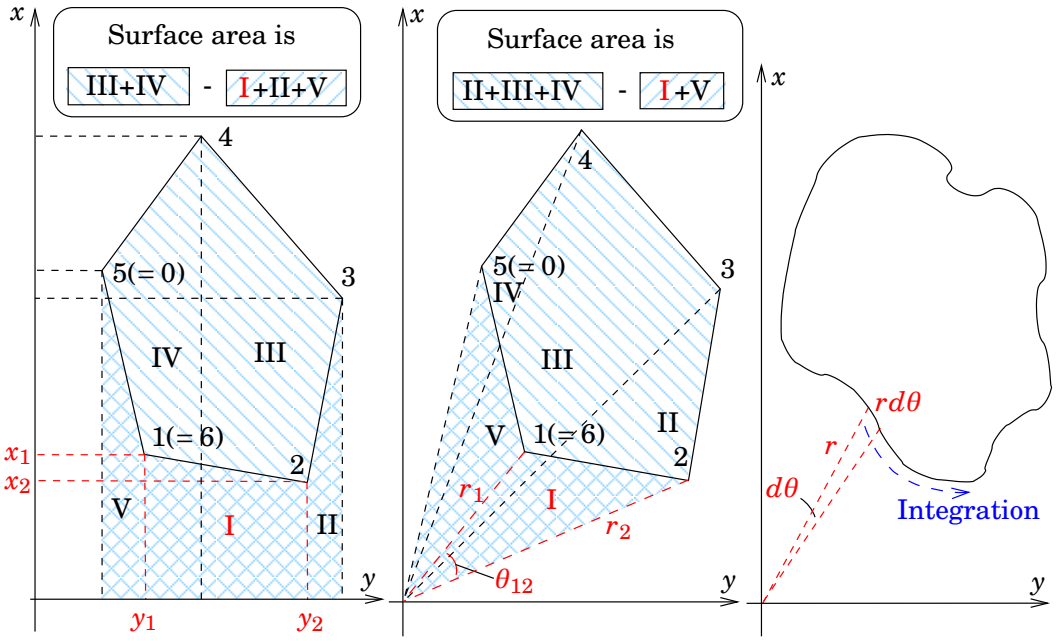
in which it is important to pay attention to algebraic signs.

Also area III is obtained in a similar way, although it is the difference between two surface areas (it is nevertheless formally a trapezoid).

$$A_{\mathrm{III}} = \frac{1}{2}(a_3 - a_2)(b_2 + b_3).$$

All equations can be made compatible by agreeing, e.g., that the b values are positive on the right side seen when moving in the direction of a, and negative on the left side. Also, the a indices are followed in numerical order, i.e., in the example case, counterclockwise. Then we obtain for the total area, by summation,

$$A = A_{\mathrm{I}} + A_{\mathrm{II}} + A_{\mathrm{III}} + \ldots$$



(a) Rectangular **(b)** In polar co-ordinates **(c)** Continuous integration

Figure 9.4. Calculating surface area.

□

and all algebraic signs, also those of the small cancelling triangles, are automatically correct.

If we have the use of co-ordinates, there are other ways of calculating surface areas. The total surface area is obtained as a sum of trapezoids (the i index is circular, i.e., $n+1$ is the same as 1):

$$A = +\frac{1}{2}\sum_{i=1}^{n}(x_{i+1}-x_i)(y_{i+1}+y_i), \tag{9.1}$$

and by interchanging x and y:

$$A = -\frac{1}{2}\sum_{i=1}^{n}(y_{i+1}-y_i)(x_{i+1}+x_i). \tag{9.2}$$

From equation 9.1 we obtain

$$A = -\frac{1}{2}\sum_{i=1}^{n}x_i(y_{i+1}+y_i)+\frac{1}{2}\sum_{i=1}^{n}x_{i+1}(y_{i+1}+y_i),$$

and by re-numbering the second term — as the i index is circular:

$$A = -\frac{1}{2}\sum_{i=1}^{n}x_i(y_{i+1}+y_i)+\frac{1}{2}\sum_{i=1}^{n}x_i(y_i+y_{i-1}) = \frac{1}{2}\sum_{i=1}^{n}x_i(y_{i-1}-y_{i+1}).$$

Similarly equation 9.2 yields

$$A = \frac{1}{2}\sum_{i=1}^{n}y_i(x_{i+1}-x_{i-1}).$$

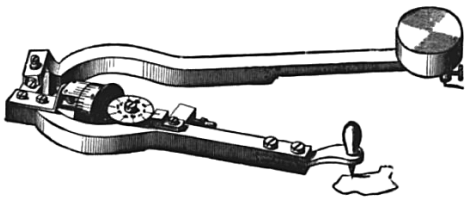


Figure 9.5. A polar planimeter from 1908 (Wikipedia). The surface area is measured by drawing along the outer edge of the figure.

These equations are known as the *shoelace formulas*[12].

If the equations 9.1 and 9.2 are added together and divided by two, we obtain

$$A = \frac{1}{2}\sum_{i=1}^{n}(x_{i+1}y_i - y_{i+1}x_i).$$

This equation calculates the surface of a polygon as the sum of *triangles extending from the origin*. It can be shown — a graphical proof is given in figure 9.6 —, that the surface area of such a triangle (example in the figure) is

$$A_{i,i+1} = \frac{1}{2}(x_{i+1}y_i - y_{i+1}x_i) = \frac{1}{2}r_i r_{i+1}\sin\theta_{i,i+1}. \tag{9.3}$$

This so-called planimeter equation 9.3 is the principle of operation of the *polar planimeter*[13]. Of course the equation can also be used directly numerically, if the figure is given in polar co-ordinates:

$$A = \frac{1}{2}\sum_{i=1}^{n} r_i r_{i+1}\sin\theta_{i,i+1}.$$

9.4 Volume calculations

The calculation of earthwork masses is explained in Kahmen and Faig (1988, sections 14.2, 14.3). Often, the case is about calculating amounts of gravel, sand and similar building materials to be moved in connection with, e.g., road building.

The means of measurement or determination are

[12] https://en.wikipedia.org/wiki/Shoelace_formula.

[13] The polar planimeter integrates mechanically the expression

$$\frac{1}{2}\oint r^2(\theta)d\theta,$$

which is the surface area of the closed figure.

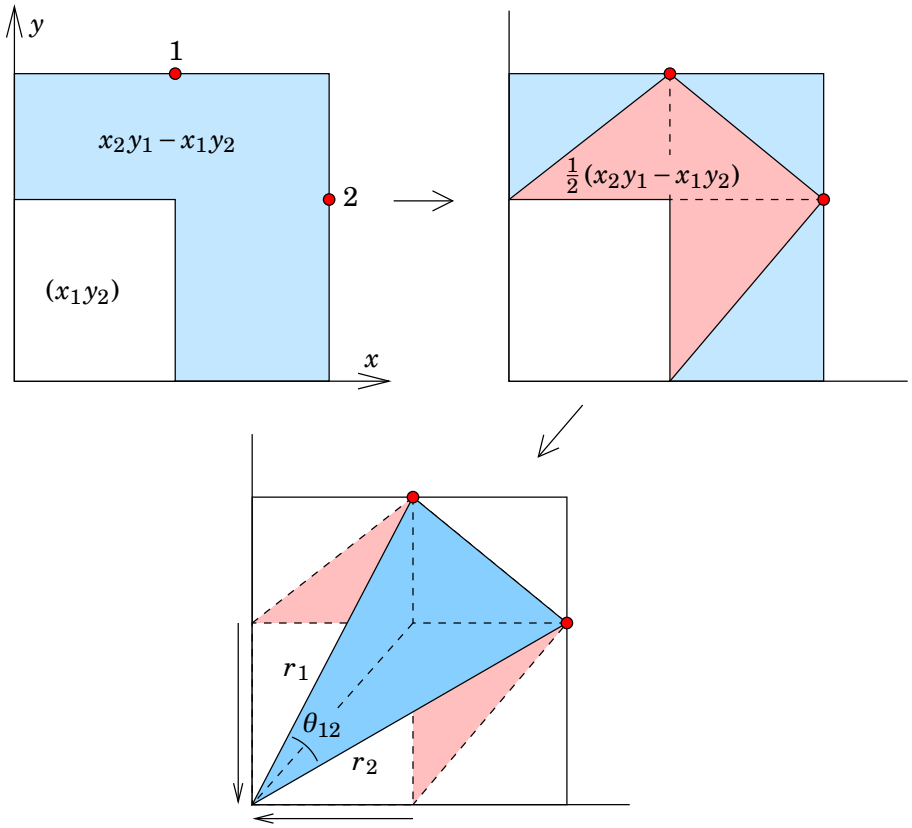


Figure 9.6. Graphical proof of the planimeter equation.

□

1. area levelling or use of a laser level (only small projects, labour intensive)
2. photogrammetry from the air, or terrestially from an aerial work platform

"pekkaniska™"

3. laser scanning, from the air or terrestrial — high productivity
4. height contours from a map. Computationally, the method is similar to surface-area calculation, section 9.3
5. a digital terrain model.

□

9.4.1 Simpson's rule and quadrature

A handy way for numerical integration, or *quadrature*, of volume from profile data is Simpson's rule (Kahmen and Faig, 1988, subsection 14.2.1)[14]. The equation is

$$V = \frac{1}{6}(F_1 + 4F_m + F_2)\,\ell, \tag{9.4}$$

in which F_1, F_2 are the surface areas of the end-point cross sections, F_m is the surface area of the midpoint cross section, and ℓ is the length of

[14]Thomas Simpson (1710 – 1761) was the son of an English weaver and an autodidact mathematician, fellow of the Royal Society. He actually didn't invent Simpson's rule, although it appeared in his textbook: the rule was known already to Johannes Kepler.

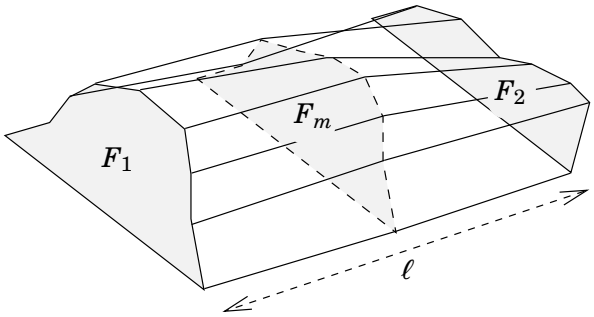


Figure 9.7. Simpson's integration rule in volume calculation.

the whole object.

Simpson's rule can be proven as follows. Let the function to be integrated be $f(x)$, and let us have at our disposal function values in the points $(-\Delta x, 0, \Delta x)$:

$$f_{-1} = f(-\Delta x),\ f_0 = f(0),\ f_1 = f(\Delta x).$$

Approximate the function f by a fourth degree polynomial:

$$\widetilde{f}(x) = a + bx + cx^2 + dx^3 + ex^4.$$

The integral over the polynomial is

$$\int_{-\Delta x}^{+\Delta x} \widetilde{f}(x)\,dx = \left[ax + \frac{1}{2}bx^2 + \frac{1}{3}cx^3 + \frac{1}{4}dx^4 + \frac{1}{5}ex^5\right]_{-\Delta x}^{+\Delta x} =$$
$$= 2a\Delta x + \frac{2}{3}c\Delta x^3 + \frac{2}{5}e\Delta x^5. \tag{9.5}$$

Write as well

$$\begin{aligned}
\widetilde{f}_{-1} &= a - b\Delta x + c\Delta x^2 - d\Delta x^3 + e\Delta x^4,\\
\widetilde{f}_0 &= a,\\
\widetilde{f}_1 &= a + b\Delta x + c\Delta x^2 + d\Delta x^3 + e\Delta x^4,
\end{aligned}$$

so that for the linear combination

$$\begin{aligned}
I = p_{-1}\widetilde{f}_{-1} + p_0\widetilde{f}_0 + p_1\widetilde{f}_1 = a(p_{-1} + p_0 + p_1) +\\
+ \left(b\Delta x + d\Delta x^3\right)(-p_{-1} + p_1) + \left(c\Delta x^2 + e\Delta x^4\right)(p_{-1} + p_1).
\end{aligned} \tag{9.6}$$

Comparing the formulas 9.5 and 9.6 shows that, to get I as close as possible to the integal 9.5, we must choose

$$\begin{aligned}
p_{-1} + p_0 + p_1 &= 2\Delta x,\\
-p_{-1} + p_1 &= 0,\\
p_{-1} + p_1 &= \frac{2}{3}\Delta x,
\end{aligned}$$

which yields

$$p_{-1} = p_1 = \frac{1}{3}\Delta x \text{ and } p_0 = \frac{4}{3}\Delta x.$$

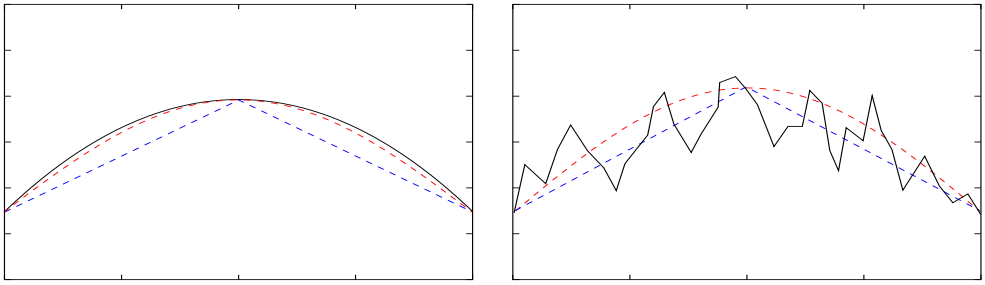


Figure 9.8. Alternatives for quadrature. Left, a mathematically well behaved function (black), Simpson (red) works best. Right, a "jagged" function. The trapezoidal rule (blue) works just as well. Realistic terrain is between these extremes.

Substituting this into equation 9.6 yields

$$I = 2a\Delta x + \frac{2}{3}c\Delta x^3, \tag{9.7}$$

and the difference with the integral is

$$\int_{-\Delta x}^{+\Delta x} \widetilde{f}(x)\,dx - I = \frac{2}{5}e\Delta x^5,$$

a fifth-degree function of the point spacing Δx. This means that by choosing Δx small enough, we can get the expression 9.7 very quickly close to the true value of the polynomial. We may write

$$I = p_{-1}\widetilde{f}_{-1} + p_0\widetilde{f}_0 + p_1\widetilde{f}_1 = \frac{1}{3}\widetilde{f}_{-1}\Delta x + \frac{4}{3}\widetilde{f}_0\Delta x + \frac{1}{3}\widetilde{f}_1\Delta x =$$
$$= \frac{1}{6}\left[\widetilde{f}_{-1} + 4\widetilde{f}_0 + \widetilde{f}_1\right]\cdot 2\Delta x.$$

Into this we substitute the true values f_{-1}, f_0, f_1 of the function, as well as $\ell = 2\Delta x$, and obtain Simpson's rule 9.4. If the function to be integrated f is not pathological — but terrain forms could well be pathological! —, then Simpson's rule will converge also very quickly to it.

9.4.2 Alternative quadrature rules

Often, a simpler rule

$$V = \frac{1}{2}(F_1 + F_2)\ell$$

(the "trapezoidal rule") works, or even

$$V = F_m\ell$$

(the "rectangle rule"). The don't converge however as beautifully as Simpson's rule: for both, the error is proportional to the point spacing, or object length, cubed,$\Delta x^3 = \ell^3$.

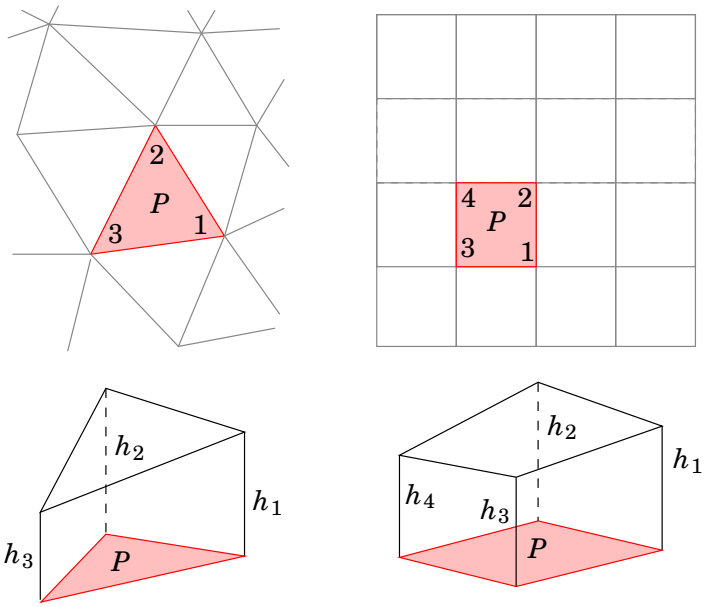


Figure 9.9. Volume calculation from digital terrain models.

If the accuracy of Simpson's rule is insufficient, because the object is too jagged or "pathological", one may achieve a better accuracy by dividing the object into slices, applying some simpler quadrature rule to each of those, and summing the contributions obtained.

Volume calculation for *triangulated-network* or *point grid* type terrain models is depicted in figure 9.9. In the case of the triangulated-network model, the volume of a surface element is evaluated by the equation

$$V = P\frac{h_1 + h_2 + h_3}{3}.$$

In the case of a point grid model, the equation to be used is

$$V = P\frac{h_1 + h_2 + h_3 + h_4}{4}.$$

Generalization: the area of a surface element P multiplied by the average height calculated from n corner points

$$\overline{h} \overset{\text{def}}{=} \frac{\sum_{i=1}^{n} h_i}{n}.$$

These equations are approximate but usually sufficient.

Self-test questions

1. Which observation techniques are available for collecting terrain point information usable for construction terrain models?

2. Which two main techniques for presentation of terrain models exist?
3. Discuss applications of terrain models.
4. Describe how the surface area of a parcel may be determined from setting-out measures of its boundary.
5. Describe how a polar planimeter works.
6. Describe how the quadrature of volumes from profile data using Simpson's rule works.

10. Three-dimensional co-ordinate reference systems

10.1 Geocentric co-ordinate reference systems

In modern geodesy, the measurement methods of satellite and space geodesy are integral parts of the global geodetic observing system[1]. Unlike traditional geodetic measurement methods, which carry out their measurements on, or close to, the Earth's surface, these measurements are *genuinely three-dimensional*, and their processing requires use of three dimensions also in computation. Additionally, the platforms of, at least, satellite measurements are in orbits around the Earth, meaning that the centre of mass of the Earth becomes naturally the origin of the co-ordinate frame used. This is why, in satellite geodesy, we use *geocentric, three-dimensional* co-ordinate reference systems. See figure 10.1. As symbols for geocentric co-ordinates, often *capital letters* are used, like X, Y, Z.

Geocentric co-ordinate reference system: The origin is in the centre of mass of the Earth, and the Z axis is directed along the rotation axis of the Earth.

There are two types of geocentric systems:

Inertial or celestial: There is no rotational motion. The directions of the axes are fixed with respect to the stars.

The X axis points (usually) to the *vernal* or *spring equinox*, the "Greenwich of the sky".

Terrestrial or co-rotating: also Earth-Centred, Earth-Fixed, or ECEF: the directions of the axes are fixed with respect to the solid, rotating Earth.

The X axis points in the direction of the *Greenwich* meridian.

Between the inertial and the terrestrial systems there is a rotation angle called *Greenwich sidereal time*. It changes rapidly with time, at the same angular rate as the rotation of the Earth with respect to the stars.

Right-handed co-ordinate frame: a corkscrew which progresses in

[1] http://www.ggos.org/.

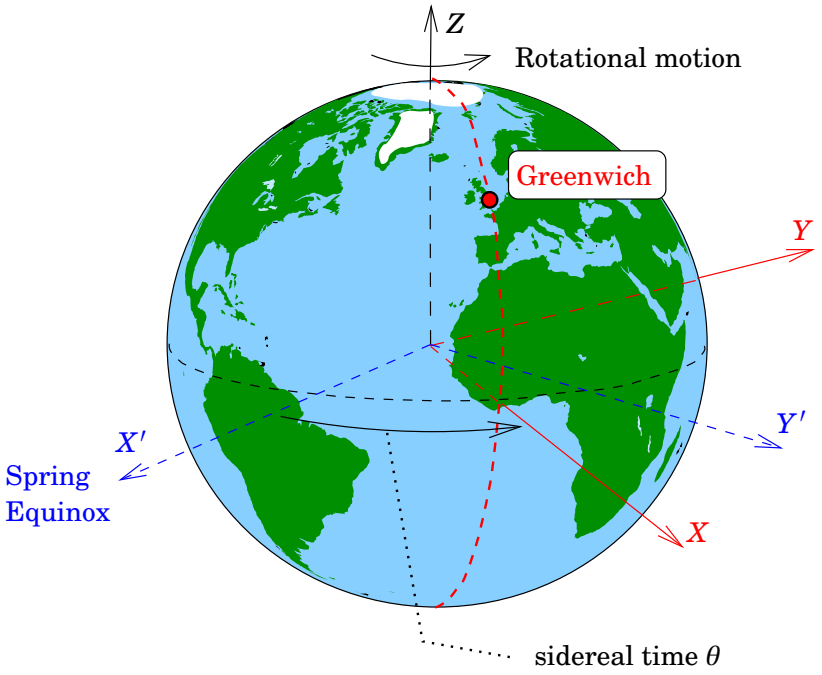


Figure 10.1. The inertial or celestial (X',Y',Z) and the terrestrial, co-rotating, or ECEF (X,Y,Z), co-ordinate reference system.

the positive x direction, turns from the y-axis direction to the z-axis direction, figure 10.2.

A geocentric system is *right-handed* if the X axis points in the direction of the intersection of the planes of the Greenwich meridian and the equator, the Z axis to the celestial North pole, and the Y axis to 90° East.

Geocentric co-ordinates may be rectangular (X,Y,Z), spherical co-ordinates, geodetic or geographical co-ordinates, or ellipsoidal co-ordinates.

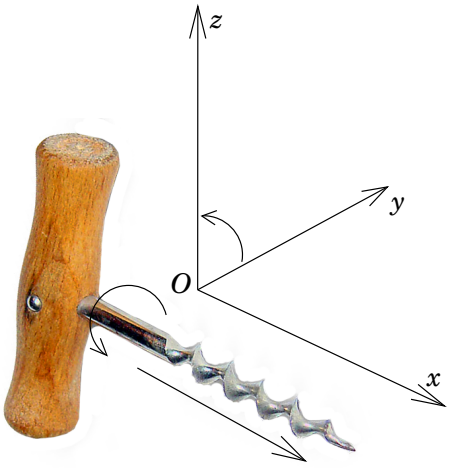


Figure 10.2. A right-handed co-ordinate frame. A $y \to z$ corkscrew (©) progresses in the $x$direction — like also a $z \to x$ corkscrew in the y direction, and an $x \to y$ corkscrew in the z direction.

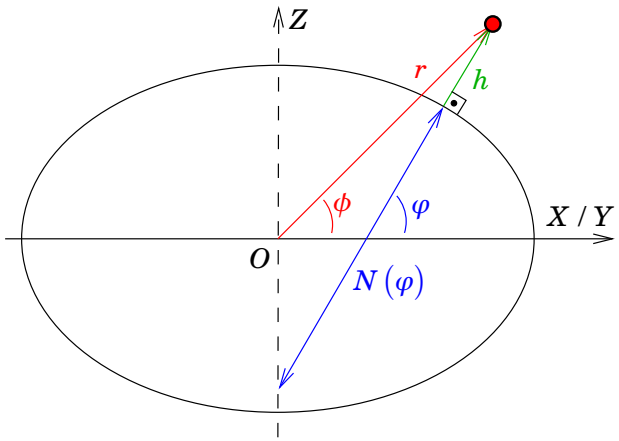


Figure 10.3. Geocentric and geodetic latitude and transversal radius of curvature.

Between them exist the following relations (figure 10.3):

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = r \begin{bmatrix} \cos\phi\cos\lambda \\ \cos\phi\sin\lambda \\ \sin\phi \end{bmatrix},$$

in which (r,ϕ,λ), i.e., the distance r from the geocentre, the *geocentric* latitude and longitude, are spherical co-ordinates, and

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} (N(\varphi)+h)\cos\varphi\cos\lambda \\ (N(\varphi)+h)\cos\varphi\sin\lambda \\ \left(\left(b^2/a^2\right)N(\varphi)+h\right)\sin\varphi \end{bmatrix},$$

in which (h,φ,λ), i.e., the height h from the reference ellipsoid and the *geodetic* latitude and longitude, are geodetic, also called geographical, co-ordinates. The quantities a and b are the semi-major and semi-minor axes of the Earth ellipsoid, in other words, the equatorial and polar radii. The transversal radius of curvature is

$$N(\varphi) = \frac{a^2}{\sqrt{a^2\cos^2\varphi + b^2\sin^2\varphi}}.$$

The third geocentric co-ordinate type, ellipsoidal co-ordinates, are sometimes used in scientific work, but in land surveying they have no practical significance, see Heiskanen and Moritz (1967, pages 39–45).

The advantage of rectangular[2] co-ordinates is that with them, calculations are easy. E.g., the distance s between two points (X_1,Y_1,Z_1) and

[2]Rectangular co-ordinates are often called *Cartesian*, after René Descartes. Strictly speaking Cartesian co-ordinates have straight co-ordinate lines, whereas rectangular co-ordinates could also be curvilinear.

(X_2, Y_2, Z_2) is simply[3]

$$s_{12} = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2 + (Z_2 - Z_1)^2}.$$

If the points are given in the form $(h_1, \varphi_1, \lambda_1)$, $(h_2, \varphi_2, \lambda_2)$, the corresponding equation won't be quite as simple!

10.2 Topocentric co-ordinates

In practical measurement work, often local or *topocentric*, three-dimensional co-ordinates[4] are used, in which the origin is the location of measurement itself, the instrument (*instrument co-oordinates*). It is natural to use spherical co-ordinates (s, α, ζ), in which s is the slant range from the instrument, α is the azimuth or horizontal direction angle, and ζ is the zenith angle. From these, easily the rectangular co-ordinates of the signal or target are calculated:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = s \begin{bmatrix} \sin\zeta \cos\alpha \\ \sin\zeta \sin\alpha \\ \cos\zeta \end{bmatrix}.$$

Today's *total stations* or electronic tacheometers can give precisely these instrument co-ordinates, either in spherical[5] (s, α, ζ) or rectangular (x, y, z) form. Conventionally we write topocentric co-ordinates in *lower case*.

In figure 10.4 are depicted both the co-ordinate axes (x, y, z) of the topocentric system, and the axes (X, Y, Z) of the geocentric system. In this figure, the signal T may be a point to be measured on the Earth's surface, but also a satellite orbiting the Earth. In any case, the measurements are obtained always first topocentrically, i.e., with respect to the plane of the local horizon (grey circle) of the observation point K.

The transformation between these two rectangular three-dimensional systems is a three-dimensional similarity of Helmert transformation, on which more in the following.

---

[3] Of course this is the straight distance in space, often passing though the solid body of the Earth. Usually we are more interested in the distance over the Earth's surface.

[4] Greek *topos* = place; cf. *utopia* = non-existent place.

[5] Strictly speaking only gyrotheodolites can provide the absolute azimuth α. For an ordinary instrument, the unknown azimuth of the zero on the horizontal circle needs to be determined separately, typically in a network adjustment or by an astronomical azimuth determination (section 10.6).

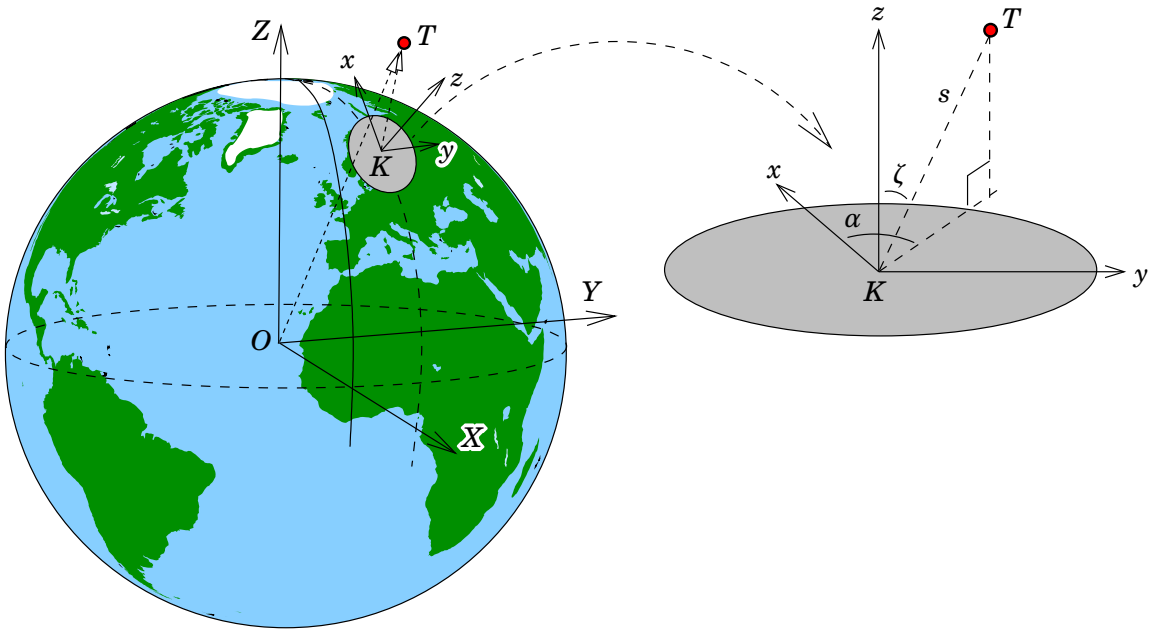


Figure 10.4. The topocentric or instrument co-ordinate frame, as well as the geocentric co-ordinate frame. The instrument is K, the geocentre O and the measured location or signal, T.

10.3 Three-dimensional transformations

A two-dimensional similarity or Helmert transformation is, if the Y axis is, as seen from the X axis, in the α direction:

$$\begin{bmatrix} X' \\ Y' \end{bmatrix} = K \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} X - X_0 \\ Y - Y_0 \end{bmatrix},$$

in which α is the rotation angle, K the scale ratio, and $\begin{bmatrix} X_0 & Y_0 \end{bmatrix}^\mathsf{T}$ are the co-ordinates of the origin of the new system, written in the old one.

The corresponding three-dimensional transformation formula is obtained by adding the Z axis and keeping it fixed:

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = K \begin{bmatrix} \cos\alpha_3 & \sin\alpha_3 & 0 \\ -\sin\alpha_3 & \cos\alpha_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X - X_0 \\ Y - Y_0 \\ Z \end{bmatrix}.$$

The 3×3 *rotation matrix* visible in the equation may be called $\overline{R}_3(\alpha_3)$.

In the same way as around the Z axis, also around the Y or X axis, rotations may take place. In that case we obtain analoguously the rotation matrices

$$\overline{R}_1(\alpha_1) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha_1 & \sin\alpha_1 \\ 0 & -\sin\alpha_1 & \cos\alpha_1 \end{bmatrix}$$

and

$$\overline{R}_2(\alpha_2) = \begin{bmatrix} \cos\alpha_2 & 0 & -\sin\alpha_2 \\ 0 & 1 & 0 \\ \sin\alpha_2 & 0 & \cos\alpha_2 \end{bmatrix}.$$

The general similarity or Helmert transformation containing all three rotations (and three translations, and a scaling) can now be expressed in the following compact form:

$$\overline{\mathbf{R}}' = K\overline{R}\left(\overline{\mathbf{R}} - \overline{\mathbf{R}}_0\right). \tag{10.1}$$

Here we write the vectors as column vectors of their *components*, i.e., geocentric co-ordinates of location, as follows:

$$\overline{\mathbf{R}}' = \begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix}, \quad \overline{\mathbf{R}} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \quad \overline{\mathbf{R}}_0 = \begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \end{bmatrix}.$$

Here, the overbar signals that we are dealing with column vectors of component values, not the vectors in space themselves. It may be left off when the context is clear[6].

Furthermore, the rotation

$$\overline{R} = \overline{R}_3(\alpha_3)\overline{R}_2(\alpha_2)\overline{R}_1(\alpha_1)$$

is the combination, or "chaining", of three rotations. Equation 10.1 is called the (three-dimensional) Helmert or similarity transformation. The elements of the size 3×3 matrix $\overline{R}$ representing the rotation of the co-ordinate frame are complicated trigonometric expressions in the angles $\alpha_1, \alpha_2, \alpha_3$ and we don't derive them here.

10.4 Transformation in case of small rotation angles

Often, the axes of two co-ordinate frames are very close to each other. In that case, rotation angles are small and one may make the approximation that $\sin\alpha \approx \alpha$ and $\cos\alpha \approx 1$. Then, the formulas become simpler. If, additionally, it may be assumed that the scale ratio K is close to unity, one may write

$$K = 1 + m,$$

[6] Note that

$$\mathbf{R} = X\mathbf{i} + Y\mathbf{j} + Z\mathbf{k} = X'\mathbf{i}' + Y'\mathbf{j}' + Z'\mathbf{k}',$$

in which $\{\mathbf{i},\mathbf{j},\mathbf{k}\}$ and $\{\mathbf{i}',\mathbf{j}',\mathbf{k}'\}$ form *orthonormal bases* of the space, for the old and the new co-ordinate frame respectively.

in which m, the *scale distortion*, is small. Additionally

$$\overline{R}_1(\alpha_1) \approx \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \alpha_1 \\ 0 & -\alpha_1 & 1 \end{bmatrix},$$

$$\overline{R}_2(\alpha_2) \approx \begin{bmatrix} 1 & 0 & -\alpha_2 \\ 0 & 1 & 0 \\ \alpha_2 & 0 & 1 \end{bmatrix},$$

$$\overline{R}_3(\alpha_3) \approx \begin{bmatrix} 1 & \alpha_3 & 0 \\ -\alpha_3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

If all α_i are small, one may furthermore assume that all $\alpha_i\alpha_j \approx 0$ if $i \neq j$. We obtain

$$\overline{R} = \overline{R}_3(\alpha_3)\overline{R}_2(\alpha_2)\overline{R}_1(\alpha_1) = \begin{bmatrix} 1 & \alpha_3 & -\alpha_2 \\ -\alpha_3 & 1 & \alpha_1 \\ \alpha_2 & -\alpha_1 & 1 \end{bmatrix} = I + \Delta\overline{R},$$

in which I is the 3×3 unit matrix, and

$$\Delta\overline{R} = \begin{bmatrix} 0 & \alpha_3 & -\alpha_2 \\ -\alpha_3 & 0 & \alpha_1 \\ \alpha_2 & -\alpha_1 & 0 \end{bmatrix}$$

is a skew-symmetric (antisymmetric) matrix: $\Delta\overline{R}^{\mathsf{T}} = -\Delta\overline{R}$.
From this follows the *co-ordinate correction formula*

$$\overline{\mathbf{R}}' - \overline{\mathbf{R}} = \left(m + \Delta\overline{R}\right)\left(\overline{\mathbf{R}} - \overline{\mathbf{R}}_0\right) = \begin{bmatrix} m & \alpha_3 & -\alpha_2 \\ -\alpha_3 & m & \alpha_1 \\ \alpha_2 & -\alpha_1 & m \end{bmatrix}\left(\overline{\mathbf{R}} - \overline{\mathbf{R}}_0\right), \quad (10.2)$$

in which $m, \alpha_1, \alpha_2, \alpha_3$ and $\overline{\mathbf{R}}' - \overline{\mathbf{R}}$ are all *small* (but $\overline{\mathbf{R}} - \overline{\mathbf{R}}_0$ is large).
The form 10.2 is the "small" form of the general Helmert transformation, between two realizations of co-ordinate reference systems that are close together, like, e.g., the different realizations of ITRS, the International Terrestrial reference System. In that case the rotation angles α_i are of order fraction of a second of arc, and the translation vector $\overline{\mathbf{R}}_0$ is under 10 cm.

10.5 The transformation between two reference ellipsoids

A classical case is the transformation between two geodetic datums which are defined on two non-geocentric reference ellipsoids. E.g., in Europe between the Hayford ellipsoid of the ED50 datum and Eastern Europe's

Krasovsky ellipsoid. A brute-force method is then to first convert geodetic co-ordinates (φ, λ, h) to rectangular (X, Y, Z), carry out a three-dimensional Helmert transformation between the two datums, and convert the result back into geodetic co-ordinates (φ, λ, h).

If the difference between the two datums is small and consists only of a shift of the reference ellipsoid's centre, it would be nice to know what relationship exists between the centre shift and the change in geodetic co-ordinates. Fortunately there is a simple equation for this. See figure 10.5.

Let the location vector of a point from the centre of one ellipsoid be $\mathbf{R}_1$, and from the other, $\mathbf{R}_2$, and the difference, in rectangular co-ordinates[7],

$$d\overline{\mathbf{R}} = \overline{\mathbf{R}}_2 - \overline{\mathbf{R}}_1 = \begin{bmatrix} X_2 - X_1 \\ Y_2 - Y_1 \\ Z_2 - Z_1 \end{bmatrix} = \begin{bmatrix} dX \\ dY \\ dZ \end{bmatrix}.$$

Around the point we define local topocentric co-ordinates (x, y, z) on a unit-vector, i.e., orthonormal, basis $\{\mathbf{N}, \mathbf{E}, \mathbf{U}\}$ ("North, East, Up"). At the location of the point, the principal radii of curvature of the reference ellipsoid are M — the meridional radius of curvature — and N — the transversal radius of curvature. Now the topocentric shifts are

$$\begin{aligned} x &= M(\varphi)\, d\varphi, \\ y &= N(\varphi) \cos\varphi d\lambda, \\ z &= dh, \end{aligned}$$

and

$$d\mathbf{R} = \mathbf{N}x + \mathbf{E}y + \mathbf{U}z = \mathbf{N}M d\varphi + \mathbf{E}N \cos\varphi d\lambda + \mathbf{U}dh.$$

In matrix notation this is

$$d\overline{\mathbf{R}} = \begin{bmatrix} dX \\ dY \\ dZ \end{bmatrix} = \begin{bmatrix} \overline{\mathbf{N}} & \overline{\mathbf{E}} & \overline{\mathbf{U}} \end{bmatrix} \begin{bmatrix} M d\varphi \\ N \cos\varphi d\lambda \\ dh \end{bmatrix} = \\ = \begin{bmatrix} N_X & E_X & U_X \\ N_Y & E_Y & U_Y \\ N_Z & E_Z & U_Z \end{bmatrix} \begin{bmatrix} M d\varphi \\ N \cos\varphi d\lambda \\ dh \end{bmatrix},$$

in which the matrix is orthogonal, in fact a rotation matrix:

$$\overline{R} = \begin{bmatrix} N_X & E_X & U_X \\ N_Y & E_Y & U_Y \\ N_Z & E_Z & U_Z \end{bmatrix} = \begin{bmatrix} -\sin\varphi\cos\lambda & -\cos\lambda & \cos\varphi\cos\lambda \\ -\sin\varphi\sin\lambda & \sin\lambda & \cos\varphi\sin\lambda \\ \cos\varphi & 0 & \sin\varphi \end{bmatrix}.$$

[7] Note that this is also the vector connecting the centres of the two reference ellipsoids!

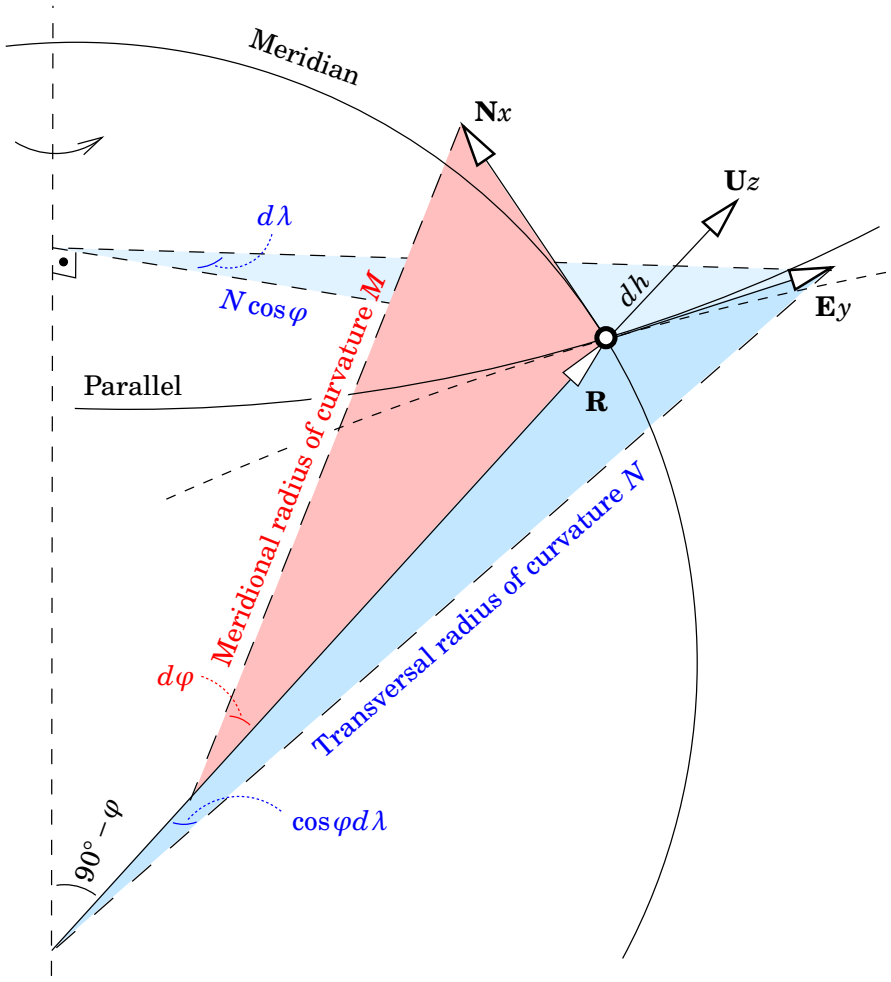


Figure 10.5. The differential connection between rectangular N,E,U co-ordinates and geodetic co-ordinates on the reference ellipsoid.

□

Inverting an orthogonal matrix is easy: $\overline{R}^{-1} = \overline{R}^{\mathsf{T}}$. I.e.,

$$\begin{bmatrix} M d\varphi \\ N \cos\varphi d\lambda \\ dh \end{bmatrix} = \begin{bmatrix} -\sin\varphi\cos\lambda & -\sin\varphi\sin\lambda & \cos\varphi \\ -\cos\lambda & \sin\lambda & 0 \\ \cos\varphi\cos\lambda & \cos\varphi\sin\lambda & \sin\varphi \end{bmatrix} \begin{bmatrix} dX \\ dY \\ dZ \end{bmatrix}.$$

Thus we may easily calculate what are the effects of shifting the centre of the reference ellipsoid on geodetic co-ordinates (φ, λ, h), evaluated on the ellipsoid:

$$\begin{bmatrix} \varphi_2 \\ \lambda_2 \\ h_2 \end{bmatrix} = \begin{bmatrix} \varphi_1 \\ \lambda_1 \\ h_1 \end{bmatrix} + \begin{bmatrix} M^{-1} & (N\cos\varphi)^{-1} & 1 \end{bmatrix} \overline{R}^{\mathsf{T}} \begin{bmatrix} X_2 - X_1 \\ Y_2 - Y_1 \\ Z_2 - Z_1 \end{bmatrix}.$$

And when the geodetic co-ordinates (φ, λ, h) change, then also the deviations of the plumb line and the heights of the geoid change, their

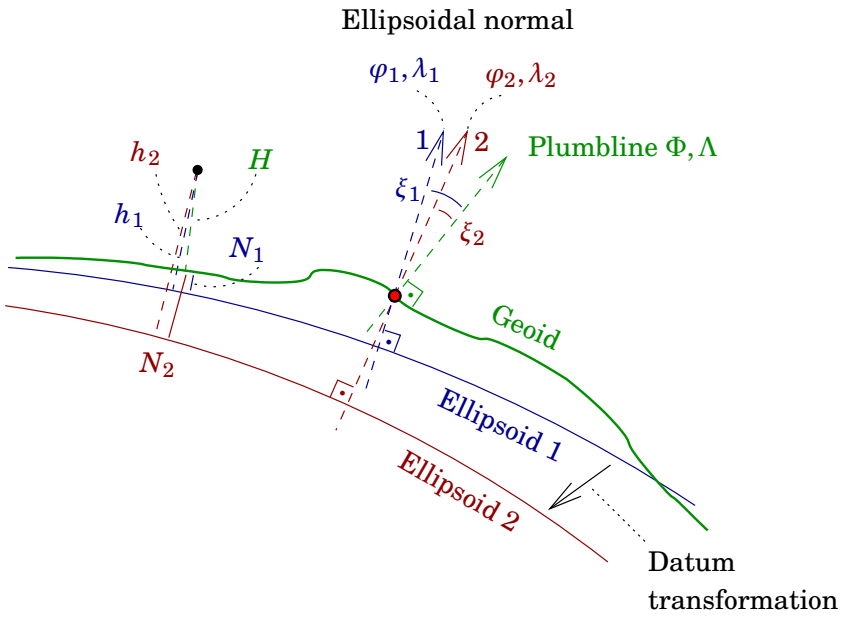


Figure 10.6. Effect of a datum transformation (shift of the centre of the reference ellipsoid) on geodetic latitude and longitude φ, λ, deviations of the plumb line ξ, η, geoid heights N and the heights h of points from the reference ellipsoid.

definitions being

$$\begin{aligned} \xi &= \Phi - \varphi, \\ \eta &= (\Lambda - \lambda)\cos\varphi, \\ N &= h - H, \end{aligned} \tag{10.3}$$

in which (Φ, Λ) are astronomically determined latitude and longitude, (ξ, η) are deviations of the plumb line in the North and East directions, h is the height above the ellipsoid, and H the height above sea level, while N is the geoid height reckoned from the reference ellipsoid. From this is obtained directly

$$\begin{bmatrix} -M(\varphi)\, d\xi \\ -N(\varphi)\, d\eta \\ dN \end{bmatrix} = \begin{bmatrix} -\sin\varphi\cos\lambda & -\sin\varphi\sin\lambda & \cos\varphi \\ -\cos\lambda & \sin\lambda & 0 \\ \cos\varphi\cos\lambda & \cos\varphi\sin\lambda & \sin\varphi \end{bmatrix} \begin{bmatrix} dX \\ dY \\ dZ \end{bmatrix},$$

because Φ, Λ and H can be calculated directly from measurements, without any use of a reference ellipsoid. See figure 10.6.

10.6 Laplace azimuth measurements

A geodetic network computed on the reference ellipsoid is oriented by astronomical observations. Two of the three orientation degrees of freedom are fixed by fixing the direction of the ellipsoidal normal in the datum point(s) to the astronomically determined direction of the plumb line.

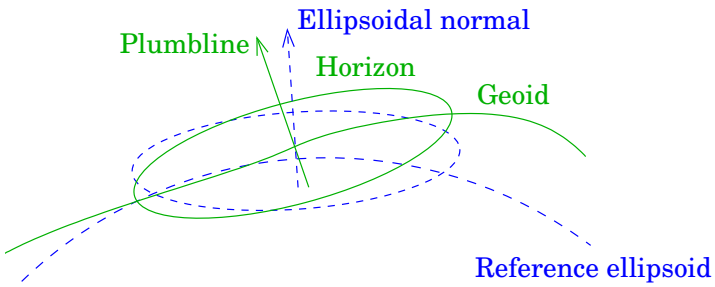


Figure 10.7. The deviation of the local plumb line from the normal on the reference ellipsoid surface.

□

That leaves a third degree of freedom, the network's orientation with respect to the local North.

The directions measured in the network are projected onto the plane of the local horizon. If one ignores the local deviations of the plumb line (i.e., one assumes that the local plane of the horizon is the same as the local tangent plane to the reference ellipsoid), one may say that the sighting directions are projected onto the reference ellipsoid. This assumption is however *not correct*. The local horizon stands perpendicular on local gravity, the direction of which differs a little from that of the normal to the reference ellipsoid. The phenomenon is known as *plumb-line deviation* (figure 10.7).

The deviations of the plumb line are ξ in the North-South direction, and η in the East-West direction. The equations were given already above, equations 10.3:

$$\begin{aligned} \xi &= \Phi - \varphi, \\ \eta &= (\Lambda - \lambda)\cos\Phi, \end{aligned}$$

in which (Φ, Λ) and (φ, λ) were also defined there. φ and λ, geographical latitude and longitude, are *computed* co-ordinates as can be found on a map, computed with respect to a certain reference ellipsoid[8].

Let the *astronomical azimuth* (absolute direction) measured with respect to the plane of the local horizon be α, and the same sighting on the reference ellipsoid, i.e., the *geodetic azimuth* $\overline{\alpha}$.

Then we may say that

$$\begin{aligned} \alpha - \overline{\alpha} &= \eta\tan\Phi + \left(\xi\sin\alpha - \eta\cos\alpha\right)\cot\zeta = \\ &= (\Lambda - \lambda)\sin\Phi + \left(\left(\Phi - \phi\right)\sin\alpha - (\Lambda - \lambda)\cos\alpha\cos\Phi\right)\cot\zeta \end{aligned} \tag{10.4}$$

[8]So this means that the deviations of the plumb line will depend on the reference ellipsoid chosen: the choice of the local reference ellipsoid or *datum* is often done so, that the square sum of plumb-line deviations is minimized over the area of interest. In other words, that the ellipsoid fits as well as possible to a level surface of the very local gravity field, the *geoid*.

in which ζ is the zenith angle. Equation 10.4 is called the *Laplace azimuth equation*. If the sighting direction is in the horizontal plane, then $\cot\zeta = 0$ and the correction above is a constant for the observation site, as the dependence on azimuth α vanishes.

Figure 10.8 below explains where both terms come from:

- The first term $\eta\tan\Phi$ comes from the projection of the direction of the celestial pole onto the local horizon being different from the projection onto the tangent plane to the reference ellipsoid. It depends on the height of the celestial pole, i.e., astronomical latitude Φ.
- The second term $\left(\xi\sin\alpha - \eta\cos\alpha\right)\cot\zeta$ comes from the difference in projections of the sighting direction onto the local horizon and onto the tangent plane of the ellipsoid. It depends on the zenith angle ζ of the sight direction, and vanishes if $\zeta = 90°$.

10.7 Traditional "2D+1D" co-ordinates

There have been for a long time in widespread use co-ordinate frames in which horizontal location and height are given separately. An example of this is the KKJ system, the Map Grid Co-ordinate System, which has long been in use in Finland but is now obsolete, and the height system N60. KKJ gives horizontal co-ordinates in the Gauss-Krüger projection on the Hayford ellipsoid, i.e., the International Ellipsoid of 1924. The co-ordinates are based on the so-called ED50 (European Datum 1950) system, which was created in 1950 by a joint adjustment of the triangulation networks of all Western European countries. This is a traditional, non-geocentric datum.

The N60 system gives orthometric heights, i.e., heights from the *geoid*, not the reference ellipsoid. The geoid (on which more will be said later, section 15.4) is an undulating reference surface similar to mean sea level. With the International Ellipsoid one must use the so-called Bomford[9] geoid model, which back in the days was determined in connection with the European Datum 1950 project.

Stating the three-dimensional location of a point in the form (x, y, H), where (x, y) is a KKJ co-ordinate pair, and H an orthometric N60 height, is problematic: the connection with the systems used by satellite positioning is complicated. Transforming the co-ordinates (x, y, H) into geocentric (X, Y, Z) co-ordinates comprises the following steps:

1. The KKJ co-ordinates (x, y) have already undergone a two-dimensional Helmert transformation, aimed at achieving an approxi-

[9] Brigadier Guy Bomford (1899 – 1996) was a gifted Brittish geodesist and student of geoid determination.

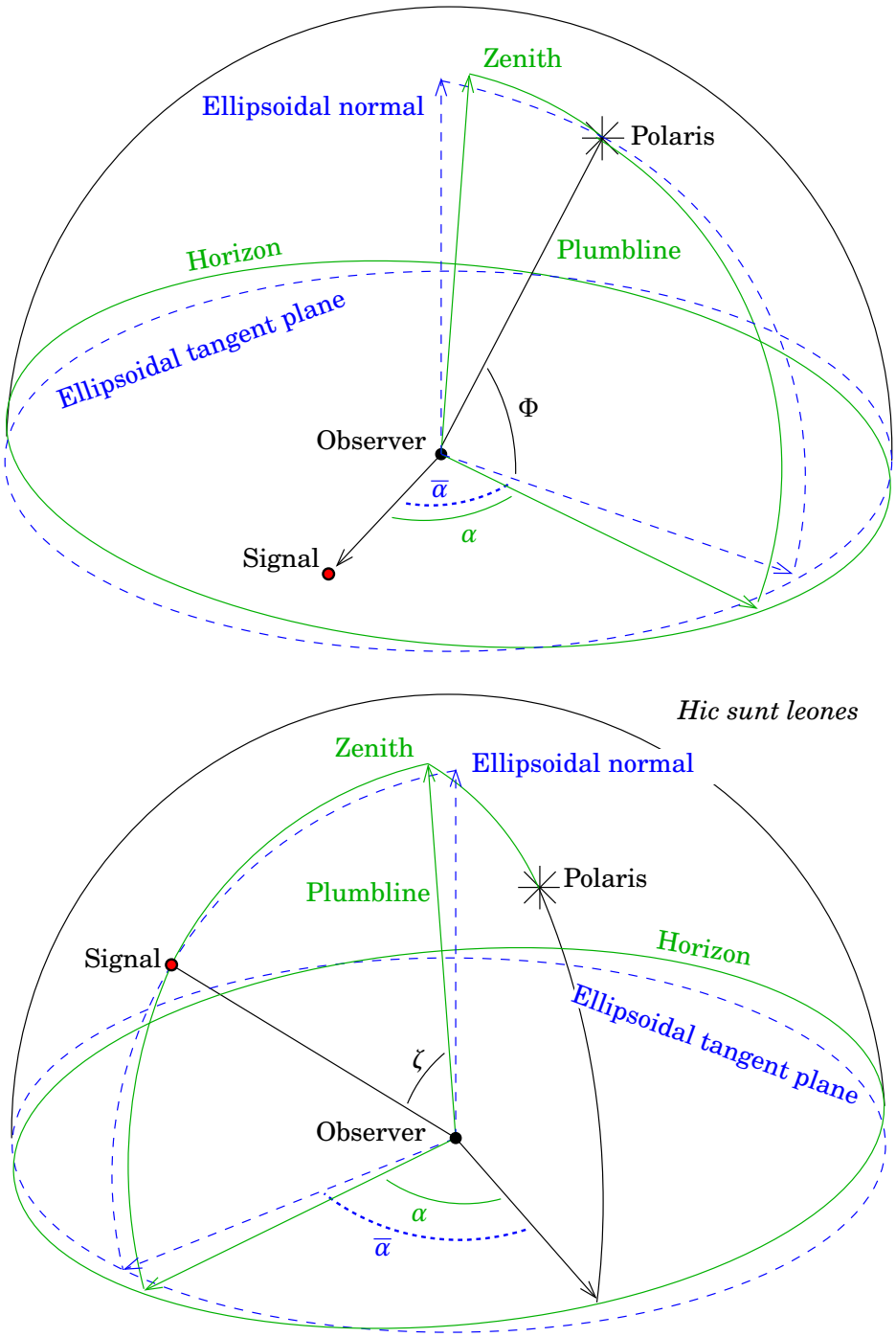


Figure 10.8. The Laplace phenomenon: the effect of the plumb-line deviation on the azimuth.

□

mate compatibility with the still older VVJ or "Helsinki System". This transformation is documented in Ollikainen (1993). The inverse of this transformation needs to be applied: $(x, y) \Longrightarrow (x', y')$.

2. The Gauss-Krüger projection is applied in the inverse direction, $(x', y') \Longrightarrow (\varphi, \lambda)$, on the International or Hayford ellipsoid.
3. In order to transform the orthometric height H into a height h from the reference ellipsoid, we need a geoid model that is compatible

with the reference ellipsoid used, i.e., Bomford's model. The geoid height $N = h - H$ is needed in every point.

4. The geodetic co-ordinates (φ, λ, h) are to be transformed into rectangular (X, Y, Z) co-ordinates. These three-dimensional co-ordinates are still in the European Datum 1950 system.
5. As the ED50 datum is non-geocentric, we still need a three-dimensional Helmert transformation to arrive at geocentric co-ordinates (X, Y, Z), see section 10.8 for details.

KKJ is on the way out in favour of new map projection systems, which are based either on the Gauss-Krüger or the UTM (Universal Transverse Mercator) projection on the GRS80 reference ellipsoid, and the EUREF-FIN datum, the ETRS89 (European Terrestrial Reference System 1989) system's Finnish national realization. As the height system, the new N2000 system is used, the connection of which to ellipsoidal heights is given by the *geoid model* FIN2005N00, see Bilker-Koivula and Ollikainen (2009). The above points 2 – 4 continue to apply, albeit with other names.

10.8 Case: the transformation between ED50 and EUREF89

This is a Helmert transformation of type 10.1, to an accurately geocentric system. Because the Hayford ellipsoid on which ED50 is based *is not* (accurately) geocentric, one must shift the origin of the co-ordinate frame from the centre of the reference ellipsoid to the centre of mass of the Earth. The shifts (translations) are of order a hundred metres, and also the rotations and the scale change are significant, see Ollikainen (1993, page 15 and table 2 on page 13):

$$\begin{bmatrix} X_2 \\ Y_2 \\ Z_2 \end{bmatrix} = (1+m) \begin{bmatrix} 1 & e_z & -e_y \\ -e_z & 1 & e_x \\ e_y & -e_x & 1 \end{bmatrix} \cdot \begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \end{bmatrix} + \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix},$$

in which, for the case EUREF89 → ED50, the transformation parameters are given in table 10.1 according to Matti Ollikainen's solution for the territory of Finland.

ED50 (European Datum 1950) is the traditional European datum on which KKJ is based; it was created well before the satellite age. As can be seen from the table, it is *non-geocentric*. EUREF89 is a modern, GNSS-based European co-ordinate reference frame. The precision figures given in the table are large because co-ordinates determined in the traditional way over a large area are just not very precise.

More recent information on matters of co-ordinates and transformations for the Finnish territory is found in Häkli et al. (2009).

Table 10.1. Transformation parameters from Ollikainen (1993) for EUREF89 → ED50.

Parameter	Value	Precision	Unit
ΔX	93.477	±3.345	m
ΔY	103.453	±5.534	m
ΔZ	123.431	±2.736	m
e_x	−0.246	±0.168	″
e_y	0.109	±0.106	″
e_z	0.068	±0.112	″
m	−2.062	±0.417	ppm

10.9 Case: The transformation between ITRF and ETRF

In Finland, a three-dimensional, satellite based, i.e., geocentric, co-ordinate reference frame called EUREF-FIN is in use. It is the national realization of ETRS89, the European Terrestrial Reference Frame, on the Finnish territory.

All geodetic satellite measurements, however, give a position solution in the same frame as that in which the GPS satellite orbital elements of the satellites are given, like ITRF2005. Then, the following transformation is needed to the corresponding ETRS89 realization, ETRF2005:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_{\text{ETRF2005}}(t) = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_{\text{ITRF2005}}(t) + \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix}^{\text{ETRF2005}}_{\text{ITRF2005}} + \\ + \begin{bmatrix} 0 & -\dot{R}_3 & \dot{R}_2 \\ \dot{R}_3 & 0 & -\dot{R}_1 \\ -\dot{R}_2 & \dot{R}_1 & 0 \end{bmatrix}^{\text{ETRF2005}}_{\text{ITRF2005}} \times (t - 1989.0) \times \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_{\text{ITRF2005}}(t),$$

in which the dot on the R parameters (Newton's dot notation) indicates derivation with respect to time. The $\dot{R}$ parameters in this equation contain the tectonic motion of the Eurasian plate.

The parameter values for the equation are found in the instructions written by the EUREF subcommission, and are presented in table 10.2.

As can be seen are the transformation parameters in this case many

Table 10.2. Transformation parameter values from the source Boucher and Altamimi 2007, tables 3 and 4.

Parameter	Value	Unit	Parameter	Value	Unit
T_1	5.6	cm	$\dot{R}_1$	0.054	10^{-3} ″/a
T_2	4.8	cm	$\dot{R}_2$	0.518	10^{-3} ″/a
T_3	−3.7	cm	$\dot{R}_3$	−0.781	10^{-3} ″/a

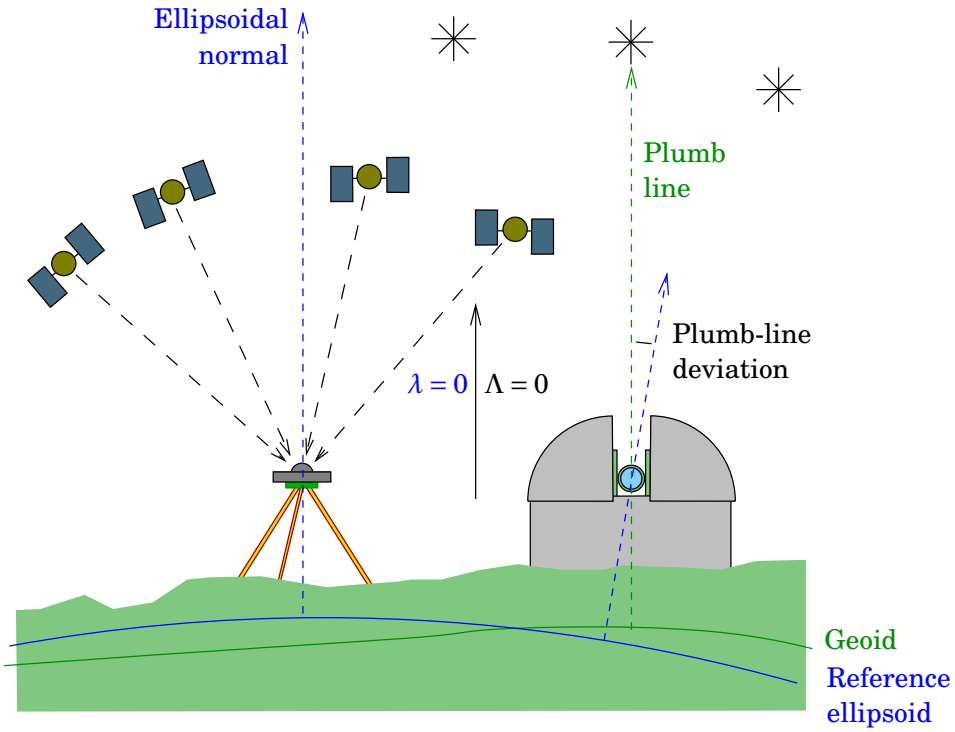


Figure 10.9. Greenwich geometry: zero longitude is a *direction*, not a *place*.

orders of magnitude smaller than in the earlier described case between EUREF-FIN and ED50. Both co-ordinate reference frames, ETRF2005 and ITRF2005, are geocentric on the centimetre level.

Self-test questions

1. Which are the two main types of geocentric co-ordinate reference systems?
2. What is sidereal time and what does it describe?
3. How many transformation parameters does a three-dimensional Helmert transformation have? ★
4. Name the parameters of the three-dimensional Helmert transformation.
5. Describe Laplace azimuth measurement and what it is used for.

Exercise 10 – 1: Greenwich: explain this

People have been taking their inexpensive hand-held GNSS receivers — mobile phones, even — to the Greenwich or zero meridian, figure 2.11, finding that *it does not show zero longitude*.

How is that possible?

There is actually a lot of good explanation on this on the Internet, including one *Journal of Geodesy* article from 2015 — and a lot of plain old

tabloid nonsense. Don't buy the nonsense. In this exercise, provide an explanation in your own words, showing that you "get" it.

Imagine that you're called up by a journalist who has heard about this, and wants your take on it as a geodesist. Your explanation — an elevator speech — should be so lucid that she "gets" it, and what's more, that when she gets home and tells her husband what she has learned, he "gets" it too...

11. Global Positioning System (GPS)

In land surveying the role of GPS, the Global Positioning System, has during the last two, three decades grown to be dominant, both in Finland and wordwide. The professional literature especially in the English language is extensive. In the Finnish language, a significant work is Poutanen (2016). In the English language a good basic work is Hofmann-Wellenhof et al. (2001).

Here we will concentrate on the GPS system, which has been fully operational for a long time. In parallel to this system, managed by the U.S. military authorities, similar systems by other countries have in recent years appeared on the scene. A special mention deserves the Russian GLONASS, which after a time of decay has now again grown back to operationality. In land surveying, satellites from both systems are already routinely used together.

The Europeans develop their own Galileo system, like also the Chinese their BeiDou or "Compass" system. The first satellites of both systems are already functioning in orbit.

These systems are, in the aggregate, called "GNSS systems", i.e., *Global Navigation Satellite Systems*. However, the relative simplicity of the way the original GPS system works makes it a suitable model for teaching the basics, and therefore we shall concentrate on it.

GPS is originally a *navigation system*. It is not the first radio navigation system using satellites: an earlier satellite navigation system was the Transit system, or NNSS (Navy Navigation Satellite System), inofficially the "Doppler positioning system". This system, which was operational 1964 – 1996, consisted of five satellites orbiting the Earth in low orbits. Geodetically useful positioning required several satellite passages overhead, amounting in practice to at least 24 hours of observations.

The GPS satellites are in much higher orbits, and anywhere on Earth, almost anytime, at least four of them are "visible". More commonly the number of visible satellites ranges from six to over ten[1]. Therefore GPS

[1]Positioning devices making use of more than one satellite system, e.g., GPS

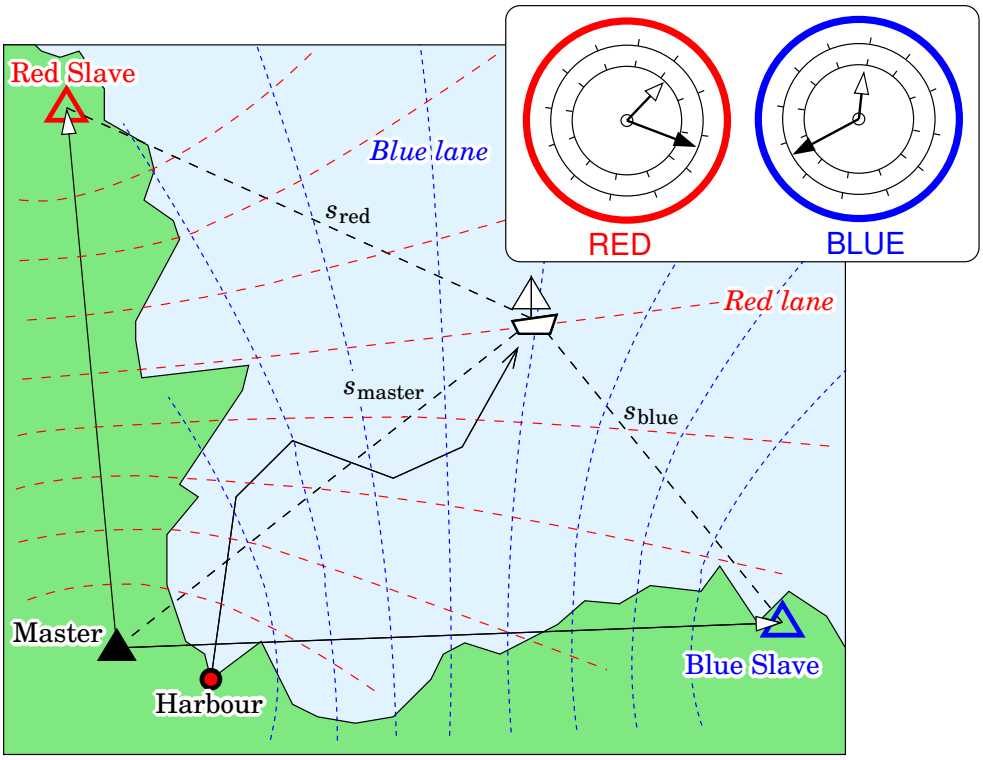


Figure 11.1. The Decca system.

positioning can be done almost instantly, within a few seconds or minutes.

11.1 Radio navigation and hyperbolic systems

Of the older terrestial radio navigation systems for seafarers may be mentioned the *Decca system*, closed down in 2000, which is an example of a *hyperbolic system*. Other systems worth mentioning are LORAN-C (no longer in use) and Omega.

Decca transmitted non-modulated carrier waves in the frequency band 70 – 130 kHz, corresponding to wavelengths 2.3 – 4.3 km. It required at least three transmitters, one "master" and at least two "slave" stations. The transmissions of the base stations were precisely synchronised, though they were all transmitting on different wavelengths.

The name "hyperbolic system" is based on the circumstance that a vessel not having a clock synchronized with the transmitters, can observe only the *difference* in reception times between the waves from two transmitters. See figure 11.1.

On the map are drawn, in two different colours, *hyperbolas*, curves of

and GLONASS, "see" even more satellites even in poor-visibility locations like the centres of big cities.

Figure 11.2. A Decca receiver. © 2005 Wikimedia Commons.

□

which the points have a *difference* in distance from the master and a certain slave that is constant. E.g., for the red hyperbolas we have

$$s_{\text{red}} - s_{\text{master}} = \text{constant}$$

and for the blue hyperbolas

$$s_{\text{blue}} - s_{\text{master}} = \text{constant}.$$

Every curve has its own constant. Unfortunately *this constant cannot be observed*, because all carrier waves look the same[2].

Therefore, an *incremental* measurement method is used. Each *lane counter* of the Decca instrument has two hands, one for the lane number, the other for the fraction within a lane. The hands move together, just like the hands of a clock.

The ship must set the lane counters to the correct starting values at a known location, e.g., the port of departure. After that, during cruising, they follow the development over time of the lane values: every lane pointer on the phase difference dial follows, how many full turns the phase pointer has made — see figures 11.1, 11.2. The precondition is, that the *radio connection with the base stations stays uninterrupted*[3].

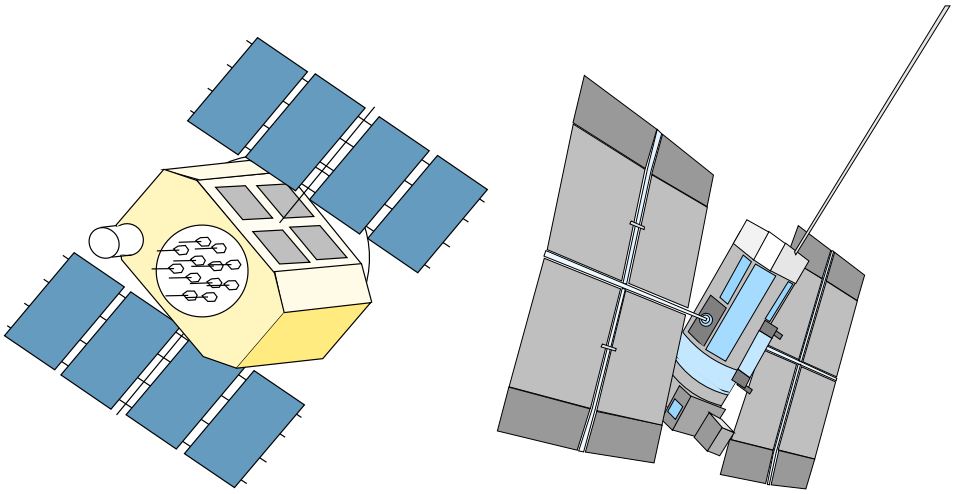
At any point in time, one may use the two[4] lane numbers and residual phase differences — fractional numbers $\in [0, 2\pi)$ — to read one's own location on a sea chart, on which the hyperbolas are pre-drawn.

---

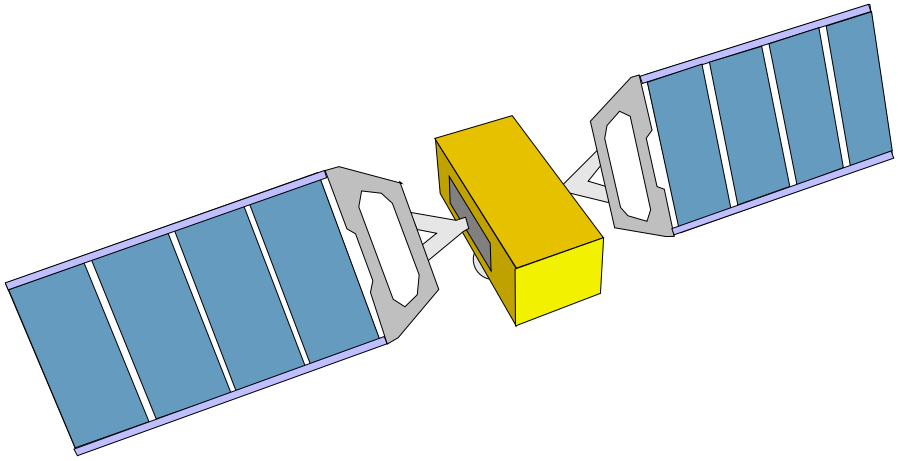
[2] In connection with the GPS system this is called the *ambiguity problem*.

[3] This is referred to as a *kinematic method*. One doesn't measure where one is, but rather follows continuously how one is moving in relation to a known starting point.

[4] Actually Decca used *three* colours: *red*, *blue* and *purple*.



(a) GPS (Block II) **(b)** GLONASS



(c) Galileo

Figure 11.3. Positioning satellites.

11.2 The GPS satellite

Characteristic of the GPS positioning system, being originally a military system, is that the *satellites are active and the users passive*. So, the positioning devices used by the users, the GPS receivers, are quiet, wheras the satellites contain radio transmitters. A GPS satellite is in a way a flying Decca base station.

A GPS satellite is big, a contraption looking like in figure 11.3a. It contains, among others, the following components:

- a precise *atomic clock*, either a caesium, a rubidium, or a hydrogen-maser clock. This clock synchronizes all signals (carrier waves and modulations) which the satellite transmits.
- *radio transmitters*. The antennas of the satellite point all the time to the Earth. Transmission power is significant only within a directional cone that contains the Earth whole; total power is about

50 W. Two carrier-wave frequencies are used, 1575.24 MHz (L_1) and 1227.60 MHz (L_2), which enables the elimination of the effect of the ionosphere. Modulated on the carrier wave are the various codes that are used in positioning, as well as a code containing orbit and other information for the users.

- *communication channels*. The satellite receives data from the GPS control centre. Besides control commands, this data contains information on the orbits, clock corrections, "health", etc. of all GPS satellites. The information is stored into the satellite's memory and transmitted forward to the users as a message modulated on the radio signal.
- *solar panels* produce the power needed by the equipment. The amount of power has grown from 400 W (Block I) to almost three kilowatts (Block IIF). At times, the satellite moves through the Earth's shadow; for this, there are batteries.
- *rocket engines* for controlling attitude and orbit, as well as a *stock of propellant* (monopropellant hydrazine). Because of orbit perturbations, orbit corrections are needed at regular intervals.
- the satellites are three-axes stabilized: the antennas are pointing to the Earth, the solar panels to the Sun. For stabilization, *reaction wheels* ("flywheels") are used.

As, during the lifetime of the GPS system, the field of electronics has seen a huge development, there are several satellite generations: "Block I", "Block II/IIA", "Block IIR", "Block IIR-M" and the newest "Block IIF" (of which the first was launched 2010), see Misra and Enge (2001). The first Block IIIA satellite is planned to be launched in 2018.

Satellites in operation today are all Block II or higher. The masses of the satellites are 845 kg (Block I), 1500 kg (Block II) and 2000 kg (Block IIR-M). The planned lifetime of the satellites — limited by the propellant stock for orbit maintenance, the diminishing power of the solar cells and batteries, as well as the development of defects in clocks and electronics in the Earth's outer radiation belt — is 4.5 (Block I), 7.5 (Block II) or today even 10 years. The satellites have regularly exceeded their planned lifetimes.

11.3 The GPS system

11.3.1 Segments

The GPS system consists of *three segments*:

Space Segment: the satellites themselves.

Control Segment: command centre, tracking stations, time synchronization, orbit determination, control.

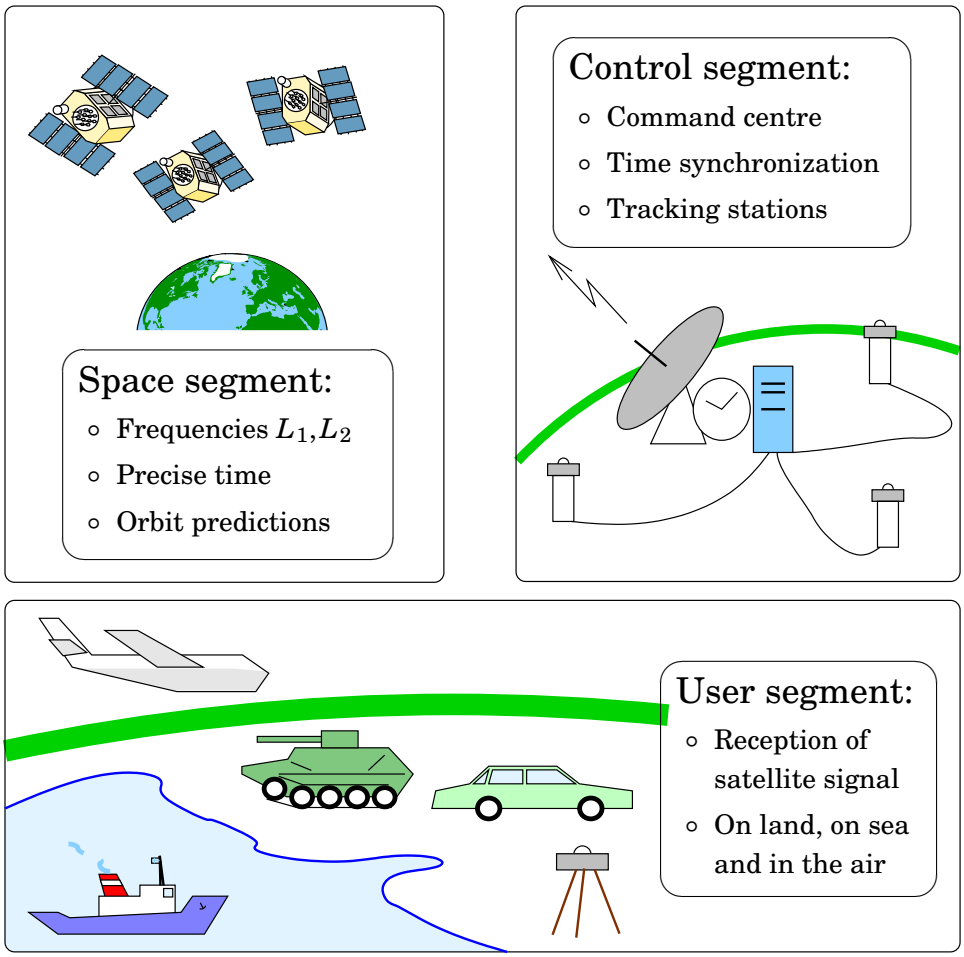


Figure 11.4. The three segments of the GPS system.

The main command centre, or Master Control Station, of the GPS system is located in Colorado Springs. There are some two dozen orbit tracking and control stations around the globe. Like the whole GPS system, also the control segment resides under the US defence department, more precisely the US Air Force.

Every control and tracking station is equipped appropriately with, i.a., a GPS receiver using a precise caesium clock.

Once every 24 hours, new orbital data — "broadcast ephemeris" — and correction information for the satellite's atomic clock is uploaded to the satellites. The satellites include these orbit and clock data into the radio signal they transmit, to be used by all users.

User Segment: all users, on land, at sea and in the air (and more and more also in space, in low Earth orbits)

See figure 11.4.

11.3.2 The constellation

The planned constellation of the GPS system consists of 24 satellites and three "active spares", which can be moved into place immediately when an active satellite breaks down. The satellites are in six different orbital planes, in each plane there are four satellites. In reality there are today over 30 satellites operating.

The height of the orbits from the Earth's surface is 20,200 km. The orbital period around the Earth is $11^{\mathrm{h}}58^{\mathrm{m}}$, i.e., in the time it takes the Earth to rotate once around her axis ($23^{\mathrm{h}}56^{\mathrm{m}}$) the satellites are seen again in the same spot in the sky as the previous day. *The GPS measurement geometry repeats every day four minutes earlier*, because the length of the day in the solar time used by our clocks is four minutes longer than the rotation period of the Earth.

The tilt of the orbital plane, or *inclination*, is[5] $i = 55°$. As a result of this, the GPS constellation's geometry isn't very strong at high latitudes, the satellites are mostly in the Southern sky.

The system has currently so complete coverage, that at least four satellites are "visible" (i.e., above an elevation angle of 15°) anywhere on Earth at any time. The number of visible satellites is almost always, and usually substantially, larger than this.

11.4 The codes in the GPS signal

Onto the carrier waves transmitted by the GPS satellites on two different frequencies, two so-called *pseudo-random codes* are modulated: the C/A code and the P code. Additionally there is still the navigation message — containing "broadcast ephemeris" and almanac data —, which is modulated onto the carrier waves as well[6]. See table 11.1.

The modulation technique used is *phase modulation*[7]: the phase of the carrier is turned "upside down", or 180°, or π, when the code switches between 0 and 1.

---

[5] The Block I satellites had a different inclination $i = 63°$. Of those satellites, none are working any more today.

[6] In connection with the GPS modernization, a frequency L_5, 1176.45 MHz, has been added to the signal. It is meant to be used by rescue services (SoL, Safety of Life). Additionally, also to the L_1 and L_2 frequencies new civilian and military codes have been added.

[7] Other modulation types in existence are *amplitude modulation* — where the strength of the carrier, the amplitude, is varied in the rhythm of the signal – and *frequency modulation* — where the frequency of the carrier is made to vary. Common radio stations use amplitude modulation, and FM stations, frequency modulation.

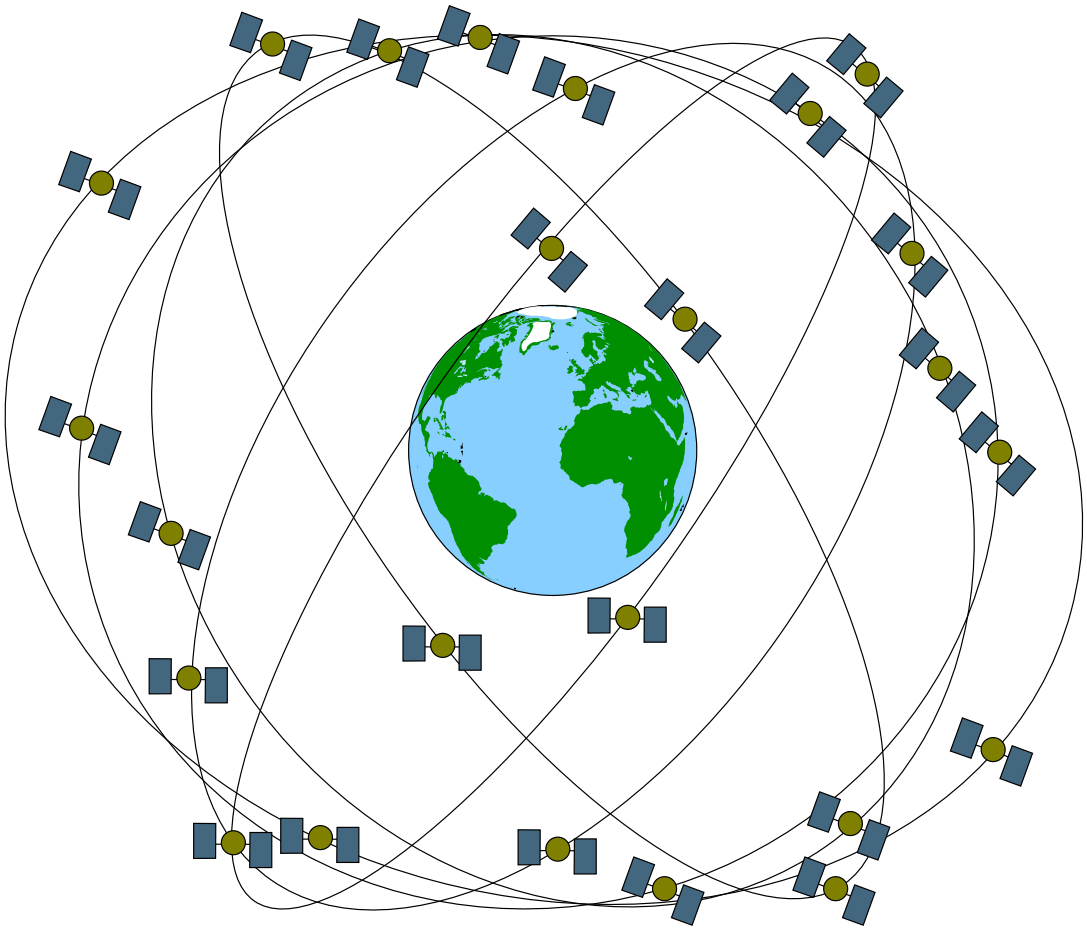


Figure 11.5. The GPS constellation. The orbits and places of satellites are realistic with respect to the Earth.

Because of the modulation, the signal transmitted by the GPS satellites is pretty broadband. The effective bandwidth is several times the bit frequency of the P code, i.e., several tens of MHz.

When already the transmission bandwidth requirement for a single satellite is this large, one might think that the bandwidth demand of the whole constellation would be huge. This is however not so: *all satellites use the same* carrier-wave frequencies L_1 and L_2. The receiver is able to

Table 11.1. Codes included in the GPS signal.

Name	Explanation	Modulation frequency	Repeat period	Carrier wave
C/A	Coarse / Acquisition	1.023 Mb/s	1 ms	L_1
P	Precise / Protected	10.23 Mb/s	1week	L_1, L_2
Y	A combination of P and the secret W code	10.23 Mb/s		L_1, L_2
-	Navigation message	50 bits/s	continuous	L_1, L_2

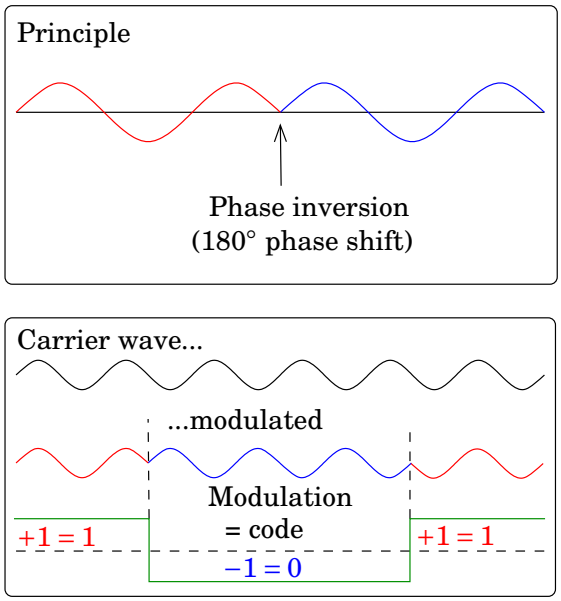


Figure 11.6. The principle of phase modulation.

separate the signals of different satellites from each other with the aid of their different *pseudo-random codes* (C/A and P). Every satellite has her own code, or "fingerprint", in the same way that in navigation at sea, every lighthouse has its own flashing sequence. The technical solution goes by the name CDMA, or "code division multiple access".

The *navigation message* is a bit stream containing, besides precise orbital information on this satellite, crude orbital information — an *almanac* — on *all* satellites. Locking on to one satellite signal is sufficient to receive the approximate orbital information for all satellites.

The pseudo-random codes are generated according to a documented mathematical recipe. The codes thus are not genuinely random: they can be exactly reconstructed using the same recipe. They behave however statistically like genuinely random bit sequences.

11.4.1 The "tree rings" of GPS

The GPS receiver's antenna on the Earth's surface receives signals from *all* the satellites that are in the sky at the moment of observation. This whole "soup" travels along the cable to the receiver's electronics[8]. Here, the first two tasks are performed:

1. Separate the signals of different satellites from each other using

[8]Generally, in the pre-amplifier of the antenna, the analogue signal is processed to bring the carrier-wave frequency down to a much lower value, without affecting the modulation ("downbanding"). This prevents crosstalk of the amplified signal back into the antenna, and makes further processing easier, like the digitization of the signal in an A-D (analogue-to-digital) converter.

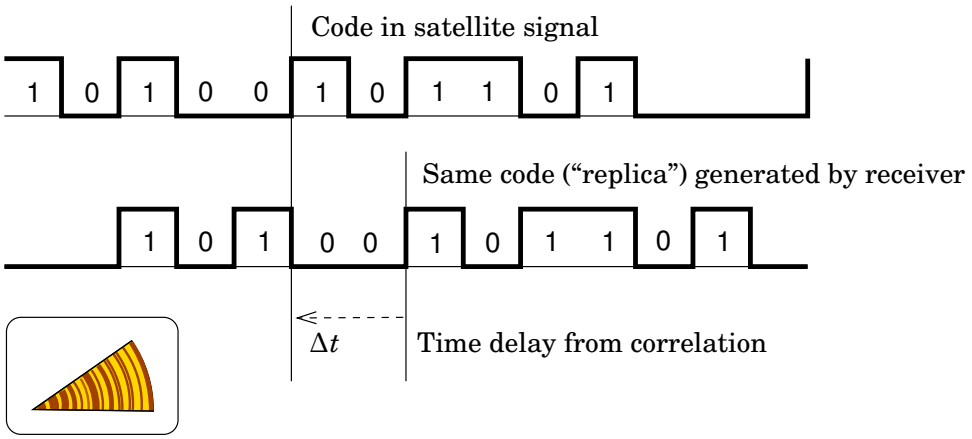


Figure 11.7. The correlation method for determining the travel time Δt of the GPS signal.

their individual pseudo-random code or “fingerprint[9]”.

2. Determine the *travel time* of the signal of each satellite from satellite to receiver.

[9]The pseudo-random codes of the different satellites have been carefully designed to be *orthogonal* to each other, i.e., the true signal of one satellite correlates as weakly as possible with the replica signal of another satellite generated in the receiver.

Tableau 11.2. How does dendrochronology work?

The method works as follows: the laboratory holds a *reference sequence* of tree rings, which has been built from partially overlapping tree-ring sequences. Rainy years show as thick, dry years as narrow rings[a]. In the reference sequence, the true, absolute year number of every ring is known.
Building a reference sequence is challenging. After doing so, however, the age of any wooden object can be determined by comparing its tree rings with the reference sequence, until the place is found where they match (*correlate*). The method works because the succession of wet and dry years is largely random.
Similar methods are in use in many fields of science: dating and correlation of ice drilling cores or geological deposits, correlation of magnetization stripes on the sea floor, etc.

[a]However, at the tree line in the mountains or the Arctic, tree ring thickness is controlled mostly by *temperature*. This circumstance has been used for reconstructing paleotemperature time series. One says that the tree-ring width is a *proxy* for temperature.

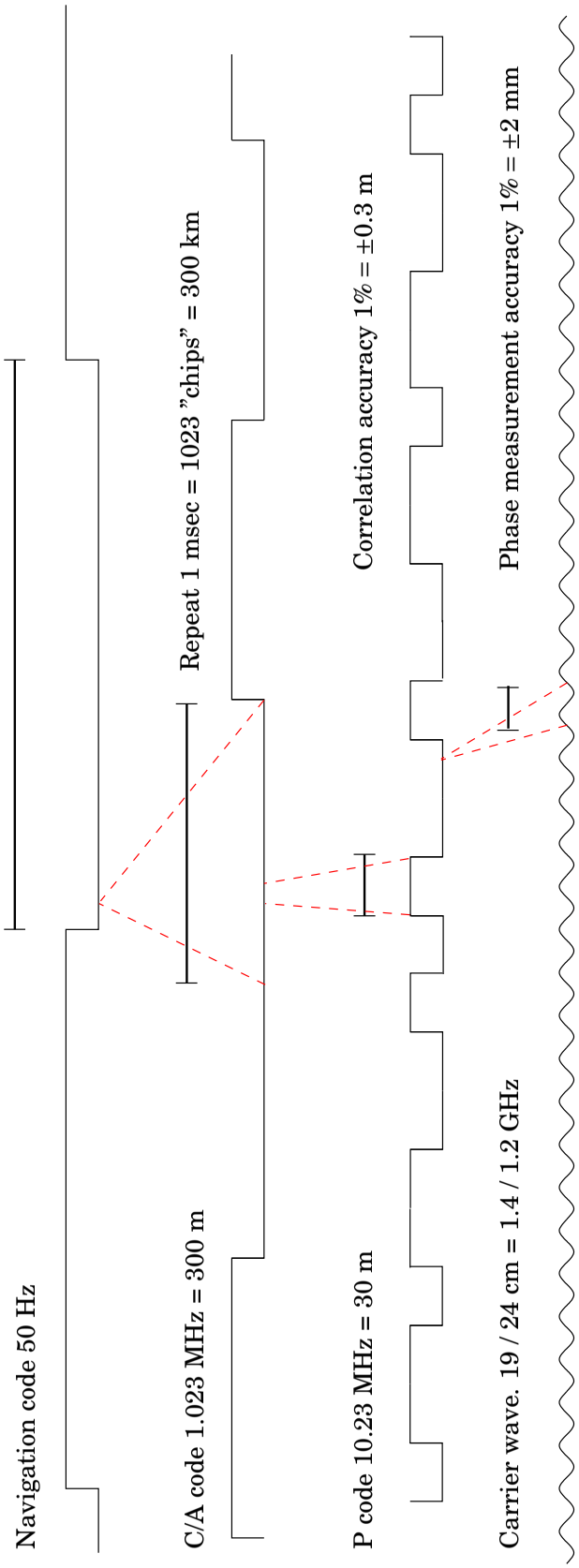


Figure 11.8. The various frequencies and effective "wavelengths" of the GPS signal.

Both tasks are carried out by a *correlation method*, which seeks the time shift Δt between the signal received from the satellite and a signal of the same form, a "replica", generated by the receiver, which would make them identical. The sequences are shifted alongside each other until a strong correlation is found, a similarity or correspondence of the pattern. The time difference obtained is essentially the GPS observable.

A suitable metaphor for the correlation method employed is the use of *year rings* to date wooden objects, or *dendrochronology*, which was already shortly explained in figure 6.7 on page 163, and on which more in the text tableau 11.2 on page 264.

In a way, also the comparison of the received GPS code with the replica code generated by the receiver is "dating": the "age" of the signal travelling from the satellite to Earth is determined...

When the correlation processing yields $\Delta t = t_{\text{rcv}} - t_{\text{xmit}}$, multiplying it with the propagation speed c of the signal will give the *pseudo-range* to the satellite, the basic observable of GPS measurement. It is called a *pseudo*-range, because it contains more than just the geometric distance, i.a., the *clock offsets*. We will return to this presently.

11.4.2 C/A code and P code

Because the one-millisecond C/A code consists only of 1023 bits, one must look at only 1023 alternative time shifts Δt. This goes very quickly. Every satellite has her own "personal" C/A code; therefore, initially there have to be made $N \times 1023$ comparisons, where N is the number of satellites.

With the C/A code one can determine the pseudo-range up to a multiple of 300 kilometres, because the code repeats every millisecond, the time in which the signal travels 300 kilometres. See figure 11.8. This is good enough if the approximate location of the receiver is already known with this accuracy. The state of modulation of the C/A code switches, if it switches, at intervals of 1μs ("chip rate"), the time in which the radio signal travels 300 m. The *accuracy of measurement* when using the C/A code is better than this: if the receiver electronics can measure the phase of the modulation with 1% accuracy, then measurement accuracy is ± 3 m.

A greater accuracy is offered by the *P code*. It too is a pseudo-random code, but its length is no less than 267 days. Every satellite uses its own, satellite specific, one week long subinterval of this period. The receiver must also in this case be able to generate a "replica" of the code. Because, however, the quantity Δt has already been obtained, using the C/A code, to millisecond accuracy, only decimals more precise than this need to be looked at.

The P-code "chip rate" or bit frequency is ten times faster than that of the

C/A code, 10.23 Mb/s, which corresponds to a traveled distance of 30 m[10], and, again assuming 1% phase measurement accuracy, an accuracy of the pseudo-range observation of ±30 cm.

The P code is kept from civilian users by encryption. The encryption is done by modulating, on top of the P code, a W code, of which the generating algorithm has not been published.

On the carrier frequency L_1 are thus modulated both the P and the C/A codes. Distinguishing between them has been made possible by using so-called *phase quadrature*: while the P code is modulated with phase shifts 0 (bit value 0) and π (bit value 1), the C/A code is similarly modulated with phase shifts $+\pi/2$ and $-\pi/2$. One speaks of "in-phase" and "quadrature" modulations.

The phase angles of the navigation message are the same as those of the C/A code. This is not a problem as their frequencies are so different: within one bit of the navigation message, duration 20 μs, fit twenty repeats of the C/A code! See figure 11.8.

11.5 GPS receivers

The receivers intended for precise geodetic work are always *dual-frequency instruments* that can measure the carrier phase of the GPS signal. Unlike with inexpensive hand-held devices, the antenna is usually separate, and connected to the receiver by a coax cable. The weak satellite signal is amplified already inside the antenna by a *pre-amplifier*[11].

The antenna may be mounted, using a standard forced-centring device, onto a geodetic tripod. In real-time mapping surveys however, a *measuring staff* is used, with the antenna screwed onto the top, and to which is attached the GPS receiver (if separate) with peripherals. The antenna has a standard 5/8 inch hole at the bottom with a similar standard screw thread (11 threads per inch) as in most geodetic instruments.

With every receiver type comes its own antenna type; the antennas and receivers of different manufacturers are not generally electrically intercompatible. However, for precise geodetic work there is the *choke-ring*

[10] This is the "effective wavelength" of the P code. It is calculated as follows:

$$\lambda_{eff} = c/f,$$

in which f is the "chip rate", a frequency-like quantity, unit s^{-1}, and c is the speed of light. So if c = 300,000,000 m and f = 10,000,000 s^{-1}, it follows that λ = 30 m.

[11] The DC voltage feed required by the pre-amplifier comes from the receiver, also through the coax cable. This complicates or prevents the mixing of receivers and antannas from different brands.

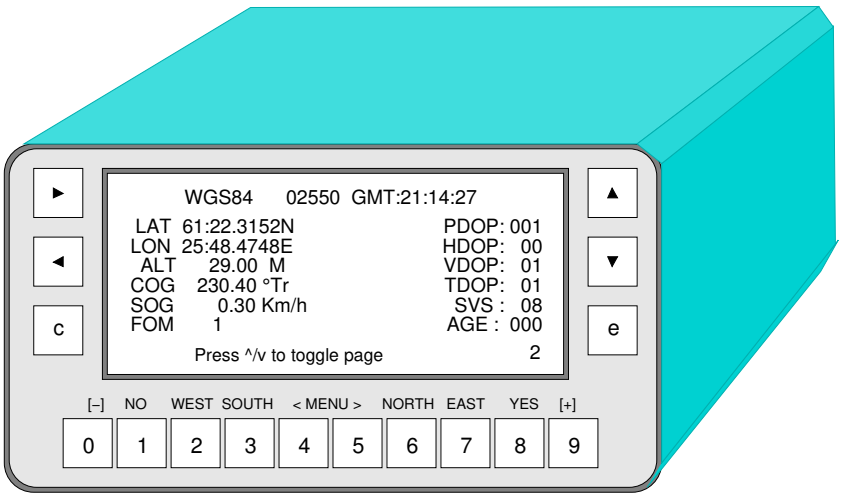


Figure 11.9. Control panel of the Ashtech Z-12.

antenna model (figure 11.10) which is available from many different manufacturers. The choke rings lessen the problem of reflection of radio waves from the Earth's surface and other surfaces, so-called *multipath*.

The *electric centre* of the antenna, the point where, in the geometric interpretation, the radio waves are apparently received, is not the same as the antenna's official reference point (ARP). It is not even unambiguously defined, but depends somewhat on the used cut-off elevation angle for the observations, see figure 11.11. As a metaphor, one may think of the apparent place of a fish under water, which also depends on the angle of view. One speaks of the *variation of the antenna's phase centre*, see Poutanen (1998, page 139).

When measurements are carried out over a relatively small area, using only one type of antenna, this variation of the phase centre vanishes from the end result, and from the position *difference* vectors computed between different points in the network. If, however, antenna types are mixed, or extensive networks measured — hundreds or thousands of kilometres across — one ought to *calibrate* the phase delay patterns of the antennas, which are fairly complicated functions of both the elevation angle η and the azimuth direction α. The calibration, in which this phase delay pattern $\Delta\phi\left(\eta,\alpha\right)$ is determined, can be carried out in the laboratory using an artificial GNSS signal source, or as a field calibration in which always two antenna types are compared to each other. Field calibration is thus always *relative*, referring to some agreed reference antenna type.

In highly precise geodynamic deformation measurements it is nowadays the practice to calibrate, not just antenna *types*, but *individual antennas*.

The radio waves transmitted by GPS satellites are clockwise *circularly polarized*; upon reflection, the polarization direction reverses to counter-

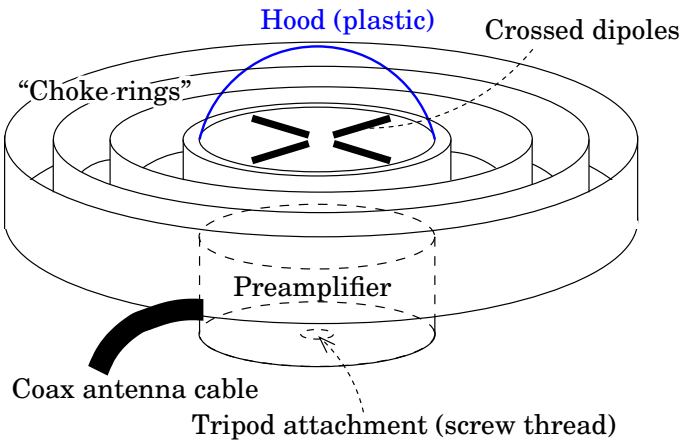


Figure 11.10. A so-called *choke-ring* GNSS antenna for precise geodetic work. The antennas provided by manufacturers are simpler and especially smaller and lighter.

clockwise. The antenna (in the example in the figure, a cross dipole) is built so, that it transfers only the clockwise polarized signal on to the receiver. In this way, the harm caused by reflections is minimized.

Technological development goes into the direction of greater integration. Today's GPS positioning devices, also the geodetic ones, as so small that they are integrated with the antenna. Because the devices are very autonomous, they don't even have a proper display screen any more.

Another development is the onmarch of *software defined receivers*: today's personal computers are beginning to be powerful enough to do the digital processing work that in today's receivers is done on the hardware level. Then, one needs in addition to a general-purpose PC, only a “dumb” analogue radio device with antenna (Lázaro, 2012).

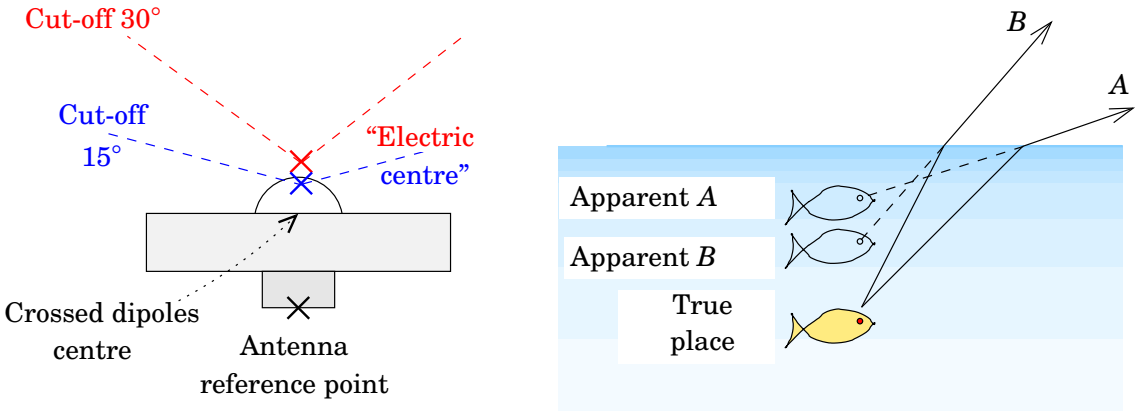


Figure 11.11. The “electric centre” of an antenna is not a self-evident thing!

11.6 The observables of GPS

A geodetic GPS receiver stores the observations it makes into its memory as a long table containing many numbers. The numbers in the table represent *distances* between the receiver and the different satellites observed. It is easy to understand why the amount of numbers becomes so large: if, e.g., the time between measurements is 30 seconds, and five satellites are visible in the local sky, and with a single-frequency receiver we observe both the C/A code and the P code, then the number of observations collected within one minute already amounts to $(60/30)\cdot 5\cdot 2 = 20$. In an hour, this means 1200 observations; if these are stored into the memory as ordinary double precision[12] real numbers (eight bytes per number), we need 9.6 kilobytes. In 24 hours, the storage requirement would then be 230.4 kB.

An often used format in international GNSS data exchange is RINEX, Receiver Independent EXchange format, Gurtner and Estey (2007). This is a text format, i.e., human-readable, in which observational data from most geodetic receiver types can be transferred, read, and processed, in a way that is independent from the manufacturer of the receiver. See the example in table 11.3 from the station Diego Garcia in the Indian Ocean.

11.6.1 Pseudo-ranges as observables

Why does one speak of *pseudo*-ranges? The prefix "pseudo" is because the observable's value is affected by, in addition to the geometric distance between satellite and receiver, also the clock errors or *offsets* Δt and ΔT, as well as the propagation delays caused by the ionospheric and tropospheric media[13]. Thus is obtained the *observation equation of pseudo-range*:

$$p = \rho + c(\Delta t - \Delta T) + d_{\text{ion}} + d_{\text{trop}}, \tag{11.1}$$

in which

p pseudo-range

ρ natural (geometric) distance. According to Pythagoras

$$\rho = \sqrt{(x-X)^2 + (y-Y)^2 + (z-Z)^2},$$

in which

[12] The "real" numbers of computers are in fact rational numbers...

[13] More precisely: the effects of free electrons and neutral molecules, respectively, on the propagation of radio waves. Therefore also the stratosphere causes "tropospheric delay".

Table 11.3. Start of a RINEX file. The device collects *five observation types*: carrier-phase observations and P-code observations on both frequencies L_1 and L_2, as well as C/A code observations on frequency L_1. The observations are stored at intervals of *30 seconds*. There are eleven satellites on the first epoch, *1. January 2000 0 : 00 : 00*, and also on the second epoch, *0 : 00 : 30*. They are all GPS satellites (G). The observation station is DGAR, Diego Garcia in the Indian Ocean (http://en.wikipedia.org/wiki/Diego_Garcia).

```
2.00           OBSERVATION DATA    G (GPS)             RINEX VERSION / TYPE
teqc  1999Oct8      gpsops              20000103 21:13:37UTCPGM / RUN BY / DATE
OSF1 V4.0 564|Alpha|cc 4.4.18.4|=+|=|                       COMMENT
DGAR                                                        MARKER NAME
30802M001                                                   MARKER NUMBER
GNOG                JPL                                     OBSERVER / AGENCY
T341U               AOA SNR-8000 ACT    3.3.32.3            REC # / TYPE / VERS
250                 AOAD/M_T                                ANT # / TYPE
  1916269.8405  6029977.3167  -801720.2273                  APPROX POSITION XYZ
        0.0814        0.0000        0.0000                  ANTENNA: DELTA H/E/N
     1     1                                                WAVELENGTH FACT L1/2
     5    L1    L2    P1    P2    C1                       # / TYPES OF OBSERV
    30.0000                                                 INTERVAL
                                                            COMMENT
This data is provided as a public service by NASA/JPL.      COMMENT
No warranty is expressed or implied regarding suitability   COMMENT
for use.  For further information, contact:                 COMMENT
Dave Stowers, NASA/JPL m/s 238-600                          COMMENT
4800 Oak Grove Drive, Pasadena CA 91109 USA                 COMMENT
                                                            COMMENT
  2000     1     1     0     0    0.0000000     GPS         TIME OF FIRST OBS
                                                            END OF HEADER
 00  1  1  0  0  0.0000000  0 11G21G23G17G30G 1G31G29G22G15G25G 3
  -6078127.503 4  -4736200.435 4  23397694.178    23397698.378    23397695.030
  -2556364.753 4  -1991958.142 4  24025055.814    24025059.840    24025056.373
  -8073501.747 5  -6291024.522 5  22565280.025    22565283.337    22565280.587
  18247234.140 4  14218628.480 4  24610505.696    24610508.842    24610505.143
  -1299479.831 4  -1012581.476 4  24824108.761    24824113.289    24824108.748
  -5233446.124 4  -4077998.775 4  24175634.461    24175638.438    24175635.537
  16878293.604 4  13151917.927 4  24427189.279    24427193.024    24427188.034
 -13489828.171 5 -10511530.918 5  22792735.726    22792739.295    22792736.451
  -4494062.929 4  -3501865.147 4  23961699.555    23961704.148    23961699.834
 -21099958.763 9 -16441519.960 9  20331187.861    20331190.408    20331187.808
 -15215098.740 5 -11855903.290 5  22202394.742    22202398.143    22202394.948
 00  1  1  0  0 30.0000000  0 11G21G23G17G30G 1G31G29G22G15G25G 3
  -6132427.986 4  -4778512.477 4  23387361.455    23387365.218    23387361.703
  -2586441.342 4  -2015394.448 4  24019332.425    24019336.115    24019332.872
  -7990741.587 5  -6226536.097 5  22581028.707    22581032.399    22581029.091
  18274808.415 4  14240114.880 4  24615752.673    24615756.082    24615752.310
  -1317133.094 4  -1026337.267 4  24820749.907    24820754.243    24820749.907
  -5259685.471 4  -4098444.960 4  24170641.011    24170645.465    24170641.809
  16938081.982 4  13198506.304 4  24438566.644    24438570.267    24438567.293
 -13548930.874 5 -10557584.968 5  22781488.710    22781492.153    22781489.545
  -4600217.585 4  -3584583.008 4  23941499.061    23941503.148    23941500.039
 -21083529.873 9 -16428718.216 9  20334314.137    20334316.731    20334314.115
 -15210926.086 5 -11852651.873 5  22203188.835    22203192.175    22203189.231
```

$\begin{bmatrix} x & y & z \end{bmatrix}^{\mathsf{T}}$ location of the satellite in space

$\begin{bmatrix} X & Y & Z \end{bmatrix}^{\mathsf{T}}$ location of the receiver in space

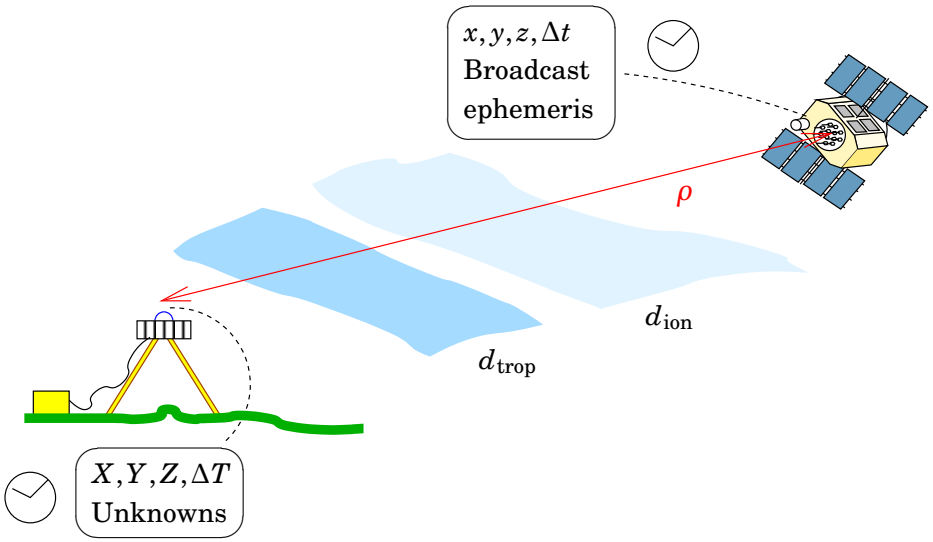


Figure 11.12. Pseudo-range observation.

c speed of light in vacuum

ΔT offset of receiver clock from GPS time (clock correction = $-\Delta T$)

Δt offset of satellite clock from GPS time

d_{ion} "ionospheric" (i.e., caused by the ionized atmosphere) propagation delay

d_{trop} "tropospheric" (i.e., caused by the neutral atmosphere) delay.

The clock offset of the satellite Δt is included in the *broadcast ephemeris* transmitted by the satellites. The receiver clock offset ΔT again remains unknown; it will have to be estimated as one unknown together with the co-ordinates. Thus we have *four* unknowns for every GPS receiver: three co-ordinates X, Y, Z and the clock offset ΔT — if we forget for a moment about the atmospheric unknowns d_{ion} and d_{trop}. For determining four unknowns, four pseudo-range observations are enough, i.e., observations to *four different satellites*. See figure 11.12.

11.6.2 The carrier wave's phase angle as observable

The "chip rate" of the C/A code, the number of bits transmitted per second, is 1.023 MHz, corresponding to a "wavelength" of 300 m, when the corresponding number for the P code, 10.23 MHz, means a wavelength of 30 m. If, instead of a modulation, we would use the *carrier wave* itself, the relevant wavelength would be 19 cm (L_1) or 24.4 cm (L_2), a distance that is more than an order of magnitude shorter. *Geodetic GPS positioning*, where dual-frequency receivers observe the phase of the carrier waves transmitted by the GPS satellites, is based on this. Electronic phase measurement is relatively easy and precise, but, as the market is

Table 11.4. Properties of carrier waves. "Factor" denotes the multiple of the base frequency (10.23 MHz).

Carrier wave	Frequency (MHz)	Wavelength (cm)	Factor
L_1	1575.42	19.0	154×
L_2	1227.60	24.4	120×

small and specialized, the prices of the devices are nevertheless as high as for other geodetic instruments: at least thousands of euros.

As is also the case with electronic range finders (distance measurement devices), phase measurement is always afflicted by an *ambiguity problem*. Of the measured phase, only the part in the interval $[0, 2\pi)$ has meaning. The corresponding pseudo-range between satellite and receiver is thus obtained only "modulo an integer number of wavelengths". If a certain pseudo-range P is compatible with a measurement done on wavelength λ, then compatible are also the pseudo-ranges $P+\lambda$, $P-\lambda$, $P+2\lambda$, $P-2\lambda$, ...

The *observation equation for carrier phase* is (as a phase angle, in radian units)

$$\overline{\phi} = 2\pi \left[\frac{\rho + c(\Delta t - \Delta T) + D_{\text{ion}} + D_{\text{trop}}}{\lambda} + N \right]$$

or (distance in metres[14])

$$\overline{P} \stackrel{\text{def}}{=} \lambda \frac{\overline{\phi}}{2\pi} = \rho + c(\Delta t - \Delta T) + D_{\text{ion}} + D_{\text{trop}} + \lambda N. \tag{11.2}$$

The observable is either the phase angle $\overline{\phi}$ or the equivalent pseudo-range $\overline{P} \stackrel{\text{def}}{=} \lambda \frac{\overline{\phi}}{2\pi}$, see figure 11.13. In this equation

D_{ion} the propagation delay of the carrier wave caused by the ionosphere (which is actually negative, $D_{\text{ion}} = -d_{\text{ion}}$)

D_{trop} the delay caused by the troposphere, $D_{\text{trop}} = d_{\text{trop}}$

λ the wavelength of the carrier wave according to the above table

N the integer unknown or *ambiguity*.

The variables $\rho \left(= \sqrt{(x-X)^2 + (y-Y)^2 + (z-Z)^2} \right)$, Δt and ΔT are the same as those for code pseudo-range, see equation 11.1. There are again four geodetic, real-valued unknowns: X, Y, Z and ΔT. But now there are also the integer values N, one for each observation, that also need to be determined in addition to the geodetic unknowns. This may seem impossible, but there are tricks for doing this, see section 12.3.

[14] Later on we'll leave the overbar off.

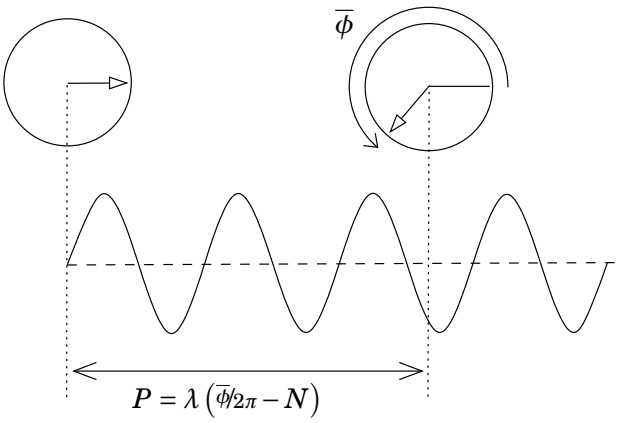


Figure 11.13. Measurement of the phase of the GPS signal's carrier wave. The measured phase angle $\overline{\phi} \in [0, 2\pi)$, the metric pseudo-range P, and the ambiguity N, in this example -2.

In equations 11.1 on page 270 and 11.2 on the previous page we have used *different symbols* for the ionospheric and troposheric delays of the carrier wave, than for the corresponding delays of code measurement, because *they are actually different*. One says that the ionosphere is *dispersive* for radio waves: the speed of propagation depends on the frequency, or equivalently, on the wavelength.

The travel speed of the code modulations is the *group velocity*, which is always less than the speed of light in vacuum. The travel speed of the

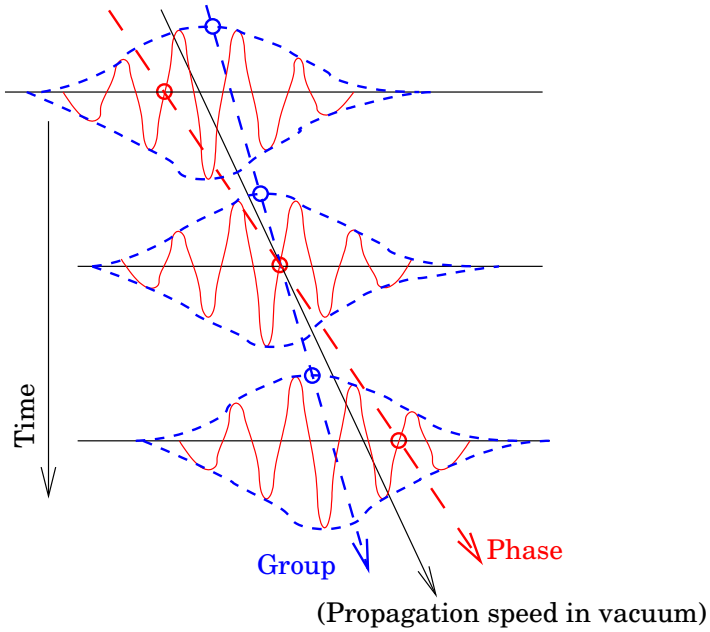


Figure 11.14. The propagation of a "wave packet" in a dispersive medium, phase and group velocity. The carrier wave travels with phase velocity, the modulations — also the pseudo-random codes of the GPS signal — travel at group velocity.

carrier phase is the *phase velocity*. In a dispersive medium these two speeds differ from each other.

11.6.3 The effects of ionosphere and troposphere

The ionosphere is a *dispersive* medium to radio waves: different frequencies propagate at different velocities. This is caused by the large number of free electrons. As a consequence of dispersion, the phase and group velocities of propagation are different. For phase propagation the index of refraction is

$$n_{\mathrm{p}} = \sqrt{1 - \frac{f_{\mathrm{p}}^2}{f^2}} \approx 1 + \frac{c_2}{f^2} + \frac{c_4}{f^4} + \frac{c_6}{f^6} + \ldots \tag{11.3}$$

In this equation the constants c_i, like also the plasma frequency f_{p}, do not depend on the frequency of the radio waves; they depend on the total electron content TEC, n_{e}, of the ionosphere. A good approximate formula, which explains 99.9% of the whole ionospheric propagation effect, is (Seeber, 1989):

$$n_{\mathrm{p}} = 1 - \frac{C}{f^2},\ C = 40.3 n_{\mathrm{e}}\,\mathrm{m^3/s^2},$$

in which the electron density n_{e} is expressed in electrons per m^3. Typical numbers are $10^2 - 10^6\,\mathrm{m}^{-3}$. The electron density varies between day and night — greater in the daytime —, with the season — greater in summer —, with solar activity, and of course with the latitude and height of the location.

The group propagation velocity, more precisely the group index of refraction, is obtained as the derivative with respect to frequency[15] of $n_{\mathrm{p}} f$:

$$n_{\mathrm{g}} = \frac{d\left(n_{\mathrm{p}} f\right)}{df} = 1 + \frac{C}{f^2},\ C \text{ the same.}$$

Because the index of refraction of the ionosphere is dependent upon the frequency f, it is possible to *eliminate* the effect of the ionosphere by combining measurements made on two different frequencies. This is the fundamental reason why the GPS system uses two different frequencies L_1 and L_2.

If we form the linear combination of code observations

$$p_3 \overset{\text{def}}{=} \frac{f_1^2 p_1 - f_2^2 p_2}{f_1^2 - f_2^2}$$

[15] when the propagation velocity is $c = c_0/n$, in which c_0 is the speed of light in vacuum, it follows that the phase propagation velocity is greater than the speed of light. The carrier phase cannot however carry information, so the directionality of time according to thermodynamics is preserved... if it were possible to move information faster than light, it would, according to special relativity, also be possible to carry information back in time!

and correspondingly

$$n_{g;3} \stackrel{\text{def}}{=} \frac{f_1^2 n_{g;1} - f_2^2 n_{g;2}}{f_1^2 - f_2^2},$$

we obtain

$$n_{g;3} = \frac{f_1^2\left(1+\frac{C}{f_1^2}\right) - f_2^2\left(1+\frac{C}{f_2^2}\right)}{f_1^2 - f_2^2} = \frac{f_1^2 - f_2^2}{f_1^2 - f_2^2} = 1,$$

from which it is seen that the effect of the ionosphere has vanished from it[16].

The *troposphere* — more precisely, the neutral atmosphere, which includes also the stratosphere, and neutral fractions of higher layers — on the other hand is not dispersive to radio waves. Its specialty is, however, a strong dependence of the index of refraction on *water vapour content*. The formula, the same formula 6.5 that also applies for electronic distance measurement, is (Rüeger, 1990, 2002):

$$N_{\mathrm{M}} = 10^6 \cdot (n_{\mathrm{M}} - 1) = \frac{77.624\,\mathrm{K/hPa}}{T}(p-e) + \frac{64.70\,\mathrm{K/hPa}}{T}\left(1+\frac{5748\,\mathrm{K}}{T}\right)e \tag{11.4}$$

in which n_{M} is the refractive index for microwaves. Here, the unit of p and e is hPa (hectopascal) or millibar. T is absolute temperature in kelvin. Note that the coefficient of the partial pressure of water vapour in this formula is at ambient temperature as much as 18 times the coefficient for air pressure p!

Because both the ionosphere and the troposphere affect the propagation of GPS radio waves, one may use GPS measurements for both ionospheric and tropospheric research. Meteorologists and climate researchers are very much interested in this, see section 17.3.

11.7 GPS measurement geometry

Let the position vector of satellite S in a geocentric frame be $\mathbf{r}^S$, and the position vector of observation station A in the same frame be $\mathbf{R}_A$, and let the distance between the two be ρ_A^S. See figure 11.15.

Then the following vector equation applies:

$$\mathbf{r}^S = \mathbf{R}_A + \mathbf{e}_A^S \rho_A^S,$$

in which $\mathbf{e}_A^S$ is the *direction* (unit vector) to satellite S seen from observation station A. The problem of GPS positioning is to compute $\mathbf{R}_A$ when ρ_A^S is given to sufficiently many satellites S.

[16] The result and its proof are the same if we take, instead of the group index of refraction, the phase one, n_{p}.

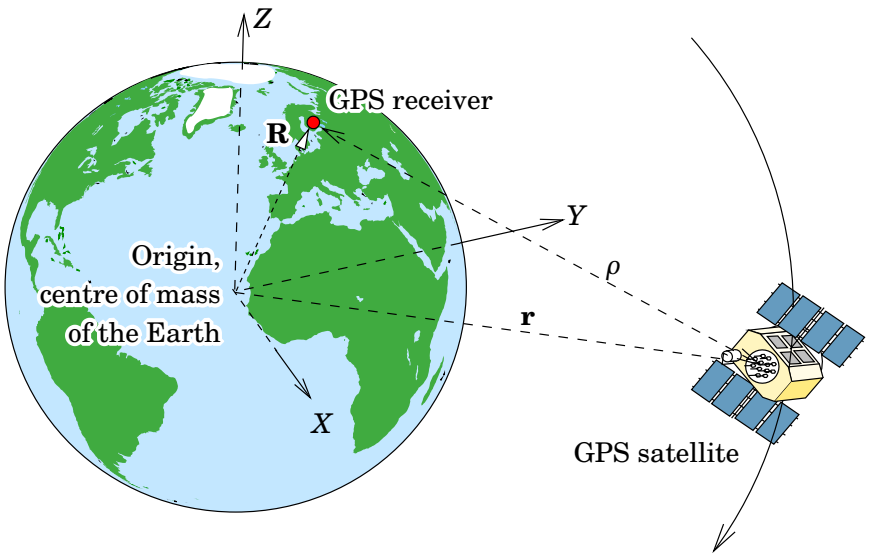


Figure 11.15. The geometry of GPS positioning.

The Pythagoras theorem gives

$$\rho_A^S = \left\| \mathbf{r}^S - \mathbf{R}_A \right\| = \sqrt{\left(x^S - X_A\right)^2 + \left(y^S - Y_A\right)^2 + \left(z^S - Z_A\right)^2}, \qquad (11.5)$$

in which

$$\mathbf{r}^S = x^S\mathbf{i} + y^S\mathbf{j} + z^S\mathbf{k},$$
$$\mathbf{R}_A = X_A\mathbf{i} + Y_A\mathbf{j} + Z_A\mathbf{k}$$

are the geocentric position vectors of satellite S and observation station A. The vectors $\mathbf{i}$,$\mathbf{j}$ and $\mathbf{k}$ are unit vectors aligned with the co-ordinate axes, i.e., $\{\mathbf{i},\mathbf{j},\mathbf{k}\}$ form an *orthonormal basis* in Euclidean space.

Often one writes the vectors as column vectors of their components, the co-ordinates of location:

$$\overline{\mathbf{r}}^S \overset{\text{def}}{=} \begin{bmatrix} x^S \\ y^S \\ z^S \end{bmatrix} \text{ and } \overline{\mathbf{R}}_A \overset{\text{def}}{=} \begin{bmatrix} X_A \\ Y_A \\ Z_A \end{bmatrix}.$$

From the observations, however, no true distances ρ are obtained, but *pseudo*-ranges p — equation 11.1 — or P — equation 11.2.

Also the atmospheric propagation delays d_{ion}, d_{trop}, D_{ion}, D_{trop} must in some way be taken into account. The alternatives are

- elimination from the observation equations — as described in subsection 11.6.3 for the ionospheric effects d_{ion}, D_{ion}
- computation using a good, externally provided atmospheric model
- modelling of the atmospheric effect using unknown parameters, to be estimated from the same observation equations — as we will describe for the tropospheric effects d_{trop},D_{trop} in subsection 17.3.1.

11.8 Measurement geometry and sensitivity of observations

Above was shown that the observables of GPS are *pseudo-ranges*, the observation equation of which looks like this:

$$p = \rho + c(\Delta t - \Delta T) + d_{\text{ion}} + d_{\text{trop}}.$$

Let us leave out the effect of the atmosphere, and assume also that the satellite's orbit — and thus, the momentaneous position vector of the satellite in space as computed from the orbit and clock time — and the satellite's clock offset Δt are *known*, i.e., already taken into account:

$$p = \rho - c\Delta T,$$

in which ρ is the geometric distance between satellite and receiver, and ΔT is the clock offset of the receiver.

Write equation 11.5 on the preceding page more simply:

$$\rho_A^S = \rho = \sqrt{(x-X)^2 + (y-Y)^2 + (z-Z)^2},$$

in which $\begin{bmatrix} x & y & z \end{bmatrix}^{\mathsf{T}}$ is the known position vector of the satellite in space, computed from the orbit information, i.e., the ephemeris. $\begin{bmatrix} X & Y & Z \end{bmatrix}^{\mathsf{T}}$ is the receiver position vector. Now

$$p = \sqrt{(x-X)^2 + (y-Y)^2 + (z-Z)^2} - c\Delta T,$$

in which there are *four unknowns*, X, Y, Z and ΔT.

For solving four unknowns, it is sufficient to have observations to four satellites. If the number of useable satellites is larger, we have redundancy and an *adjustment problem*.

Question: How do small "perturbances" in the location of the receiver impact a certain measurement p?

Answer: Look at the *place* of the satellite in the sky. See figure 11.16. Let the satellite's direction vector as seen from the observation site be $\mathbf{e}$. This is a unit vector: its length is

$$\|\mathbf{e}\| = \sqrt{e_1^2 + e_2^2 + e_3^2} = 1.$$

If the satellite's place in the sky is azimuth α, elevation η, then

$$\mathbf{e} \overset{\text{def}}{=} e_1 \cdot \mathbf{N} + e_2 \cdot \mathbf{E} + e_3 \cdot \mathbf{U}$$

and the column vector of components is

$$\bar{\mathbf{e}} \overset{\text{def}}{=} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} = \begin{bmatrix} \cos\alpha\cos\eta \\ \sin\alpha\cos\eta \\ \sin\eta \end{bmatrix} = \begin{bmatrix} \frac{x-X}{\rho} \\ \frac{y-Y}{\rho} \\ \frac{z-Z}{\rho} \end{bmatrix}.$$

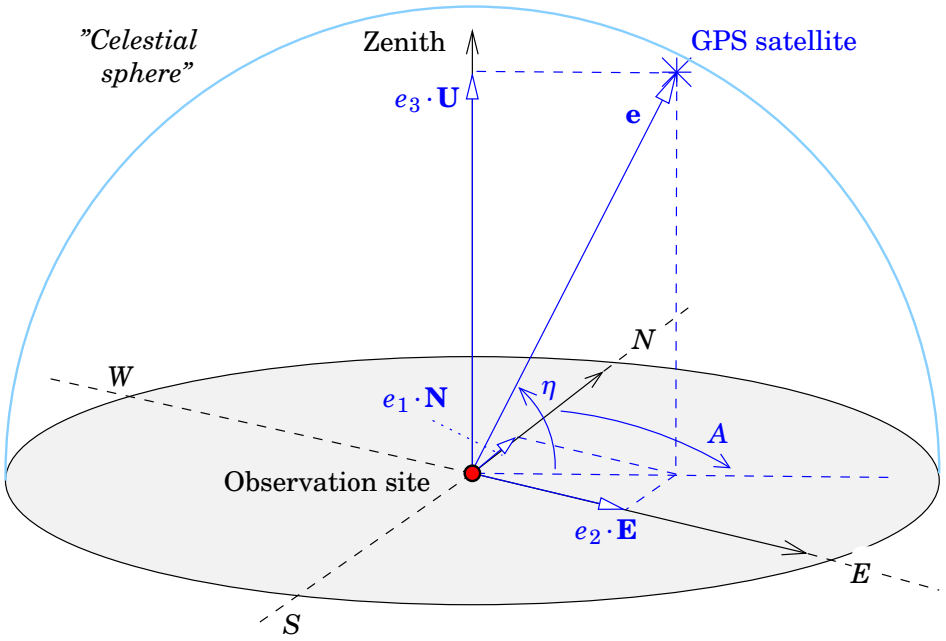


Figure 11.16. The geometry between a GPS satellite and an observation site. $\{\mathbf{N}, \mathbf{E}, \mathbf{U}\}$ is the local-horizon orthonormal basis ("North, East, Up").

□

Here, $\{\mathbf{N}, \mathbf{E}, \mathbf{U}\}$ is the basis of orthonormal unit vectors in the local-horizon system ("North, East, Up").

We carry out a *sensitivity analysis*. In what way do small co-ordinate shifts ΔX, ΔY or ΔZ in the location of the observation site affect the observable p?

- If a lot, then the observation p will help in determining the unknown in question, i.e., the co-ordinate.
- If not at all, then the unknown in question cannot be determined using observation p.
- The greater the *sensitivity* of the observations, the better will be the precision of the solution of the unknowns.

Intuitively: the observation p is most affected, in a ratio of one on one, by a shift in observation site location along the satellite's direction vector $\mathbf{e}$. Shifts in observation site location perpendicular to this direction vector have no effect at all.

Equation:

$$\text{"Effect"} = \langle \Delta\mathbf{R} \cdot \mathbf{e} \rangle = \Delta X e_1 + \Delta Y e_2 + \Delta Z e_3.$$

This intuitive result may also be derived more formally by *linearization*. See tableau 11.5 on the next page and section 13.6.

Tableau 11.5. A more exact derivation of the influence formula by means of linearization.

Choose for the observation station $\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$ an approximate location $\begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \end{bmatrix}$ and an approximate clock offset $\Delta T = 0$. Then we can construct an approximate pseudo-range observation

$$p_0 = \sqrt{(x - X_0)^2 + (y - Y_0)^2 + (z - Z_0)^2}.$$

Carry out a Taylor series expansion in the neighbourhood of the approximate location, around the value p_0. The first, linear terms yield

$$p \approx p_0 + \frac{\partial p}{\partial X}(X - X_0) + \frac{\partial p}{\partial Y}(Y - Y_0) + \frac{\partial p}{\partial Z}(Z - Z_0) - c\Delta T \Longrightarrow$$
$$\Longrightarrow \Delta p = p - p_0 \approx \frac{\partial p}{\partial X}\Delta X + \frac{\partial p}{\partial Y}\Delta Y + \frac{\partial p}{\partial Z}\Delta Z - c\Delta T. \tag{11.6}$$

Here, the coefficients, *partial derivatives*, are obtained as follows:

$$\frac{\partial p}{\partial X} = -\frac{x - X}{\rho}, \quad \frac{\partial p}{\partial Y} = -\frac{y - Y}{\rho}, \quad \frac{\partial p}{\partial Z} = -\frac{z - Z}{\rho}. \tag{11.7}$$

The values for the coefficients are evaluated at the approximate location $\begin{bmatrix} X_0 & Y_0 & Z_0 \end{bmatrix}^T$ instead of at the true but unknown location. This suffices, because in formula 11.6 the values $\Delta p, \Delta X, \Delta Y, \Delta Z$ are *small differences* between the true (p, X, Y, Z) and approximate values (p_0, X_0, Y_0, Z_0).
Note that the partial derivatives 11.7 are *direction cosines* that together describe the direction to the satellite in the sky as seen from the observation station, projected onto the co-ordinate axes X, Y and Z.

11.8.1 DOP quantities and observation equations

The measure of the quality of the measurement geometry of the GPS satellites in the local sky is DOP, Dilution of Precision. Using the above geometric sensitivity analysis, we may calculate various variants of DOP. A *larger* DOP number means a *poorer* measurement geometry!

DOP characterizes the quality of the satellite *geometry*, i.e., how much worse or better one may get the co-ordinates resolved from "standard quality" measurements, due to the poorer or better geometry of the satellites on offer in the sky. This is valuable information in the planning phase of measurement work.

One rule of thumb is, that the measurement geometry is acceptable if GDOP $< 7 \ldots 10$, depending on the intended use.

If we now describe the "variation" of the co-ordinates of the observation site by a (small) correction quantity $\Delta\overline{\mathbf{R}} = \begin{bmatrix} \Delta X & \Delta Y & \Delta Z \end{bmatrix}^T$, and the "variation" of the receiver clock by an (also small) correction quantity

Table 11.6. The different variants of the DOP quantity.

Acronym	Name	Quantity characterized
GDOP	Geometric DOP	Place and time
PDOP	Position DOP	Place
HDOP	Horizontal DOP	Horizontal location
VDOP	Vertical DOP	Height
TDOP	Time DOP	Time

ΔT, we may express the dependence of the observable $p^{(i)}$ on these altogether four unknowns like this:

$$\Delta p^{(i)} = \begin{bmatrix} e_1^{(i)} & e_2^{(i)} & e_3^{(i)} & -c \end{bmatrix} \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \\ \Delta T \end{bmatrix} =$$

$$= \begin{bmatrix} \cos\alpha^i \cos\eta^i & \sin\alpha^i \cos\eta^i & \sin\eta^i & -c \end{bmatrix} \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \\ \Delta T \end{bmatrix},$$

in which α^i and η^i are the azimuth and elevation of satellite i in the local sky. This way of writing is called *linearization*.

This equation may be understood as an *observation equation*. If the equation is written symbolically, as is the practice in geodesy, in the form[17]

$$\underline{\ell} + \underline{\mathrm{v}} = A\widehat{\mathrm{x}}, \tag{11.8}$$

then the elements of the vector of unknowns $\widehat{\mathrm{x}}$ — the estimators — are $\Delta\widehat{X}, \Delta\widehat{Y}, \Delta\widehat{Z}$, and $\widehat{\Delta T}$, the vector of observations $\underline{\ell}$ is made up of values $\Delta\underline{p}^{(i)}$, and the *design matrix* is

$$A = \begin{bmatrix} \cos\alpha^1 \cos\eta^1 & \sin\alpha^1 \cos\eta^1 & \sin\eta^1 & -c \\ \cos\alpha^2 \cos\eta^2 & \sin\alpha^2 \cos\eta^2 & \sin\eta^2 & -c \\ \vdots & \vdots & \vdots & \vdots \\ \cos\alpha^i \cos\eta^i & \sin\alpha^i \cos\eta^i & \sin\eta^i & -c \\ \vdots & \vdots & \vdots & \vdots \\ \cos\alpha^n \cos\eta^n & \sin\alpha^n \cos\eta^n & \sin\eta^n & -c \end{bmatrix}. \tag{11.9}$$

This is in fact the linearized version of the original observation equation 11.1.

[17] The residuals $\underline{\mathrm{v}}$ are needed to reconcile the observations $\underline{\ell}$, which contain measurement uncertainty, with each other when there are more observations than unknowns. See section 13.4.

This design matrix *contains everything we know about the GPS measurement geometry*[18]. From this, all DOP quantities may be calculated, without using a single real observation — it suffices that the places of the satellites in the sky can be computed. The size of the matrix is $n \times 4$: n rows and four columns, where n is the number of satellites available for use.

The situation is the same as in the case of reconnaissance of a terrestrial geodetic network: the quality of the network can already be judged based on point locations and planned measurement geometry, before even a single measurement has been carried out. This is a great tool for planning.

We will discuss more about observation equations and least-squares adjustment in section 13.4. Here we do not even try to compute a least-squares solution. We only look into the *precision* of the unknowns $\begin{bmatrix} \Delta X & \Delta Y & \Delta Z & \Delta T \end{bmatrix}^\mathsf{T}$ to be computed!

We assume, for this computation, that *all observations are equally precise* — their precision may be assumed 1 — and that *they are statistically independent of each other*. Then, the following simple calculation is valid. It gives a picture of the *role of the GPS measurement geometry* in the final precision of the measurement results. Other factors, like the technical capability of the receiver and antenna used, duration of measurement, the atmosphere, etc., can be looked at separately.

11.8.2 Error ellipsoids for presentation of measurement precision

From the design matrix A we may construct the *normal matrix*, or *weight matrix of the unknowns* as follows:

$$P_{xx} \stackrel{\text{def}}{=} A^\mathsf{T} A. \tag{11.10}$$

Expressed in the satellite positions in the sky, i.e., azimuths a^i and elevation angles η^i, the result is according to equation 11.9 given in table 11.7.

The inverse of the weight matrix P_{xx}, $Q_{xx} \stackrel{\text{def}}{=} P_{xx}^{-1}$, is the *weight-coefficient matrix*:

$$Q_{xx} = \begin{bmatrix} q_{xx} & q_{xy} & q_{xz} & q_{xt} \\ q_{yx} & q_{yy} & q_{yz} & q_{yt} \\ q_{zx} & q_{zy} & q_{zz} & q_{zt} \\ q_{tx} & q_{ty} & q_{tz} & q_{tt} \end{bmatrix}.$$

This matrix, like the weight matrix P_{xx} or the design matrix A, still describes exclusively the *geometry of measurement site and satellites*, and nothing else.

[18] As does also a so-called *sky plot*, e.g., http://www.insidegnss.com/auto/marapr09-borre.pdf.

Table 11.7. The weight or normal matrix of the GPS unknowns. The summation $\sum$ is understood to be over all satellites, $i = 1, \ldots, n$.

$$P_{\mathrm{xx}} = \begin{bmatrix} \sum_i \cos^2\alpha^i \cos^2\eta^i & \sum_i \sin\alpha^i \cos\alpha^i \cos^2\eta^i & \sum \cos\alpha^i \sin\eta^i \cos\eta^i & -c\sum_i \cos\alpha^i \cos\eta^i \\ \sum_i \sin\alpha^i \cos\alpha^i \cos^2\eta^i & \sum_i \sin^2\alpha^i \cos^2\eta^i & \sum_i \sin\alpha^i \sin\eta^i \cos\eta^i & -c\sum_i \sin\alpha^i \cos\eta^i \\ \sum_i \cos\alpha^i \sin\eta^i \cos\eta^i & \sum_i \sin\alpha^i \sin\eta^i \cos\eta^i & \sum_i \sin^2\eta^i & -c\sum_i \sin\eta^i \\ -c\sum_i \cos\alpha^i \cos\eta^i & -c\sum_i \sin\alpha^i \cos\eta^i & -c\sum_i \sin\eta^i & nc^2 \end{bmatrix}. \tag{11.11}$$

Now, the *variance matrix* of the vector of unknowns, or *solution*, $\hat{\mathrm{x}} = \begin{bmatrix} \Delta\widehat{X} & \Delta\widehat{Y} & \Delta\widehat{Z} & \widehat{\Delta T} \end{bmatrix}^{\mathsf{T}}$, is

$$\Sigma_{\mathrm{xx}} = \sigma^2 Q_{\mathrm{xx}} = \sigma^2 \left[\begin{array}{ccc|c} q_{xx} & q_{xy} & q_{xz} & q_{xt} \\ q_{yx} & q_{yy} & q_{yz} & q_{yt} \\ q_{zx} & q_{zy} & q_{zz} & q_{zt} \\ \hline q_{tx} & q_{ty} & q_{tz} & q_{tt} \end{array}\right].$$

The constant σ^2 is called the *variance of unit weight*. Its square root, the *mean error of unit weight* σ, is the mean error, assumed constant, of one observable, i.e., one pseudo-range.

The variance matrix of the co-ordinate solution is a 3×3 element sized *submatrix* Σ_{rr} of the variance matrix of the unknowns. Its elements are

$$\Sigma_{\mathrm{rr}} = \sigma^2 \begin{bmatrix} q_{xx} & q_{xy} & q_{xz} \\ q_{yx} & q_{yy} & q_{yz} \\ q_{zx} & q_{zy} & q_{zz} \end{bmatrix}. \tag{11.12}$$

The DOP quantities are calculated directly from the weight-coefficient matrix Q:

$$\begin{aligned} \mathrm{PDOP} &= \sqrt{q_{xx} + q_{yy} + q_{zz}}, \\ \mathrm{HDOP} &= \sqrt{q_{xx} + q_{yy}}, \\ \mathrm{VDOP} &= \sqrt{q_{zz}}, \\ \mathrm{TDOP} &= \sqrt{q_{tt}}, \\ \mathrm{GDOP} &= \sqrt{q_{xx} + q_{yy} + q_{zz} + q_{tt}} = \sqrt{\mathrm{PDOP}^2 + \mathrm{TDOP}^2}. \end{aligned}$$

The mean errors of the co-ordinates are obtained as the square roots of the diagonal elements of the variance matrix 11.12:

$$\begin{aligned} \sigma_X &= \sigma\sqrt{q_{xx}}, \\ \sigma_Y &= \sigma\sqrt{q_{yy}}, \\ \sigma_Z &= \sigma\sqrt{q_{zz}}. \end{aligned}$$

The familiar point mean error in the plane is now related directly to the HDOP quantity:

$$\underbrace{\sigma_P}_{\text{precision}} \overset{\text{def}}{=} \sqrt{\sigma_X^2 + \sigma_Y^2} = \underbrace{\sigma}_{\text{instrumental technology etc.}} \cdot \underbrace{\mathrm{HDOP}}_{\text{geometry}}.$$

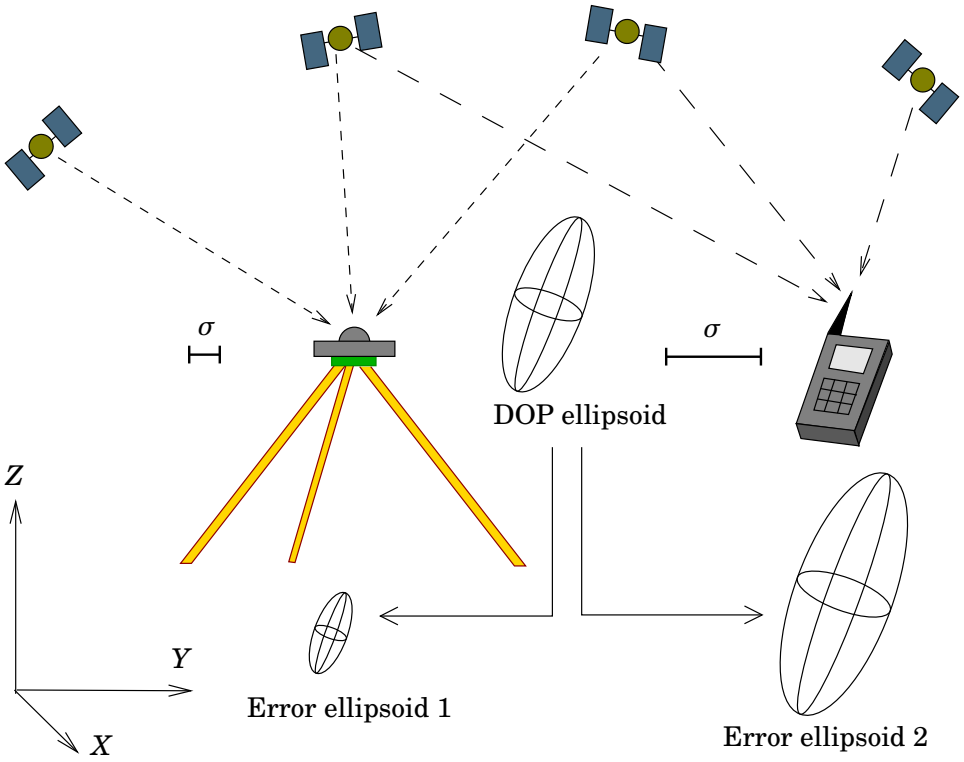


Figure 11.17. The connection between DOP ellipsoid and error ellipsoid, and the mean error of unit weight σ. The DOP ellipsoid characterizes only the effect of the geometry, whereas the error ellipsoid depends also on the precision of measurement, i.e., the device type.

The co-ordinate variance matrix can be graphically represented by a three-dimensional *error ellipsoid*. The error ellipsoid around a measurement point visualizes the uncertainty of the location based on the above definitions. A similar DOP ellipsoid is obtained by leaving off the constant σ: it has the same shape as the error ellipsoid, but it has no metric size — note that the elements of matrix Q, like those of the design matrix A, are dimensionless.

In the general case it is not so simple to calculate the parameters of the ellipsoid from the matrix elements. Let's look at a simpler special case, which nevertheless is practically relevant. If the measurement geometry is symmetric — i.e., the satellites and their elevation angles are evenly distributed by azimuth, around the horizon —, it follows that the DOP ellipsoid will be oriented along the co-ordinate axes: $q_{xy} = q_{xy} = q_{yz} = 0$ (and $q_{xx} = q_{yy}$!), and the matrix is

$$\Sigma_{\mathrm{rr}} = \sigma^2 \begin{bmatrix} q_{xx} & 0 & 0 \\ 0 & q_{yy} & 0 \\ 0 & 0 & q_{zz} \end{bmatrix}.$$

In this special case the axes of the DOP ellipsoid are pointing along the

local co-ordinate axes, and the longest axis points in the vertical direction. In a practical measurement situation, the variance matrix of the co-ordinates is often close to this. The longest axis of the error ellipsoid is almost always close to the vertical, which tells us that the height is more weakly determined than the horizontal location[19]. In this case the co-ordinate mean errors are

$$\sigma_X = \sigma_Y = \sigma \cdot \frac{1}{2}\sqrt{2} \cdot \mathrm{HDOP},$$
$$\sigma_Z = \sigma \cdot \mathrm{VDOP},$$

based on the above definitions. The lengths of the semi-axes of the DOP ellipsoid are $\sqrt{q_{xx}}$, $\sqrt{q_{yy}}$ and $\sqrt{q_{zz}}$. In the symmetric case, we thus have

$$q_{xx} = q_{yy} = \frac{1}{2}\mathrm{HDOP}^2$$

in the horizontal plane, and always

$$q_{zz} = \mathrm{VDOP}^2$$

in the vertical direction.

11.8.3 DOP and measurement planning

In general it can be said, that the *measurement geometry* is better if HDOP and VDOP (and thus also PDOP) are smaller. This again requires that

- a sufficient number of satellites is above the horizon and observable from the measurement site, and
- they are positioned evenly around the sky.

This of course works only, if one has a sufficiently free view from the measurement site up to the sky. In practice, there will always be obstacles, which should be mapped on a *horizon plot* for planning the measurement. Many planning programmes let the user draw a horizon plot and take it into account when calculating DOP.

---

[19] Reasons for this are:

- Only satellites above the horizon contribute to the determination of height, i.e., this is *extrapolation*. In the horizontal plane again, there are satellites both in the East and in the West, both in the North and in the South which contribute to the positioning: *interpolation*.

  Also because of the same non-symmetry, the vertical location unknown Z and the clock unknown ΔT "compete" for the same information when they are estimated together from the same observation data.
- Also the uncertainty in the signal delays caused by the atmosphere affects mostly in the vertical direction, whereas — also due to the horizontal stratification of the atmosphere — the situation in the horizontal plane is more symmetric.

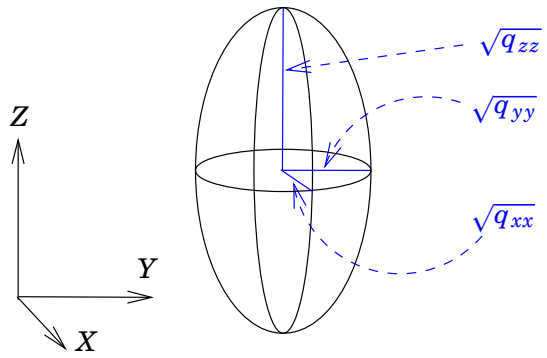


Figure 11.18. The DOP ellipsoid of GPS positioning, assuming its principal axes are in the same directions as the co-ordinate axes.

The situation improves if the receivers to be used can make use of both the GPS and the GLONASS system: then, more satellites are seen, and a good positioning geometry is more easily achieved.

Of course, the GPS positioning geometry is only one factor among others. Other factors in achievable measurement precision are, e.g.:

- the capabilities of receiver and antenna
- local disturbances, like *multipath*, reflections off the ground and objects nearby
- the activity of the Sun, ionospheric conditions
- the measurement mode: static or kinematic, absolute (e.g., "Precise Point Positioning", PPP) or relative, etc.
- in relative GNSS measurement, the distance between measurement points or from the base station. Geodetic measurements are (almost) always relative and are carried out as network measurements.
- in static GNSS measurement, the duration of measurement, the number of measurement epochs. Geodetic base network measurements are almost always static, despite the method being time consuming, because of its robustness. Only in local measurements, like detail surveys, the faster kinematic technique (RTK, Real-Time Kinematic) is used.
- etcetera.

11.8.4 Example 1: an azimuthally symmetric geometry

Assume that *the satellites are evenly distributed around the sky*, i.e., according to azimuth α, for every elevation angle η. Then, in the matrix 11.11 above

1. almost all non-diagonal elements vanish, because they contain either $\sin\alpha$ or $\cos\alpha$, or even $\sin\alpha\cos\alpha$. Only $\sum_i \sin\eta^i$ does not vanish.

2.

$$\sum_{i=1}^{n} \cos^2\alpha^i \cos^2\eta^i = \frac{1}{2}\sum_{i=1}^{n} \cos^2\eta^i,$$
$$\sum_{i=1}^{n} \sin^2\alpha^i \cos^2\eta^i = \frac{1}{2}\sum_{i=1}^{n} \cos^2\eta^i.$$

Therefore, the normal matrix $P = A^\mathsf{T}A$ becomes an almost diagonal matrix that would be relatively easy to invert — although we will not even try:

$$P_{xx} = A^\mathsf{T}A = \begin{bmatrix} \frac{1}{2}\sum_{i=1}^{n}\cos^2\eta^i & 0 & 0 & 0 \\ 0 & \frac{1}{2}\sum_{i=1}^{n}\cos^2\eta^i & 0 & 0 \\ 0 & 0 & \sum_{i=1}^{n}\sin^2\eta^i & -c\sum_{i=1}^{n}\sin\eta^i \\ 0 & 0 & -c\sum_{i=1}^{n}\sin\eta^i & nc^2 \end{bmatrix}.$$

Transform now the observation equation 11.8 and the design matrix 11.9 in the following way:

$$\underline{\ell} = A\Lambda\Lambda^{-1}\widehat{\mathrm{x}} + \underline{\mathrm{v}} = \widetilde{A}\widetilde{\mathrm{x}} + \underline{\mathrm{v}},$$

in which

$$\widetilde{\mathrm{x}} = \begin{bmatrix} \Delta\widehat{X} \\ \Delta\widehat{Y} \\ \Delta\widehat{Z} \\ \widetilde{\Delta T} \end{bmatrix} = \Lambda^{-1}\widehat{\mathrm{x}} = \begin{bmatrix} \Delta\widehat{X} \\ \Delta\widehat{Y} \\ \Delta\widehat{Z} \\ \widehat{\Delta T} - \frac{\sum\sin\eta}{nc}\Delta\widehat{Z} \end{bmatrix} \tag{11.13}$$

and

$$\widetilde{A} = A\Lambda = \begin{bmatrix} \cos\alpha^1\cos\eta^1 & \sin\alpha^1\cos\eta^1 & \sin\eta^1 - \frac{1}{n}\sum\sin\eta & -c \\ \cos\alpha^2\cos\eta^2 & \sin\alpha^2\cos\eta^2 & \sin\eta^2 - \frac{1}{n}\sum\sin\eta & -c \\ \vdots & \vdots & \vdots & \vdots \\ \cos\alpha^i\cos\eta^i & \sin\alpha^i\cos\eta^i & \sin\eta^i - \frac{1}{n}\sum\sin\eta & -c \\ \vdots & \vdots & \vdots & \vdots \\ \cos\alpha^n\cos\eta^n & \sin\alpha^n\cos\eta^n & \sin\eta^n - \frac{1}{n}\sum\sin\eta & -c \end{bmatrix},$$

because[20]

$$\Lambda = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & +\frac{\sum\sin\eta}{nc} & 1 \end{bmatrix}, \Lambda^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{\sum\sin\eta}{nc} & 1 \end{bmatrix}$$

Now we get a clean diagonal matrix:

$$\widetilde{P_{xx}} = \widetilde{A}^\mathsf{T}\widetilde{A} = \begin{bmatrix} \frac{1}{2}\sum_{i=1}^{n}\cos^2\eta^i & 0 & 0 & 0 \\ 0 & \frac{1}{2}\sum_{i=1}^{n}\cos^2\eta^i & 0 & 0 \\ 0 & 0 & \sum_{i=1}^{n}\left(\sin\eta^i - \frac{1}{n}\sum_{j=1}^{n}\sin\eta^j\right)^2 & 0 \\ 0 & 0 & 0 & nc^2 \end{bmatrix}.$$

[20] Verify that $\Lambda\Lambda^{-1} = \Lambda^{-1}\Lambda = I$!

Let us write the equation for the error or visual ellipsoid of $\widetilde{P}$:

$$\widetilde{\mathrm{x}}^{\mathsf{T}}\widetilde{P_{\mathrm{xx}}}\widetilde{\mathrm{x}} = 1,$$

in which $\widetilde{\mathrm{x}}$ as in equation 11.13. The result is

$$p_{11}\Delta\widehat{X}^2 + p_{22}\Delta\widehat{Y}^2 + \widetilde{p}_{33}\Delta\widehat{Z}^2 + p_{44}\widetilde{\Delta T}^2 = \frac{\Delta\widehat{X}^2}{q_{xx}} + \frac{\Delta\widehat{Y}^2}{q_{yy}} + \frac{\Delta\widehat{Z}^2}{\widetilde{q}_{zz}} + \frac{\widetilde{\Delta T}^2}{\widetilde{q}_{tt}} = 1,$$

in which are immediately seen the elements of the variance matrix of the unknowns

$$Q_{\mathrm{xx}} = P_{\mathrm{xx}}^{-1} = \left[A^{\mathsf{T}}A\right]^{-1}$$

according to the definition:

$$q_{xx} = q_{yy} = \frac{2}{\sum_i \cos^2\eta^i},$$

from which

$$\mathrm{HDOP} = \sqrt{q_{xx} + q_{yy}} = \frac{2}{\sqrt{\sum_i \cos^2\eta^i}}.$$

Similarly, after a bit of reorganizing,

$$\widetilde{q}_{zz} = \frac{1}{\widetilde{p_{33}}} = \frac{1}{\sum_{i=1}^{n}\left(\sin\eta^i - \frac{1}{n}\sum_{j=1}^{n}\sin\eta^j\right)^2} = \frac{n}{n\sum_i \sin^2\eta^i - \left(\sum_i \sin\eta^i\right)^2},$$

and VDOP is its quare root[21].

11.8.5 Example 2: a singular case

Look again at the design matrix A, equation 11.9, and write it into the form

$$A = \begin{bmatrix} \left[\mathbf{e}^{(1)}\right]^{\mathsf{T}} & -c \\ \left[\mathbf{e}^{(2)}\right]^{\mathsf{T}} & -c \\ \vdots & \vdots \\ \left[\mathbf{e}^{(n)}\right]^{\mathsf{T}} & -c \end{bmatrix},$$

assuming there are n satellites. If the satellites are all on the same circle, then the unit direction vector in the direction of satellite i is

$$\mathbf{e}^{(i)} = a\mathbf{e}_0 + b_i\mathbf{e}_1 + c_i\mathbf{e}_2$$

in which the values b_i and c_i satify the condition $b_i^2 + c_i^2 = 1 - a^2$, for all satellites $i = 1,\ldots,n$. Here, $a\mathbf{e}_0$ is the vector from the observer to the centre of the circle. Thus there are only three independent vectors $\mathbf{e}^{(i)}$ when four are needed.

[21]Note that if η is constant — i.e., all $\eta^i, i = 1\ldots n$, are the same — then the denominator vanishes! So, the determination of height by GPS requires that there are satellites on different elevations in the sky.

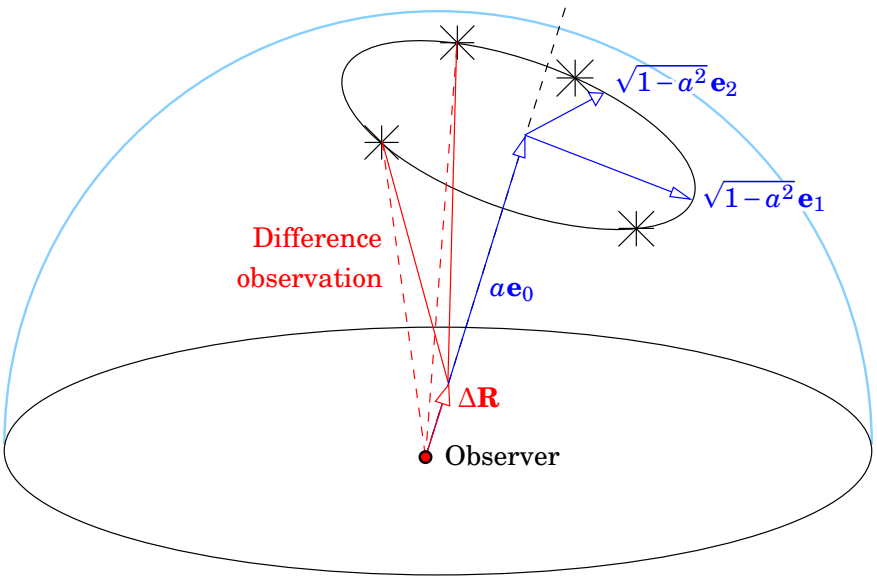


Figure 11.19. The circle singularity or "dangerous circle" for GPS.

See figure 11.19. The situation is also geometrically clear: if the observation site is shifted along the direction of the vector $\mathbf{e}_0$, any difference between pseudo-ranges from two different satellites will remain unchanged. What is going on here is, that the receiver's clock unknown ΔT and the component of the observation site's location in the $\mathbf{e}_0$ direction (i.e., the *projection* onto the $\mathbf{e}_0$ direction) *cannot be separated from each other* in this geometry.

This is the circle singularity or "dangerous circle" in the case of GPS positioning, compare subsection 5.8.1. GPS positioning is in fact a three-dimensional resection!

11.8.6 Computing example for DOP quantities

See figure 11.20.

Let one satellite be in the zenith ($\eta = 90°$) and three satellites at an elevation angle $\eta = 30°$ at azimuths $\alpha = 0°, 120°, 240°$. Compute first the design matrix A according to formula 11.9. Given is

$$\begin{aligned} \alpha_1 &= 0°, & \eta_1 &= 90°, \\ \alpha_2 &= 0°, & \eta_2 &= 30°, \\ \alpha_3 &= 120°, & \eta_3 &= 30°, \\ \alpha_4 &= -120°, & \eta_4 &= 30°. \end{aligned}$$

Numeric values are obtained by remembering that

$$\begin{aligned} &\sin(90°) = 1, && \cos(90°) = 0, \\ &\sin(30°) = 1/2, && \cos(30°) = \tfrac{1}{2}\sqrt{3}, \\ &\sin(120°) = -\sin(-120°) = \tfrac{1}{2}\sqrt{3}, && \cos(120°) = \cos(-120°) = 1/2. \end{aligned}$$

The result is

$$A = \begin{bmatrix} 0 & 0 & 1 & -c \\ \frac{1}{2}\sqrt{3} & 0 & 1/2 & -c \\ -\frac{1}{2}\cdot\frac{1}{2}\sqrt{3} & \left(\frac{1}{2}\sqrt{3}\right)^2 & 1/2 & -c \\ -\frac{1}{2}\cdot\frac{1}{2}\sqrt{3} & -\left(\frac{1}{2}\sqrt{3}\right)^2 & 1/2 & -c \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & -c \\ \frac{1}{2}\sqrt{3} & 0 & 1/2 & -c \\ -\frac{1}{4}\sqrt{3} & \frac{3}{4} & 1/2 & -c \\ -\frac{1}{4}\sqrt{3} & -\frac{3}{4} & 1/2 & -c \end{bmatrix}.$$

Next we compute the weight matrix of the unknowns, equation 11.10:

$$P_{\mathrm{xx}} = A^{\mathsf{T}}A = \begin{bmatrix} 9/8 & 0 & 0 & 0 \\ 0 & 9/8 & 0 & 0 \\ 0 & 0 & 3/4 & -\frac{3}{2}c \\ 0 & 0 & -\frac{3}{2}c & 4c^2 \end{bmatrix}.$$

Inverting this matrix would yield Q, the weight-coefficient matrix of the unknowns. Here, we invert the matrix *only partially*:

$$Q_{\mathrm{xx}} = P_{\mathrm{xx}}^{-1} = \begin{bmatrix} 8/9 & 0 & 0 & 0 \\ 0 & 8/9 & 0 & 0 \\ 0 & 0 & \multicolumn{2}{c}{\multirow{2}{*}{$\begin{bmatrix} 3/4 & -\frac{3}{2}c \\ -\frac{3}{2}c & 4c^2 \end{bmatrix}^{-1}$}} \\ 0 & 0 & & \end{bmatrix} = \begin{bmatrix} q_{xx} & & & \\ & q_{yy} & & \\ & & q_{zz} & q_{zt} \\ & & q_{tz} & q_{tt} \end{bmatrix}. \tag{11.14}$$

From this we read directly, that the weight coefficients of the co-ordinates X and Y are

$$q_{xx} = q_{yy} = 8/9 = 0.889\ldots,$$

and thus

$$\mathrm{HDOP} = \sqrt{q_{xx} + q_{yy}} = \sqrt{16/9} = 4/3 = 1.333\ldots.$$

From formula 11.14 we see, that the $\widehat{Z}$ co-ordinate and the clock unknown $\widehat{\Delta T}$ are "entangled" with each other ($q_{zt} \neq 0$) and calculating their weight coefficients is not attempted here[22].

---

[22] In fact, if instead of the unknown $\widehat{\Delta T}$ we consider the unknown $c\widehat{\Delta T}$ — a recommended trick, as now, the *conditioning* of the matrix is better, i.e., it is less sensitive to perturbations and the rounding errors of calculation have less impact —, the sub-matrix to be inverted is

$$\begin{bmatrix} 3/4 & -3/2 \\ -3/2 & 4 \end{bmatrix}^{-1} = \begin{bmatrix} 16/3 & 2 \\ 2 & 1 \end{bmatrix},$$

and

$$Q_{\mathrm{xx}} = \begin{bmatrix} 8/9 & & & \\ & 8/9 & & \\ & & 16/3 & 2 \\ & & 2 & 1 \end{bmatrix}.$$

This is now the weight-coefficient matrix (i.e., up to a scale factor, the variance matrix) of the vector of unknowns $\begin{bmatrix} \widehat{X} & \widehat{Y} & \widehat{Z} & c\widehat{\Delta T} \end{bmatrix}^{\mathsf{T}}$.

In this result is also seen, how $\widehat{Z}$ and $\widehat{\Delta T}$ "compete" for the same information: $\mathrm{VDOP} = \sqrt{q_{zz}} = \sqrt{16/3} \approx 2.309$, when, without the clock unknown, it would be $\mathrm{VDOP} = \sqrt{4/3} \approx 1.155$.

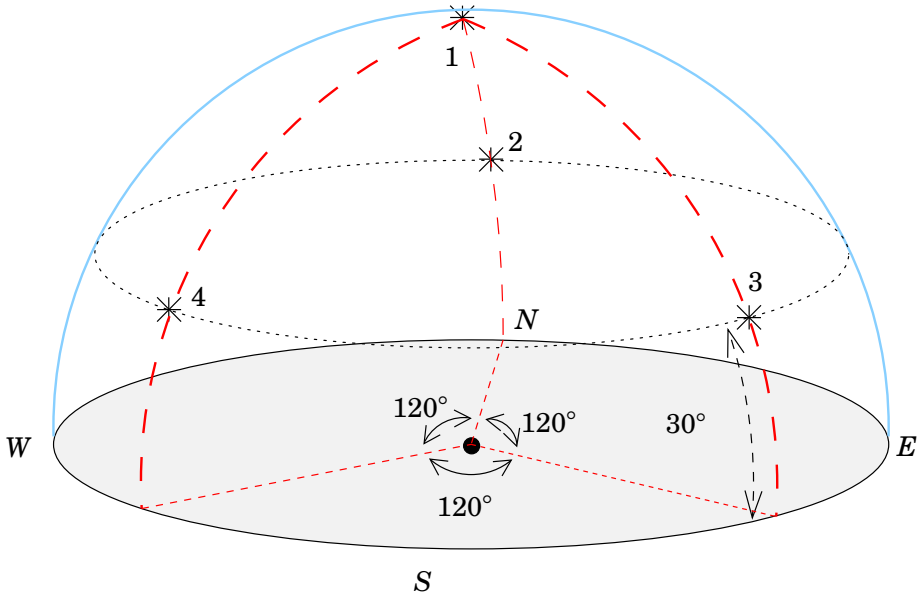


Figure 11.20. Calculation example of DOP quantities.

11.9 The orbits of GPS satellites

The orbit of GPS satellites in the Earth's gravitational field is approximately an ellipse satisfying Kepler's laws. In practice, it is almost a circle, the radius[23] of which is 26,560 km, and the orbital period $11^{\mathrm{h}}58^{\mathrm{m}}$. The tilt angle or the orbital planes with respect to the equator, the *inclination*, is $i = 55°$, meaning that at the latitude of Finland, the GPS satellites will never pass through the zenith. However, due to their great height, the satellites are also visible "over the North pole" in the Northern half of the sky, albeit very low. See figure 11.21.

We need *six orbital elements* for describing a satellite orbit. As orbital elements we may choose, e.g., the three components of place

$$\mathbf{r}(t_0) = x(t_0)\cdot\mathbf{i} + y(t_0)\cdot\mathbf{j} + z(t_0)\cdot\mathbf{k}$$

and the three components of velocity

$$\dot{\mathbf{r}}(t_0) = \dot{x}(t_0)\cdot\mathbf{i} + \dot{y}(t_0)\cdot\mathbf{j} + \dot{z}(t_0)\cdot\mathbf{k}$$

at a certain time t_0, using Newton's dot notation for the time derivative, and the vectors $\{\mathbf{i},\mathbf{j},\mathbf{k}\}$ forming an orthonormal basis. From these one may calculate place and velocity

$$\mathbf{r}(t) = x(t)\cdot\mathbf{i} + y(t)\cdot\mathbf{j} + z(t)\cdot\mathbf{k},$$
$$\dot{\mathbf{r}}(t) = \dot{x}(t)\cdot\mathbf{i} + \dot{y}(t)\cdot\mathbf{j} + \dot{z}(t)\cdot\mathbf{k},$$

[23] So: the distance from the Earth's surface is about 26,560 km − 6378 km = 20,182 km, where we have used 6378 km for the Earth's radius.

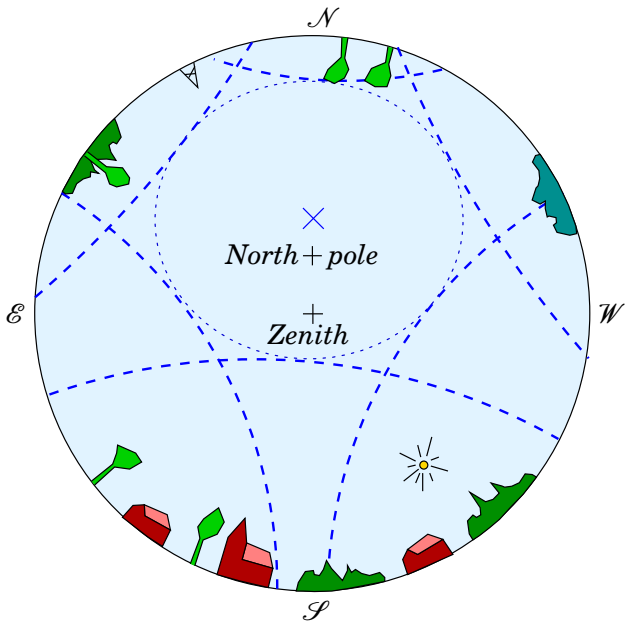


Figure 11.21. The six orbital planes of GPS satellites in the Helsinki sky. Note that inside the oval around the zenith, GPS satellites will never be found, although they are also visible in the Northern sky, though very close to the horizon.

for some later moment t, by just calculating, step by small step, forward in time, correcting both the velocity using the gravitation equation, and the place using the velocity, figure 11.22. We know the attractional field of the Earth as an equation: the acceleration caused by the attraction is computable when we know the place in space of the satellite.

The geometry of a satellite orbit is normally described using the *six Kepler orbital elements*[24], Ω, i, ω, a, e and ν, see figure 16.13. More details are given in Poutanen (1998) page 97 and in section 16.7. Between the Kepler elements and the position and velocity vector representation described above, there is a one-on-one correspondence:

$$\text{Kepler} : \{\Omega, i, \omega, a, e, \nu\} \Longleftrightarrow \{\mathbf{r}(t_0), \dot{\mathbf{r}}(t_0)\}.$$

This means that, from given Kepler elements, we may *calculate* the *position* of the satellite in space, as well as its *velocity*. All GPS computation software packages know how to do this.

11.9.1 The navigation message transmitted by the satellites

As already noted, all GPS satellites transmit a navigation message modulated upon the carrier wave of their radio signal. The modulation frequency of the nav message is 50 Hz, so every second contains 50 bits.

[24] So *every* satellite has six Kepler orbital elements that describe the shape, size and orientation of the orbit of that satellite, as well as the location of the satellite in its orbit, in space.

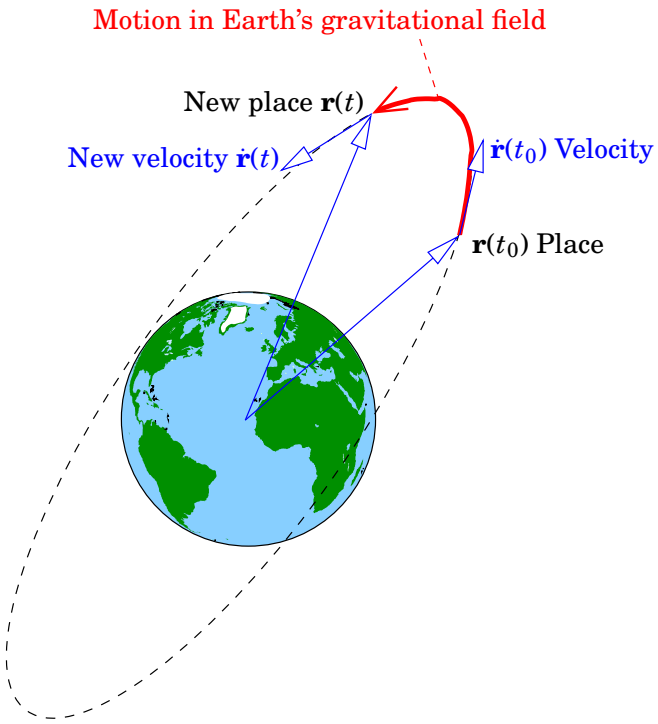


Figure 11.22. Satellite orbital motion described by position and velocity vectors.

The whole navigation message consists of 25 "packets" (*frames*), each of which contains 1500 bits and transmission of which lasts 30 seconds. Thus, the total length is 37,500 bits, and the duration of the transmission is 12.5 minutes.

When a GPS receiver is switched on for the first time, the search for satellites starts. Immediately when the first satellite is "caught" (*lock-on*), the reading of the navigation message starts. Lock-on may easily last for several minutes, especially if the approximate location fed to the receiver is seriously wrong, or the instrument has been transported between continents. After that, however, the finding of the other satellites generally proceeds apace.

The navigation message is uploaded to the satellites by the Control Segment, typically once every 24 hours. After that, the information is, as part of the signal of the GPS satellites, available to all users of the GPS system. The navigation message consists of the following three parts:

1. Information related to time keeping, like the clock corrections of the satellites; the "health" information on the satellites, i.e., how good quality the positioning signal and orbital data transmitted by the satellite are; freshness of the nav message.
2. The satellite's orbital information (*broadcast ephemeris*), i.e., orbital information disseminated by the satellite itself by radio. These ephemeris are computed by the U.S. military authorities and are

based on observation data continuously produced by a global network of tracking stations. The orbital elements of all satellites computed from the observations are uploaded to the satellites by the control segment, typically once per 24 hours. From the satellite's memory, they are then transmitted, modulated onto the radio signal, as a bit stream to all users.

The orbital elements are the Kepler elements augmented by coefficients, nine in total, describing the perturbations caused by the Earth's flattening, both short period (half the satellite orbital period) and secular (growing linearly with time), which for GPS satellites need to be taken into account. The origin of the theory used is the classical article by Yoshihide Kozai (Kozai, 1959).

Broadcast ephemeris are used in navigation applications and in real-time positioning. Using them is practical also in GPS surveying and relative positioning in relatively small areas.

From the ephemeris, every satellite's position in space at the moment of observation is computed, so that they may be used as "beacons" for the determination of the location of the ground station, as well as the velocity of the satellite[25]. More is said about the computing methods used in Poutanen (1998) pages 98 – 104.

3. The *almanac* for all satellites[26]. The purpose of the almanac is to provide approximate orbital elements for all satellites, sufficient for planning measurement campaigns and helping the receiver in finding satellites. The validity of the almanac is many weeks.

   The almanac contains also a crude global ionosphere model.

11.9.2 Precise ephemeris

Precise ephemeris, i.e., precise orbital elements, may be obtained by the user some time afterward directly through the Internet. They are computed and distributed by a service called the International GNSS Service (IGS), see section 11.10. A standard format named SP3 ("Standard Product 3") is used for data distribution, originally designed by the US National Geodetic Survey.

The SP3 ephemeris file contains the orbital data in the form of three-dimensional vectors of position and velocity,

$$\begin{bmatrix} x(t_i) \\ y(t_i) \\ z(t_i) \end{bmatrix}, \begin{bmatrix} \dot{x}(t_i) \\ \dot{y}(t_i) \\ \dot{z}(t_i) \end{bmatrix},$$

[25] Knowing the velocity of the satellite would not be necessary for this, but it is needed to calculate the *Doppler shift* of the signal frequency. The receiver must know the Doppler shift of every satellite in order to lock on, and remain locked on, to every satellite signal.

[26] *Almanac* is apparently not a word of Arabic origin, though it looks like it is.

Table 11.8. Precise ephemeris in the original SP3 format. Satellite numbers, vectors of place, velocity vectors, clock correction and clock drift, date and time, etc. Example © U.S. National Geodetic Survey.

```
V1994 12 17  0  0  0.00000000      96     d ITR92 FIT  NGS
##  779 518400.00000000   900.00000000 49703 0.0000000000000
+   25     1  2  4  5  6  7  9 12 14 15 16 17 18 19 20 21 22
+         23 24 25 26 27 28 29 31  0  0  0  0  0  0  0  0  0
+          0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
+          0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
+          0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
++         7  6  5  5  5  5  5  5  5  6  5  5  5  5  6  5  5
++         5  5  6  5  5  5  5  5  0  0  0  0  0  0  0  0  0
++         0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
++         0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
++         0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
%c cc cc ccc ccc cccc cccc cccc cccc ccccc ccccc ccccc ccccc
%c cc cc ccc ccc cccc cccc cccc cccc ccccc ccccc ccccc ccccc
%f  0.0000000  0.000000000  0.00000000000  0.000000000000000
%f  0.0000000  0.000000000  0.00000000000  0.000000000000000
%i    0    0    0    0      0      0      0      0         0
%i    0    0    0    0      0      0      0      0         0
/* CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
/* CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
/* CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
/* CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
*  1994 12 17  0  0  0.00000000
P  1  16258.524750  -3529.015750 -20611.427050    -62.540600
V  1  -6560.373522  25605.954994  -9460.427179     -0.024236
P  2 -21998.652100  -8922.093550 -12229.824050   -131.326200
V  2  -9852.750736 -12435.176313  25738.634180     -0.029422
P  4 -26019.547600   4809.810900  -2508.578200      3.544600
V  4   2559.038002  -3340.527442 -31621.490838      0.016744
   *
   *
   *
P 29  -1638.431050 -24391.479200  10455.312650      3.690300
V 29   5754.005457 -12065.761570 -27707.056273      0.003537
P 31   6265.255800 -25687.986950   -753.359000     70.830800
V 31   3053.344058    -63.091750  31910.454757      0.033749
*  1994 12 17  0 15  0.00000000
P  1  15716.820135  -1169.850490 -21281.578766    -62.542746
V  1  -5439.955846  26738.341429  -5409.793390     -0.023226
P  2 -22813.261065  -9927.616864  -9816.490189   -131.328686
V  2  -8178.974330  -9924.329320  27813.754308     -0.025238
   *
   *
```

for every epoch t_i. The velocity vector is optional. The data is tabulated at time intervals of 15 minutes, i.e., $t_{i+1} - t_i = 15^{\mathrm{m}}$. From these, the place $\mathbf{r}(t)$ and velocity $\dot{\mathbf{r}}(t)$ is interpolated to the moment of measurement t by

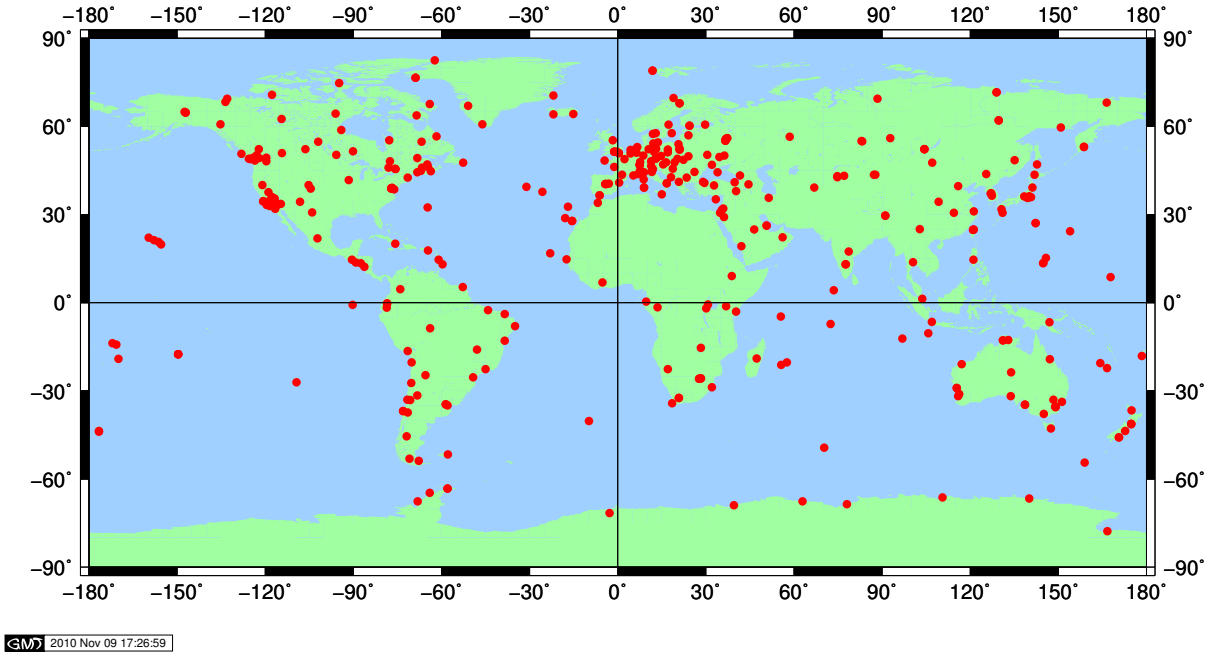


Figure 11.23. The tracking stations of the IGS, situation 2010 (data ľ IGS).

Lagrange[27] interpolation.

The current version is SP3-c, which allows also the distribution of orbital information on GLONASS satellites. It is a *text format*, i.e., human-readable.

The best-known source has been the International GNSS Service already from 1992. The precise orbits produced by them are published on the Internet a couple of weeks after the time of observation.

Recently, also "rapid orbits" solutions have begun to be produced. These are almost as precise as precise ephemeris, but are turned out faster.

The newest "ultra rapid" solutions are satellite orbital predictions twenty-four hours into the future, which can thus be used in real-time applications.

Precise ephemeris are, unlike broadcast ephemeris, disseminated over the Internet and *not via GPS satellites*. The above organizations are independent from the United States military authorities.

Unlike broadcast ephemeris, precise ephemeris are very close to the *true* orbits of the satellites, where they actually *were* at that moment. Broadcast ephemeris are *predictions* and therefore less accurate.

[27] Joseph-Louis Lagrange (1736 – 1813) was a French mathematician, astronomer, developer of classical mechanics, one of the 72 names on the Eiffel Tower.

11.10 The International GNSS Service

The International GNSS Service (IGS) was established in 1990 by the IAG (International Association of Geodesy), and it became an official service of the IAG in 1994. The main purpose of the service is to produce precise orbital ephemeris in support of *geodynamics* research, the study of the motions of the solid Earth. However, its products are used much more broadly, in many fields of geophysics.

The activities of IGS are led by a Central Bureau, currently at JPL (Jet Propulsion Laboratory) in the United States. In 2015 the IGS used globally observations from some 500 GNSS stations to compute its orbital ephemeris. The number has grown only slowly over recent years.

The computation as such is carried out by seven different computing centres; the orbital data is available for use a couple of weeks after the time of measurement. In the computed data are included also the clock correction parameters for the satellites, and separately the Earth's rotation parameters (EOP), i.e, polar motion, variations in the length of the day (LoD), and precession and nutation.

The IGS Central Bureau Information Service is found at the address http://igscb.jpl.nasa.gov.

Self-test questions

1. Explain how a hyperbolic positioning system like Decca functions.

Exercise 11 – 1: Calculation of DOP quantities

In this exercise you are going to write a piece of software to calculate the various DOP quantities of a GPS measurement geometry, when the positions of the satellites in the sky, their azimuths α and elevation angles η, are given.

You may use your preferred rapid prototyping language: MatLab™, GNU Octave, Scilab, R; even Excel™. And don't bother with inputting from files, just put the satellite positions into the source.

1. For an arbitrary number of satellites, write, or adapt, code to build the design matrix A and the normal or weight matrix $P = A^{\mathsf{T}}A$ (one could also build, from the vector of observations $\underline{\ell}$, the right-hand side of the system of normal equations, $A^{\mathsf{T}}\underline{\ell}$, but we *don't need it here*. The beauty of DOP is that we can use it in measurement planning, before any real observations are available.)
2. Make your software output to the screen the various DOP quantities.
3. Now assume you have *five satellites*, one more than the minimum to make positioning possible. Play with the satellite positions $(\alpha_i, \eta_i), i =$

Table 11.9. DOP calculation script

```
% Dilution of Precision (DOP):
% Part 1. Run the program. What does the error message tell you?
% What can you say about the value of VDOP?
% Places of satellites, azimuth A and elevation h in the sky (below).
% Part 2. Change the elevation of the first satellite 30 -> 60 degrees.
% Run again. Note down the HDOP and VDOP values.
% Why is the run now successful?
A1 = 0; h1 = 30;
A2 = 90; h2 = 30;
A3 = 180; h3 = 30;
A4 = 270; h4 = 30;
conv = pi/180;
% Design matrix A (below)
% Part 3. Modify the program to include a fifth satellite, place in the sky
% A5 = 0, h5 = 45.
A = [cos(A1*conv)*cos(h1*conv), sin(A1*conv)*cos(h1*conv), sin(h1*conv), -1;
cos(A2*conv)*cos(h2*conv), sin(A2*conv)*cos(h2*conv), sin(h2*conv), -1;
cos(A3*conv)*cos(h3*conv), sin(A3*conv)*cos(h3*conv), sin(h3*conv), -1;
cos(A4*conv)*cos(h4*conv), sin(A4*conv)*cos(h4*conv), sin(h4*conv), -1];
N = A'*A;
Ninv = inv(N)
HDOP = sqrt(Ninv(1,1) + Ninv(2,2));
VDOP = sqrt(Ninv(3,3));
% Part 4. Add to the program the evaluation and output of PDOP
% PDOP = Position Dilution of Precision. See lecture notes
% for definition.
fprintf(1, 'HDOP = %20.10f\n', HDOP);
fprintf(1, 'VDOP = %20.10f\n\n', VDOP);
% Part 5. Play around with the five satellite places in the sky, in order
% to minimize PDOP.
```

1,...,5, in order to minimize one of the DOPs, e.g., the PDOP. What is the best geometry of five satellites you found?

4. You cheated, didn't you? You cannot observe GPS satellites that are below the horizon. So, introduce the constraint $\eta > 5°$.
5. After PDOP, try to minimize HDOP, and VDOP.

What did you learn?

12. Processing GPS observations

12.1 Forming difference observations

In geodetic GPS measurement and GPS surveying applications, commonly one wishes to measure the difference in location between two points. The distance between the points may be, e.g., 100 km or 1000 km. This is a much shorter distance than that to the GPS satellites, which orbit at a height of some 20,000 km. See figure 12.1.

Seen from the satellite, the angle α separating the observation sites is very small, in the example case (Helsinki and Sodankylä) only 2°! For this reason many errors will be partly common to the two sites, similar and approximately equal in magnitude. The satellite clock error or *clock offset* is even identical. The effect of orbit error is roughly the same due to the geometry, the errors caused by ionosphere and troposphere are also similar, due to the similarity in geometry and the long-range *spatial correlation*[1] of atmospheric conditions. One should however remember

[1]By this is meant that conditions change only slowly with place. Helsinki and

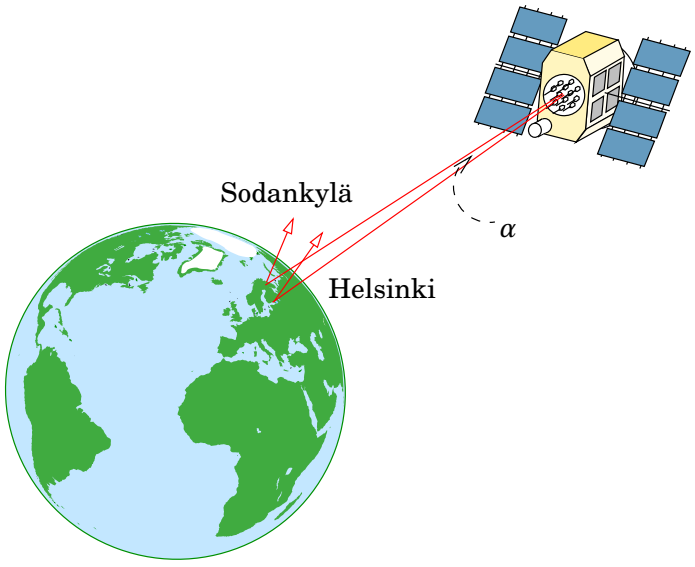
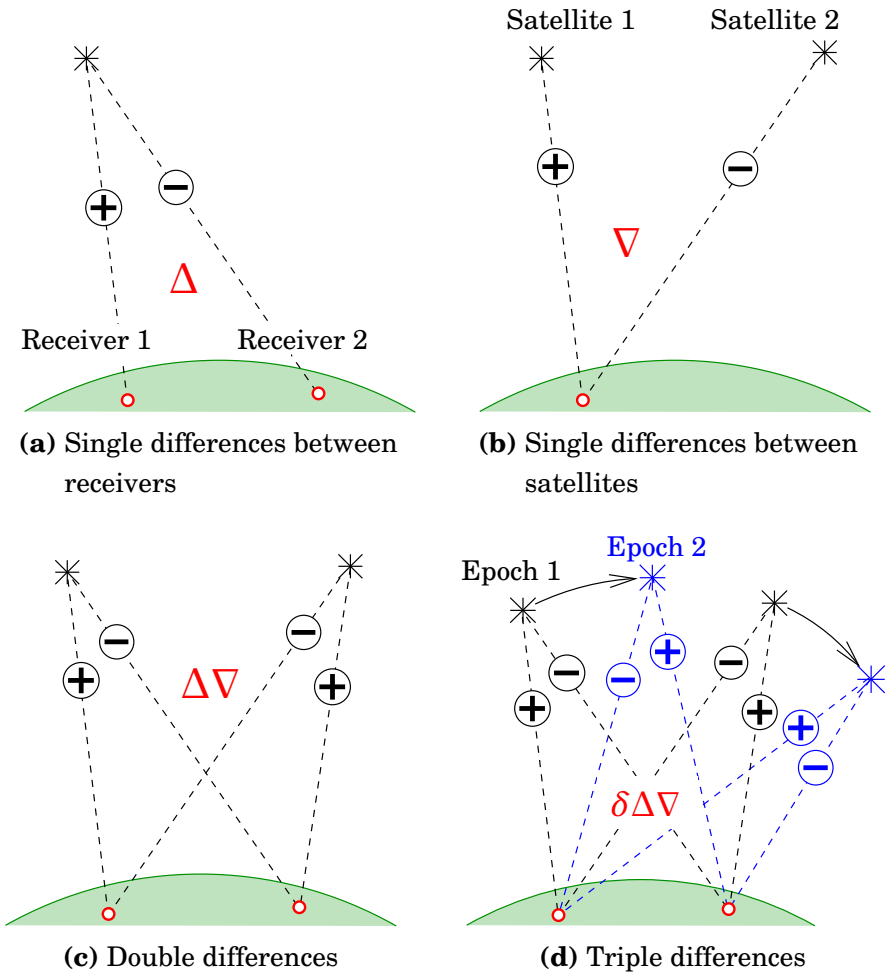


Figure 12.1. "Common-mode" error assumption.



(a) Single differences between receivers

(b) Single differences between satellites

(c) Double differences

(d) Triple differences

Figure 12.2. Forming various difference observations and the symbols used for them.

that the *difference in directions of the plumb line* between Helsinki and Sodankylä is already 7°, so the difference in elevation angle of a satellite above the local horizon may amount to this much.

Based on the "common error" assumption, we form *differences* between observations from two sites to one satellite. In these differences, many errors vanish entirely or are materially reduced. Forming the difference is straightforward: subtract two simultaneously made raw observations from each other, each lifted from an observation file looking like table 11.3.

The differences can be *single* — either between two receivers or between two satellites, in which cases are used the visually appropriate symbols Δ or ∇ —, *double*, or *triple*. See figure 12.2.

Sodankylä lie practically in the same climate zone, and if there is a high or low pressure zone over Northern Europe, it will undoubtedly affect both Helsinki and Sodankylä. The *synoptic scale* (http://en.wikipedia.org/wiki/Synoptic_scale_meteorology) of weather phenomena is of order 1000 km.

Table 12.1. The effect of forming difference observations from GPS observations on the magnitude of various errors.

Error source \ type of difference	∇	Δ	$\Delta\nabla$	$\delta\Delta\nabla$
Satellite orbit $\{\mathbf{r}^S, \dot{\mathbf{r}}^S\}$	-	$\Downarrow$[a]	$\Downarrow$	$\Downarrow\Downarrow$[b]
Satellite clock Δt	-	0[c]	0	0
Receiver clock ΔT	0	-	0	0
Ionosphere $d_{\text{ion}}, D_{\text{ion}}$	-	$\Downarrow$	$\Downarrow$	$\Downarrow\Downarrow$
Troposphere $d_{\text{trop}}, D_{\text{trop}}$	-	$\Downarrow$	$\Downarrow$	$\Downarrow\Downarrow$
Ambiguities N	-	-	-	0[d]

[a] The error is reduced substantially, especially for small distances between measurement points.

[b] The error is diminished even more strongly.

[c] The error is completely eliminated.

[d] Only zero if there isn't a so-called *cycle slip*.

The *influence* of forming the various difference types on the magnitude of errors — the interesting thing here! — has been catalogued in table 12.1.

12.1.1 Example: single difference

We explain with equations how a single difference is calculated from original observations, and why some systematic errors are eliminated altogether, while some others are substantially reduced.

12.1.1.1 One receiver (observer) A, two satellites S, T:

$$\nabla^{ST} p_A = p_A^S - p_A^T,$$
$$\nabla^{ST} P_A = P_A^S - P_A^T.$$

Write equation 11.1:

$$p_A^S = \rho_A^S + c\left(\Delta t^S - \Delta T_A\right) + d_{\text{ion},A}^S + d_{\text{trop},A}^S,$$
$$p_A^T = \rho_A^T + c\left(\Delta t^T - \Delta T_A\right) + d_{\text{ion},A}^T + d_{\text{trop},A}^T.$$

Here it has been taken into account that, of the clock offsets, Δt is *satellite specific*, ΔT again is *observer*, i.e., *receiver, specific*. Subtraction yields the difference quantity

$$\nabla^{ST} p_A = \nabla^{ST} \rho_A + c\nabla^{ST} \Delta t + \nabla^{ST} d_{\text{ion},A} + \nabla^{ST} d_{\text{trop},A},$$

in which the definitions

$$\begin{aligned}\nabla^{ST}\rho_A &= \rho_A^S - \rho_A^T,\\ \nabla^{ST}\Delta t &= \Delta t^S - \Delta t^T,\\ \nabla^{ST} d_{\text{ion},A} &= d_{\text{ion},A}^S - d_{\text{ion},A}^T,\\ \nabla^{ST} d_{\text{trop},A} &= d_{\text{trop},A}^S - d_{\text{trop},A}^T.\end{aligned}$$

Here, the clock offset of receiver A, ΔT_A, has vanished, because, being a receiver property, it is the same for different satellites and cancels out in calculating the difference observation between satellites.

This is important in practice, because receiver clocks are usually based on inexpensive quartz oscillators, the drift of which can be considerable.

The same equation applies also for the carrier phase:

$$\nabla^{ST} P_A = \nabla^{ST}\rho_A + c\nabla^{ST}\Delta t + \nabla^{ST} D_{\text{ion},A} + \nabla^{ST} D_{\text{trop},A} + \lambda\nabla^{ST} N_A,$$

in which

$$\nabla^{ST} N_A = N_A^S - N_A^T$$

is the difference in ambiguities between satellites S and T. Without a *cycle slip* occurring, it will be constant in time.

□ *12.1.1.2 Two receivers A,B, one satellite S:*

$$\begin{aligned}\Delta_{AB} p^S &= p_A^S - p_B^S,\\ \Delta_{AB} P^S &= P_A^S - P_B^S.\end{aligned} \tag{12.1}$$

Here, in the same way, the satellite clock offset Δt drops out: it is a property of the satellite, not the receiver, and vanishes when one calculates the difference quantity between two different receivers with the same satellite.

Also, the influence of orbit errors, ionosphere and troposphere is reduced substantially: *for short distances between receivers* we have, e.g.,

$$\begin{aligned}\left|\Delta_{AB}\rho^S\right| &= \left|\rho_A^S - \rho_B^S\right| \ll \left|\rho_A^S\right| \approx \left|\rho_B^S\right|,\\ \Delta_{AB} d_{\text{ion}}^S &= d_{\text{ion},A}^S - d_{\text{ion},B}^S \approx 0,\\ \Delta_{AB} D_{\text{ion}}^S &= D_{\text{ion},A}^S - D_{\text{ion},B}^S \approx 0,\\ \Delta_{AB} d_{\text{trop}}^S &= d_{\text{trop},A}^S - d_{\text{trop},B}^S \approx 0,\\ \Delta_{AB} D_{\text{trop}}^S &= D_{\text{trop},A}^S - D_{\text{trop},B}^S \approx 0,\end{aligned}$$

because

1. The measurement geometry is almost the same in A as in B, see figure 12.1.
2. Atmospheric conditions don't change much between points A and B, i.e., the measurement rays traverse nearly the same air mass.
3. The elevation angle in the sky of satellite S seen from A is nearly the same as seen from B.

Table 12.2. Summary of GPS observables and difference quantities. In the equations is used the notations $d = d_{\text{ion}} + d_{\text{trop}}$, $D = D_{\text{ion}} + D_{\text{trop}}$. Note the consistent use of deltas and nablas.

Pseudo-range p	Carrier phase ϕ; equivalent pseudo-range P
$p_A^S = \rho_A^S + c(\Delta t^S - \Delta T_A) + d_A^S$	$P_A^S = \rho_A^S + c(\Delta t^S - \Delta T_A) + D_A^S + \lambda N_A^S$
Single difference, between satellites:	
$\nabla^{ST} p_A = \nabla^{ST} \rho_A + c\nabla^{ST} \Delta t + \nabla^{ST} d_A$	$\nabla^{ST} P_A = \nabla^{ST} \rho_A + c\nabla^{ST} \Delta t + \nabla^{ST} D_A + \lambda \nabla^{ST} N_A$
Single difference, between receivers:	
$\Delta_{AB} p^S = \Delta_{AB} \rho^S + c\Delta_{AB} \Delta T + \Delta_{AB} d^S$	$\Delta_{AB} P^S = \Delta_{AB} \rho^S + c\Delta_{AB} \Delta T + \Delta_{AB} D^S + \lambda \Delta_{AB} N^S$
Double difference:	
$\nabla^{ST} \Delta_{AB} p = \nabla^{ST} \Delta_{AB} \rho + \nabla^{ST} \Delta_{AB} d$	$\nabla^{ST} \Delta_{AB} P = \nabla^{ST} \Delta_{AB} \rho + \nabla^{ST} \Delta_{AB} D + \lambda \cdot \nabla^{ST} \Delta_{AB} N$
Triple difference:	
$\delta_{12} \nabla^{ST} \Delta_{AB} p = = \delta_{12} \nabla^{ST} \Delta_{AB} \rho + \delta_{12} \nabla^{ST} \Delta_{AB} d$	$\delta_{12} \nabla^{ST} \Delta_{AB} P = = \delta_{12} \nabla^{ST} \Delta_{AB} \rho + \delta_{12} \nabla^{ST} \Delta_{AB} D + \lambda \cdot (\text{cycle slips})$

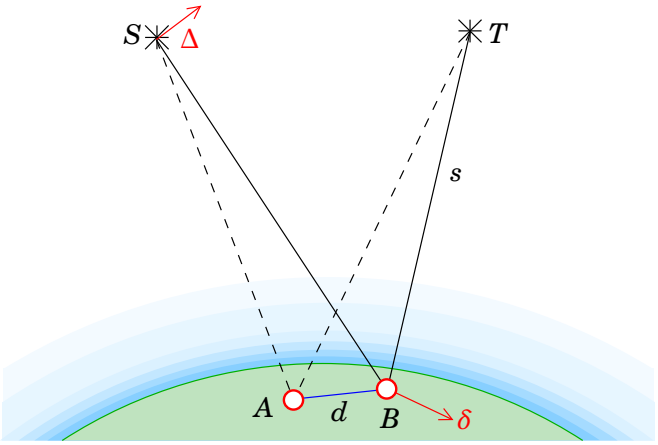


Figure 12.3. Double difference, short distance between GPS receivers.

12.1.2 Other difference quantities

In the same way, by combining the operations described above, we may calculate also double and triple differences. The formulas look complicated but the process is straightforward, see the summary in table 12.2. The equations in the table are directly derived from the original observation equations 11.1 and 11.2 by addition and subtraction.

12.2 Relative (static) GPS

The difference in location, or *vector*, $\mathbf{R}_{AB} = \mathbf{R}_B - \mathbf{R}_A$ between two observation sites A and B may be solved more precisely than the absolute location of either site $\mathbf{R}_A, \mathbf{R}_B$ with respect to the centre of mass of the Earth. See figure 11.15. The reason for this is the cancellation or near-cancellation from simultaneous observations of various error sources, which are almost the same in both points. For computation are used *difference observations* $\Delta_{AB}p, \Delta_{AB}P$ between the two observation sites, in which this cancellation already occurs. As was shown already earlier in figure 12.1, the places of the satellites in the sky are nearly the same seen from observation sites A and B, and also the atmosphere above A and B is surely rather similar.

How much, e.g., the *orbit error* of satellites S and T affects the determination of the vector AB? See figure 12.3. An order-of-magnitude rule of thumb says, that the positioning error δ caused by orbit error in point B relative to point A is

$$\delta \approx \frac{d}{s}\Delta,$$

in which Δ is the assumed orbit error. This is only a crude estimate. We know that $s \gtrsim 20{,}000\,\mathrm{km}$.

The values given in table 12.3 for the orbit error, 1 m and 2 cm, corre-

Table 12.3. Relation between orbit error, length of vector, and positioning error.

Vector length d (km)	Orbit error Δ (m)	Positioning error δ (mm)
1	1	0.05
10	1	0.5
100	1	5
1000	1	50
1	0.02	0.001
10	0.02	0.01
100	0.02	0.1
1000	0.02	1

spond to the accuracies of today's *broadcast*[2] and *precise* ephemeris. The conclusion is, that

> *In GPS surveying work in a small area* $(1-100\,\mathrm{km})$ *the orbit may generally be assumed known.*

In geodetic work, first one computes double differences $\nabla^{ST}\Delta_{AB}P$ from observations in points A and B. As we are dealing with carrier-phase observations, the ambiguities or integer unknowns $\nabla^{ST}\Delta_{AB}N$ must first be resolved. After that, a *vector*

$$\overline{\mathbf{R}}_{AB} = \begin{bmatrix} \Delta X_{AB} \\ \Delta Y_{AB} \\ \Delta Z_{AB} \end{bmatrix} = \begin{bmatrix} X_B - X_A \\ Y_B - Y_A \\ Z_B - Z_A \end{bmatrix}$$

between the points is computed from the observations. This is where the name *relative GPS* (or GNSS) *measurement* comes from.

This description may be easily generalized to cover the measurement of several points, a *geodetic network*.

12.3 Fixing ambiguities

Resolving the integer unknowns or *ambiguities* is a precondition for using GPS carrier-wave observations. There are several methods for this.

[2]This is a crude estimate. The quality of broadcast ephemeris has improved since the early days of GPS, slowly but surely. Other GNSS systems, like GLONASS, Galileo and BeiDou, perform at about the same level, or perhaps a little poorer (Montenbruck et al., 2015).

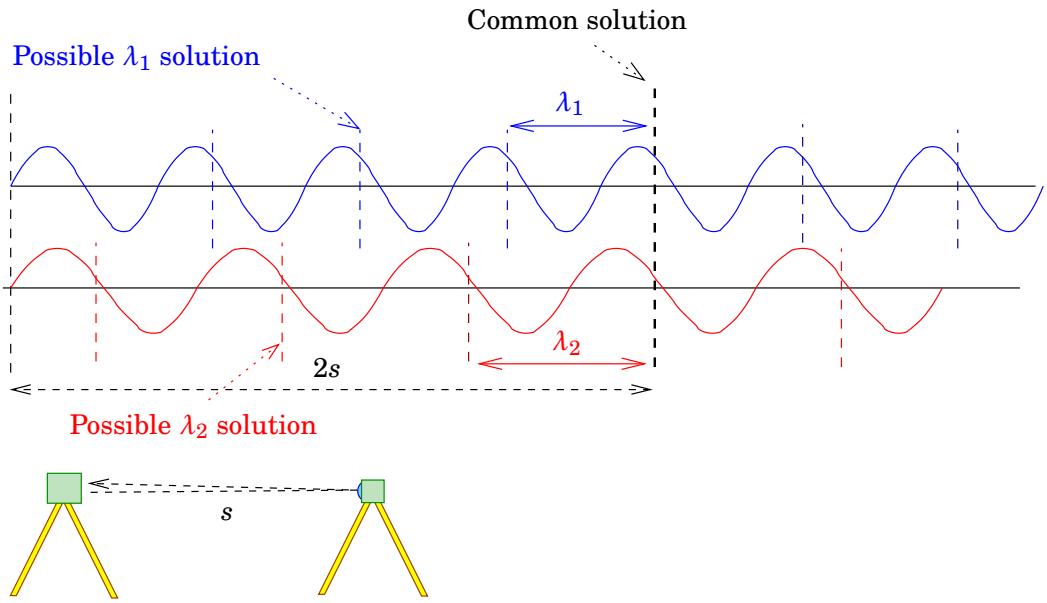


Figure 12.4. One-dimensional ambiguity resolution in the case of a distance measurement instrument.

1. *Distance measurement equipment* resolves the integers by measuring at several different wavelengths. In figure 12.4, as earlier in figure 6.10, is shown how this puzzle can be solved.

   A GPS satellite transmits on two frequencies L_1 and L_2. When there are two frequencies, they can be combined in a way which makes it easier to solve the ambiguities over short distances. Calculate the *phase difference* $\phi_w = \phi_1 - \phi_2$ between the phase measurements at L_1 and L_2. It is like using a carrier wave with a frequency $f_w = f_1 - f_2 = 347.82\,\text{MHz}$, and a corresponding "wavelength" $\lambda_w = c/f = 86\,\text{cm}$. This method is called *wide-laning*.

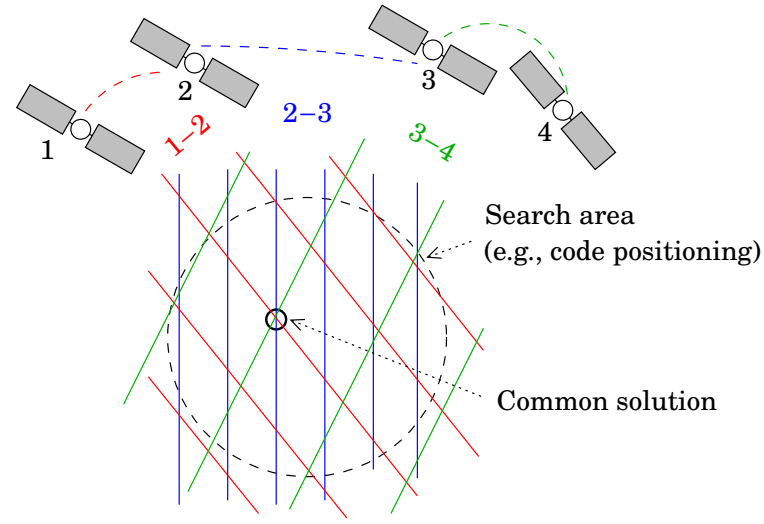
   The wide-lane solution only works over short distances, because otherwise the difference in ionospheric influence between the two measurement sites grows too large and shuffles the deck.

   The code observation yields a pseudo-range already at the metre precision level, after which the wide-laning method yields the ambiguities, and the phase difference ϕ_w yields an "ambiguity free" pseudo-range at a precision level of centimetres.

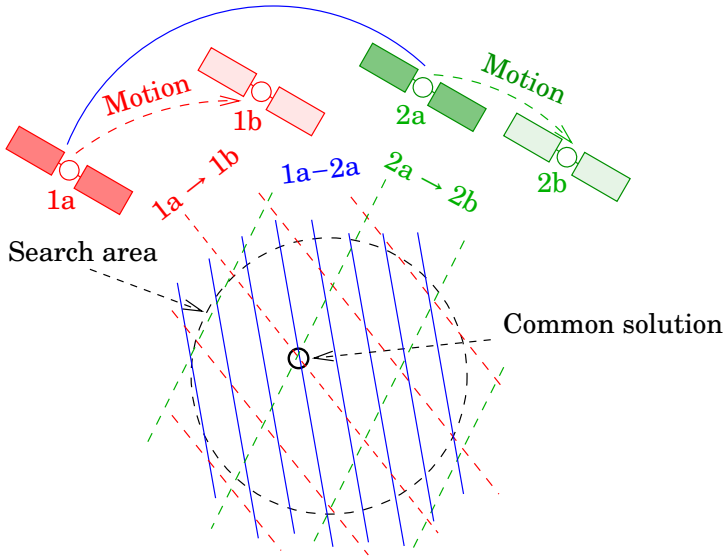
2. Use many satellites. At the moment, over 30 GPS satellites orbit the Earth, at any moment, of these, 6 – 10 are visible. This is a more complicated method, because, unlike in distance measurement, the geometry is three-dimensional. Efficient algorithms for this exist.

3. Use the same satellites for a longer time. Because the GPS measurement geometry has time to change, we obtain more conditions.

See figure 12.5.



(a) Three-dimensional



(b) Three-dimensional plus time

Figure 12.5. Various ambiguity-resolution methods used in GPS computation.

In recent years, other global positioning systems have appeared besides GPS. The Russian GLONASS — in which every satellite has its own carrier transmission frequency, so-called *frequency division multiple access* (FDMA), which complicates ambiguity resolution — has, after a long period of decay, started to actively increase again the number of operational satellites, the upcoming European Galileo system has lauched several test satellites and launches more all the time, and also the Chinese are coming with their Beidou-2 system. Joint use of the systems in the same receiver is technically complicated but promises a very fast and reliable resolution of ambiguities. Relief may come from *software defined GNSS receiver*s, in which all processing work after the antenna and analogue electronics is implemented digitally in software inside an

off-the-shelf computer.

12.4 Real-time positioning

The static method described above is based on *post-processing*. For geodetic use this is usually unproblematic. The use of precise satellite orbits — essential if one wants geodetic precision for long vectors — imposes also a certain waiting time: a couple of weeks in the case of precise ephemeris.

Sometimes, however, we need the co-ordinates of new points immediately, and there are also situations in which this would be useful or handy. Then, we speak of *real-time*[3] positioning. *Navigation* is a broad field of application. tosiaikainen

Real-timeness can be implemented by transferring the observations made in point A on the fly to point B, e.g., by radio.

12.4.1 Differential GPS (DGPS)

Differential GPS is a real-time positioning method based on *code observations*, which makes use of a reference or *base station*. It thus is a *relative* measurement between base station A and moving receiver, or *rover*, B. As always with measurements between two receivers, the error sources are the same or nearly the same in both receivers: the orbit errors and clock offsets of the satellite, as well as the effect of the atmosphere, cancel out in the difference measurement between A and B, either completely or nearly so. tukiasema

In the DGPS method, it is not the raw observations in A that are transferred to B — that would be an excessive amount of information to transfer. Instead, from the observations is first calculated the *difference between measured and computed pseudo-range*. One may compute, for every observation p_A, the geometric distance $\rho_A^{(0)}$ between observation site A and satellite S. This "reference distance" is based on the known location of point A and the orbital data transmitted by the satellite. Then, one obtains the *pseudo-range offset* for each satellite in the sky:

$$dp_A \stackrel{\text{def}}{=} p_A - \rho_A^{(0)}.$$

The orbital data from the same satellite are also available to the moving point B, which may itself compute from these the same $\rho_A^{(0)}$ — after all, the location of point A is known. So, the *information content* of the

[3]The formal definition of real-timeness is *a guaranteed latency*. It may be long, as long as it is guaranteed. The latency or response time is the time that elapses from the measurement event to the availability for use of the measurement values.

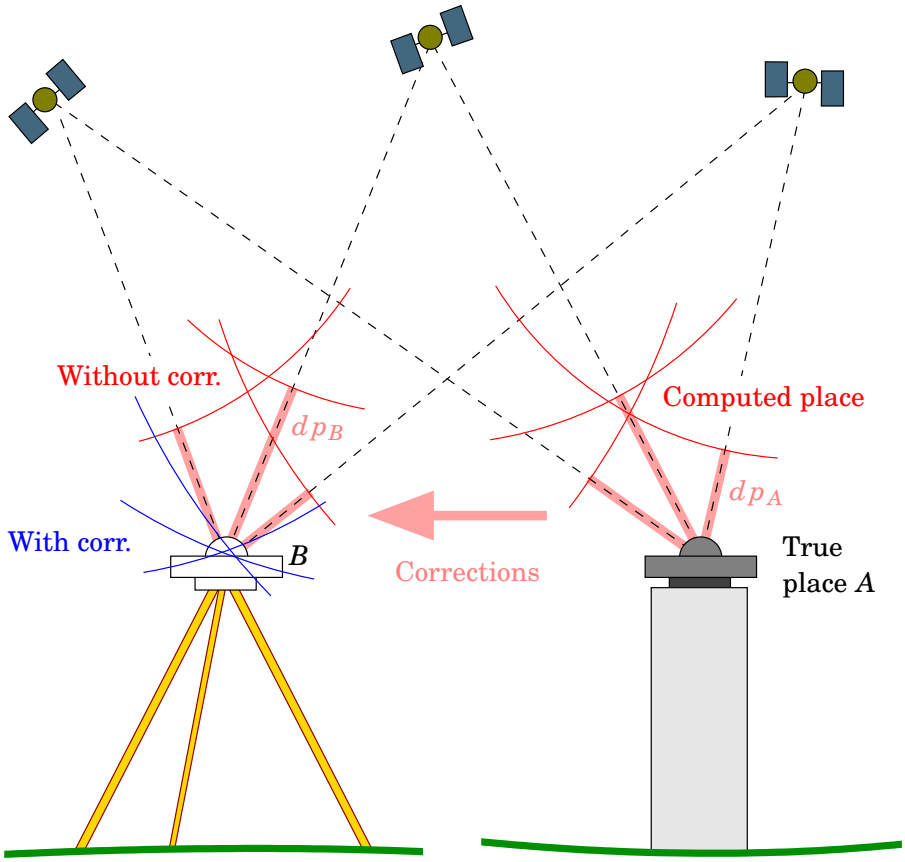


Figure 12.6. Principle of operation of the DGPS method, somewhat simplified.

□

offsets dp_A *is the same* as that of the full measurements p_A, and the offsets may replace them in the dissemination.

Using pseudo-range *offsets* has the following advantages:

- The numerical values are much smaller. The offsets were of order ±100 m back when Selective Availability (SA, an artificial reduction in accuracy of the disseminated orbital and clock information) was still on. When, in 2000, SA was switched off, the magnitude of the offsets dropped to the level of ±5 m. Both orders of magnitude are a fraction of the size of the observables themselves, thousands of kilometres.
- The values change more slowly. They crawl over the course of hours in a way which looks random. Extrapolation over several seconds or minutes into the future works better than with raw observations.

For both of these reasons, the communications bandwidth needed is much less, and the following channels are sufficient:

- mobile telephony. Modern network data connections (3G, 4G) are fast indeed. The “mobile Internet”
- a radio modem (short distances)

- for navigation at sea, long-wave radio.

The values of the offsets vary also slowly as a function of *place*. Therefore we may, if the distance AB is suitably short, 100 – 1000 km, reconstruct with sufficient accuracy

$$dp_B \approx dp_A + c(-\Delta T_A + \Delta T_B),$$

in which $\Delta T_A, \Delta T_B$ are the receiver clock offsets.

In addition to the original, single observations, we may also construct *difference observations* $\nabla^{ST} p_B$ between two satellites S and T:

$$\left.\begin{array}{l} dp_B^S \approx dp_A^S + c(-\Delta T_A + \Delta T_B) \\ dp_B^T \approx dp_A^T + c(-\Delta T_A + \Delta T_B) \end{array}\right\} \Rightarrow \nabla^{ST} dp_B = dp_B^T - dp_B^S \approx dp_A^T - dp_A^S,$$

a difference between disseminated quantities, from which are eliminated the clock offsets of both receivers. Now we may calculate the single difference to the two satellites in point B:

$$\nabla^{ST} \rho_B^{(0)} = \nabla^{ST} p_B - \nabla^{ST} dp_B,$$

a purely geometric quantity between the location of the point and the locations in space of the satellites S and T *computed from the orbital data*, all known to the user. From these, the unknown location of point B may be computed, without any influence from the satellites' orbit errors or clock offsets. Three difference observations, i.e., four satellites, are sufficient.

Differential GPS positioning and navigation ("DGPS") is commonly used to refer to precisely this real-time method using the *modulations* of the GPS signal, the C/A and P codes.

12.4.2 Kinematic (real-time) positioning ("RTK")

If we use carrier-phase measurements in real time, we speak of RTK, real-time kinematic, measurement. With this technique, precision is much better, even if only over short distances. Unlike DGPS, this is an *incremental* method: the measurement must start from a *point with known co-ordinates*, and preferably also close on a known point — just in case. The co-ordinates of the points measured inbetween are obtained in relation to these known points.

See figure 12.7. When using *double differences*, between two satellites and the base station and rover, the only unknowns are the rover's three co-ordinates X, Y and Z, the location of the rover in three-dimensional space $\mathbb{R}^3$. All the possible locations of the rover that are compatible with the double-difference observation form now a *bundle of hyperboloids of revolution*. The distance separating the hyperboloids corresponds to one wavelength. In the figure, the hyperboloids are drawn in cross section as

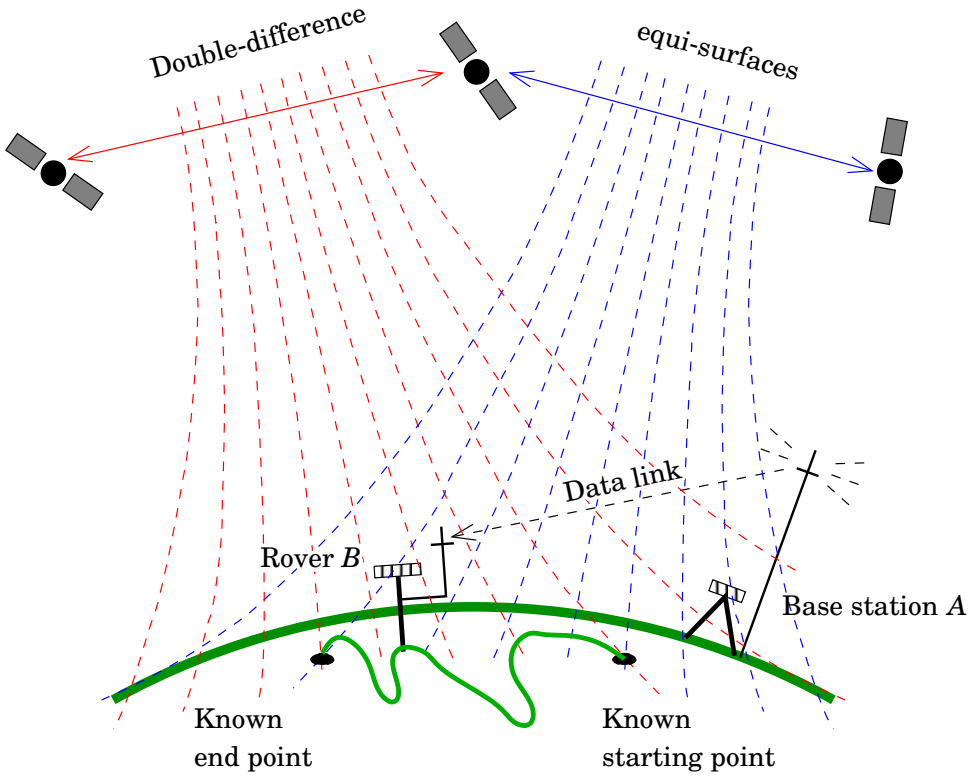


Figure 12.7. Principle of operation of the RTK method. Compare with the Decca figure 11.1!

□

curves of different colours, corresponding to two different satellite pairs. As always with GPS measurement, the minimum number of satellites is four.

On known point A is installed a permanent (or semi-permanent) GPS reference or base station. Let the momentaneous location of the rover be B. The double differences are

$$\nabla^{ST}\Delta_{AB}P = P_A^S - P_B^S - P_A^T + P_B^T .$$

Substituting equation 11.2 on page 273 into this yields

$$\nabla^{ST}\Delta_{AB}P = \nabla^{ST}\Delta_{AB}\rho + \nabla^{ST}\Delta_{AB}\left(D_{\text{ion}} + D_{\text{trop}}\right) + \lambda \cdot \nabla^{ST}\Delta_{AB}N .$$

Let us forget the atmospheric terms for a moment:

$$\nabla^{ST}\Delta_{AB}P = \nabla^{ST}\Delta_{AB}\rho + \lambda \cdot \nabla^{ST}\Delta_{AB}N .$$

If the *rover*'s antenna is set up on a point with known co-ordinates[4] B, *all co-ordinates are known*:

- The co-ordinates of satellites S and T can be computed from the orbital ephemeris[5].

[4] Or the co-ordinates of the rover's starting point are *determined* before starting to move, the so-called on-the-fly method. See the next subsection.

[5] The accuracy of the ephemeris is not a limiting factor, as in the RTK technique, the distance AB is always short.

- The co-ordinates of stations A and B were assumed known.

Therefore the *geometric* double difference for that moment, $\nabla^{ST}\Delta_{AB}\rho(t_0)$, is *computable*. After that, also the *ambiguities*[6] may be resolved from the observations:

$$\nabla^{ST}\Delta_{AB}N = \frac{\nabla^{ST}\Delta_{AB}P(t_0) - \nabla^{ST}\Delta_{AB}\rho(t_0)}{\lambda}.$$

After this, we take the rover away from the known point and go measure. We move in the terrain and measure a suitable number of unknown points — but *in such a way that the connection to the satellites is not interrupted*. Then namely also the values of the ambiguities $\nabla\Delta N_{AB}^{ST}$ do not change, and from the measurements $\nabla^{ST}\Delta_{AB}P(t_P)$ at terrain point P we may straightforwardly calculate the *geometric* double difference

$$\nabla^{ST}\Delta_{AB}\rho(t_P) = \nabla^{ST}\Delta_{AB}P(t_P) - \lambda\cdot\nabla^{ST}\Delta_{AB}N\,.$$

This is the essence of *kinematic GPS positioning*. The correct place is obtained immediately, even with millimetres precision. Of course only in relation to the reference station, not absolutely. Therefore the precise geodetic determination of the location of the reference or base station is essential.

RTK works best over short distances, from hundreds of metres to tens of kilometres. The real-time nature requires use of a data communications link between base station and rover. The possible data communications solutions are in principle the same as in the case of DGPS.

12.4.3 On-the-fly (OTF) ambiguity resolution

Above it was assumed that the moving receiver, the rover, starts from a point with co-ordinates known to geodetic precision. Such a point may however be *created* also "on the fly", by staying on the starting point for so long, that the satellite geometry changes and the resolving, i.e., fixing, the ambiguities becomes successful. The link with the base station has to be open, and the signal connection with the satellites visible commonly with the base station must be uninterrupted. lennossa

For solving the location with GPS, generally four satellites are sufficient. The RTK technique, however, requires at least a fifth satellite, in order to resolve the ambiguities. Without redundancy, the ambiguities $\nabla^{ST}\Delta_{AB}N$ for every double-difference observation would be freely

[6] These values must thus be integers or close to integers, in which case they may be rounded, resulting in the "fix" solution. If they are not near integer values, they may *not* be rounded. Then one obtains the, generally weaker, "float" solution. This may happen if, e.g., there is an error in the given co-ordinates of A or B. Or if the effect of the atmosphere is too strong after all, or the distance AB too long.

chooseable, and one could compute, from the three values $\nabla^{ST}\Delta_{AB}\rho$ thus obtained, a completely fantasy vector solution $\mathbf{R}_{AB}$, without there being generated any contradictions!

And the more satellites are available, the quicker the ambiguity resolution will proceed. This is why people are so interested in devices that are able to use at the same time the signals from both the GPS satellites and the Russian GLONASS satellites.

Over short distances (less than a km) the number of satellites effectively doubles, because one can use the difference between the L_1 and L_2 frequencies, the "wide lane", the effective wavelength of which is 86 cm.

Today's RTK devices are able to intelligently use many known points around a measurement area. Before and after the survey, these points are visited, and the instrument forms, using the known and measured point locations, a local *transformation formula*. Using the formula, all measured points are transformed to the same system in which the known points have been given. This is a handy but also dangerous property: the accuracy of the transformed co-ordinates can not be better than the interior accuracy of this local system. If it is, e.g., the old system KKJ based on traditional measurement techniques (see subsection 3.2.1), one loses the major advantage of GPS surveying, its superior geometric accuracy!

12.4.4 Network-mode real-time services

In recent years, many *base-station networks* have been built, in order to offer differential GPS (DGPS) and real-time kinematic (RTK) support services. The main problem with using a single base station is, that the corrections disseminated by it are good close to the base station, but deteriorate quickly as the distance of the rover from the base station grows.

It is intuitively clear, that the corrections change *only slowly*, and almost linearly, with place. The correction for the satellite clock offset is even constant. The satellites are so high, that even over an area the size of Europe, the *geometry* is kind of the same in different parts of the continent. Also the atmosphere is usually rather similar everywhere within a small area. This opens up the possibility of *interpolation*, and the base-station networks do precisely this.

Realistic alternatives for disseminating correction signals:

1. geostationary communications satellites. Their advantage is the homogeneous coverage of large areas, their disadvantage, the satellites' low elevation angle at Finnish latitudes.
2. using the mobile Internet through the mobile telephony network. The Internet is not real time, but in practice fast enough for it not to matter. Advantages are
   (a) an easy way to *charge* for the service

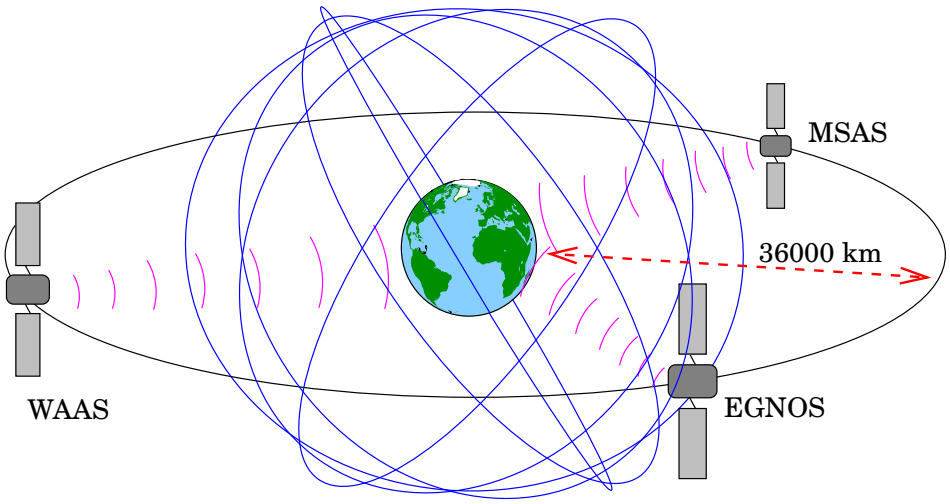


Figure 12.8. Satellite-Based Augmentation Systems (SBAS).

(b) the possibility to supply, to the location of the receiver, *tailored* corrections — the "virtual base station" idea

(c) the nowadays large communications bandwidth, i.e., data transfer capacity, at little cost.

Using the services in network mode requires also *software support*. The geometric aspect of interpolating the corrections is easy, the problem is formed by the *modelling* of the propagation delay by the *atmosphere*, mostly the ionosphere. In order to achieve a good accuracy, the base-station network must be sufficiently dense. This complex problem field is the subject of active research.

12.5 SBAS (satellite based augmentation systems)

For this type of system, also the name wide-area differential GPS (WADGPS) is used. Typical for the systems is, that they use geostationary satellites to disseminate differential corrections for GPS positioning. Globally, three intercompatible systems exist:

- WAAS (Wide Area Augmentation Service, USA)
- EGNOS (European Geostationary Navigation Overlay System)
- MSAS (Multi-functional Satellite Augmentation System, Japanese).

These are already in widespread use, though their development continues. The services are continent-wide and are based on the simultaneous use of many GPS base stations in network mode. In this way, accurate differential corrections are obtained for the area covered by the base stations. The signal structure and frequencies used are the same as for the GPS satellites, which makes it relatively easy to modify an existing GPS

receiver for SBAS use.

Part of the system is RIMS, the Remote Integrity Monitoring System, which sounds a warning if the quality of positioning cannot be guaranteed. This is important in *safety critical* applications — SoL, "Safety of Life" — like positioning aircraft during approach and landing.

12.6 Real-time satellite positioning support services in Finland

12.6.1 The Finnish @FOKUS DGPS service

This differential GPS service, operated by Indagon Oy[7], may be used all over Finland. Precision varies with distance to the nearest base station, the promised precision is 0.6 – 2 m. The network comprises 18 base stations all over Finland. The DGPS corrections conform to the generally used standard RTCM-SC104.

The corrections are disseminated over the Internet according to the NTRIP[8] standard. The name of the system is @FOKUS.

The DGNSS service of the Finnish Transport Agency This service, operated by the Finnish Transport Agency (formerly by the Finnish Maritime Administration) sends correction messages by radio on long wavelengths, frequencies 287.5 – 314.5 kHz. The service covers the sea areas of the Baltic Sea, and in Finland also the lake area of Saimaa. The user community consists of seafarers. The service is free of charge. https://goo.gl/0vodzr.

12.6.2 VRSnet, or GNSSnet, or Trimnet

This is the real-time kinematic (RTK) network service maintained by Geotrim Oy, originally planned for use by the Finnish National Land Survey, at address http://vrsnet.fi. The technology is called network RTK, i.e., virtual reference station RTK, and it is based on generating, for every user, a computational "virtual" base station close to him, for which are generated correction data in the RTCM-SC104 format, which can be used by devices of any type.

The network covers all of Finland with over a hundred base stations (situation 2016) and supports also the use of GLONASS. The corrections are disseminated commercially over the mobile Internet: a cluster of servers is located at Geotrim headquarters in Vantaa, to which the user can log in over the Internet. The corrections obtained are tailored for

[7] http://indagon.com/fi/project/fokus/.

[8] "Networked Transport of RTCM via Internet Protocol", http://en.wikipedia.org/wiki/NetworkedTransportofRTCMviaInternetProtocol.

each user; precision obtainable is of order centimetre in the horizontal plane, a bit poorer vertically.

12.6.3 SmartNet

SmartNet, operated and maintained by Leica Oy, is a real-time kinematic (RTK) base station network, http://www.smartnet-eu.com/. The technology is otherwise similar to that of VRSnet. In Finland there are currently well over a hundred base stations.

12.6.4 The experimental service of the National Land Survey

This service of the Finnish National Land Survey's Geospatial Research Institute, which is at the moment experimental and free of charge but which requires registration, offers both differential GNSS and network RTK. The base stations are the twenty new FinnRef stations (http://euref-fin.fgi.fi/fgi/en/positioning-service).

12.6.5 Archive data service

These network RTK service providers, in Finland and abroad, usually supply under agreement also data in the so-called RINEX[9] format for post-processing. Typically, the data is thinned out, from a data rate of one Hz — one measurement event per second, collecting data from each visible satellite — to, e.g., 1/30 Hz, i.e., one measurement event every 30 seconds.

12.6.6 The GDGPS system

A global real-time DGPS service using the Internet as its distribution channel. The system was developed and is being operated by NASA's Jet Propulsion Laboratory. http://www.gdgps.net/. As part of the service, it offers on-line Precise Point Positioning (PPP) using uploaded RINEX files.

Self-test questions

1. What is the "common mode" error assumption?

Exercise 12 – 1: Geodetic GPS positioning

Here, we will determine a precise vector between two stations by processing "in the cloud". GNSS data from many hundreds of continuously operating stations are available on-line, as are services for using these data for geodetic position computation. We are going to exploit this for a test computation.

[9] RINEX, Receiver Independent EXchange format, is a practical, device independent text format that even a human being can read. http://epic.awi.de/29985/1/Gur2007a.pdf. For all receiver brands, there is a conversion program.

1. Get the RINEX data files for the stations Sodankylä (SODA) and Metsähovi (METS), see figure 12.1. As sources you may use the web site of the FGI[10], which however requires registration, or the SOPAC web site[11] of the Scripps Institution of Oceanography in San Diego, California, or the EUREF data centre[12] at the Royal Observatory in Brussels, Belgium. Download data for some single day[13] — the same for both stations — of which the size should not exceed 5 MB.
2. The data downloaded may be *Hatanaka compressed*[14]. This is a compression technique specifically for RINEX data, exploiting the similarity of the data for successive epochs. You need to get the conversion program named CRX2RNX as a binary for your operating system, and convert the data to uncompressed RINEX. The latter is human-readable and even clear. Also the Hatanaka compressed file is human-readable, and the compression technique used is clearly visible.

   Remember that the data may also be compressed by a standard method, like zip or gzip or Unix compress (`*.Z`) ... which you first have to expand. On Windows, the utility 7-zip may be useful.

   Also, on Windows, you may run into the line-ending problem: Unix text files end their lines with a line-feed (LF) only, Windows uses carriage-return (CR) followed by line-feed (LF). The text editor Notepad++ may be useful.
3. Now, you can upload your RINEX files to the cloud. Two alternatives — choose one for this exercise:
   (a) AUSPOS, an Australian government service[15]: It tends to be a bit slow, often the result comes overnight, in the form of an extensive report. Publically available data from nearby IGS stations is used in the positioning computation, as shown in a map.
   (b) APPS, a U.S. government service of the Jet Propulsion Laboratory — the headquarters of the global International GNSS Service (IGS, subsection 11.10) , and yes, they too use the data from the IGS network (figure 11.23) as a reference in the computations. You have to upload METS and SODA separately,

[10] http://euref-fin.fgi.fi/fgi/fi/paikannuspalvelu/rinex-palvelu.

[11] http://sopac.ucsd.edu/dataBrowser.shtml.

[12] http://www.epncb.oma.be/_networkdata/stationlist.php.

[13] Suggestion: use your last year's birthday!

[14] http://sopac.ucsd.edu/hatanaka.shtml.

[15] http://www.ga.gov.au/scientific-topics/positioning-navigation/geodesy/auspos.

one file at a time. The location is
http://apps.gdgps.net/apps_file_upload.php.

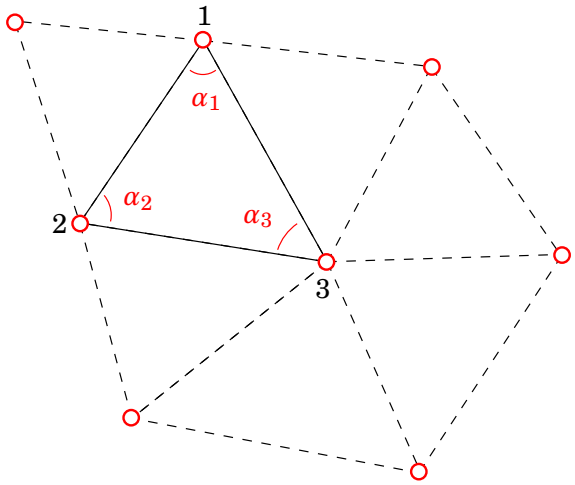
4. The AUSPOS results come by email.
   (a) Read the results carefully. What other stations were included in the computation, and in what reference frame is the result expressed?
   (b) Look up the geocentric Cartesian co-ordinate solution X, Y, Z.
   (c) The geodetic co-ordinates, and their precisions ("Positional Uncertainty").
   (d) Other interesting stuff. How is the tropospheric propagation delay modelled? Yes, they estimate dry and wet tropospheric zenith propagation delays as well as horizontal gradients!
   (e) Note the use of a geoid model, for obtaining heights over mean sea level. How good do you think it is?
   (f) Note ambiguity resolution.
5. The APPS results appear online.
   (a) For APPS, go to the summary file (`*.sum`) and look up the following things:
   (b) The geocentric Cartesian co-ordinate solution X, Y, Z, and their uncertainties ("sigmas"). How does this precision concept differ from that of AUSPOS?
   (c) The geodetic co-ordinates Lat, East_Lon and Height, and their sigmas. Compare the height sigma with the others.
   (d) The other interesting headers. How is the troposphere modelled here? Compare with AUSPOS.

13. Adjustment calculus in geodesy

13.1 Why adjustment?

In geodesy, as in science in general, we know that *all measurements are wrong*. Therefore we collect always more measurements than the strict minimum, so as to be able to judge at least somewhat realistically the uncertainties in the measurement results. The practice is called *redundancy*.

A good example is measurements in a *geodetic network*. In a triangulation network, the directions of the triangle points to other triangle points are measured.



From the direction measurements to neighbouring points made in points 1, 2, 3, *angles* $\alpha_1, \alpha_2, \alpha_3$ are calculated. For the angles, the *triangle condition*[1] applies:

$$\alpha_1 + \alpha_2 + \alpha_3 = 180^\circ.$$

The triangle condition makes the following checks possible:

[1] On the curved surface of the Earth, the sum is not exactly 180° but a little larger, the so-called *spherical excess*. This is an example of non-Euclidean geometry.

1. The measurements from which the angles $\alpha_1, \alpha_2, \alpha_3$ were calculated do not contain *gross* errors. E.g., if we obtain for the sum of angles $\underline{\alpha}_1 + \underline{\alpha}_2 + \underline{\alpha}_3 = 173^\circ.6742$, we may immediately conclude that the must be at least one gross error in the measurement set, because the accuracies of the instruments used are of order fraction of a degree.
2. The amount by which the sum obtained differs from the theoretical value 180° allows us to infer the precision of measurement. E.g., if we obtain as the sum $\underline{\alpha}_1 + \underline{\alpha}_2 + \underline{\alpha}_3 = 179^\circ.9958$, then the *closing error* $\underline{\Delta} = \underline{\alpha}_1 + \underline{\alpha}_2 + \underline{\alpha}_3 - 180^\circ = -0^\circ.0042$, and the inference is, that the precision of the measurement method used is of order several thousandths of a degree.

If we do excess measurements in this way, we need a method for removing the, small, contradictions between these measurements, and reconciling them with each other. A brute-force trick would be to just throw measured value $\underline{\alpha}_3$ away, and compute a replacement value, guaranteed to be compatible, of $\underline{\alpha}_3 = 180^\circ - \underline{\alpha}_1 - \underline{\alpha}_2$. However, one may justifiably ask, why $\underline{\alpha}_3$ rather than $\underline{\alpha}_1$ or $\underline{\alpha}_2$? Such arbitrariness is unacceptable. The right solution is *network adjustment*.

In the simple triangle case we divide the closing error *equally* among the angles: the adjusted angle values will be democratically

$$\widehat{\alpha}_1 = \underline{\alpha}_1 - \frac{\underline{\Delta}}{3},$$
$$\widehat{\alpha}_2 = \underline{\alpha}_2 - \frac{\underline{\Delta}}{3},$$
$$\widehat{\alpha}_3 = \underline{\alpha}_3 - \frac{\underline{\Delta}}{3},$$

after which $\widehat{\alpha}_1 + \widehat{\alpha}_2 + \widehat{\alpha}_3 = 180^\circ$ exactly.

If we however know that the angle α_3 was measured, e.g., *twice* (and the value $\underline{\alpha}_3$ is the average of these measurements) but the angles α_1 and α_2 only once using the same instrument, we may take this into account by adjusting in the following way:

$$\widehat{\alpha}_1 = \underline{\alpha}_1 - 2/5\,\underline{\Delta},$$
$$\widehat{\alpha}_2 = \underline{\alpha}_2 - 2/5\,\underline{\Delta},$$
$$\widehat{\alpha}_3 = \underline{\alpha}_3 - 1/5\,\underline{\Delta},$$

in which still $\underline{\Delta} = \underline{\alpha}_1 + \underline{\alpha}_2 + \underline{\alpha}_3 - 180^\circ$. In this case one speaks of *weighting* the measurements. The measured value $\underline{\alpha}_3$ is given a *double weight* — and thus a halved correction — compared to the measurements $\underline{\alpha}_1, \underline{\alpha}_2$.

A large weight means a small correction, and vice versa. See figure 13.1. The weight ratios are $1:1:2$, the correction ratios are the reverse $2:2:1$, and the sum of the correction ratios is 5. This is how the above correction coefficients ("weight coefficients") $2/5, 2/5, 1/5$ are obtained.

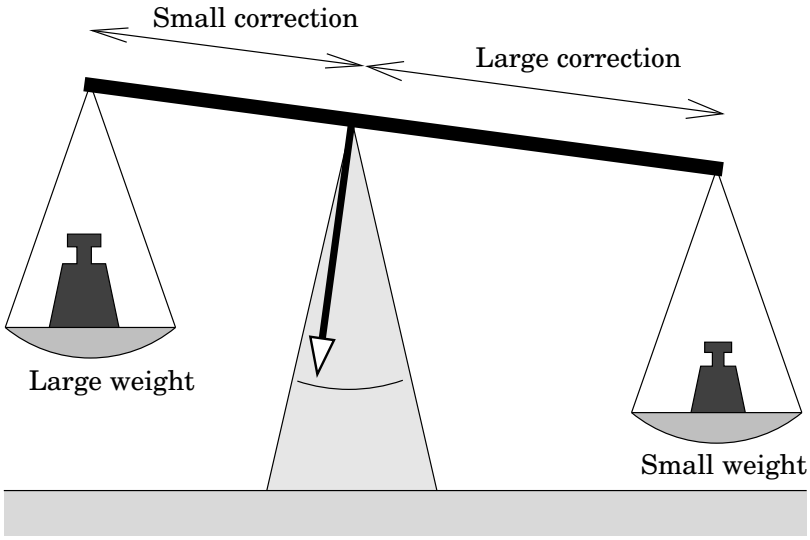


Figure 13.1. Metaphor: a large weight means a small correction and vice versa.

The adjustment of a realistic, complex triangulation network (or any geodetic network) is mathematically much more complicated, but this is the basic idea.

It is commonly assumed that the random measurement errors of the observations are distributed according to the bell curve named after C. F. Gauss (figure 2.5 on page 32), i.e., they are *normally distributed*. At least in that case, the theoretically best solution is given by least-squares adjustment. The simplest example of this is computing the *average*.

13.2 The average

Assume that the same quantity has been observed n times — so, we have a *stochastic quantity* $\underline{\ell}$ —, the observation values being $\ell_i, i = 1,\ldots,n$. The observations have a statistical expectancy μ, and they all have the same standard deviation or *mean error*[2] σ. The average of the observations is[3]

$$\overline{\ell} = \frac{1}{n}\left(\underline{\ell}_1 + \underline{\ell}_2 + \ldots + \underline{\ell}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\underline{\ell}_i.$$

One can show that, as an estimator of μ, this is the "best" possible linear combination of observations. One can also show that this linear combi-

[2]Expressed more theoretically, if the expectancy operator is $E\{\cdot\}$, we may write: $E\{\underline{\ell}\} = \mu$ and $\sigma^2 = E\left\{\left(\underline{\ell} - \mu\right)^2\right\}$.

[3]Here, the values $\underline{\ell}_i$ are written as *stochastic*, because the formation of the average may be *repeated*, to form different realizations of the stochastic quantity $\overline{\ell}$.

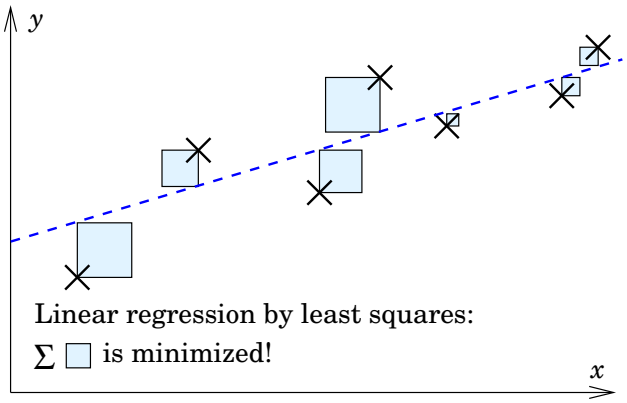


Figure 13.2. The idea of linear regression.

nation minimizes the square sum of the residuals[4]

$$\underline{v}_i = \overline{\ell} - \underline{\ell}_i,$$

i.e.,

$$\sum_{i=1}^{n} \underline{v}_i^2 = \underline{v}_1^2 + \underline{v}_2^2 + \ldots \underline{v}_n^2 = \min.$$

This property is the origin of the name "least-squares method".

We may also *estimate* the standard deviation or mean error of a single observation, σ, by the sample mean error

$$\widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \underline{v}_i^2}{n-1}}.$$

From this again follows as the quality, or uncertainty, measure for the average $\overline{\ell}$, *its* mean error (standard deviation) estimate, $\widehat{\sigma}_n = \frac{\widehat{\sigma}}{\sqrt{n}}$. This value, which describes the uncertainty of the average, thus becomes the smaller, the longer the series of measurement values is, i.e., the larger n.

If we know the mean error σ of a single observation *a priori*, we may also directly use the formula $\sigma_n = \frac{\sigma}{\sqrt{n}}$, the result of which is not an estimate but a computed value.

13.3 Linear regression

In linear regression, two parameters a and b are estimated when given are observations $\underline{y}$ that depend linearly on the argument x:

$$\underline{y}_i + \underline{v}_i = \widehat{a} + \widehat{b} x_i,$$

[4]The residual $\overline{\ell} - \underline{\ell}_i$ of an observation is *not* the same as the (opposite of the) *error* $\underline{\ell}_i - \mu$ of that observation! The residual is computable from the observations, the error is not.

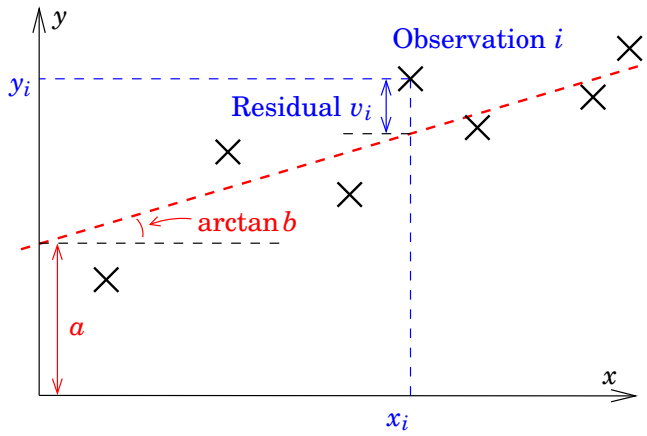


Figure 13.3. Linear regression, definitions of quantities.

in which $\underline{v}_i$ is again the *residual* of observation i. The parameters a and b describe a *straight line* that runs as well as possible — i.e., with as small residuals as possible — through the measured "point cloud" $\left(x_i, \underline{y}_i\right)$. See figures 13.2, 13.3.

Linear regression is a *least-squares method*, see figure 13.2. The sum of the squared residuals is minimized.

The least-squares solution is

$$\widehat{a} = \frac{\sum \underline{y} - \widehat{b} \sum x}{n}, \quad \widehat{b} = \frac{n \sum \left(x\underline{y}\right) - \sum x \sum \underline{y}}{n \sum \left(x^2\right) - \left(\sum x\right)^2},$$

in which is used the notation

$$\sum (\cdot) \stackrel{\text{def}}{=} \sum_{i=1}^{n} (\cdot),$$

a summation over all n points, or co-ordinate pairs, $(x_i, \underline{y}_i)$.

The "hat notation" $(\widehat{a}, \widehat{b})$ is an often used way to designate *estimators*.

13.4 The theory of least-squares adjustment

13.4.1 Calculating the solution from the observations

Presumably the first to use the method of least squares was C. F. Gauss, although Adrien-Marie Legendre[5] has also been claimed to be the inventor of the method. Gauss also carried out extensive geodetic network computations[6] in Hannover, using his method.

[5] Adrien-Marie Legendre (1752 – 1833) was a French mathematician, one of the 72 names on the Eiffel Tower.

[6] The numerical work of the network adjustment was carried out by an army of manual computers under Gauss' command. Back then, a "computer" was a human being!

In astronomy, the first application of the method was computing the orbits of asteroids and comets from observations. This, and the adjustment of geodetic networks, were special cases of situations that occur all the time in the life of an observer:

- We have available a body of observations, and we wish to compute from it certain interesting unknowns, in a way which
  - treats all observations as equally valuable
  - makes the deviations of the computed values for the unknowns of interest from their "true values" as small as possible.
- In addition, it would still be desirable that
  - any gross errors still hiding out in the observations are found and removed.

For this is used the *parametric form* of the least-squares adjustment method, which is based on the formation of *observation equations*.

13.4.2 The observation equations

Forming the observation equations is done as follows.

We write all observations as *linear*[7] functions of all unknowns:

$$\begin{aligned} \underline{\ell}_1 + \underline{v}_1 &= a_{11}\widehat{x}_1 + a_{12}\widehat{x}_2 + \dots a_{1m}\widehat{x}_m, \\ \underline{\ell}_2 + \underline{v}_2 &= a_{21}\widehat{x}_1 + a_{22}\widehat{x}_2 + \dots a_{2m}\widehat{x}_m, \\ &\dots \\ \underline{\ell}_n + \underline{v}_n &= a_{n1}\widehat{x}_1 + a_{n2}\widehat{x}_2 + \dots a_{nm}\widehat{x}_m, \end{aligned} \tag{13.1}$$

if there are n observations $\underline{\ell}_i$, n residuals $\underline{v}_i$, and m unknowns $\widehat{x}_j$.

The system of equations can be conveniently written in the form of a *matric equation*

$$\underline{\ell} + \underline{\mathrm{v}} = A\widehat{\mathrm{x}}, \tag{13.2}$$

in which

$$\underline{\ell} = \begin{bmatrix} \underline{\ell}_1 \\ \underline{\ell}_2 \\ \vdots \\ \underline{\ell}_n \end{bmatrix}, A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}, \widehat{\mathrm{x}} = \begin{bmatrix} \widehat{x}_1 \\ \widehat{x}_2 \\ \vdots \\ \widehat{x}_m \end{bmatrix}, \underline{\mathrm{v}} = \begin{bmatrix} \underline{v}_1 \\ \underline{v}_2 \\ \vdots \\ \underline{v}_n \end{bmatrix}.$$

The matrix A is *rectangular*, i.e., $n > m$, it is taller than it is wide. There are more observations, i.e., equations, than there are unknowns: *redundancy*. The observations $\underline{\ell}$, the unknowns $\widehat{\mathrm{x}}$ and the residuals $\underline{\mathrm{v}}$ are so-called *abstract vectors*, elements of an abstract vector space:

$$\begin{aligned} \underline{\ell}, \underline{\mathrm{v}} &\in \mathbb{R}^n, \\ \widehat{\mathrm{x}} &\in \mathbb{R}^m. \end{aligned}$$

[7] Often, the observation equations of real life are not linear. Then, usually *linearization* is possible. See section 13.6.

Often, one may assume that all observations $\underline{\ell}_i, i = 1, \ldots n$ have the same mean error σ (and a similarly shaped statistical distribution), and that the observations are *statistically independent of each other*, which also means that they do not intercorrelate. This assumption is referred to as *i.i.d.* — "*independent, identically distributed*".

13.4.3 The normal equations

From the matric equation 13.2 we compute the *least-squares solution* by first multiplying from the left with the matrix A^T, A's transpose:

$$A^\mathsf{T} A \widehat{\mathrm{x}} = A^\mathsf{T} \underline{\ell} + A^\mathsf{T} \underline{\mathrm{v}}.$$

Set[8]

$$A^\mathsf{T} \underline{\mathrm{v}} = 0,$$

yielding for the *least-squares solution* $\widehat{\mathrm{x}}$:

$$\left[A^\mathsf{T} A\right] \widehat{\mathrm{x}} = A^\mathsf{T} \underline{\ell}. \tag{13.3}$$

Now we have a system of m equations, and m unknowns in vector $\widehat{\mathrm{x}}$; the coefficient matrix $A^\mathsf{T} A$ is *square*.

13.4.4 Solving the normal equations

The solution, or *estimator*, is obtained, e.g., in the following way:

$$\widehat{\mathrm{x}} = \left[A^\mathsf{T} A\right]^{-1} A^\mathsf{T} \underline{\ell},$$

assuming that the matrix $A^\mathsf{T} A$ can actually be *inverted*, i.e., it is not singular.

The equations 13.3 are known as the *normal equations*. Here it is assumed that all observations have the same mean error σ (i.e., the same *variance* σ^2), and that they do not intercorrelate.

Of course, solving the system of equations by traditional means, without matrices, to find the elements of $\widehat{\mathrm{x}}$, i.e., the unknowns, is also a readily useable method: write the normal equations 13.3 in the following way, which is well suited to computer coding:

$$\sum_{i=1}^{m} \left[\sum_{j=1}^{n} a_{jk} a_{ji} \right] \widehat{x}_i = \left[\sum_{j=1}^{n} a_{jk} \underline{\ell}_j \right], k = 1, \ldots, m.$$

[8] If one writes $\eta \overset{\text{def}}{=} A\xi$, the following holds:

$$\left\langle \eta \cdot \underline{\mathrm{v}} \right\rangle = \left\langle (A\xi) \cdot \underline{\mathrm{v}} \right\rangle = \xi^\mathsf{T} A^\mathsf{T} \underline{\mathrm{v}} = 0$$

for an *arbitrary vector* ξ; we say that the sub-space of vectors $A\xi$ of the space of observables (the "solution space", spanned by the columns of A) is *perpendicular* upon the sub-space of residuals. This is where the name "normal equations" comes from.

This is a system of m linear equations in m unknowns $\widehat{x}_i$, for the solution of which numerical standard methods and software libraries are on offer. The greatest challenge is usually finding suitable observation equations 13.1 in a concrete measurement situation.

The solutions for the average and linear regression presented above, sections 13.2 on page 321 and 13.3 on page 322, are special cases of the general adjustment solution, as we shall show.

13.4.5 Assessing the precision

When one computes the least-squares solution with the equation

$$\widehat{\mathrm{x}} = \left[A^{\mathsf{T}}A\right]^{-1} A^{\mathsf{T}}\underline{\ell},$$

one can, with the *propagation law of variances*, also obtain immediately the precision of estimator $\widehat{\mathrm{x}}$.

Assume that the variance matrix of the observations $\underline{\ell}$ is $\mathrm{Var}\left\{\underline{\ell}\right\} = \sigma^2 I$, i.e., the observations do not correlate with each other and are all equally precise.

If we call

$$L \overset{\text{def}}{=} \left[A^{\mathsf{T}}A\right]^{-1} A^{\mathsf{T}},$$

we obtain, based on the linear dependence,

$$\begin{aligned}\mathrm{Var}\left\{\widehat{\mathrm{x}}\right\} = \Sigma_{\mathrm{xx}} &= L\,\mathrm{Var}\left\{\underline{\ell}\right\}L^{\mathsf{T}} = \\ &= \left[A^{\mathsf{T}}A\right]^{-1} A^{\mathsf{T}}\cdot\sigma^2 I\cdot A\left[A^{\mathsf{T}}A\right]^{-1} = \sigma^2\cdot\left[A^{\mathsf{T}}A\right]^{-1}.\end{aligned}$$

This interesting result tells us, that the matric quantity $\left[A^{\mathsf{T}}A\right]^{-1}$ represents the *propagation of the mean error* σ of the observations into the variances of the end result of the adjustment $\widehat{\mathrm{x}}$.

The matrix $P_{\mathrm{xx}} \overset{\text{def}}{=} \left[A^{\mathsf{T}}A\right]$ is called the *weight matrix* of the unknowns, and its inverse, the matrix $Q_{\mathrm{xx}} \overset{\text{def}}{=} \left[A^{\mathsf{T}}A\right]^{-1}$, is called the *weight-coefficient matrix* of the unknowns (Baarda, 1981).

13.5 Examples of the least-squares method

Both the average and the linear regression are nice examples of least-squares adjustment methods: practical and still relatively simple. In the following, we go through them step by step, showing how they are special cases of the general least-squares method.

13.5.1 The average as a least-squares method

Observe the same quantity x directly n times:

$$\begin{aligned} \underline{x}_1 + \underline{v}_1 &= \widehat{x} \\ \underline{x}_2 + \underline{v}_2 &= \widehat{x} \\ &\vdots \\ \underline{x}_n + \underline{v}_n &= \widehat{x} \end{aligned} \tag{13.4}$$

Here, $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ are individual observation values, $\widehat{x}$ is an estimator of the unknown quantity x, and $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_n$ are the residuals of the observations.

The secret of formulating a suitable adjustment method is: find the *standard form* of the system of observation equations,

$$\underline{\ell} + \underline{\mathrm{v}} = A\widehat{\mathrm{x}}.$$

In this case, success means choosing

$$\underline{\ell} = \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \\ \vdots \\ \underline{x}_n \end{bmatrix}, \quad \underline{\mathrm{v}} = \begin{bmatrix} \underline{v}_1 \\ \underline{v}_2 \\ \vdots \\ \underline{v}_n \end{bmatrix}, \quad \widehat{\mathrm{x}} = \begin{bmatrix} \widehat{x} \end{bmatrix}, \quad \text{and} \quad A = \left.\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}\right\} n \text{ times.}$$

Verify that this really agrees with equations 13.4.

The normal equations are

$$A^{\mathsf{T}} A\widehat{\mathrm{x}} = A^{\mathsf{T}} \underline{\ell}$$

in which the "normal matrix"

$$A^{\mathsf{T}} A = \overbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}}^{n \text{ times}} \left.\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}\right\} n \text{ times} = n$$

and the "right-hand side vector"

$$A^{\mathsf{T}} \underline{\ell} = \overbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}}^{n \text{ times}} \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \\ \vdots \\ \underline{x}_n \end{bmatrix} = \sum_{i=1}^{n} \underline{x}_i.$$

So, the solution is immediately obtained as follows:

$$n\widehat{\mathrm{x}} = \sum_{i=1}^{n} \underline{x}_i \quad \Longrightarrow \quad \widehat{\mathrm{x}} = \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i,$$

the classical formula for the average!

Also deriving the *precision*, or mean error, of the solution is not too hard. Let the mean errors of all the observations $\underline{x}_i, i = 1,\ldots,n$, be the same σ (a precondition for using the formula for the average), then the *variance matrix* of the whole observation vector $\underline{\ell}$ is

$$\mathrm{Var}\left\{\underline{\ell}\right\} = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix},$$

a matrix of size $n \times n$. According to subsection 13.4.5 the variance matrix of the vector of unknowns[9] is

$$\mathrm{Var}\left\{\widehat{\mathrm{x}}\right\} = \sigma^2 \left[A^\top A\right]^{-1} = \frac{1}{n}\sigma^2,$$

and the mean error of the unknown is the square root of this:

$$\sigma_{\widehat{\mathrm{x}}} = \frac{\sigma}{\sqrt{n}},$$

just as it should be.

13.5.2 Linear regression as a least-squares method

Write the observation equations of linear regression into the form

$$\underline{y}_i + \underline{v}_i = \widehat{a} + \widehat{b}x_i,$$

in which every pair $\left(x_i, \underline{y}_i\right), i = 1,\ldots,n$ is one observation, and the coefficients $\widehat{a}, \widehat{b}$ of the straight line to be fitted are to be determined.

If we call

$$\underline{\ell} = \begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_n \end{bmatrix}, \widehat{\mathrm{x}} = \begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix}$$

and again (essential!) the design matrix

$$A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix},$$

one may write this system into the form of observation equations[10]

$$\underline{\ell} + \underline{\mathrm{v}} = A\widehat{\mathrm{x}}.$$

---

[9] The "vector" $\widehat{\mathrm{x}}$ has here only one element, as has also the "matrix" $\mathrm{Var}\left\{\widehat{\mathrm{x}}\right\}$.

[10] The use of the letter x may be confusing here, and it has nothing to do with the x_i!

In the normal equations

$$A^{\mathsf{T}}A\widehat{\mathrm{x}} = A^{\mathsf{T}}\underline{\ell}$$

the normal matrix is

$$A^{\mathsf{T}}A = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix},$$

and

$$A^{\mathsf{T}}\underline{\ell} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \underline{y}_i \\ \sum_{i=1}^{n} x_i \underline{y}_i \end{bmatrix},$$

the right-hand side vector. From this, the solution is obtained by inversion of the above 2×2 matrix. However, this may also be done simply by elimination and back substitution. The system of equations is

$$\begin{aligned} n \cdot \widehat{a} + \left(\sum_{i=1}^{n} x_i \right) \cdot \widehat{b} &= \sum_{i=1}^{n} \underline{y}_i, \\ \left(\sum_{i=1}^{n} x_i \right) \cdot \widehat{a} + \left(\sum_{i=1}^{n} x_i^2 \right) \cdot \widehat{b} &= \sum_{i=1}^{n} x_i \underline{y}_i. \end{aligned}$$

Subtract the first equation from the second after multiplication by the factor $\frac{1}{n}\sum_{i=1}^{n} x_i$, yielding

$$\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right) \widehat{b} = \sum_{i=1}^{n} x_i \underline{y}_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \underline{y}_i,$$

from which

$$\widehat{b} = \frac{n \sum_{i=1}^{n} x_i \underline{y}_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \underline{y}_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}.$$

Now $\widehat{a}$ is obtained by back substitution:

$$\widehat{a} = \frac{1}{n} \left(\sum_{i=1}^{n} \underline{y}_i - \left(\sum_{i=1}^{n} x_i \right) \cdot \widehat{b} \right).$$

The expressions found are equivalent to the formulas given in section 13.3.

13.5.3 Computation example of linear regression

Given are the measurement results in table 13.1. A graphical presentation of the measurements is given in figure 13.4.

Table 13.1. Measurement results for linear regression.

i	1	2	3	4	5	$\sum$
x_i	1.51	2.44	3.34	4.41	5.05	16.75
$\underline{y}_i$	2.32	3.12	3.57	3.93	4.15	17.09
x_i^2	2.28	5.95	11.16	19.45	25.50	64.34
$x_i \underline{y}_i$	3.50	7.61	11.92	17.33	20.96	61.32

In the table are precomputed the needed sums

$$\sum_{i=1}^{n} x_i, \quad \sum_{i=1}^{n} \underline{y}_i, \quad \sum_{i=1}^{n} x_i^2 \quad \text{and} \quad \sum_{i=1}^{n} x_i \underline{y}_i.$$

From this

$$\widehat{b} = \frac{5 \cdot 61.32 - 16.75 \cdot 17.09}{5 \cdot 64.34 - 16.75^2} = \frac{20.3425}{41.1375} = 0.495,$$
$$\widehat{a} = \frac{1}{5}\left(17.09 - 16.75 \cdot \widehat{b}\right) = 1.76.$$

This solution has been plotted into figure 13.4.

The normal matrix, or *weight matrix of the unknowns*, is obtained as follows:

$$P_{xx} = A^{\mathsf{T}} A = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = \begin{bmatrix} 5.00 & 16.75 \\ 16.75 & 64.34 \end{bmatrix}$$

and its inverse is

$$Q_{xx} = \left[A^{\mathsf{T}} A\right]^{-1} = \begin{bmatrix} 1.564 & -0.4072 \\ -0.4072 & 0.1215 \end{bmatrix},$$

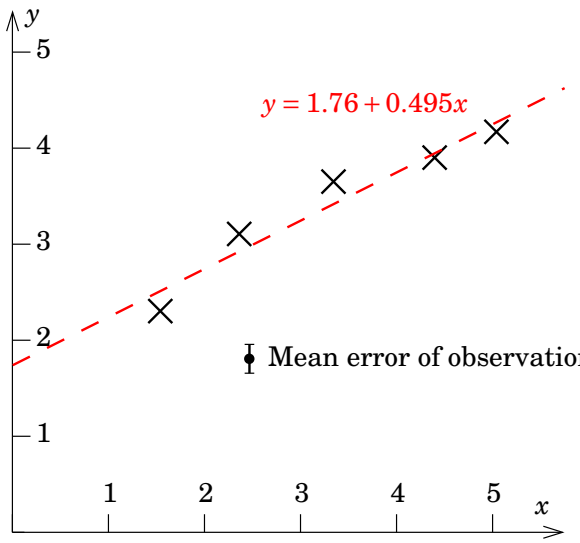


Figure 13.4. Computation example of linear regression.

the so-called *weight-coefficient matrix*. From this is obtained the variance matrix

$$\Sigma_{xx} = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix} = \sigma^2 Q_{xx} = \sigma^2 \begin{bmatrix} 1.564 & -0.4072 \\ -0.4072 & 0.1215 \end{bmatrix},$$

from which are obtained the mean errors σ_a and σ_b as the square roots of the diagonal elements:

$$\widehat{a} = 1.76 \pm 1.25\sigma,$$
$$\widehat{b} = 0.495 \pm 0.349\sigma.$$

Here, σ is the *a priori* (given in advance) mean error of a single $\underline{y}$ value, the *mean error of unit weight*.

13.6 Linearization of geodetic models

In geodesy, as more generally in science, there often are relationships between two quantities that behave *non-linearly*. Examples of this are the relationship between observables and unknowns, or between co-ordinates in two different co-ordinate frames.

Many theories, however, like the least-squares adjustment method, are based on *linear* equations, the mathematics of which is essentially simpler. Also the law of propagation of errors (variances) applies only to linear relationships between quantities.

In practice, often a formally non-linear relationship, e.g., between point co-ordinates and the measured direction to a point, is *almost* linear within the uncertainty area of the point location. Geodetic measurements are exceptionally precise: the location uncertainty of a point may be mere centimetres when the distance between points can be hundreds of metres or kilometres. In that case one may look, instead of at the original quantities, at the relationship between *small differences* of these quantities — which will be almost linear. The matter will be demonstrated using a Taylor series expansion.

13.6.1 The scalar case

Generally if we have two quantities between which exists a functional relationship

$$y = f(x),$$

we may linearize by choosing an *approximate value* x_0 and *expanding* the function into a *Taylor series* in the neighbourhood of the approximate value. We obtain

$$y = f(x_0) + \left.\frac{df}{dx}\right|_{x=x_0} (x - x_0) + \ldots,$$

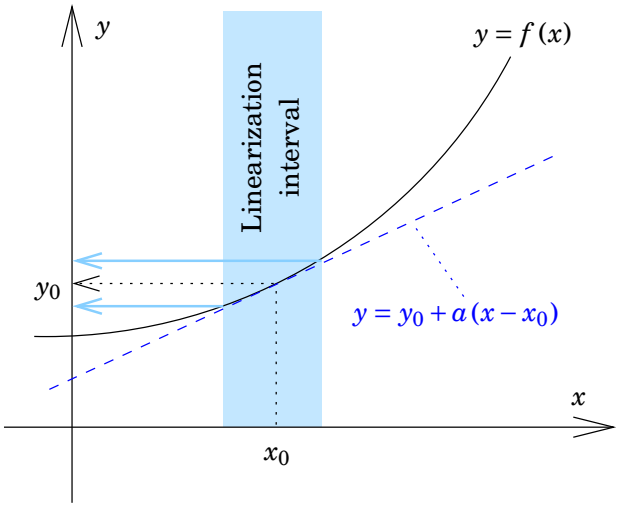


Figure 13.5. One-dimensional mapping and linearization.

i.e.,

$$y - y_0 \approx a\,(x - x_0), \tag{13.5}$$

in which $y_0 \stackrel{\text{def}}{=} f\,(x_0)$ and $a = \left.\frac{df}{dx}\right|_{x=x_0}$. This may be written into the form

$$\Delta y = a \Delta x,$$

which is often abbreviated to

$$y = ax,$$

as long as it is remembered that x, y are "linearized quantities", i.e., difference quantities reckoned from the approximate values x_0, y_0.

13.6.2 The vector case

If there are two vectorial quantities,

$$\mathrm{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} \in \mathbb{R}^n \quad \text{and} \quad \mathrm{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_m \end{bmatrix} \in \mathbb{R}^m,$$

between which there exists a functional relationship

$$\mathrm{y} = F\,(\mathrm{x}) = F\,(x_1, x_2, \ldots, x_n),$$

i.e.,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} F_1\,(\mathrm{x}) \\ F_2\,(\mathrm{x}) \\ \vdots \\ F_m\,(\mathrm{x}) \end{bmatrix} = \begin{bmatrix} F_1\,(x_1, x_2, \ldots, x_n) \\ F_2\,(x_1, x_2, \ldots, x_n) \\ \vdots \\ F_m\,(x_1, x_2, \ldots, x_n) \end{bmatrix},$$

it gets complicated. Also in this case, we can choose a *vector* of approximate values

$$\mathrm{x}_0 = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ \dots \\ x_n^{(0)} \end{bmatrix},$$

and a corresponding vector of approximate values $\mathrm{y}_0 \overset{\text{def}}{=} F(\mathrm{x}_0)$, after which again

$$\begin{aligned} \mathrm{y} = \mathrm{y}_0 &+ \left.\frac{\partial F(x_1, x_2, \dots, x_n)}{\partial x_1}\right|_{\mathrm{x}=\mathrm{x}_0} \left(x_1 - x_1^{(0)}\right) + \\ &+ \left.\frac{\partial F(x_1, x_2, \dots, x_n)}{\partial x_2}\right|_{\mathrm{x}=\mathrm{x}_0} \left(x_2 - x_2^{(0)}\right) + \cdots \\ &\cdots + \left.\frac{\partial F(x_1, x_2, \dots, x_n)}{\partial x_n}\right|_{\mathrm{x}=\mathrm{x}_0} \left(x_n - x_n^{(0)}\right), \end{aligned}$$

i.e.,

$$\begin{aligned} y_i = y_i^{(0)} &+ \left.\frac{\partial F_i(x_1, x_2, \dots, x_n)}{\partial x_1}\right|_{\mathrm{x}=\mathrm{x}_0} \left(x_1 - x_1^{(0)}\right) + \\ &+ \left.\frac{\partial F_i(x_1, x_2, \dots, x_n)}{\partial x_2}\right|_{\mathrm{x}=\mathrm{x}_0} \left(x_2 - x_2^{(0)}\right) + \cdots \\ &\cdots + \left.\frac{\partial F_i(x_1, x_2, \dots, x_n)}{\partial x_n}\right|_{\mathrm{x}=\mathrm{x}_0} \left(x_n - x_n^{(0)}\right), \\ &i = 1, \dots, m. \end{aligned}$$

In this equation there are m different rows, and in every row there are n different (linear) terms. As a summary of this system of equations, we write the following matric equation:

$$\mathrm{y} = \mathrm{y}_0 + A(\mathrm{x} - \mathrm{x}_0) + \dots,$$

in which the matrix A is

$$A = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} & \dots & \frac{\partial F_1}{\partial x_n} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} & \dots & \frac{\partial F_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial x_1} & \frac{\partial F_m}{\partial x_2} & \dots & \frac{\partial F_m}{\partial x_n} \end{bmatrix}.$$

This matrix is the so-called *matrix of Jacobi*[11] of the vector mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ between the two abstract vector spaces $\mathbb{R}^n$ and $\mathbb{R}^m$. The matrix describes *locally*, i.e., in the neighbourhood of point $\mathrm{x} = \mathrm{x}_0$, in

[11] Carl Gustav Jacob Jacobi, 1804 – 1851, was a Jewish German mathematician, University of Königsberg 1827 – 1842.

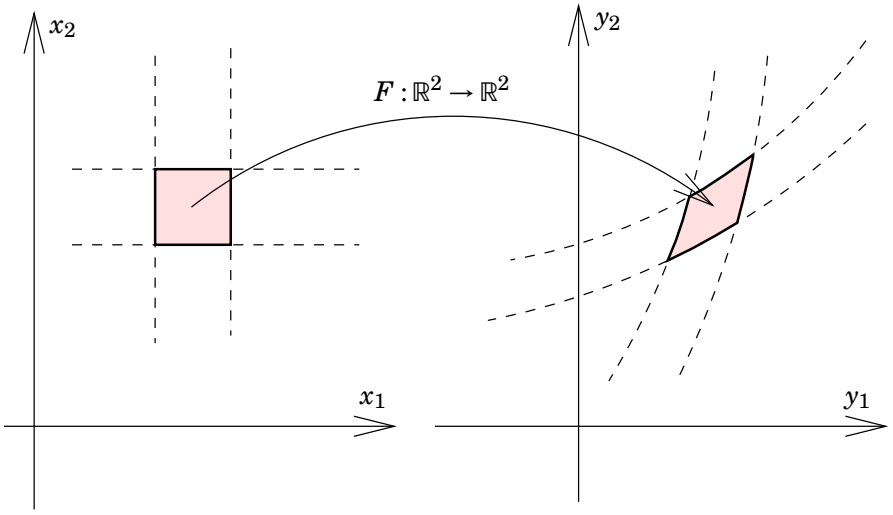


Figure 13.6. A two-dimensional mapping.

which way small "perturbances" in the x vector propagate into the y vector:

$$\Delta \mathrm{y} \stackrel{\text{def}}{=} \mathrm{y} - \mathrm{y}_0 \approx A\,(\mathrm{x} - \mathrm{x}_0) = A\,\Delta \mathrm{x},$$

when defining $\Delta \mathrm{x} = \mathrm{x} - \mathrm{x}_0$ and $\Delta \mathrm{y} = \mathrm{y} - \mathrm{y}_0$. So, between the difference quantities $\Delta \mathrm{x}, \Delta \mathrm{y}$, the mapping is locally linear. This is referred to as *linearization*.

In the general case, $m \neq n$. In the special case where $m = n$, we may think that the mapping F has an inverse mapping $G = F^{-1}$, for which

$$\mathrm{x} = G\,(\mathrm{y}).$$

Locally, in the neighbourhood of the approximate point x_0, we may say of this: if the matrix A is *singular*, i.e., its determinant $\det A = 0$, this means that the mapping F does not locally (i.e., in a suitably small neightbourhood of point x_0) have an inverse mapping. This again means that there may be many (in fact, infinitely many) different values x having the same image $\mathrm{y} = F\,(\mathrm{x})$. On the other hand, if $\det A \neq 0$, such an inverse mapping does (in a sufficiently small neighbourhood of approximate point x_0) exist.

Interpretation: the determinant $\det A$ describes in which way *volumes* are mapped under the vector mapping F: if, e.g., $n = m = 2$, it describes how the surface area of a small square in the $\mathbb{R}^n$ space is mapped to the surface area of a parallellogram in the $\mathbb{R}^m$ space, i.e., the ratio of these two surface areas. If $n = m = 3$, it similarly describes the ratio of the volumes of a small cube in $\mathbb{R}^n$ space and of its corresponding parallellepiped in $\mathbb{R}^m$ space. If the ratio is zero, then apparently the square is "squeezed" to a line segment, and the cube to a parallellogram, and the mapping is thus singular.

13.6.3 Linearization of observation equations

Let us consider as an example the functional relationship between unknowns x and observables ℓ, which in a realistic observation geometry is rarely linear. We have to *linearize*: let the non-linear observation equations be

$$\underline{\ell} + \underline{\mathrm{v}} = F\left(\widehat{\mathrm{x}}\right), \tag{13.6}$$

in which $F(\cdot)$ is a multidimensional, usually non-linear, "observation function".

The models are linearized again by expanding them into a Taylor series around roughly estimated solution co-ordinates ("approximate values"), and using only the first-degree terms of the series. If the approximate co-ordinates used are not good enough, we end up computing the solution *iteratively*.

Choose the *approximate values* x_0 and compatibly ℓ_0 so that the following applies for them:

$$\ell_0 = F(\mathrm{x}_0). \tag{13.7}$$

So, if the number of unknowns is m and the number of observations n:

$$\ell_i^{(0)} = F_i\left(x_1^{(0)}, x_2^{(0)}, \ldots x_{m-1}^{(0)}, x_m^{(0)}\right), i = 1 \ldots n.$$

This is subtracted from equation 13.6, and we do a Taylor series expansion, retaining only the linear terms:

$$\begin{aligned}\left(\underline{\ell}_i - \ell_i^{(0)}\right) + \underline{v}_i = F_i\left(\widehat{x}_1, \widehat{x}_2, \ldots \widehat{x}_m\right) - F_i\left(x_1^{(0)}, x_2^{(0)}, \ldots x_m^{(0)}\right) \approx \\ \approx \sum_{j=1}^{m} \left.\frac{\partial F_i}{\partial x_j}\right|_{x_j = x_j^{(0)}} \left(\widehat{x}_j - x_j^{(0)}\right) + \ldots\end{aligned}$$

Call

$$A_{ij} \stackrel{\text{def}}{=} \left.\frac{\partial F_i}{\partial x_j}\right|_{x_j = x_j^{(0)}}, \quad i = 1 \ldots n, j = 1 \ldots m, \tag{13.8}$$

the elements of the so-called *second-order design matrix*. The matrix itself is then

$$A = \left.\begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} & \cdots & \frac{\partial F_1}{\partial x_m} \\ \frac{\partial F_2}{\partial x_1} & \frac{\partial F_2}{\partial x_2} & \cdots & \frac{\partial F_2}{\partial x_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_n}{\partial x_1} & \frac{\partial F_n}{\partial x_2} & \cdots & \frac{\partial F_n}{\partial x_m} \end{bmatrix}\right|_{x_1 = x_1^{(0)}, x_2 = x_2^{(0)}, x_m = x_m^{(0)}, \ldots}.$$

If we call

$$\begin{aligned}\left(\underline{\ell} - F(\mathrm{x}_0)\right) &\stackrel{\text{def}}{=} \Delta\underline{\ell} \\ \left(\widehat{\mathrm{x}} - \mathrm{x}_0\right) &\stackrel{\text{def}}{=} \Delta\widehat{\mathrm{x}}\end{aligned}$$

("replacement" or "linearized" observables and unknowns), we obtain for the linearized observation equations:

$$\Delta\underline{\ell} + \underline{\mathrm{v}} = A\Delta\widehat{\mathrm{x}}. \tag{13.9}$$

The least-squares solution to be computed here *minimizes* the square sum of residuals $\underline{\mathrm{v}}^{\mathsf{T}} Q_{\ell\ell}^{-1} \underline{\mathrm{v}}$, wherefore it is called the least-squares method. The matrix $Q_{\ell\ell}$ is the *weight-coefficient matrix*[12] characterizing the precision and possible statistical interdependence (correlation) of the observations, see section 13.7.

From equation 13.9, often the Δ are left off for the sake of simplicity. The Δ quantities are typically *much smaller* than the "whole" quantities. Therefore the numerics work well even if the elements of the A would not be exact.

Equation 13.7 however must always be calculated *precisely*, i.e., with a sufficient number of decimals.

13.7 Propagation of variances

If the stochastic quantity $\underline{y}$ is a linear function of the stochastic quantity $\underline{x}$, i.e.,

$$\underline{y} = L\underline{x},$$

we may also write

$$\sigma_y = L\sigma_x,$$

in which σ_x, σ_y are the *mean errors* of quantities x and y. We may also write

$$E\left\{\underline{y}\right\} = E\left\{L\underline{x}\right\} = LE\left\{\underline{x}\right\}$$

("propagation of expectancies"), in which $E\{\cdot\}$ is the expectancy operator. The expectancy is a *linear* operator.

If we define the variance as follows:

$$\mathrm{Var}\left\{\underline{x}\right\} = \sigma_x^2 \overset{\mathrm{def}}{=} E\left\{\left(\underline{x} - E\left\{\underline{x}\right\}\right)^2\right\},$$

it follows that

$$\sigma_y^2 = L^2\sigma_x^2.$$

This is the *law of propagation of variances* for simple stochastic quantities.

If the stochastic quantities

$$\underline{\mathrm{x}} = \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \\ \cdots \\ \underline{x}_n \end{bmatrix}, \quad \underline{\mathrm{y}} = \begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \cdots \\ \underline{y}_m \end{bmatrix}$$

[12] Generally it is written $\Sigma_{\ell\ell} = \sigma^2 Q_{\ell\ell}$, in which $\Sigma_{\ell\ell}$ is the *variance matrix* and σ the so-called *mean error of unit weight*.

have several components — i.e., they are abstract vectorial quantities —, it holds, if $\underline{y} = L\underline{x}$, that

$$E\left\{\underline{y}\right\} = LE\left\{\underline{x}\right\} \tag{13.10}$$

and

$$\mathrm{Var}\left\{\underline{y}\right\} = L\,\mathrm{Var}\left\{\underline{x}\right\}L^{\mathsf{T}}, \tag{13.11}$$

in which now L and the variances are *matrices*.

$$L = \begin{bmatrix} L_{11} & L_{12} & \cdots & L_{1n} \\ L_{21} & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \vdots \\ L_{m1} & \cdots & \cdots & L_{mn} \end{bmatrix}$$

is an $m \times n$ size matrix,

$$\mathrm{Var}\left\{\underline{x}\right\} = \Sigma_{xx} = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1x_2} & \cdots & \sigma_{x_1x_n} \\ \sigma_{x_2x_1} & \sigma_{x_2}^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{x_nx_1} & \cdots & \cdots & \sigma_{x_n}^2 \end{bmatrix}$$

is a square matrix of size $n \times n$, and

$$\mathrm{Var}\left\{\underline{y}\right\} = \Sigma_{yy} = \begin{bmatrix} \sigma_{y_1}^2 & \sigma_{y_1y_2} & \cdots & \sigma_{y_1y_m} \\ \sigma_{y_2y_1} & \sigma_{y_2}^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{y_my_1} & \cdots & \cdots & \sigma_{y_m}^2 \end{bmatrix}$$

is a square matrix of size $m \times m$. Here, the variances are

$$\sigma_{x_i}^2 = \mathrm{Var}\left\{\underline{x}_i\right\} = E\left\{\left(\underline{x}_i - E\left\{\underline{x}_i\right\}\right)^2\right\}$$

and the *covariances*

$$\sigma_{x_ix_j} = \mathrm{Cov}\left\{\underline{x}_i, \underline{x}_j\right\} = E\left\{\left(\underline{x}_i - E\left\{\underline{x}_i\right\}\right)\left(\underline{x}_j - E\left\{\underline{x}_j\right\}\right)\right\},$$

and similarly for the components of $\underline{y}$.

Equation 13.11 is called the *general law of propagation of variances*. It is a generalization of equation 2.4, already derived in subsection 2.4.6 on page 36, for an arbitrary number of variables. The linearity property expressed by formula 13.10 may be obtained by *linearization* if needed, as discussed earlier.

13.8 The forward geodetic problem as an example of error propagation

As an application of the law of propagation of variances, we may look at the *forward geodetic problem*, in which the known uncertainties of direction and distance measurement are *propagated* into co-ordinate uncertainties of an unknown point.

The forward geodetic problem: given measurements s, α as well as the co-ordinates x_P, y_P of the starting point P, determine the co-ordinates of the unknown point

$$x = x_P + s \cdot \cos\alpha,$$
$$y = y_P + s \cdot \sin\alpha.$$

The problem is solved in the following way. Take approximate values $s^{(0)}, \alpha^{(0)}$:

$$s = s^{(0)} + \Delta s,$$
$$\alpha = \alpha^{(0)} + \Delta\alpha,$$

and write a Taylor series expansion:

$$x \approx x_P + s^{(0)}\cos\alpha^{(0)} + \Delta s\cos\alpha^{(0)} + s^{(0)}\frac{\partial\cos\alpha}{\partial\alpha}\Delta\alpha =$$
$$= \overbrace{x_P + s^{(0)}\cos\alpha^{(0)}}^{x^{(0)}} + \overbrace{\begin{bmatrix} \cos\alpha^{(0)} & -s^{(0)}\sin\alpha^{(0)} \end{bmatrix}\begin{bmatrix} \Delta s \\ \Delta\alpha \end{bmatrix}}^{\Delta x},$$

and in the same way

$$y \approx y^{(0)} + \overbrace{\begin{bmatrix} \sin\alpha^{(0)} & s^{(0)}\cos\alpha^{(0)} \end{bmatrix}\begin{bmatrix} \Delta s \\ \Delta\alpha \end{bmatrix}}^{\Delta y}.$$

Now we have, dropping, but remembering, the approximation labels (0), and turning the vectors x and y into random or *stochastic* quantities:

$$\underline{\mathrm{y}} \overset{\mathrm{def}}{=} \begin{bmatrix} \Delta\underline{x} \\ \Delta\underline{y} \end{bmatrix}, \quad \underline{\mathrm{x}} \overset{\mathrm{def}}{=} \begin{bmatrix} \Delta\underline{s} \\ \Delta\underline{\alpha} \end{bmatrix},$$

and

$$L = \begin{bmatrix} \cos\alpha & -s\sin\alpha \\ \sin\alpha & s\cos\alpha \end{bmatrix}$$

as well as

$$\mathrm{Var}\{\underline{\mathrm{x}}\} = \begin{bmatrix} \sigma_s^2 & 0 \\ 0 & \sigma_\alpha^2 \end{bmatrix},$$

and the above equations may now be written as

$$\underline{\mathrm{y}} = L\underline{\mathrm{x}}.$$

The variance matrix is

$$\mathrm{Var}\{\underline{\mathrm{y}}\} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} = L\,\mathrm{Var}\{\underline{\mathrm{x}}\}L^\mathsf{T} =$$
$$= \begin{bmatrix} \cos\alpha & -s\sin\alpha \\ \sin\alpha & s\cos\alpha \end{bmatrix}\begin{bmatrix} \sigma_s^2 & 0 \\ 0 & \sigma_\alpha^2 \end{bmatrix}\begin{bmatrix} \cos\alpha & \sin\alpha \\ -s\sin\alpha & s\cos\alpha \end{bmatrix},$$

and the elements are

$$\sigma_x^2 = \sigma_s^2 \cos^2\alpha + \sigma_\alpha^2 s^2 \sin^2\alpha,$$
$$\sigma_y^2 = \sigma_s^2 \sin^2\alpha + \sigma_\alpha^2 s^2 \cos^2\alpha,$$
$$\sigma_{xy} = \cos\alpha \sin\alpha \left(\sigma_s^2 - s^2\sigma_\alpha^2\right),$$

computed using the law 13.11 of propagation of variances[13].

By substitution still

$$\cos\alpha = \frac{x - x_P}{s},$$
$$\sin\alpha = \frac{y - y_P}{s},$$

an alternative form is obtained:

$$\sigma_x^2 = \mathrm{Var}\left\{\Delta\underline{x}\right\} = \left(\frac{x - x_P}{s}\right)^2 \sigma_s^2 + (y - y_P)^2 \sigma_\alpha^2,$$
$$\sigma_y^2 = \mathrm{Var}\left\{\Delta\underline{y}\right\} = \left(\frac{y - y_P}{s}\right)^2 \sigma_s^2 + (x - x_P)^2 \sigma_\alpha^2,$$
$$\sigma_{xy} = \mathrm{Cov}\left\{\Delta\underline{x}, \Delta\underline{y}\right\} = \left[\left(\frac{\sigma_s}{s}\right)^2 - \sigma_\alpha^2\right](x - x_P)(y - y_P), \qquad (13.12)$$

in which

$$x - x_P = s\cos\alpha,$$
$$y - y_P = s\sin\alpha.$$

This is how mean errors of the observations σ_s, σ_α are converted to co-ordinate mean errors σ_x, σ_y. The precision is affected both by the observational precisions σ_s, σ_α and by the geometry α, s.

The *error ellipse* is the statistical area of uncertainty of the location solution (x, y) of the point. This can be used in statistical testing.

As a measure of point precision, there exists a suitable quantity that is independent of the directions of the co-ordinate axes. The error ellipse is really a *visual representation of the variance matrix*. The variance

[13] If we express the variance of direction α in gon, we may substitute into all formulas

$$\sigma_\alpha^2 = \left(\frac{\sigma_\alpha\,[\mathrm{g}]}{\rho}\right)^2,$$

in which ρ is the size of a radian in the degree unit in question, in this case $\rho = 63.661977236758$. Similarly if one uses seconds of arc: then

$$\sigma_\alpha^2 = \left(\frac{\sigma_\alpha\,[''\,]}{\rho}\right)^2,$$

in which now $\rho = 57.295779513 \times 60 \times 60 = 206264.806247$.

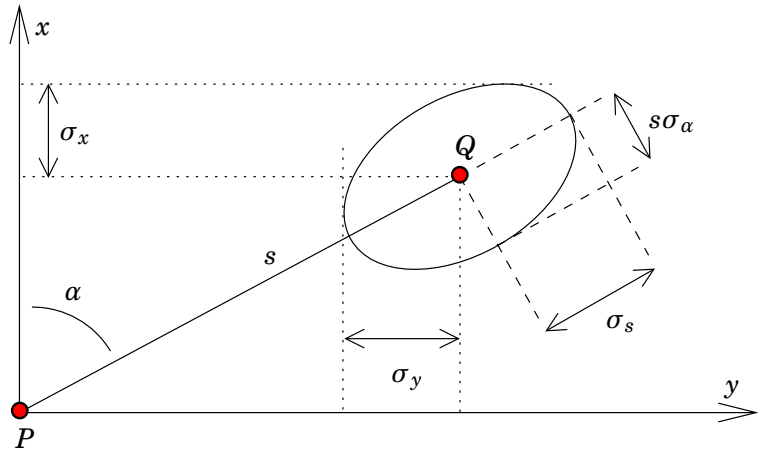


Figure 13.7. Quantities related to the error ellipse.

matrix of the co-ordinates of point P or of the co-ordinate differences of two points P,Q, i.e., (x,y), may be written as

$$\Sigma = \mathrm{Var}\begin{bmatrix} \underline{x} \\ \underline{y} \end{bmatrix} = \begin{bmatrix} \mathrm{Var}\{\underline{x}\} & \mathrm{Cov}\{\underline{x},\underline{y}\} \\ \mathrm{Cov}\{\underline{x},\underline{y}\} & \mathrm{Var}\{\underline{y}\} \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}.$$

The *invariants* of this matrix are its eigenvalues and -vectors: the solutions of the eigenvalue problem $(\Sigma - \lambda I)\mathbf{x} = 0$ i.e., $(\lambda_i, \mathbf{x}_i), i = 1,2$. If we rotate the co-ordinate axes so, that they are oriented along the main axes of the ellipse, we obtain

$$\Sigma = \begin{bmatrix} s^2\sigma_\alpha^2 & 0 \\ 0 & \sigma_s^2 \end{bmatrix}$$

and obviously $\lambda_1 = s^2\sigma_\alpha^2$ and $\lambda_2 = \sigma_s^2$.

Remark — more generally, one solves the determinant equation

$$\det\begin{bmatrix} \Sigma_{11} - \lambda & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} - \lambda \end{bmatrix} = 0.$$

This is the so-called *characteristic polynomial*:

$$(\Sigma_{11} - \lambda)(\Sigma_{22} - \lambda) - \Sigma_{12}^2 = 0,$$

i.e.,

$$\lambda^2 - (\Sigma_{11} + \Sigma_{22})\lambda + \left(\Sigma_{11}\Sigma_{22} - \Sigma_{12}^2\right) = 0,$$

a standard-issue quadratic equation. The eigenvalues are now

$$\begin{aligned} \lambda_{1,2} &= \frac{1}{2}\left[\Sigma_{11} + \Sigma_{22} \pm \sqrt{(\Sigma_{11} + \Sigma_{22})^2 - 4\left(\Sigma_{11}\Sigma_{22} - \Sigma_{12}^2\right)}\right] = \\ &= \frac{1}{2}\left[\Sigma_{11} + \Sigma_{22} \pm \sqrt{(\Sigma_{11} - \Sigma_{22})^2 + 4\Sigma_{12}^2}\right] = \\ &= \frac{1}{2}(\Sigma_{11} + \Sigma_{22}) \pm \sqrt{\left[\frac{1}{2}(\Sigma_{11} - \Sigma_{22})\right]^2 + \Sigma_{12}^2}, \end{aligned}$$

and the semi-major and semi-minor axes of the error ellipse are $\sqrt{\lambda_1}, \sqrt{\lambda_2}$.

Also the *directions* of the axes may be determined: look at the linear combination of co-ordinates

$$z(\theta) = x \sin\theta + y \cos\theta,$$

which is a function of the direction angle θ.

The propagation law of variances yields now

$$\mathrm{Var}\left\{\underline{z}\right\} = \Sigma_{11} \sin^2\theta + \Sigma_{22} \cos^2\theta + 2\Sigma_{12} \sin\theta \cos\theta.$$

The axes of the visual ellipse are *stationary values* of this function of θ,

$$\frac{d}{d\theta} \mathrm{Var}\left\{\underline{z}\right\} = 0.$$

By differentiation

$$2 \sin\theta \cos\theta (\Sigma_{11} - \Sigma_{22}) + 2\left(\cos^2\theta - \sin^2\theta\right)\Sigma_{12} = 0$$

or

$$\sin 2\theta (\Sigma_{11} - \Sigma_{22}) + 2 \cos 2\theta \cdot \Sigma_{12} = 0$$

and

$$\theta = \frac{1}{2} \arctan\left(-\frac{2\Sigma_{12}}{\Sigma_{11} - \Sigma_{22}}\right) + k \cdot 100\,\mathrm{gon} =$$

$$= \arctan\left(-\frac{\Sigma_{12}}{\Sigma_{12} + \sqrt{\left[\frac{1}{2}(\Sigma_{11} - \Sigma_{22})\right]^2 + \Sigma_{12}^2}}\right) + k \cdot 100\,\mathrm{gon},$$

using the *half-angle formula* for the arc tangent[14].

One obtains

$$\lambda_1 + \lambda_2 = \Sigma_{11} + \Sigma_{22} = \mathrm{Var}\left\{\underline{x}\right\} + \mathrm{Var}\left\{\underline{y}\right\} = \sigma_x^2 + \sigma_y^2 \tag{13.13}$$

and

$$\lambda_1 \lambda_2 = \Sigma_{11}\Sigma_{22} - \Sigma_{12}^2 = \det\Sigma = \sigma_x^2 \sigma_y^2 - \sigma_{xy}^2 \tag{13.14}$$

(in which σ_{xy}^2 is computed by formula 13.12). The quantities 13.13, 13.14 are *invariants* — i.e., always the same, no matter how the co-ordinate axes are oriented — and especially the quantity 13.13, which is called the *point variance* of point P, σ_P^2, is a suitable measure of point precision:

$$\sigma_P^2 = \sigma_x^2 + \sigma_y^2.$$

The *point mean error* σ_P is the square root of this point variance.

[14]This is how one avoids division by zero in the edge case $\Sigma_{11} = \Sigma_{22}$.

13.9 Observables and observation equations in practice

Here we present the observables of classical geodesy: measurements of horizontal directions, zenith angles and slant ranges, in the form of linearized observation equations. The GPS observables presented already above, equations 11.1 and 11.2, are special cases of slant-range measurement.

13.9.1 Slant-range measurement

Let the co-ordinates of instrument and signal be (x_1, y_1, z_1) and (x_2, y_2, z_2). Then, the functional model, equation 13.6, is

$$s = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}. \tag{13.15}$$

Assume first that point 1, co-ordinates $\begin{bmatrix} x_1 & y_1 & z_1 \end{bmatrix}^\mathsf{T}$, is known. Then, the abstract vector of unknowns x consists only of the co-ordinates of the second point, $\mathrm{x} = \begin{bmatrix} x_2 & y_2 & z_2 \end{bmatrix}^\mathsf{T}$. The abstract vector of observations ℓ of the observations has only one element, $\ell = [s]$, and the design matrix A (equation 13.8) consists of the partial derivatives of this one observable with respect to all unknowns x_2, y_2, z_2:

$$A = \begin{bmatrix} \dfrac{\partial s}{\partial x_2} & \dfrac{\partial s}{\partial y_2} & \dfrac{\partial s}{\partial z_2} \end{bmatrix}.$$

The elements are calculated by differentiating equation 13.15, e.g.,

$$\frac{\partial s}{\partial x_2} = \frac{x_2 - x_1}{s}$$

and so on. The end result is

$$A = \begin{bmatrix} \dfrac{x_2 - x_1}{s} & \dfrac{y_2 - y_1}{s} & \dfrac{z_2 - z_1}{s} \end{bmatrix}.$$

Note that the elements of the matrix A are the components of a unit vector! By using the horizontal direction angle (azimuth) α and the zenith angle ζ we may write

$$A = \begin{bmatrix} \cos\alpha_{12} \sin\zeta_{12} & \sin\alpha_{12} \sin\zeta_{12} & \cos\zeta_{12} \end{bmatrix},$$

or symbolically

$$A = \begin{bmatrix} \mathbf{e}_{12}^\mathsf{T} \end{bmatrix}$$

in which $\mathbf{e}_{12}$ is the *direction vector* between points 1 and 2, for which holds

$$\|\mathbf{e}_{12}\| = 1.$$

The general case, where both points are unknown, is discussed next. The vector of unknowns is formed

$$\mathrm{x} = \left[\begin{array}{ccc|ccc} x_1 & y_1 & z_1 & x_2 & y_2 & z_2 \end{array}\right]^\mathsf{T},$$

and the vector of observations is again $\ell = [s]$. The design matrix A, equation 13.8, is

$$A = \left[\begin{array}{ccc|ccc} \frac{\partial s}{\partial x_1} & \frac{\partial s}{\partial y_1} & \frac{\partial s}{\partial z_1} & \frac{\partial s}{\partial x_2} & \frac{\partial s}{\partial y_2} & \frac{\partial s}{\partial z_2} \end{array} \right].$$

Based on the previous, the end result is

$$\begin{aligned} A &= \left[\begin{array}{ccc|ccc} -\frac{x_2-x_1}{s} & -\frac{y_2-y_1}{s} & -\frac{z_2-z_1}{s} & \frac{x_2-x_1}{s} & \frac{y_2-y_1}{s} & \frac{z_2-z_1}{s} \end{array} \right] = \\ &= \left[\begin{array}{cc} \mathbf{e}_{21}^{\mathsf{T}} & \mathbf{e}_{12}^{\mathsf{T}} \end{array} \right] = \left[\begin{array}{cc} -\mathbf{e}_{12}^{\mathsf{T}} & \mathbf{e}_{12}^{\mathsf{T}} \end{array} \right]. \end{aligned}$$

The linearized observation equation is now (stochastic):

$$\left[\Delta \underline{s} + \underline{v}\right] = \left[\begin{array}{cc} -\mathbf{e}_{12}^{\mathsf{T}} & \mathbf{e}_{12}^{\mathsf{T}} \end{array} \right] \left[\begin{array}{c} \Delta\widehat{\mathbf{r}}_1 \\ \Delta\widehat{\mathbf{r}}_2 \end{array} \right],$$

in which

$$\begin{aligned} \mathbf{e}_{12} &= \left[\begin{array}{ccc} \frac{x_2-x_1}{s} & \frac{y_2-y_1}{s} & \frac{z_2-z_1}{s} \end{array} \right]^{\mathsf{T}}, \\ \Delta\widehat{\mathbf{r}}_i &= \left[\begin{array}{ccc} \Delta\widehat{x}_i & \Delta\widehat{y}_i & \Delta\widehat{z}_i \end{array} \right]^{\mathsf{T}}, i = 1,2, \\ \Delta\widehat{x}_i &= \widehat{x}_i - x_i^{(0)}, i = 1,2, \\ \Delta\underline{s} &\stackrel{\text{def}}{=} \underline{s} - s^{(0)}. \end{aligned}$$

Here, the superscript (0) identifies an approximate value.

For approximate values, the *functional model* applies exactly:

$$s^{(0)} = f\left(\mathrm{x}^{(0)}\right) = \sqrt{\left(x_2^{(0)} - x_1^{(0)}\right)^2 + \left(y_2^{(0)} - y_1^{(0)}\right)^2 + \left(z_2^{(0)} - z_1^{(0)}\right)^2}.$$

13.9.2 Azimuth measurement

If the azimuth or horizontal direction α is measured between unknown points 1 and 2:

$$\alpha = \alpha_{12} = \arctan\left(\frac{y_2 - y_1}{x_2 - x_1}\right) + k\pi,$$

then $\mathrm{x} = \left[\begin{array}{cc|cc} x_1 & y_1 & x_2 & y_2 \end{array} \right]^{\mathsf{T}}$.

The design matrix is obtained again using the *chain rule*, e.g., (abbreviating $x_{12} = x_2 - x_1$ etc.):

$$\begin{aligned} \frac{\partial \arctan\left(\frac{y_{12}}{x_{12}}\right)}{\partial x_1} &= \frac{\partial \arctan\left(\frac{y_{12}}{x_{12}}\right)}{\partial\left(\frac{y_{12}}{x_{12}}\right)} \frac{\partial\left(\frac{y_{12}}{x_{12}}\right)}{\partial x_{12}} \frac{\partial x_{12}}{\partial x_1} = \\ &= \frac{1}{1 + \left(\frac{x_{12}}{y_{12}}\right)^2} \cdot \left(-\frac{y_{12}}{x_{12}^2}\right) \cdot (-1) = \\ &= \frac{y_{12}}{x_{12}^2 + y_{12}^2} = \frac{y_{12}}{\rho^2} = -\frac{\sin\alpha_{21}}{\rho}. \end{aligned}$$

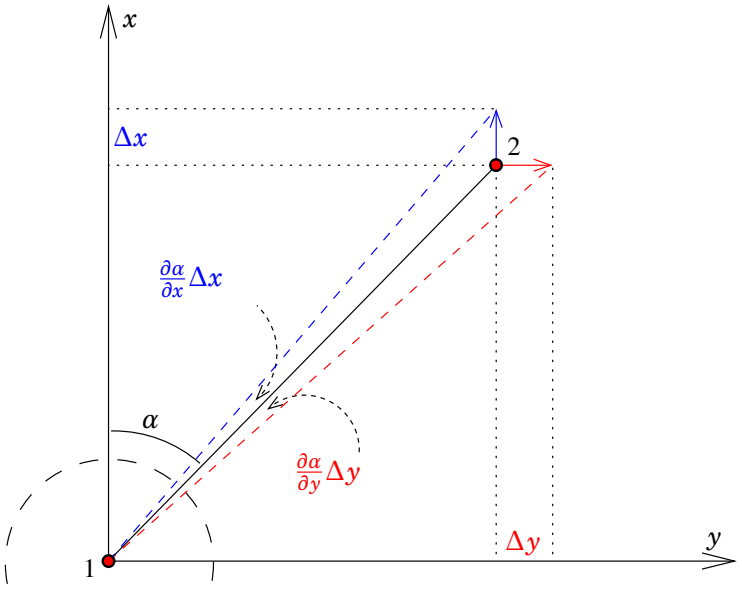


Figure 13.8. The geometry of azimuth measurement (seen from above) and the design matrix.

The whole matrix:

$$A = \left[-\frac{\sin\alpha_{21}}{\rho} \quad +\frac{\cos\alpha_{21}}{\rho} \;\middle|\; -\frac{\sin\alpha_{12}}{\rho} \quad +\frac{\cos\alpha_{12}}{\rho} \right] =$$
$$= \left[+\frac{y_2-y_1}{\rho^2} \quad -\frac{x_2-x_1}{\rho^2} \;\middle|\; -\frac{y_2-y_1}{\rho^2} \quad +\frac{x_2-x_1}{\rho^2} \right],$$

in which $\rho = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$, the distance between instrument and signal, projected onto the horizontal plane.

From this, the linearized observation equation is obtained:

$$\Delta\underline{\alpha} + \underline{v} = \left[+\frac{y_2-y_1}{\rho^2} \quad -\frac{x_2-x_1}{\rho^2} \;\middle|\; -\frac{y_2-y_1}{\rho^2} \quad +\frac{x_2-x_1}{\rho^2} \right] \left[\begin{array}{c} \Delta\widehat{x}_1 \\ \Delta\widehat{y}_1 \\ \hline \Delta\widehat{x}_2 \\ \Delta\widehat{y}_2 \end{array}\right].$$

13.9.3 Zenith-angle measurement

Measured is the zenith angle between points 1 and 2,

$$\zeta = \arccos\frac{z_2-z_1}{s} = \arctan\left(\frac{\pi}{2} - \frac{z_2-z_1}{\rho}\right),$$

in which again $\rho = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$.

Now $\mathrm{x} = \left[x_1 \quad y_1 \quad z_1 \;\middle|\; x_2 \quad y_2 \quad z_2 \right]^\mathsf{T}$ and

$$A = \left[\frac{\partial\zeta}{\partial\rho}\cdot\frac{\partial\rho}{\partial x_1} \quad \frac{\partial\zeta}{\partial\rho}\cdot\frac{\partial\rho}{\partial y_1} \quad \frac{\cos\zeta_{21}}{s} \;\middle|\; \frac{\partial\zeta}{\partial\rho}\cdot\frac{\partial\rho}{\partial x_2} \quad \frac{\partial\zeta}{\partial\rho}\cdot\frac{\partial\rho}{\partial y_2} \quad \frac{\cos\zeta_{12}}{s} \right] =$$
$$= \left[\frac{\cos\zeta_{21}}{\rho}(-\cos\alpha_{21}) \quad \frac{\cos\zeta_{21}}{\rho}\sin\alpha_{21} \quad \frac{\rho}{s^2} \;\middle|\; \frac{\cos\zeta_{12}}{\rho}(-\cos\alpha_{12}) \quad \frac{\cos\zeta_{12}}{\rho}\sin\alpha_{12} \quad -\frac{\rho}{s^2} \right] =$$
$$= \left[-\frac{x_2-x_1}{s^2\tan\zeta_{12}} \quad +\frac{y_2-y_1}{s^2\tan\zeta_{12}} \quad \frac{\rho}{s^2} \;\middle|\; +\frac{x_2-x_1}{s^2\tan\zeta_{12}} \quad -\frac{y_2-y_1}{s^2\tan\zeta_{12}} \quad -\frac{\rho}{s^2} \right].$$

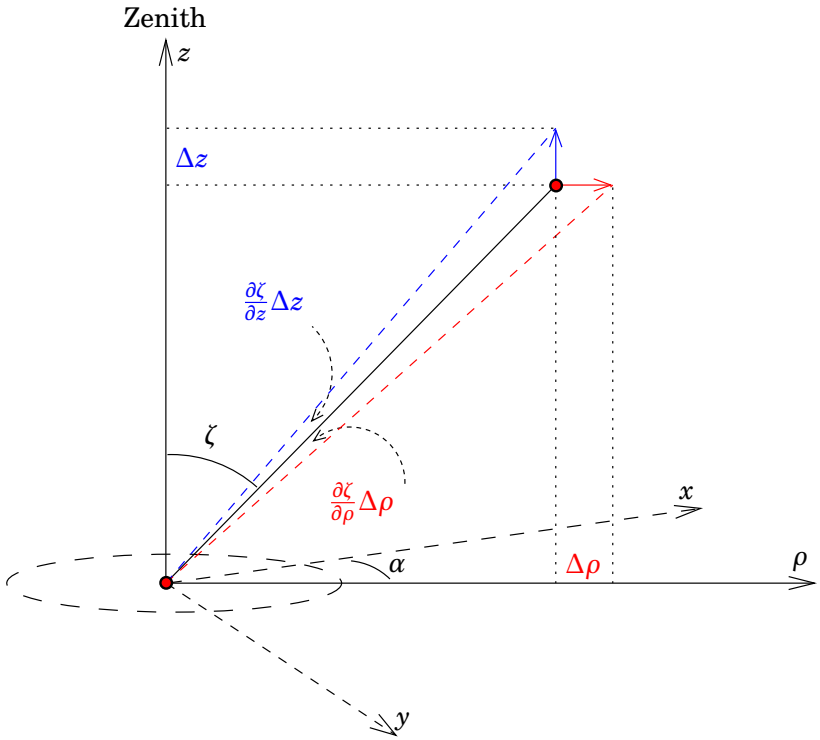


Figure 13.9. The geometry of zenith-angle measurement.

Here the partial derivative $\frac{\partial\zeta}{\partial\rho}$ has been computed in cylindrical co-ordinates (ρ,α,z). Then $\zeta=\zeta(\rho,z)$.

Now the linearized observation equation is

$$\Delta\underline{\zeta}+\underline{v}=\left[\begin{array}{ccc|ccc} -\frac{x_2-x_1}{s^2\tan\zeta_{12}} & +\frac{y_2-y_1}{s^2\tan\zeta_{12}} & \frac{\rho}{s^2} & +\frac{x_2-x_1}{s^2\tan\zeta_{12}} & -\frac{y_2-y_1}{s^2\tan\zeta_{12}} & -\frac{\rho}{s^2} \end{array}\right]\left[\begin{array}{c} \Delta\widehat{x}_1 \\ \Delta\widehat{y}_1 \\ \Delta\widehat{z}_1 \\ \hline \Delta\widehat{x}_2 \\ \Delta\widehat{y}_2 \\ \Delta\widehat{z}_2 \end{array}\right].$$

13.10 Helmert transformation in the plane

13.10.1 Theory

In the plane, if more than two points are given in both co-ordinate frames, it is possible to derive the unknown parameters of a Helmert transformation between the two frames. We start from equation 3.1 for a *single point*:

$$\left[\begin{array}{c} x \\ y \end{array}\right]=\left[\begin{array}{c} x_0 \\ y_0 \end{array}\right]+K\left[\begin{array}{cc} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{array}\right]\left[\begin{array}{c} u \\ v \end{array}\right].$$

Writing $K \stackrel{\text{def}}{=} 1+m$, with m the *scale distortion*, assumed small, this becomes, for also small rotation angles θ:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} \approx \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + (1+m) \begin{bmatrix} 1 & -\theta \\ \theta & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \approx$$
$$\approx \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 0 & -\theta \\ \theta & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} m & 0 \\ 0 & m \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix},$$

from which

$$\begin{bmatrix} x' \\ y' \end{bmatrix} - \begin{bmatrix} x \\ y \end{bmatrix} \approx \begin{bmatrix} x_0 - \theta y + mx \\ y_0 + \theta x + my \end{bmatrix},$$

which can be rearranged into

$$\begin{bmatrix} x'-x \\ y'-y \end{bmatrix} = \begin{bmatrix} 1 & 0 & x & -y \\ 0 & 1 & y & x \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ m \\ \theta \end{bmatrix}.$$

This is recognized as a a set of two observation equations, with the observation vector, the vector of unknowns, and the design matrix being, respectively,

$$\ell = \begin{bmatrix} x'-x \\ y'-y \end{bmatrix}, \quad \mathrm{x} = \begin{bmatrix} x_0 \\ y_0 \\ m \\ \theta \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 & x^{(0)} & -y^{(0} \\ 0 & 1 & y^{(0)} & x^{(0)} \end{bmatrix},$$

where $x^{(0)}, y^{(0)}$ are *approximate values* of the co-ordinates of the point in the (u,v) frame[15].

For multiple points given in the two frames, we obtain

$$\begin{bmatrix} x'_i - x_i \\ y'_i - y_i \end{bmatrix} = \begin{bmatrix} 1 & 0 & +x_i^{(0)} & -y_i^{(0)} \\ 0 & 1 & +y_i^{(0)} & +x_i^{(0)} \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \\ m \\ \theta \end{bmatrix},$$

with i the point number. We see that, if the number of available points n exceeds 2, there will be *redundancy*: more observation equations — $2n$ — than there are unknowns — 4. Written stochastically, we have now, with residuals

$$\begin{bmatrix} \underline{x}'_i - \underline{x}_i \\ \underline{y}'_i - \underline{y}_i \end{bmatrix} + \begin{bmatrix} \underline{v}_{2i-1} \\ \underline{v}_{2i} \end{bmatrix} = \begin{bmatrix} 1 & 0 & +x_i^{(0)} & -y_i^{(0)} \\ 0 & 1 & +y_i^{(0)} & +x_i^{(0)} \end{bmatrix} \begin{bmatrix} \widehat{x}_0 \\ \widehat{y}_0 \\ \widehat{m} \\ \widehat{\theta} \end{bmatrix}, i = 1, \ldots, n. \tag{13.16}$$

[15]Use of approximate values is OK here, as both m and θ are assumed small.

13.10.2 Propagation of uncertainty

With more than two common points between the (x', y') and (x, y) frames, it will be possible not only to derive the Helmert transformation parameters, but also how the point uncertainties propagate into those of the transformation parameters. See, e.g., (Kahmen and Faig, 1988, pages 253 – 255).

We start from the Helmert transformation equation rewritten as observation equations, 13.16. It is assumed that the co-ordinate values $x_i^{(0)}, y_i^{(0)}$ available for the points are good enough, and that the scale distortion m and the rotation angle θ of the axes are both small. This is typically the case as many, also local, co-ordinate reference frames are approximately of the correct scale and oriented correctly to the North.

The above observation equations have $2n$ rows, where n is the number of points: $i = 1, \ldots, n$. The design matrix A looks like

$$A = \begin{bmatrix} 1 & 0 & +x_i^{(0)} & -y_i^{(0)} \\ 0 & 1 & +y_i^{(0)} & +x_i^{(0)} \end{bmatrix}.$$

If we assume, that the co-ordinates of the given points are all equally precise, we may derive the *normal matrix*, or weight matrix of the unknowns, P_{xx} as follows ($\sum$ means summation over all given points, all sub- and superscripts dropped):

$$P_{\mathrm{xx}} = A^{\mathsf{T}} A = \begin{bmatrix} n & 0 & \sum x & -\sum y \\ 0 & n & \sum y & \sum x \\ \sum x & \sum y & \sum \left(x^2 + y^2\right) & 0 \\ -\sum y & \sum x & 0 & \sum \left(x^2 + y^2\right) \end{bmatrix}$$

The significance of this is that the variance matrix of the unknowns $\widehat{\mathrm{x}} = \begin{bmatrix} \widehat{x}_0 & \widehat{y}_0 & \widehat{m} & \widehat{\theta} \end{bmatrix}^{\mathsf{T}}$ is precisely

$$\mathrm{Var}\left\{\widehat{\mathrm{x}}\right\} \stackrel{\mathrm{def}}{=} \begin{bmatrix} \sigma_{x_0}^2 & \sigma_{x_0 y_0} & \sigma_{x_0 m} & \sigma_{x_0 \theta} \\ \sigma_{x_0 y_0} & \sigma_{x_0}^2 & \sigma_{y_0 m} & \sigma_{y_0 \theta} \\ \sigma_{x_0 m} & \sigma_{y_0 m} & \sigma_m^2 & \sigma_{m\theta} \\ \sigma_{x_0 \theta} & \sigma_{y_0 \theta} & \sigma_{m\theta} & \sigma_\theta^2 \end{bmatrix} = \sigma^2 \cdot Q_{\mathrm{xx}},$$

in which $Q_{\mathrm{xx}} = P_{\mathrm{xx}}^{-1}$ is the weight-coefficient matrix of the unknowns, σ is the mean error of unit weight, in this case, the assumed precision of one point co-ordinate. Here, $\sigma_{x_0}^2 = \mathrm{Var}\left\{\widehat{x}_0\right\}$ etc. are variances, $\sigma_{x_0 y_0} = \mathrm{Cov}\left\{\widehat{x}_0, \widehat{y}_0\right\}$ etc., covariances.

If now $\sum x = \sum y = 0$, in other words, x, y are *barycentric co-ordinates*, we obtain

$$P_{\mathrm{xx}} = \begin{bmatrix} n & 0 & 0 & 0 \\ 0 & n & 0 & 0 \\ 0 & 0 & \sum \left(x^2 + y^2\right) & 0 \\ 0 & 0 & 0 & \sum \left(x^2 + y^2\right) \end{bmatrix}$$

and, with $Q_{xx} = P_{xx}^{-1}$,

$$\mathrm{Var}\{\hat{\mathrm{x}}\} = \mathrm{Var}\left\{\begin{bmatrix}\hat{x}_0\\ \hat{y}_0\\ \hat{m}\\ \hat{\theta}\end{bmatrix}\right\} = \sigma^2 Q_{xx} = \sigma^2 \begin{bmatrix}\frac{1}{n} & 0 & 0 & 0\\ 0 & \frac{1}{n} & 0 & 0\\ 0 & 0 & \frac{1}{\sum(x^2+y^2)} & 0\\ 0 & 0 & 0 & \frac{1}{\sum(x^2+y^2)}\end{bmatrix}.$$

Now the variances of the transformation parameters are:

$$\mathrm{Var}\{\hat{x}_0\} = \sigma_{x_0}^2 = \sigma^2 (Q_{xx})_{11} = \frac{\sigma^2}{n}.$$
$$\mathrm{Var}\{\hat{y}_0\} = \sigma_{y_0}^2 = \sigma^2 (Q_{xx})_{22} = \frac{\sigma^2}{n}.$$
$$\mathrm{Var}\{\hat{m}\} = \sigma_m^2 = \sigma^2 (Q_{xx})_{33} = \frac{\sigma^2}{\sum\left(x^2+y^2\right)},$$
$$\mathrm{Var}\{\hat{\theta}\} = \sigma_\theta^2 = \sigma^2 (Q_{xx})_{44} = \frac{\sigma^2}{\sum\left(x^2+y^2\right)},$$

Also, the parameters do not statistically correlate with each other: the covariances between them vanish.

Now, the actual observation equations that you have to solve are 13.16, where the quantities $\underline{v}_{2i-1}, \underline{v}_{2i}$ are the important *residuals*, containing valuable quality-control information.

Exercise 13 – 1: Helmert transformation parameter estimation

We have a set of points given in two different co-ordinate frames[16]:

Point	x'	y'	x	y
10	6697976.388	27427023.03	6698108.117	3427192.039
36	6700867.976	27428566.410	6700999.695	3428735.464
714	6701212.704	27424871.393	6701344.461	3425040.439
717	6696502.735	27428579.172	6696634.432	3428748.182
17_vara	6697821.437	27424568.639	6697953.182	3424737.628
35_vara	6709919.416	27434433.390	6710051.148	3434602.545
2061	6694497.478	27432539.402	6694629.164	3432708.409
2062	6704170.468	27432163.151	6704302.189	3432332.245
2063	6703595.075	27426736.815	6703726.833	3426905.879

With this data, do the following operations:

1. Calculate (solve) the four Helmert transformation parameters[17] x_0, y_0, m, θ. For those who want to avoid coding work there is an

[16] The points are in the municipality of Porvoo, the co-ordinate frames are ETRS-GK27 and Porvoo's own old system.

[17] We leave off the stochastic underlines here, as these parameter values are just single realizations of the corresponding stochastic variables.

on-line Helmert transformation solver in Excel at http://www.engineeringsurveyor.com/software/transformations.xls. (Remember to choose the Helmert tab, not the Affine tab!). Courtesy of the Engineering Surveyor and Mark Adams.

2. Calculate the parameters x'_0, y'_0, m', θ' of the *inverse transformation* — i.e., swap columns x', y' with columns x, y .
3. Verify that the scales $K = m + 1, K' = m' + 1$ are each other's inverses, i.e., $K' = 1/K$, and that the θ rotation parameters each other's opposites, i.e., $\theta' = -\theta$.
4. Look at the *residuals*. How precise were the input co-ordinates of the points?
5. How many *observations* are there (Hint: the number of observations equals the number of residuals)? How many *unknowns*? What is the number of *degrees of freedom* b (the difference between the number of observations and the number of unknowns, the *redundancy*)?
6. Estimate $\widehat{\sigma} \stackrel{\text{def}}{=} \sqrt{\frac{\sum_{i=1}^{2n} \underline{v}_i^2}{b}}$, where b is the number of degrees of freedom, $\underline{v}_i$ the residuals, and $2n$ the number of observations. The estimated quantity σ is called the *mean error of unit weight*, the typical precision of a point co-ordinate[18].
7. Repeat the calculation after you changed one x value by adding one metre. Note the effect on the residuals. Would you be able to *identify* this co-ordinate as the one containing the gross error? Re-compute also σ. What do you see?

[18] More correctly, this only holds true if one of the co-ordinate sets given, (x', y') or (x, y), is assumed exact. Otherwise it is the typical precision of a co-ordinate *difference*.

14. Statistical methods in geodesy

14.1 The method of least squares

Explaining the methods of least squares is simplest if one assumes, that all observables are stochastic quantities that are *normally distributed* (figure 2.5), both individually and *together*: they form a so-called multinormal distribution. If the observables are statistically independent of each other — e.g., if they were produced by independent measurement processes — this is automatically the case.

Above was presented the method of least squares as a way to minimize the random errors in estimated quantities, e.g., point co-ordinates. In chapter 13 the *parametric adjustment method* was presented, in which observations are expressed as functions of the unknowns. The alternative, adjustment by conditions, is suitable, e.g., for computing traverses.

Here, the parametric method will be discussed in a bit more detail.

Let the *observations*, as a vector $\underline{\ell}$, be linear functions of the *unknowns* x[1]:

$$\begin{array}{ccccccc} \underline{\ell} & = & A & \mathrm{x} & + & \underline{\mathrm{n}} \\ \begin{bmatrix} \ \end{bmatrix} & = & \begin{bmatrix} \ \end{bmatrix} & \begin{bmatrix} \ \end{bmatrix} & + & \begin{bmatrix} \ \end{bmatrix} \\ [n] & = & \left[\ n \times m\ \right] & [m] & + & [n] \end{array}$$

Here, the observations, elements of the vector $\underline{\ell}$, are stochastic quantities. Assume that they are normally distributed around the "true" value of the observed quantity. Then, the elements of the vector of *observation errors* $\underline{\mathrm{n}}$ are also normally distributed[2].

[1] Here are presented graphically also the sizes of vectors and matrices. n is the number of observations, m the number of unknowns.

[2] Often, they are also assumed to be statistically independent from each other, meaning that their random variations happen independently of each other. However, neither the elements of solution vector $\widehat{\mathrm{x}}$ nor those of the vector of residuals $\underline{\mathrm{v}}$ will be statistically independent of each other.

In this, rather general, case we may compute the *least-squares solution* in the following simple way:

$$\begin{array}{ccccc} \widehat{\mathrm{x}} & = & \left[A^{\mathsf{T}}Q_{\ell\ell}^{-1}A\right]^{-1} & \left[A^{\mathsf{T}}Q_{\ell\ell}^{-1}\right] & \underline{\ell} \\ \left[\ \right] & = & \left[\ \right]^{-1} & \left[\ \ \right] & \left[\ \right] \\ [m] & = & [m\times m]^{-1} & [m\times n] & [n] \end{array} \tag{14.1}$$

in which $Q_{\ell\ell}$ is the *weight-coefficient matrix* of the observations ($[n\times n]$):

$$Q_{\ell\ell} = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ q_{n1} & q_{n2} & \cdots & q_{nn} \end{bmatrix}.$$

From this, the variance matrix of the observations is obtained as follows:

$$\mathrm{Var}\left\{\underline{\ell}\right\} \overset{\text{def}}{=} \Sigma_{\ell\ell} = \sigma^2 Q_{\ell\ell} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix},$$

i.e.,

$$\begin{aligned} \sigma_i^2 &= \mathrm{Var}\left\{\underline{\ell}_i\right\} = E\left\{\left(\underline{\ell}_i - E\left\{\underline{\ell}_i\right\}\right)^2\right\} = \sigma^2 q_{ii}, \\ \sigma_{ij} &= \mathrm{Cov}\left\{\underline{\ell}_i, \underline{\ell}_j\right\} = E\left\{\left(\underline{\ell}_i - E\left\{\underline{\ell}_i\right\}\right)\left(\underline{\ell}_j - E\left\{\underline{\ell}_j\right\}\right)\right\} = \sigma^2 q_{ij}. \end{aligned}$$

Here, σ is the variance of unit weight.

The variance matrix of the solution is obtained though propagation of variances: let

$$\widehat{\mathrm{x}} = L\underline{\ell}$$

in which

$$L \overset{\text{def}}{=} \left[A^{\mathsf{T}}Q_{\ell\ell}^{-1}A\right]^{-1}\left[A^{\mathsf{T}}Q_{\ell\ell}^{-1}\right].$$

Then

$$\begin{aligned} Q_{\mathrm{xx}} &= LQ_{\ell\ell}L^{\mathsf{T}} = \left[A^{\mathsf{T}}Q_{\ell\ell}^{-1}A\right]^{-1}A^{\mathsf{T}}Q_{\ell\ell}^{-1} \times Q_{\ell\ell} \times Q_{\ell\ell}^{-1}A\left[A^{\mathsf{T}}Q_{\ell\ell}^{-1}A\right]^{-1} = \\ &= \left[A^{\mathsf{T}}Q_{\ell\ell}^{-1}A\right]^{-1}, \end{aligned}$$

by suitable elimination. So, the variance matrix of the solution vector $\widehat{\mathrm{x}}$, $\Sigma_{\mathrm{xx}} = \sigma^2 Q_{\mathrm{xx}}$, is obtained in any case as a side product of computing the solution, equation 14.1!

14.2 The residuals from the adjustment

The least-squares estimators of the observations $\widehat{\ell}$ and unknowns $\widehat{\mathrm{x}}$ are *connected* to each other though the functional model

$$\widehat{\ell} = A\widehat{\mathrm{x}},$$

and from the original observations

$$\underline{\ell} = A\mathrm{x} + \underline{\mathrm{n}}$$

one computes the so-called *residuals*[3]:

$$\underline{\mathrm{v}} \overset{\text{def}}{=} \widehat{\ell} - \underline{\ell} = A\widehat{\mathrm{x}} - \underline{\ell} = A\left(\widehat{\mathrm{x}} - \mathrm{x}\right) - \underline{\mathrm{n}}.$$

Residuals are central in the quality control of geodetic network solutions.

1. The size of the residuals tells something about contradictions present in the network solution, possible gross errors, or even model errors.
   (a) The size of the residual of a certain observation can tell whether there may be a gross error hiding out in this observation.
   (b) The network must be *reliable*, i.e., there has to be redundancy, an overdetermination by the observational material. E.g., all kinds of closing errors offer possibilities for testing.
   Without redundancy, the residuals may be very small indeed, but this means nothing!

An often used form of the *observation equations* is

$$\underline{\ell} + \underline{\mathrm{v}} = A\widehat{\mathrm{x}}.$$

The residuals of a least-squares adjustment have four nice properties, here given without proof:

1. The quadratic form $\underline{\mathscr{E}} = \underline{\mathrm{v}}^{\mathsf{T}} Q_{\ell\ell}^{-1} \underline{\mathrm{v}}$, the square sum of residuals, is *minimized* — this is what the methods of least squares got its name from. In fact, the square root of this is the *norm* of the vector of residuals $\underline{\mathrm{v}}$, or its length, in the $Q_{\ell\ell}$ metric, which is thus minimized: $\left\|\underline{\mathrm{v}}\right\|_Q^2 = \underline{\mathscr{E}}$.
2. The variance $\Sigma_{\lambda\lambda}$ of an *arbitrary linear combination* $\underline{\lambda} = \Lambda\widehat{\mathrm{x}}$ of the unknowns $\widehat{\mathrm{x}}$ (and its mean error $\sqrt{\Sigma_{\lambda\lambda}}$) is minimized.
3. The adjusted observables $\widehat{\ell}$ and the residuals $\underline{\mathrm{v}}$ are mutually *orthogonal* in the $Q_{\ell\ell}$ metric, i.e., if as the scalar product is defined $\langle \mathrm{a} \cdot \mathrm{b} \rangle_Q \overset{\text{def}}{=} \mathrm{a}^{\mathsf{T}} Q_{\ell\ell}^{-1} \mathrm{b}$:

$$\left\langle \widehat{\ell} \cdot \underline{\mathrm{v}} \right\rangle_Q = \left\langle A\widehat{\mathrm{x}} \cdot \underline{\mathrm{v}} \right\rangle_Q = \left(A\widehat{\mathrm{x}}\right)^{\mathsf{T}} Q_{\ell\ell}^{-1} \underline{\mathrm{v}} = \widehat{\mathrm{x}}^{\mathsf{T}} A^{\mathsf{T}} Q_{\ell\ell}^{-1} \underline{\mathrm{v}} = 0,$$

[3]The vector $\underline{\mathrm{v}}$ of residuals is not the same as the vector of observation errors, or "noise", $\underline{\mathrm{n}}$! The residual is the difference between original observation and adjusted observation, i.e., a *correction*. However, not even an adjusted observation — or unknown — is the "truth". The truth is not knowable, and the values of the elements of the vector $\underline{\mathrm{n}}$, unlike the values of the elements of vector $\underline{\mathrm{v}}$, cannot be computed.

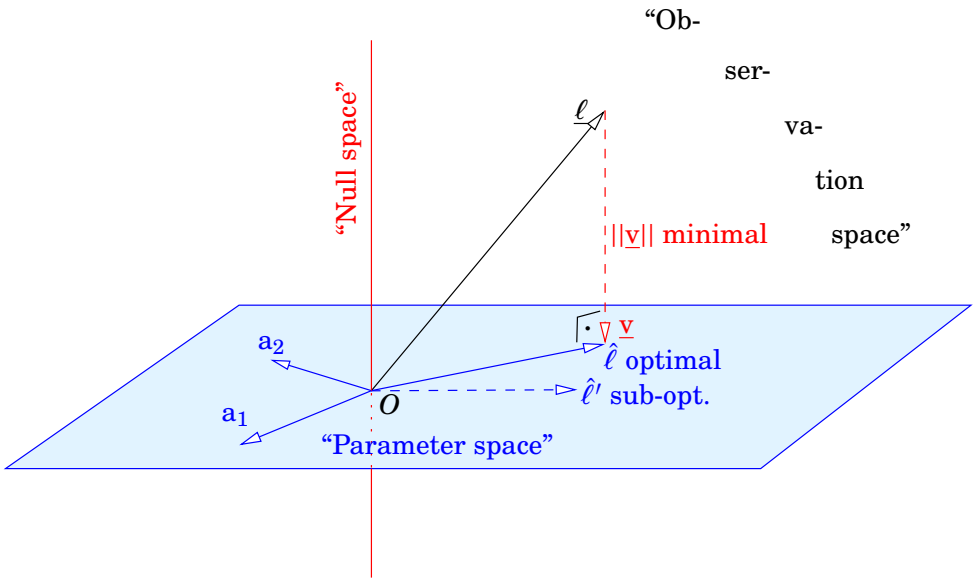


Figure 14.1. Least-squares adjustment as an orthogonal projection.

because

$$\mathrm{a}_i^{\mathsf{T}} Q_{\ell\ell}^{-1} \underline{\mathrm{v}} = 0,\ i = 1 \ldots m,$$

i.e., the vector of residuals is orthogonal to all columns a_i of the design matrix A.

Figure 14.1 gives a geometrical interpretation: the unknowns are those coefficients in the linear combination of the columns of the A matrix that minimize the norm of the vector of residuals.

4. The covariance matrix between the unknowns $\hat{\mathrm{x}}$ and the residuals $\underline{\mathrm{v}}$ vanishes: $\Sigma_{\mathrm{xv}} = \sigma^2 Q_{\mathrm{xv}} = 0$. So, they *don't correlate* with each other.

Because

$$\underline{\ell} = A\hat{\mathrm{x}} - \underline{\mathrm{v}},$$

it follows, based on the law of propagation of variances and the above mentioned property 4, that

$$Q_{\ell\ell} = A Q_{\mathrm{xx}} A^{\mathsf{T}} + Q_{\mathrm{vv}} \quad \Longrightarrow \quad Q_{\mathrm{vv}} = Q_{\ell\ell} - A Q_{\mathrm{xx}} A^{\mathsf{T}},$$

a useful equation for computing the *weight-coefficient matrix* — and the variance matrix — *of the vector of residuals* $\Sigma_{\mathrm{vv}} = \sigma^2 Q_{\mathrm{vv}}$.

14.3 Testing and hypotheses for testing

The observational material may contain *gross errors*. In a real-life adjustment calculus we must be able to say, based on our knowledge of the statistical distribution of the observations

1. something about the possible occurrence of gross errors

2. How large gross errors would have to be, so that they can be noticed and removed.

Finding gross errors belongs to the field of *statistical testing*.

Gross errors that are found

- are removed from the observation set, and the measurements in question are repeated. After the fact, this is laborious and costly, wherefore at least part of statistical testing is done already in the field; or
- are simply left out. This assumes that the measurement was originally planned redundantly, i.e., so many measurements have been made that one can afford to leave a (small) fraction of them out.

Statistical testing always requires the formulation of *hypotheses*. One of the hypotheses is always the

Null hypothesis: All the measurements in the network are correct, there are no gross errors in them. This hypothesis is called H_0, the *null hypothesis*.

Additionally there must always be at least one

Alternative hypothesis: The network contains some gross error, or some combination of gross errors, or a specific gross error. This hypothesis is called H_a, the *alternative hypothesis*.

Generally we wish to know, or make a judgement on, two matters:

1. Are there *generally* any gross errors left in this observation set?
2. Is *this specific observation* in error?

Figure 14.2 shows the role of testing in the whole planning and measurement process.

In the following sections, these questions will be discussed separately.

14.4 Overall validation

Firstly we choose the following alternative hypothesis:

> H_a: somewhere in the measurement material (we don't know yet where) there is a gross error.

This kind of hypothesis can be tested using the so-called χ^2 test. The method, and the tables belonging to it, can be found in statistics textbooks and on the Internet. The quantity to be tested is precisely the length of the vector of residuals in the $Q_{\ell\ell}$ metric, its *norm* squared:

$$\underline{\mathscr{E}} = \underline{v}^T \Sigma_{\ell\ell}^{-1} \underline{v} = \frac{1}{\sigma^2} \left(\underline{v}^T Q_{\ell\ell}^{-1} \underline{v} \right) = \frac{1}{\sigma^2} \left\| \underline{v} \right\|_{Q_{\ell\ell}}^2 .$$

The quantity is distributed according to the χ^2_{n-m} distribution, i.e., the χ^2 distribution with $n-m$ degrees of freedom (figure 14.3, the number of

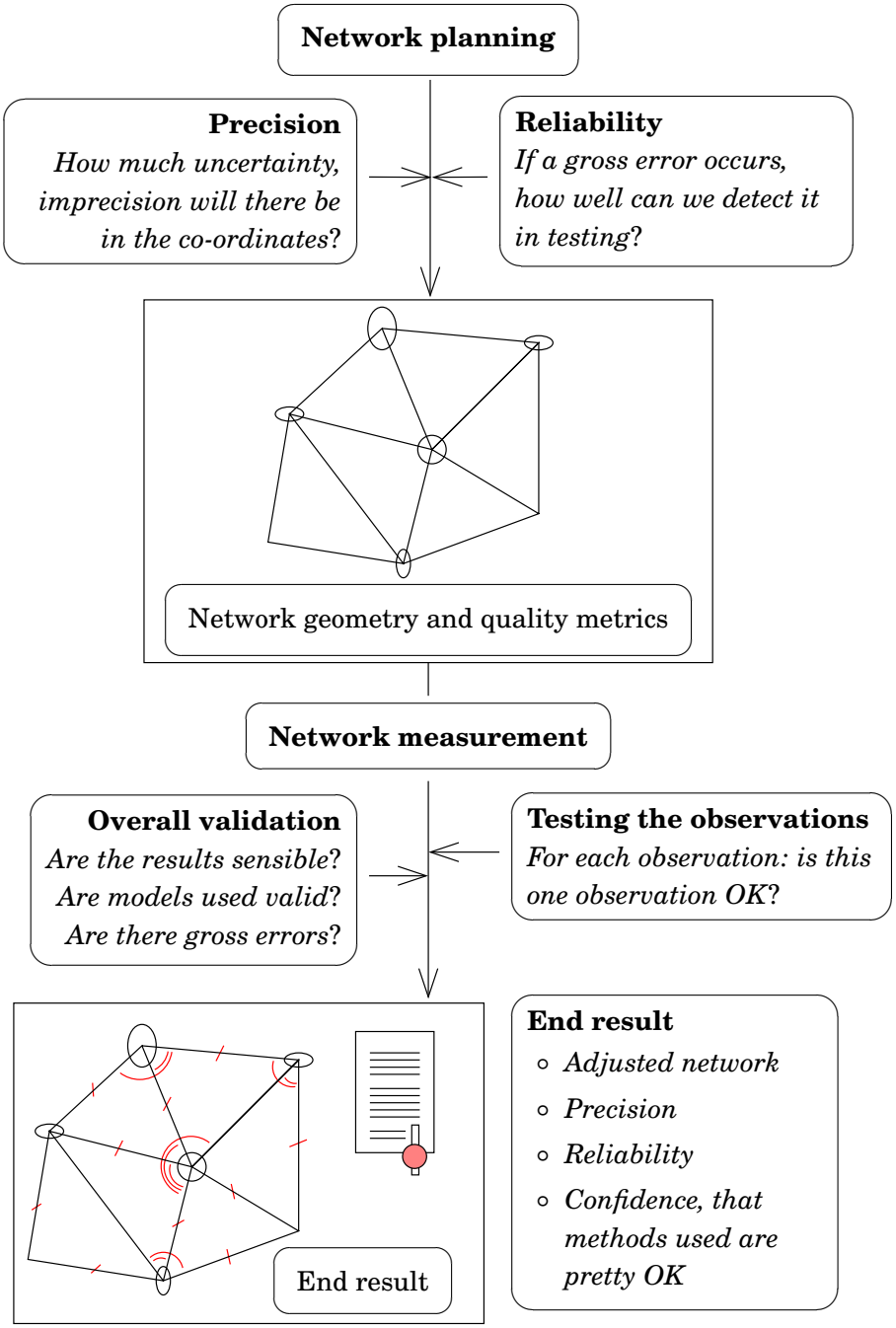


Figure 14.2. The planning and measurement process.

degrees of freedom is the difference between the number of observations and the number of unknowns, also known as the *redundancy* $n - m$). By testing this quantity, one may infer whether the material contains "some" gross error or not[4], without yet stating in which observation it

[4]The χ^2 test cannot distinguish between actual gross errors, and the possible unsuitability of the functional model $\ell = Ax$ applied to those measurements. If it happens that the χ^2 test rejects the null hypothesis, but all observations appear to be okay, it might be that there is a problem with the functional model: some systematic effect may have been overseen.

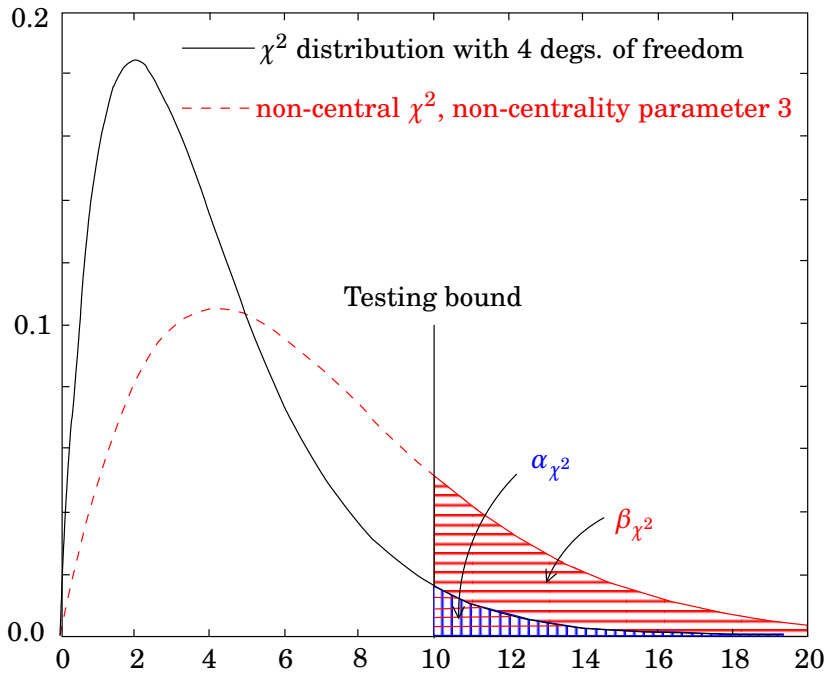


Figure 14.3. The χ^2 distribution with four degrees of freedom.

might be found.

The tables used by the χ^2test are found in the literature and on the Internet. Matlab contains ready routines for applying the method.

The quantity[5] $\underline{\mathscr{E}}$ is distributed according to χ^2_{n-m} only in the case that the material contains no gross errors, i.e., the *null hypothesis* H_0 applies. Then, the expectancy of $\underline{\mathscr{E}}$ is

$$E\left\{\underline{\mathscr{E}}|H_0\right\} = E\left\{\chi^2_{n-m}\right\} = n-m.$$

Assume now instead however, that the observations contain one or more gross errors, taken together $\nabla\ell$; the *alternative hypothesis* H_a applies. The effect of this error vector on the *residuals* is ∇v. In this case, the distribution of $\underline{\mathscr{E}}$ is the so-called *non-central* χ^2, in figure 14.3 the red dashed curve.

The interesting question is now, how large the effect of $\nabla\ell$ on ∇v is going to be. We may hope that it will be large, because then, the network is *reliable*[6]. Generally not the whole of $\nabla\ell$ propagates into ∇v; the adjustment conveys part into the effect of a gross error on the vector of unknowns[7], magnitude ∇x. See also section 14.8.

[5] In Baarda's terminology: "shifting variate".

[6] "Interior reliability".

[7] This is called the *exterior reliability*. A small effect means a large exterior reliability.

In this situation, the residual is $\underline{\mathrm{v}} + \nabla\mathrm{v}$. Written out into terms, the quantity $\underline{\mathscr{E}}$ becomes

$$\underline{\mathscr{E}} = \left\|\underline{\mathrm{v}}\right\|^2 + \|\nabla\mathrm{v}\|^2 = \underline{\mathrm{v}}^{\mathsf{T}}\Sigma_{\ell\ell}^{-1}\underline{\mathrm{v}} + \underline{\mathrm{v}}^{\mathsf{T}}\Sigma_{\ell\ell}^{-1}\nabla\mathrm{v} + \nabla\mathrm{v}^{\mathsf{T}}\Sigma_{\ell\ell}^{-1}\underline{\mathrm{v}} + \nabla\mathrm{v}^{\mathsf{T}}\Sigma_{\ell\ell}^{-1}\nabla\mathrm{v}.$$

The expectancy of $\underline{\mathscr{E}}$ becomes

$$\begin{aligned} E\left\{\underline{\mathscr{E}}|H_{\mathrm{a}}\right\} &= E\left\{\underline{\mathrm{v}}^{\mathsf{T}}\Sigma_{\ell\ell}^{-1}\underline{\mathrm{v}}\right\} + E\left\{\underline{\mathrm{v}}^{\mathsf{T}}\right\}\Sigma_{\ell\ell}^{-1}\nabla\mathrm{v} + \nabla\mathrm{v}^{\mathsf{T}}\Sigma_{\ell\ell}^{-1}E\left\{\underline{\mathrm{v}}\right\} + \nabla\mathrm{v}^{\mathsf{T}}\Sigma_{\ell\ell}^{-1}\nabla\mathrm{v} = \\ &= E\left\{\chi^2_{n-m}\right\} + 0 + 0 + \nabla\mathrm{v}^{\mathsf{T}}\Sigma_{\ell\ell}^{-1}\nabla\mathrm{v} = \\ &= (n-m) + \lambda, \end{aligned}$$

in which $\lambda \overset{\text{def}}{=} \nabla\mathrm{v}^{\mathsf{T}}\Sigma_{\ell\ell}^{-1}\nabla\mathrm{v}$ is called the *non-centrality parameter* of the χ^2 distribution. It describes how much the effect of the assumed gross error on the residuals, $\nabla\mathrm{v}$, extends outside the uncertainty area of the observations as described by the matrix $\Sigma_{\ell\ell}$.

A quadratic quantity is always positive. Therefore, the χ^2 test is *one-sided*, unlike the later to be presented test for the normal distribution. So, because $\underline{\mathscr{E}}$ a quadratic quantity, every gross error — and even systematic errors, i.e., errors in the functional model used — will tend to increase it. *Each and every error makes χ^2 larger*, and makes noticing the error more likely. This makes the χ^2 test such a useful overall test.

Actually the χ^2 test validates a lot more than just the observations. It assures that

1. The observation set does not contain any (large) gross errors.
2. The functional model used (the observation equations) is valid with sufficient accuracy.
3. The assumed mean errors of the observations (and the possible assumption of non-correlatedness) are realistic.

14.5 Locating gross errors

If we have inferred that the material presumably contains one or more gross errors, we want next to find out which observations are under suspicion. Let us assume for simplicity, that a gross error occurs only in one observation, although there might well be errors in several observations simultaneously.

The simplest way to search for gross errors, or rather, to do *outlier detection*, is to *look at the residuals*.

Table 14.1. Rejection bounds for significance levels in a two-sided test based on the *standard normal distribution*, i.e., mean error $\sigma = 1$ and expectancy $\mu = 0$.

Significance level (%)	Rejection bound
95	1.96
97.5	2.24
99	2.57
99.9	3.29

Let the vector of residuals be

$$\underline{\mathrm{v}} = \begin{bmatrix} \underline{v}_1 \\ \underline{v}_2 \\ \underline{v}_3 \\ \vdots \\ \underline{v}_i \\ \vdots \\ \underline{v}_{n-1} \\ \underline{v}_n \end{bmatrix}.$$

The element $\underline{v}_i$ is the residual of observation number i, i.e., $\underline{\ell}_i$. Its variance is

$$\sigma^2_{v_i} = \sigma^2 [Q_{\mathrm{vv}}]_{ii}$$

and the mean error of $\underline{v}_i$ is the square root of this.

Assume that the residuals $\underline{v}_i$ are *normally distributed*. Then, we may *test* every observation $i = 1, \ldots, n$:

$$\begin{aligned} |\underline{v}_i| > 1.96\sqrt{\sigma^2_{v_i}} \quad &\Longrightarrow \quad \underline{\ell}_i \text{ is probably in error} \\ |\underline{v}_i| \leq 1.96\sqrt{\sigma^2_{v_i}} \quad &\Longrightarrow \quad \underline{\ell}_i \text{ is presumably correct.} \end{aligned}$$

This two-sided test based on the normal distribution uses a *significance level* of 95%: *Even if there is no gross error*, there nevertheless is a probability of 100%−95%, i.e., 5%, that, based on the test, observation $\underline{\ell}_i$ will be rejected.

Table 14.1 gives a list of the rejection bounds for different significance levels in the two-sided test based on the standard normal distribution.

The method described here works correctly only if the observations do not correlate with each other, i.e., the matrix $Q_{\ell\ell}$ is a diagonal matrix. If it is not, the literature offers an adapted[8] testing method called *data snooping* (Baarda, 1968).

[8]The trick is simply that, instead of the residuals $\underline{\mathrm{v}}$, *weighted residuals* $\underline{\mathrm{w}} \stackrel{\mathrm{def}}{=} Q^{-1}_{\ell\ell}\underline{\mathrm{v}}$ and their variances are used.

Table 14.2. Example of linear regression.

i	1	2	3	4	5	$\sum$
x_i	1.51	2.44	3.34	4.41	5.05	16.75
$\underline{y}_i$	2.32	3.12	3.57	3.93	4.15	17.09
$\widehat{a}+\widehat{b}x_i$	2.51	2.97	3.41	3.94	4.26	
$\underline{v}_i$	+0.19	−0.15	−0.16	+0.01	+0.11	0
$\underline{v}_i^2$	0.0361	0.0225	0.0256	0.0001	0.0121	0.0964

14.6 Computation example: linear regression

Let us return to the linear regression example already used in section 13.5.3, see table 14.2.

Recall that the least-squares solution found was

$$\widehat{a} = 1.76 \pm 1.25\sigma,$$
$$\widehat{b} = 0.495 \pm 0.349\sigma.$$

We compute the function values $\widehat{a}+\widehat{b}x_i$ of the fitted line, as well as its residuals $\underline{v}_i = \left(\widehat{a}+\widehat{b}x_i\right) - \underline{y}_i$. The condition $\sum_{i=1}^{n}\underline{v}_i = 0$ is a good check.

If the observations $\underline{y}_i$ have a variance matrix $\Sigma_{\ell\ell} = \sigma^2 I$, then the "shifting variate" is

$$\underline{\mathscr{E}} = \underline{\mathrm{v}}^{\mathsf{T}}\Sigma_{\ell\ell}^{-1}\underline{\mathrm{v}},$$

in which $\underline{\mathrm{v}}$ is the vector formed by the residuals v_i.We obtain

$$\underline{\mathscr{E}} = \frac{\sum_{i=1}^{n}\underline{v}_i^2}{\sigma^2}.$$

If it is given *a priori* that $\sigma = \pm 0.15$, it follows that

$$\underline{\mathscr{E}} = \frac{0.0964}{0.0225} = 4.28.$$

The quantity $\underline{\mathscr{E}}$ is distributed according to χ_3^2: there are $n = 5$ observations and $m = 2$ unknowns (a and b), so the *number of degrees of freedom* (redundancy) is $n-m = 3$. According to the table, the probability that under the null hypothesis $\chi_3^2 > 4.642$ is 20%, so the value 4.28 is at least on a significance level of 80% fully acceptable.

The logic is, that if we search for a gross error in observation number i, we look for the linear combination of residuals in which the error shows clearest. We calculate the *orthogonal projection* of $\underline{\mathrm{v}}$ (in the $Q_{\ell\ell}$ metric) in the direction of the assumed gross error $\mathrm{e}_i \overset{\text{def}}{=} \begin{bmatrix} 0 & 0 & 0 & \cdots & 1 & \cdots & 0 & 0 \end{bmatrix}^{\mathsf{T}}$ (where the one is in place i) :

$$\underline{w}_i \overset{\text{def}}{=} \langle \mathrm{e}_i, \underline{\mathrm{v}} \rangle = \mathrm{e}_i^{\mathsf{T}} Q_{\ell\ell}^{-1}\underline{\mathrm{v}}.$$

Together the components $\underline{w}_i$ form the vector $\underline{\mathrm{w}}$ and they are optimally suited for discerning gross errors, or "outliers".

Table 14.3. Values of the χ^2_3distribution.

x	$\int_0^x \chi^2_3(\xi)d\xi$	$\int_x^\infty \chi^2_3(\xi)d\xi$
4.642	0.80	0.20
6.251	0.90	0.10
7.815	0.95	0.05
9.837	0.98	0.02
11.345	0.99	0.01
12.838	0.995	0.005
14.796	0.998	0.002
16.266	0.999	0.001

Next, the individual residuals are tested. Compute first the weight-coefficient matrix of the vector of residuals by the following formula:

$$Q_{vv} = Q_{\ell\ell} - AQ_{xx}A^T,$$

in which

$$Q_{xx} = \begin{bmatrix} 1.564 & -0.4072 \\ -0.4072 & 0.1215 \end{bmatrix}$$

was already computed in section 13.5.3, $Q_{\ell\ell} = I$, and

$$A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} 1 & 1.51 \\ 1 & 2.44 \\ 1 & 3.34 \\ 1 & 4.41 \\ 1 & 5.05 \end{bmatrix}.$$

After laborious calculation (Matlab!), we obtain the matrix Q_{vv}, shown in table 14.4 on the next page. Of this matrix, mostly the *diagonal elements* are interesting:

$$\sigma_{v_1} = \sigma\sqrt{[Q_{vv}]_{11}} = 0.15 \cdot 0.623 = 0.0935,$$
$$\sigma_{v_2} = \sigma\sqrt{[Q_{vv}]_{22}} = 0.1255,$$

etc. (Remember $\sigma = 0.15$.) See table 14.4 on the following page[9].

As can be seen, all observations are acceptable, with the exception of y_1, which, on the 95% significance level, is barely rejected (rejection bound 1.96). However, already on a significance level of 97.5%, it too is accepted.

[9]Note how the mean errors of the residuals are *systematically smaller* than the mean errors of the observations $\sigma = \pm 0.15$, especially close to the edges! With a large number of points, this phenomenon vanishes and we may write $Q_{vv} \approx Q_{\ell\ell}$. This is often done in any case. Then, gross errors in the edge points won't be noticed sufficiently well.

Table 14.4. Example of linear regression. Computing the residuals, their variance-covariance matrix, and normalized residuals.

i	1	2	3	4	5
v_i	+0.19	−0.15	−0.16	+0.01	+0.11
Q_{vv}	+0.3887	−0.4032	−0.2019	+0.0375	+0.1807
	−0.4032	+0.6998	−0.2006	−0.0821	−0.0112
	−0.2019	−0.2006	+0.8007	−0.1978	−0.1969
	+0.0375	−0.0821	−0.1978	+0.6646	−0.4178
	+0.1807	−0.0112	−0.1969	−0.4178	+0.4502
σ_{v_i}	0.0935	0.1255	0.1342	0.1223	0.1006
$\frac{\lvert v_i \rvert}{\sigma_{v_i}}$	2.03	1.20	1.19	0.08	1.09

Next, we add to the observed value y_3 a simulated *gross error* +1.0.

Now, as the least-squares solution we obtain the result of table 14.5:

$$\widehat{b} = \frac{5 \cdot 64.66 - 16.75 \cdot 18.09}{5 \cdot 64.34 - 16.75^2} = \frac{20.2925}{41.1375} = 0.493.$$

$$\widehat{a} = \frac{1}{5}\left(18.09 - 16.75 \cdot \widehat{b}\right) = 1.97,$$

In table 14.5 the σ_{v_i} values have not changed.

Compute the "shifting variate"

$$\underline{\mathscr{E}} = \frac{\sum_{i=1}^{n} \underline{v_i^2}}{\sigma^2} = \frac{1.1973}{0.0225} = 53.21!$$

There is something very, very rotten going on here...

Now look at table 14.5. the largest testing value by far, 7.08, is seen for the erroneous observation 3. But also observations 1 and 5 are rejected at the 95% significance level! For this reason, one should be careful. *Based on the test, one should reject only one observation at a time*, after which the whole least-squares computation should be repeated.

14.7 Significance level of the test

When we test a certain alternative hypothesis against the null hypothesis using, e.g., a normally distributed testing quantity or variate, one must choose a suitable *rejection bound*. If the variate to be tested exceeds this bound, H_0 is rejected and H_a accepted. Choosing the rejection bound is an important strategic decision.

See figure 14.5. In the figure, the rejection bound chosen is $h = 2.5\sigma$: if the testing variate exceeds 2.5 times its own mean error σ, the null hypothesis H_0 is *rejected* and the alternative hypothesis H_a is *accepted*.

Now, the strategy may lead to two types of error:

Table 14.5. Example of linear regression. A simulated gross error in point 3; original data, linear regression, residuals, testing.

i	1	2	3	4	5	$\sum$
x_i	1.51	2.44	3.34	4.41	5.05	16.75
$\underline{y}_i$	2.32	3.12	4.57	3.93	4.15	18.09
x_i^2	2.28	5.95	11.16	19.45	25.50	64.34
$x_i \underline{y}_i$	3.50	7.61	15.26	17.33	20.96	64.66
$\widehat{a} + \widehat{b} x_i$	2.71	3.17	3.62	4.14	4.46	
$\underline{v}_i$	+0.39	+0.05	−0.95	+0.21	+0.31	0.01
$\underline{v}_i^2$	0.1521	0.0025	0.9025	0.0441	0.0961	1.1973
σ_{v_i}	0.0935	0.1255	0.1342	0.1223	0.1006	
$\frac{\lvert \underline{v}_i \rvert}{\sigma_{v_i}}$	4.17	0.40	7.08	1.72	3.08	
Rejection?	*		**		*	

ensimmäisen lajin virhe

1. The null is rejected although it is valid. This is called an *error of the first kind*[10]. The probability of this error happening is the size of the vertically hatched (blue) area. In case of the normal distribution, it amounts to $\alpha = 1.24\%$ (two-sided), if the rejection

[10] Also *rejection error*.

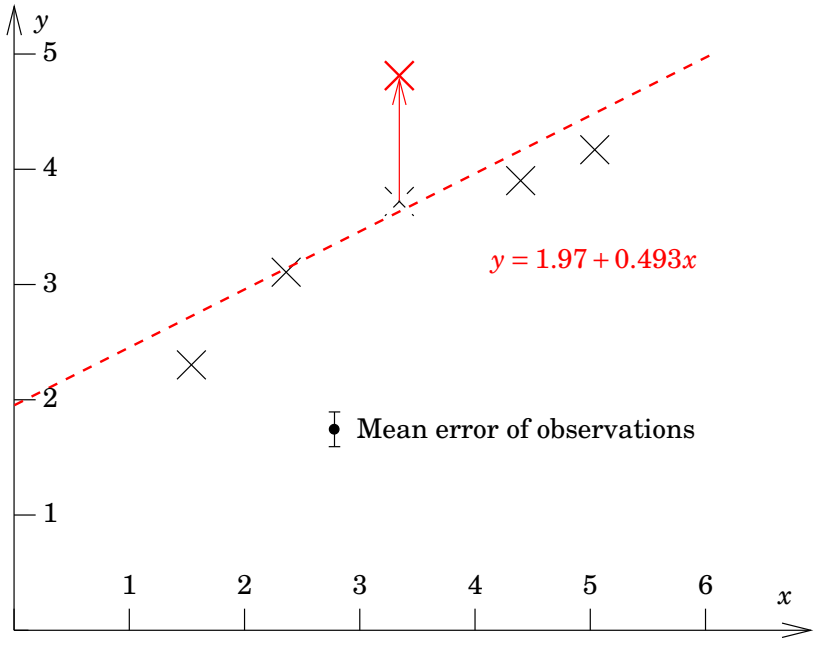


Figure 14.4. Example of linear regression, observation 3 contains a simulated gross error.

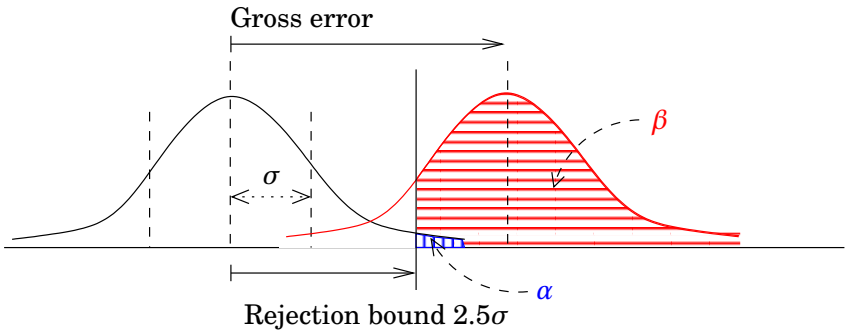


Figure 14.5. Statistical testing based on the normal distribution.

bound is $h = 2.5\sigma$. The quantity α is called the *significance level* of the test. merkitsevyystaso

2. The null is accepted although there is a gross error, i.e., H_0 is false, H_a is true. This is an *error of the second kind*[11]. The probability of it happening depends on the size k, of the gross error, more precisely, on the size of the difference $k-h/\sigma$. toisen lajin virhe

   Its opposite is called the *power* β[12] of the test. In the figure it is the size of the horizontally hatched (red) area. erotuskyky

Choosing the *testing strategy*, i.e., h, is thus always a *compromise*. It depends on what are the relative costs of errors of the first and of the second kind — including non-monetary "costs". $h = 3\sigma$ is an often used empirical solution — the "three-sigma rule".

Note also, that, in testing the body of observations, there is a *link* between the overall validation test and the per-observation tests! The link is through the significance levels, see figure 14.6: If the significance level of the χ^2 test is α_{χ^2} , and that of the test for a single observation is α, the connection is

$$1 - \alpha_{\chi^2} = (1-\alpha)^{n-m},$$

[11] Also *acceptance error*.

[12] I.e., the probability of an error of the second kind, if there is indeed a gross error in the observation, is $1-\beta$, or $100\% - \beta$.

Table 14.6. Rejection bound and significance level of a test in case of the normal distribution.

$\frac{h}{\sigma}$	α, % (one sided)	α, % (two sided)
2.0	2.28	4.56
2.5	0.62	1.24
3.0	0.13	0.27
3.5	0.02	0.05

Table 14.7. Rejection bound of test, assumed size of gross error, and corresponding power. Normal distribution.

$\frac{k-h}{\sigma}$	$\frac{k}{\sigma}(h=2.5)$	β, %
0.5	3.0	69.1
1.0	3.5	84.1
1.5	4.0	93.3
2.0	4.5	97.7
2.5	5.0	99.4
3.0	5.5	99.9
3.5	6.0	99.98

with $n-m$ the number of degrees of freedom[13]. In other words, the joint probability that all observations individually pass their tests, must be the same as the probability of passing the overall validation. Only on that condition may it be expected that, if the common χ^2 test finds something "rotten", then also the tests for the individual observations will point to the "guilty" observation. After removing or correcting this observation, the procedure is repeated, until the χ^2 test passes[14].

14.8 Reliability

14.8.1 Principle

The *reliability* of a measurement network is the property that gross errors are found easily, and are found even if they are relatively small.

Reliability corresponds to the network being "strong"; it is however not the same kind of strength as that, which gives the best possible precision.

See figure 14.7. From points A,B and C are measured the directions to a fourth point. Drawn are error ellipses for three different cases:

I when the point is far from the points A,B

II when the point is in a location where the directions to points A and B are perpendicular to each other, and

III when the point is between points A and B.

As can be seen, the *most precise result is obtained in case III*. The error ellipse is the smallest.

[13] This procedure is similar to the so-called *Bonferroni correction*, https://en.wikipedia.org/wiki/Bonferroni_correction.

[14] . . . if it still doesn't pass, perhaps the other models used should be checked, such as the assumed precisions of the observations, etc.

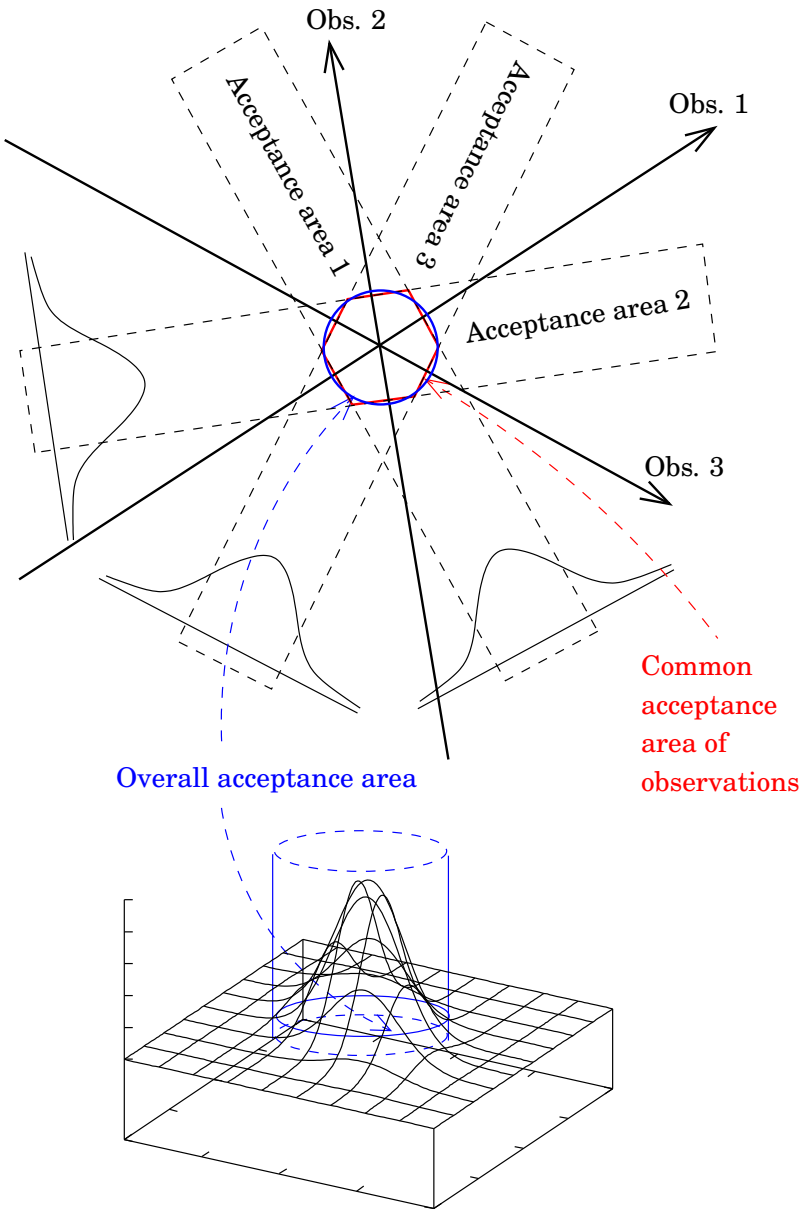


Figure 14.6. Harmonization of the significance levels of the overall validation and per-observation tests.

However, *reliability is poor* (non-existent) in case III. If the measurement made from point C contains a gross error (dashed line), we obtain in case III still a seemingly good — precise — *but erroneous* result. See the dashed error ellipse.

In the cases I and II, a gross error in the observation from C will produce a contradictory result. It is not possible to find a location for the target point that would be compatible with the direction measurements from all three points. This is a *good thing*, because it enables the detection of the gross error. The network is then called *reliable*.

In network planning, attention must be paid to reliability, of course in addition to precision. The network must be designed redundantly, i.e.,

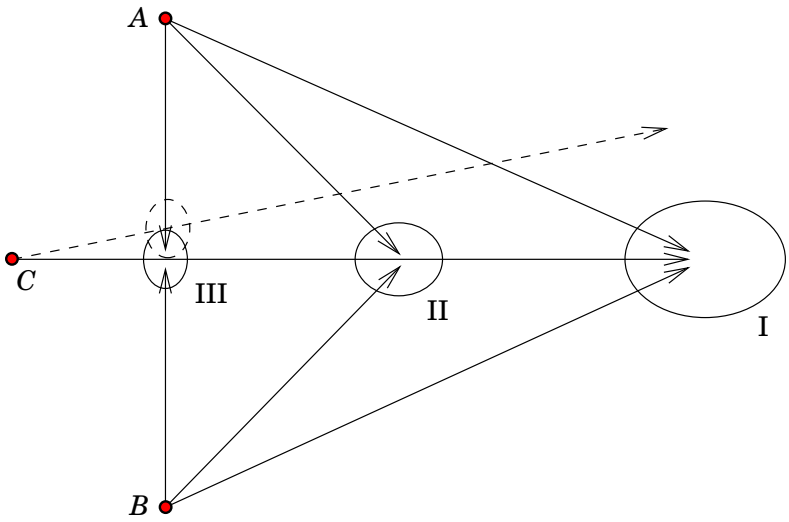


Figure 14.7. An example of reliability.

it must contain measurements that check each other. Common sense helps a lot here. There exist also mathematical and software tools for evaluating the reliability of a network. Always one must ask "what if... this or that observation were in error... would I notice?"

14.8.2 Example

In this example, the observation points A, B, C are located on the edge of a circle, and point P, the direction to which is being measured, is located near the origin (centre point of the circle) O. See figure 14.8.

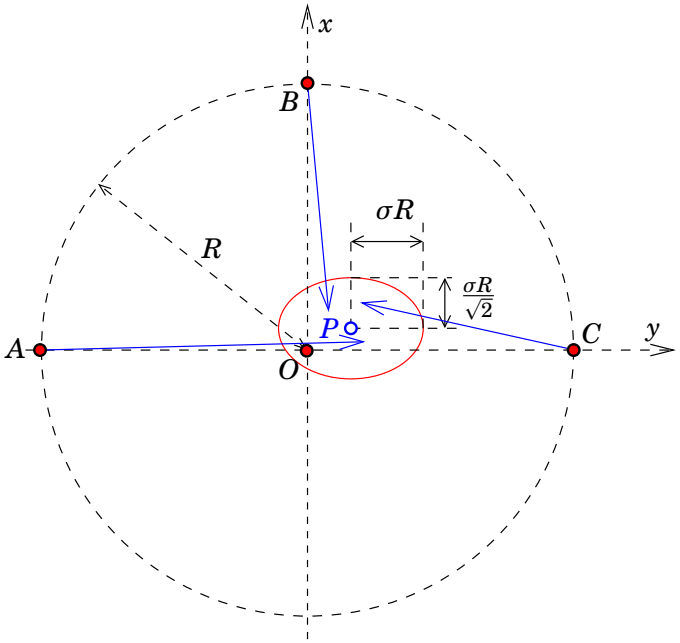


Figure 14.8. Another example of reliability.

The observation equations are obtained by looking at the geometry:

$$\begin{bmatrix} \underline{t}_{AP} - t_{AO} \\ \underline{t}_{BP} - t_{BO} \\ \underline{t}_{CP} - t_{CO} \end{bmatrix} + \begin{bmatrix} \underline{v}_1 \\ \underline{v}_2 \\ \underline{v}_3 \end{bmatrix} = \frac{1}{R} \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \widehat{x}_P \\ \widehat{y}_P \end{bmatrix}.$$

Symbolically:

$$\underline{\ell} + \underline{\mathrm{v}} = A\widehat{\mathrm{x}},$$

in which

$$A = \frac{1}{R} \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 0 \end{bmatrix},$$

$$\underline{\ell} = \begin{bmatrix} \underline{\ell}_1 \\ \underline{\ell}_2 \\ \underline{\ell}_3 \end{bmatrix} = \begin{bmatrix} \underline{t}_{AP} - t_{AO} \\ \underline{t}_{BP} - t_{BO} \\ \underline{t}_{CP} - t_{CO} \end{bmatrix}.$$

The least-squares solution is

$$\widehat{\mathrm{x}} = \left[A^{\mathsf{T}}A\right]^{-1} A^{\mathsf{T}}\underline{\ell} = R \cdot \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \underline{\ell}_3 - \underline{\ell}_1 \\ -\underline{\ell}_2 \end{bmatrix} = R \cdot \begin{bmatrix} \frac{1}{2}\left(\underline{\ell}_3 - \underline{\ell}_1\right) \\ -\underline{\ell}_2 \end{bmatrix}.$$

From this are obtained the residuals:

$$\underline{\mathrm{v}} = A\widehat{\mathrm{x}} - \underline{\ell} = \begin{bmatrix} -\frac{1}{2}\left(\underline{\ell}_3 - \underline{\ell}_1\right) \\ \underline{\ell}_2 \\ \frac{1}{2}\left(\underline{\ell}_3 - \underline{\ell}_1\right) \end{bmatrix} - \begin{bmatrix} \underline{\ell}_1 \\ \underline{\ell}_2 \\ \underline{\ell}_3 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}\underline{\ell}_1 - \frac{1}{2}\underline{\ell}_3 \\ 0 \\ -\frac{1}{2}\underline{\ell}_1 - \frac{1}{2}\underline{\ell}_3 \end{bmatrix}.$$

Note 1: as can be seen, the observation $\underline{\ell}_2$ has vanished from the residuals! If $\underline{\ell}_2 = \underline{t}_{BP} - t_{BO}$ contains a gross error, we are never going to notice it as an overly large residual.

Note 2: from the residuals it cannot be seen whether a gross error is from observation $\underline{\ell}_1$ or from observation $\underline{\ell}_3$. In the residuals, their coefficients are identical.

Next, compute the shifting variate:

$$\underline{\mathscr{E}} = \underline{\mathrm{v}}^{\mathsf{T}} \Sigma_{\ell\ell}^{-1} \underline{\mathrm{v}}.$$

Here, $\Sigma_{\ell\ell}$ is the variance matrix of the observations. Assume that the observations do not correlate with each other and that their mean error is σ. Then

$$\Sigma_{\ell\ell} = \sigma^2 \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}.$$

We obtain (H_0 is the null hypothesis):

$$\underline{\mathscr{E}} | H_0 = \frac{1}{\sigma^2} \sum_{i=1}^{3} \underline{v}_i^2 = \frac{1}{2\sigma^2} \left(\underline{\ell}_1 + \underline{\ell}_3\right)^2.$$

Note that, because the mean errors of both $\underline{\ell}_1$ and $\underline{\ell}_3$ are σ and they do not correlate, the mean error of the sum $\underline{\ell}_1 + \underline{\ell}_3$ is $\sigma\sqrt{2}$ and its variance $2\sigma^2$. The number of degrees of freedom is 1 and the variate $\underline{\mathscr{E}}$ is distributed according to χ^2_1, as should be the case according to theory.

By comparing the value $\underline{\mathscr{E}}$ computed from the observations with the values from the χ^2_1table, one can test, whether the observations might contain a gross error. If all observations are free of gross errors, the expectancy of $\underline{\mathscr{E}}$ is 1.

Nevertheless, as already pointed out above, we cannot observe any gross errors in $\underline{\ell}_2$ at all. We say[15], that the measurement geometry is *reliable* for observations $\underline{\ell}_1$ and $\underline{\ell}_3$, but *unreliable* for observation $\underline{\ell}_2$. If observation $\underline{\ell}_2$ contained a gross error of size ∇, it would slip *in its entirety* into the co-ordinate $\widehat{y}_P$ as an error $R \cdot \nabla$! We also say[16], that the measurement geometry is *unreliable* for unknown $\widehat{y}_P$, but *reliable* for unknown $\widehat{x}_P$.

A sufficiently large gross error ∇ in observations $\underline{\ell}_1$ or $\underline{\ell}_3$ again would be detected as an overly large value for the shifting variate (alternative hypothesis H_a):

$$\underline{\mathscr{E}}|H_a = \frac{1}{2\sigma^2}\left(\underline{\ell}_1 + \underline{\ell}_3 + \nabla\right)^2,$$

the expectancy of which is $1 + \frac{1}{2}(\nabla/\sigma)^2$. If $\nabla \gg \sigma$, this would be detected with considerable confidence.

Note reliability has nothing to do with *precision*! The precision of the unknowns $\widehat{\mathrm{x}} = \begin{bmatrix} \widehat{x}_P & \widehat{y}_P \end{bmatrix}^\mathsf{T}$ is described by their variance matrix

$$\mathrm{Var}\left\{\widehat{\mathrm{x}}\right\} = \sigma^2\left[A^\mathsf{T} Q_{\ell\ell}^{-1} A\right]^{-1} = \sigma^2 R^2 \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix},$$

i.e., the mean error of $\widehat{x}_P$ is $\frac{1}{2}\sigma R\sqrt{2}$, and that of $\widehat{y}_P$ is σR, and they are uncorrelated with each other.

However, a good mean error gives no solace if the co-ordinate solution $\widehat{y}_P$ contains a gross error...

14.9 The meaning of redundancy

Even though the reliability of a measurement network would be good, we may still ask, is it easy to identify in which observation the gross error has occurred? It this is not easy, we end up measuring all suspect observations again, or thow them out. This is not good.

[15] This is called interior reliability.

[16] So-called exterior reliability.

From the viewpoints both of good reliability and identifiability of gross errors, the *degree of redundancy* of a geodetic measurement network should not be too low. If the number of observations is n and the number of unknowns m, then the number of conditions, or *degrees of freedom*, is $n-m$. The degree of redundancy is then $(n-m)/n$. It is often stated as a percentage. E.g., linear regression through five points: $n=5$, $m=2$, i.e., a degree of redundancy of $3/5 = 60\%$. On the other hand, a levelling line of ten points between two known points: $n=11$, $m=10$, the degree of redundancy being $1/11 = 9\%$ — weak, but unfortunately common. By measuring in both directions we obtain $n=22$, $m=10$, i.e., a degree of redundancy of $12/22 \approx 55\%$, which is already good. A good rule of thumb is, that a degree of redundancy of 50% is desirable.

14.10 Deformation analysis

Literature:

FIG Commission 6 (1998, pages 191–256)

Cooper (1987, pages 331–352)

Vaníček and Krakiwsky (1986, pages 611–659)

Deformation analysis is one practical application of statistical testing. The null hypothesis H_0 in these tests is, that no observable deformation has happened. There may be many different alternative hypotheses H_a: from the hypothesis that some unspecified deformation took place, to many concrete hypotheses about the precise nature of the deformation sought for.

Deformation analysis is also an application that involves the time dimension: compared are measurements collected at two or more measurement epochs. The deformations studied may be natural, like deformations in the Earth's crust brought about by tectonic movements or by varying glacial loads; or they may be man-made, like the subsidence caused by mineral extraction (oil, natural gas, irrigation water, ...). The object of study may be the Earth's crust in an area, or a building or other man-made structure like a reservoir dam. The possibilities are very broad.

14.10.1 Height deformation analysis

One-dimensional or *height deformation analysis* studies vertical movement, e.g., using levelling. The simplest case is that in which the same levelling line or network of n points has been measured twice:

$$\underline{H}_i(t_1), i = 1,\ldots,n,$$
$$\underline{H}_i(t_2), i = 1,\ldots,n,$$

and the corresponding variance matrices of the heights are available: $\Sigma(t_1)$ and $\Sigma(t_2)$.

Clearly the comparison is possible only, if both measurements are first reduced to the same reference or *datum point*. E.g., choose the first network point as the datum point:

$$H_1^{(1)}(t_1) = H_1^{(1)}(t_2) \quad (= \text{some known value}).$$

After this, the variance matrices for both measurement times or *epochs* are only of size $(n-1)\times(n-1)$, because now point 1 is *known* and no longer has (co-)variances.

$$\Sigma^{(1)}(t_1) = \begin{bmatrix} \sigma_{22}^{(1)} & \sigma_{23}^{(1)} & \cdots & \sigma_{2n}^{(1)} \\ \sigma_{32}^{(1)} & \sigma_{33}^{(1)} & \cdots & \sigma_{3n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n2}^{(1)} & \sigma_{n3}^{(1)} & \cdots & \sigma_{nn}^{(1)} \end{bmatrix},$$

and the same for $\Sigma^{(1)}(t_2)$. Here

$$\begin{aligned} \sigma_{ii}^{(k)} &= \mathrm{Var}\left\{H_i^{(k)}\right\}, \\ \sigma_{ij}^{(k)} &= \mathrm{Cov}\left\{H_i^{(k)}, H_j^{(k)}\right\}. \end{aligned}$$

Now, calculate the height displacements between the two measurement epochs and their variances, assuming that the measurements made at times t_1 and t_2 are statistically independent of each other:

$$\Delta\underline{H}_i^{(1)} \stackrel{\text{def}}{=} \underline{H}_i^{(1)}(t_2) - \underline{H}_i^{(1)}(t_1),\ i = 2,\ldots,n;$$

$$\Sigma_{\Delta H \Delta H}^{(1)} = \Sigma^{(1)}(t_1) + \Sigma^{(1)}(t_2).$$

After this it is intuitively clear[17] that the following quantity, the *shifting variate*, has the
χ^2_{n-1} distribution:

$$\underline{\mathscr{E}} = \left[\Delta\underline{\mathrm{H}}^{(1)}\right]^{\mathsf{T}} \left[\Sigma_{\Delta H \Delta H}^{(1)}\right]^{-1} \Delta\underline{\mathrm{H}}^{(1)}.$$

Here

$$\Delta\underline{\mathrm{H}}^{(1)} = \begin{bmatrix} \Delta\underline{H}_2^{(1)} \\ \Delta\underline{H}_3^{(1)} \\ \vdots \\ \Delta\underline{H}_n^{(1)} \end{bmatrix}$$

is the (abstract) vector of height differences.

Statistical testing for deformation is based on this $\underline{\mathscr{E}}$ variate.

[17] ... provided that they are both normally distributed.

14.10.2 Horizontal deformation analysis

In two dimensions we proceed in the same way as in the one-dimensional case, except that

1. it is tempting to write the plane co-ordinates as *complex numbers*, and
2. there are now *two* datum points, the co-ordinates of which are considered identical for both epochs.

So, if there are n points, then the size of the variance matrix is now $(n-2)\times(n-2)$. Also the variance matrix is now complex valued, and *Hermitian*, i.e., its transpose is its complex conjugate.

The testing variate is again the shifting variate

$$\underline{\mathscr{E}} = \left[\underline{\mathbf{d}}^{(AB)}\right]^{\dagger}\left[\boldsymbol{\Sigma}_{dd}^{(AB)}\right]^{-1}\underline{\mathbf{d}}^{(AB)},$$

where $\underline{\mathbf{d}}$ is the complex vector of all co-ordinate differences, or *displacement vector*:

$$\underline{\mathbf{d}}^{(AB)} = \begin{bmatrix} \underline{x}_3^{(AB)}(t_2) - \underline{x}_3^{(AB)}(t_1) + i\left(\underline{y}_3^{(AB)}(t_2) - \underline{y}_3^{(AB)}(t_1)\right) \\ \underline{x}_4^{(AB)}(t_2) - \underline{x}_4^{(AB)}(t_1) + i\left(\underline{y}_4^{(AB)}(t_2) - \underline{y}_4^{(AB)}(t_1)\right) \\ \cdots \\ \underline{x}_n^{(AB)}(t_2) - \underline{x}_n^{(AB)}(t_1) + i\left(\underline{y}_n^{(AB)}(t_2) - \underline{y}_n^{(AB)}(t_1)\right) \end{bmatrix}.$$

AB is the chosen datum or starting point for both epochs t_1 and t_2. The other points are numbered $3,4,\ldots,n$. The symbol † signifies both transposition and complex conjugate, the so-called *Hermitian*[18] *conjugate*:

$$A^{\dagger} \overset{\text{def}}{=} \overline{A^{\mathsf{T}}} = \overline{A}^{\mathsf{T}}.$$

Warning

In Cooper's book (Cooper, 1987, page 335) there is an *error* under equation (9.52), the correct equation is (inverse, not transpose):

$$\Omega = \widehat{\mathbf{d}}^{\mathrm{t}} Q_{\mathrm{d}}^{-1} \widehat{\mathbf{d}}.$$

14.10.3 Example

Let the adjusted co-ordinates $\underline{\mathbf{x}}_i(t_1), i = 1,\ldots,4$ of the deformation network from the first measurement epoch be the following[19] (unit metre, point number i):

[18] Charles Hermite (1822 – 1901) was a French mathematician.

[19] These are *only* the co-ordinates of the points to be tested. They are assumed to be connected, for both epochs, to the same two datum points outside the area, which are assumed motionless.

i	$\underline{x}_i(t_1)$	$\underline{y}_i(t_1)$
1	1234.123	2134.453
2	2224.045	2034.487
3	2232.495	975.456
4	1148.865	879.775

and the co-ordinates of the second measurement epoch $\underline{\mathbf{x}}_i(t_2), i = 1,\ldots,4$ be the following:

i	$\underline{x}_i(t_2)$	$\underline{y}_i(t_2)$
1	1234.189	2134.485
2	2224.004	2034.433
3	2232.451	975.497
4	1148.929	879.766

so we are computing the inter-epoch differences vector $\underline{\mathbf{d}}$:

i	$\Delta\underline{x}_i$	$\Delta\underline{y}_i$
1	−0.066	−0.032
2	+0.041	+0.054
3	+0.044	−0.041
4	−0.064	+0.009

Using real numbers, with the definition

$$\underline{\mathrm{d}} \stackrel{\text{def}}{=} \underline{\mathrm{x}}(t_2) - \underline{\mathrm{x}}(t_1) = \begin{bmatrix} \underline{x}_1(t_2) - \underline{x}_1(t_1) \\ \underline{x}_2(t_2) - \underline{x}_2(t_1) \\ \underline{x}_3(t_2) - \underline{x}_3(t_1) \\ \underline{x}_4(t_2) - \underline{x}_4(t_1) \\ \hline \underline{y}_1(t_2) - \underline{y}_1(t_1) \\ \underline{y}_2(t_2) - \underline{y}_2(t_1) \\ \underline{y}_3(t_2) - \underline{y}_3(t_1) \\ \underline{y}_4(t_2) - \underline{y}_4(t_1) \end{bmatrix} = \begin{bmatrix} \Delta\underline{x}_1 \\ \Delta\underline{x}_2 \\ \Delta\underline{x}_3 \\ \Delta\underline{x}_4 \\ \hline \Delta\underline{y}_1 \\ \Delta\underline{y}_2 \\ \Delta\underline{y}_3 \\ \Delta\underline{y}_4 \end{bmatrix},$$

we find by computation

$$\underline{\mathrm{d}}^{\mathrm{T}}\underline{\mathrm{d}} = \sum_{i=1}^{4} \left[\underline{x}_i(t_2) - \underline{x}_i(t_1)\right]^2 + \sum_{i=1}^{4} \left[\underline{y}_i(t_2) - \underline{y}_i(t_1)\right]^2 = 0.017771\,\mathrm{m}^2.$$

Similarly with complex numbers, with the definition

$$\underline{\mathbf{d}} \stackrel{\text{def}}{=} \begin{bmatrix} \underline{\mathbf{z}}_1(t_2) - \underline{\mathbf{z}}_1(t_1) \\ \underline{\mathbf{z}}_2(t_2) - \underline{\mathbf{z}}_2(t_1) \\ \underline{\mathbf{z}}_3(t_2) - \underline{\mathbf{z}}_3(t_1) \\ \underline{\mathbf{z}}_4(t_2) - \underline{\mathbf{z}}_4(t_1) \end{bmatrix} \stackrel{\text{def}}{=} \begin{bmatrix} \Delta\mathbf{z}_1 \\ \Delta\mathbf{z}_2 \\ \Delta\mathbf{z}_3 \\ \Delta\mathbf{z}_4 \end{bmatrix}$$

we obtain similarly by computation

$$\underline{\mathbf{d}}^{\dagger}\underline{\mathbf{d}} = \sum_{i=1}^{4} \left[\overline{\underline{\mathbf{z}}}_i(t_2) - \overline{\underline{\mathbf{z}}}_i(t_1)\right]\left[\underline{\mathbf{z}}_i(t_2) - \underline{\mathbf{z}}_i(t_1)\right] = 0.017771\,\mathrm{m}^2.$$

Here, $\underline{\mathbf{z}}_i \overset{\text{def}}{=} \underline{x}_i + i\underline{y}_i$, and $\overline{\underline{\mathbf{z}}}_i \overset{\text{def}}{=} \underline{x}_i - i\underline{y}_i$ is its complex conjugate.

Let the precisions (mean co-ordinate errors) of the co-ordinates $\underline{x}_i(t_1)$ and $\underline{y}_i(t_1)$ of the first epoch be $\sigma_1 = \pm 5\,\mathrm{cm}$, and the precisions of the co-ordinates $\underline{x}_i(t_2), \underline{y}_i(t_2)$ of the second measurement $\sigma_2 = \pm 1\,\mathrm{cm}$. The variance matrices of the co-ordinate vectors are thus $\Sigma_1 = \sigma_1^2 I$ and $\Sigma_2 = \sigma_2^2 I$.

1. Compute the mean error σ of a single co-ordinate difference $\Delta\underline{x} = \underline{x}(t_2) - \underline{x}(t_1)$. Propagation of variances yields

   $$\sigma^2 = \sigma_1^2 + \sigma_2^2 = (25+1)\,\mathrm{cm}^2 = 26\,\mathrm{cm}^2.$$

   Now the variance matrix of co-ordinate *differences* is

   $$\Sigma = \Sigma_1 + \Sigma_2 = \sigma^2 I,$$

   with $\sigma = \sqrt{26}\,\mathrm{cm} = 5.1\,\mathrm{cm} = 0.051\,\mathrm{m}$.

2. Compute the deformation's testing variate, the shifting variate

   $$\underline{\mathscr{E}} = \underline{\mathrm{d}}^{\mathsf{T}}\Sigma^{-1}\underline{\mathrm{d}} = \frac{\underline{\mathrm{d}}^{\mathsf{T}}\underline{\mathrm{d}}}{\sigma^2}.$$

   Here $\underline{\mathrm{d}} = \underline{\mathrm{x}}_2 - \underline{\mathrm{x}}_1$ is the displacement vector, the abstract vector of co-ordinate differences between the epochs. Because we assume that both co-ordinate sets are given in the same, common datum, the definition points of which nevertheless *don't* belong to the set $1-4$, we may assume that all co-ordinates are free. In that case the number of degrees of freedom is $h = 2n = 8$, where n is the number of points. The variance matrix of the components of the displacement vector, or co-ordinate differences, $\underline{\mathrm{d}}$ is $\sigma^2 I$.

   Answer:

   $$\underline{\mathscr{E}} = \frac{1}{0.0026\,\mathrm{m}^2}\left(\underline{\mathrm{d}}^{\mathsf{T}}\underline{\mathrm{d}}\right) = \frac{0.017771\,\mathrm{m}^2}{0.0026\,\mathrm{m}^2} = 6.835.$$

3. The quantity $\underline{\mathscr{E}}$ is distributed according to the χ_8^2 distribution. If the limit value of this distribution for a significance level of 95% is 15.51 (cf. Cooper (1987) page 355), has in this case a deformation probably taken place?

   Answer: No, it has not. $6.835 < 15.51$.

4. If, however, the assumed precisions were $\sigma_1 = \sigma_2 = \pm 1\mathrm{cm}$, would then, at a significance level of 95%, probably a deformation have taken place?

Answer: Yes, it would. $\sigma^2 = (1+1)\,\mathrm{cm}^2 = 0.0002\,\mathrm{m}^2$ and

$$\underline{\mathscr{E}} = \frac{1}{0.0002\,\mathrm{m}^2}\left(\underline{\mathrm{d}}^{\mathsf{T}}\underline{\mathrm{d}}\right) = \frac{0.017771\,\mathrm{m}^2}{0.0002\,\mathrm{m}^2} = 88.9 > 15.51.$$

15. Gravity in geodesy

15.1 Measuring gravity

Gravitation is a fundamental force of the universe. It is an attraction acting between all celestial bodies. According to Newton's law of gravitation, the force is proportional to the masses of both bodies, and inversely proportional to the square of the distance between the bodies.

Galileo Galilei (1564 – 1642) was the first to show experimentally, that all bodies fall equally fast: their acceleration of free fall is the same, independently of their mass m.

This may be understood in this way, that as the mass m grows, the gravitational force F grows, but also the *inertia* of the body m, appearing in the formula $F = ma$, grows, and the acceleration a remains unchanged. One says that the "heavy mass" of the body is the same as its "inertial mass".

However, a broader way to look at this, which makes more theoretical sense and is more general, is, that there is no difference in principle between gravity and all kinds of *pseudo-forces* — like the centrifugal force — caused by irregular motion. Already the Hungarian baron and geophysicist Loránd Eötvös[1] did many very precise tests to investigate if any difference could be found between "heavy" and "inertial" mass, e.g., in connection with material composition. The answer was consistently "no[2]".

From this, Albert Einstein (1879 – 1955) concluded logically, that gravity is a *geometric* property of space-time connected with its curvature, and developed his famous field equations[3] which link together the curvature tensor of space-time and the energy-momentum tensor of the matter contained in space-time. Locally, e.g., inside a small, closed elevator, it is not

[1] Baron Loránd Eötvös de Vásárosnamény (1848 – 1919) was a Hungerian geophysicist and student of gravity.

[2] Eötvös' tests have been repeated with even much greater accuracy. The answer did not change.

[3] The theory is known as general relativity.

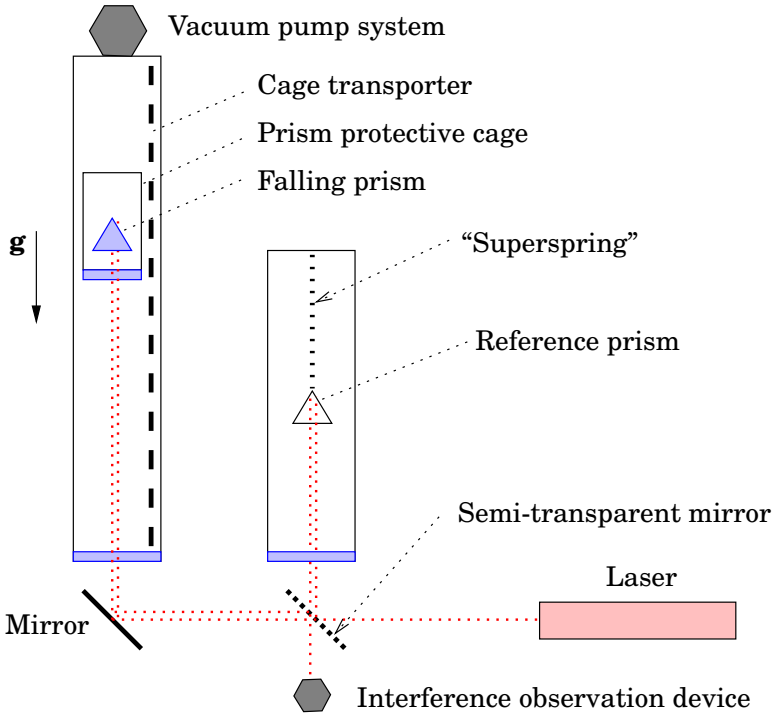


Figure 15.1. An absolute or ballistic gravimeter.

possible to know if the "gravity" inside the elevator is the result of the Earth's attraction, or of the acceleration produced by a rocket engine in the floor of the elevator!

The measurement and study of the acceleration of gravity is a specialization called gravimetry.

Gravity is the *acceleration of free fall*, and it is expressed in the SI unit $\mathrm{m/s^2}$. On the Earth's surface, gravity is about $9.8\,\mathrm{m/s^2}$. In gravimetry, however, smaller units are needed, and thus was invented the milligal (mGal) and the microgal (μGal). In very precise work, still the nanogal (nGal) is encountered.

Unit	In SI units ($\mathrm{m/s^2}$)	As fraction of gravity (about!)
mGal	10^{-5}	10^{-6}
μGal	10^{-8}	10^{-9}
nGal	10^{-11}	10^{-12}

For measuring gravity, instruments called *gravimeters* have been built. An ordinary field gravimeter is in principle just an extremely sensitive spring balance (figure 15.2). Its measurement precision may be 0.01 – 0.1 mGal. Furthermore there exist *ballistic* gravimeters, which measure interferometrically the acceleration of a falling body, figure 15.1.

Ballistic gravimeters are *absolute*. Field or spring gravimeters are not absolute: the have a *drift*. This means that the measurement values produced by the same acceleration of gravity change slowly over time.

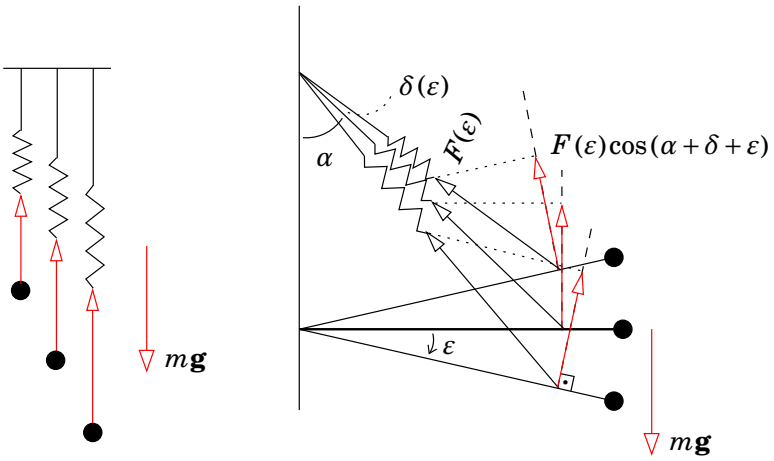


Figure 15.2. Principle of operation of relative or spring gravimeter: increasing sensitivity through a diagonal solution, *astatization*.

Therefore field measurements are planned to always start from a known point and end on a known point. The measurement values are adjusted between the end points in proportion to time according to the drift thus determined.

15.2 Gravity and geopotential

15.2.1 The gradient of a scalar field

In figure 15.3 is seen how the forms of the terrain are depicted on a map by height contours. In the figure, as an example could have served, instead of the terrain height $h(x,y)$, any scalar function of two variables. In the figure have been drawn as arrows the *gradient* of the height field,

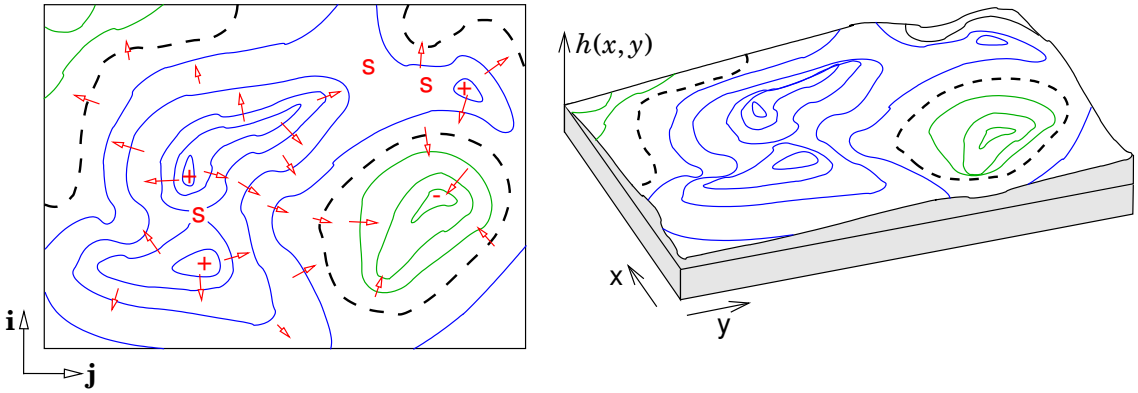


Figure 15.3. The terrain height $h(x,y)$ depicted by height contours, and height gradients (arrows). "+", "–", local maximum, minimum, "S" saddle point. On the right, the terrain in perspective.

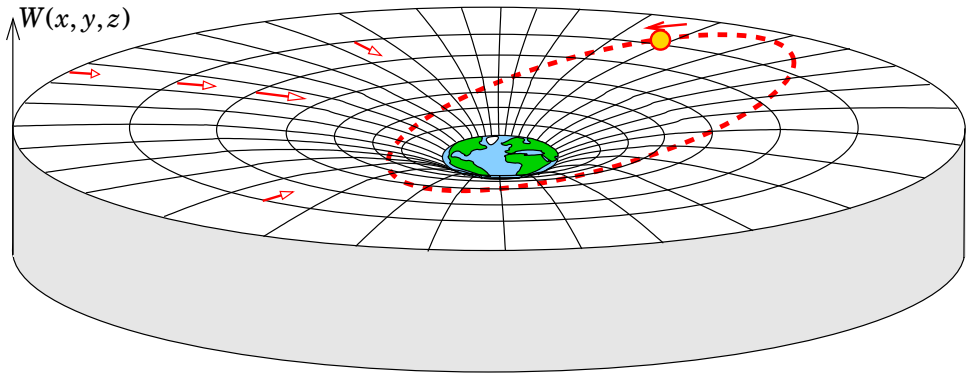


Figure 15.4. A geopotential table. Tables like this can be found in science museums. The surface of the table describes the gravity potential of the Earth, albeit only in two dimensions. The arrows again depict the *gradient* of the geopotential, i.e., the slope of the table surface.

On the geopotential table, a glass marble can be made to orbit around the "Earth" in an elliptical Kepler orbit, if the figure of the surface is sufficiently realistic, i.e., in agreement with Newton's gravitation formula.

□

the vector field

$$\mathbf{v}(x,y) = \frac{\partial h(x,y)}{\partial x}\mathbf{i} + \frac{\partial h(x,y)}{\partial y}\mathbf{j} \stackrel{\text{def}}{=} \operatorname{grad} h = \nabla h,$$

in which $\mathbf{i}$ and $\mathbf{j}$ are unit vectors in the x and y co-ordinate directions. This is the vector valued field, the value of which in every point (x,y) consists of two components, the partial derivatives of the height field with respect to the co-ordinates, in this point, $\frac{\partial h}{\partial x}$ and $\frac{\partial h}{\partial y}$.

The gradient vector describes the *slope* of the Earth's surface: it is the longer, the steeper the slope of the Earth's surface is. And the direction of the vector is of course the direction to which the terrain is sloping[4].

The gradient is always perpendicular to the height contour, which is a set of points having the same height value, an *equi-value curve*. Along it, height is constant.

Like the height field $h(x,y)$, we may also visualize the *geopotential* $W(x,y,z)$ in three-dimensional space, with "height contours" or *equipotential surfaces*, and a three-dimensional gradient. On the equipotential surfaces, the value of the geopotential is constant.

Figure 15.4 is a similar visualization of the geopotential by the curved surface of a table. The distance of the surface from the floor corresponds to the geopotential, i.e., the energy level of an object on the surface. The circles drawn on the table visualize the equipotential surfaces (in reality

[4] Actually the arrows are drawn in the "wrong" direction, i.e., the direction into which the Earth's surface is going down. Thus they describe the vector field $-\mathbf{v}$.

three-dimensional) of the Earth's gravity field, and the curves radiating outward from the Earth, the "lines of force" along which the gradient vector of the potential — i.e., the gravity vector — is everywhere directed.

15.2.2 Normal gravity and disturbing potential

Gravity consists of *two parts*:

- the attraction by the Earth's masses
- the centrifugal (pseudo-)force caused by the Earth's rotation.

The contribution of centrifugal force to all of gravity is less than one percent, the same order of magnitude as the difference in gravity between equator and poles.

The gravity field of the Earth contains all kinds of irregular variations from place to place. Most of the gravity field can however be described as the *field of an ellipsoid of revolution*. This mathematically defined, regular model field, in which the flattening and rotational motion of the Earth have been taken into account, is called the *normal field*.

The lines of force and level surfaces (equipotential surfaces) of the normal gravity field are depicted in figure 15.5. The *reference ellipsoid* is one equipotential surface of the normal gravity field, in the same way as the *geoid* (section 15.4) is an equipotential surface of the *true* gravity field.

The *potential* of the normal gravity field, the *normal potential*, is written with the symbol $U(x,y,z)$. *Normal gravity* itself is the *gradient* of this potential. The gravity vector is the gradient of the geopotential $W(x,y,z)$[5]:

$$\mathbf{g} = \nabla W = \mathrm{grad} W = \frac{\partial W}{\partial x}\mathbf{i} + \frac{\partial W}{\partial y}\mathbf{j} + \frac{\partial W}{\partial z}\mathbf{k},$$

in which $\mathbf{i},\mathbf{j},\mathbf{k}$ are the unit vectors in the x, y, and z directions.

In the same way, also the normal gravity vector

$$\boldsymbol{\gamma} = \nabla U = \mathrm{grad} U = \frac{\partial U}{\partial x}\mathbf{i} + \frac{\partial U}{\partial y}\mathbf{j} + \frac{\partial U}{\partial z}\mathbf{k}$$

is the gradient of the normal gravity potential U.

By subtracting from the true gravity potential the normal potential, the *disturbing potential* is obtained:

$$T \stackrel{\mathrm{def}}{=} W - U.$$

[5]The name of the symbol ∇ is *nabla*. It may be an old Greek word for a phoenician harp, the shape of which it resembles.

Table 15.1. Normal potential and normal gravity according to the GRS80 ellipsoid (Heikkinen (1981), simplified). In the equations, 9.78… is normal gravity itself, and 0.0000030877… the vertical gradient of normal gravity on the equator on the surface of the reference ellipsoid, according to the normal model. Units m, $\mathrm{m/s^2}$ and $\mathrm{m^2/s^2}$

$$
\begin{aligned}
U &= 62636860,8500+ \\
&+ \left(-9,78032677 - 0,05163075\sin^2\varphi - 0,00022761\sin^4\varphi - 0,00000123\sin^6\varphi\right) h + \\
&+ \left(+0,01543899 - 0,00002195\sin^2\varphi - 0,00000010\sin^4\varphi\right)\cdot 10^{-4}\cdot h^2 + \\
&+ \left(-0,00002422 + 0,00000007\sin^2\varphi\right)\cdot 10^{-8}\cdot h^3 + 0,00000004\cdot 10^{-12}\cdot h^4 + \ldots \\
\frac{\partial U}{\partial h} &= -9,78032677 - 0,05163075\sin^2\varphi - 0,00022761\sin^4\varphi - 0,00000123\sin^6\varphi + \\
&+ \left(+0,03087798 - 0,00004390\sin^2\varphi - 0,00000020\sin^4\varphi\right)\cdot 10^{-4}\cdot h + \\
&+ \left(-0,00007265 + 0,00000021\sin^2\varphi\right)\cdot 10^{-8}\cdot h^2 + 0,00000015\cdot 10^{-12}\cdot h^3 + \ldots
\end{aligned}
$$

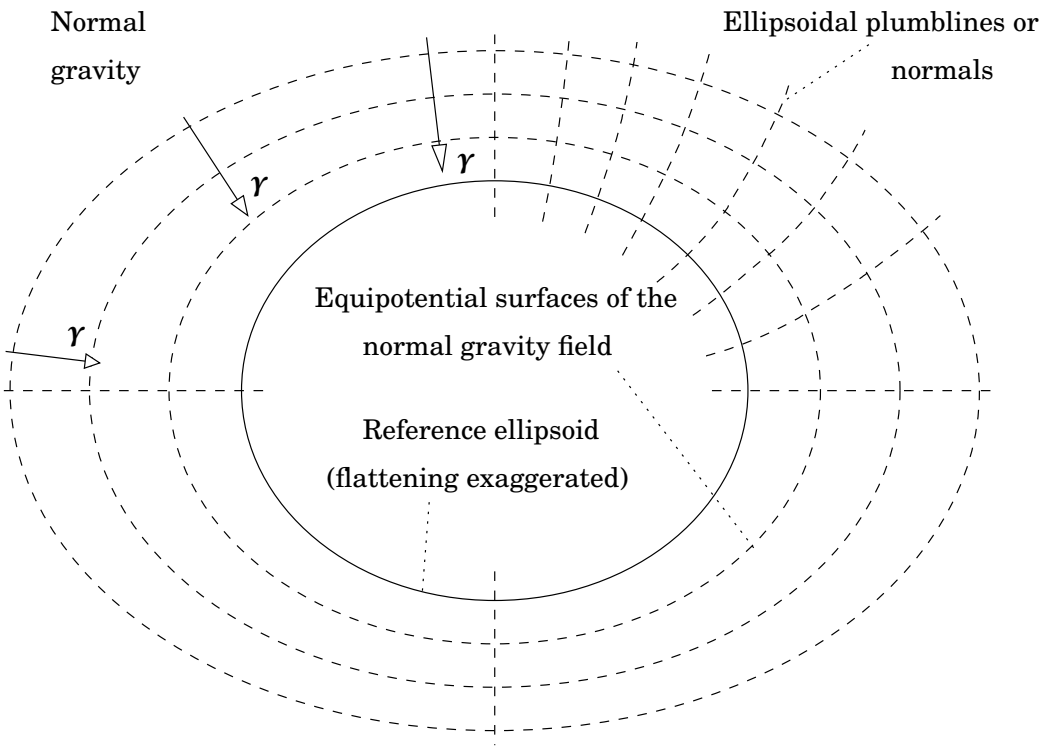


Figure 15.5. The normal gravity field of the Earth.

The *magnitude* of normal gravity is designated by the symbol $\gamma \overset{\text{def}}{=} \left\|\boldsymbol{\gamma}\right\|$, in the same way as the magnitude of true gravity $g \overset{\text{def}}{=} \|\mathbf{g}\|$. Because the two vectors have nearly identical directions — straight down — we may also write

$$g = -\frac{\partial W}{\partial h}, \gamma = -\frac{\partial U}{\partial h}.$$

Normal gravity can be computed exactly if we know, for a point P, the geodetic latitude φ_P and the height from the reference ellipsoid h_P:

$$\gamma_P = \gamma\left(\varphi_P, h_P\right).$$

Normal gravity, like true gravity, diminishes quickly when moving upward. The rate of diminishing is about 0.3 mGal for every metre. The dependence on latitude is much weaker.

15.2.3 Distance between equipotential surfaces

Because the normal gravity field is meant to be an idealized representation of the true gravity field, the equipotential surfaces of both for the same potential W = constant and U = constant, for the same constant value, run close to each other. See figure 15.7. The points P and Q lie on the same plumb line: P lies on the surface $W = W_P$ of the W field, whereas Q lies on the corresponding surface $U = U_Q = W_P$ of the U field. So:

$$W_P = U_Q$$

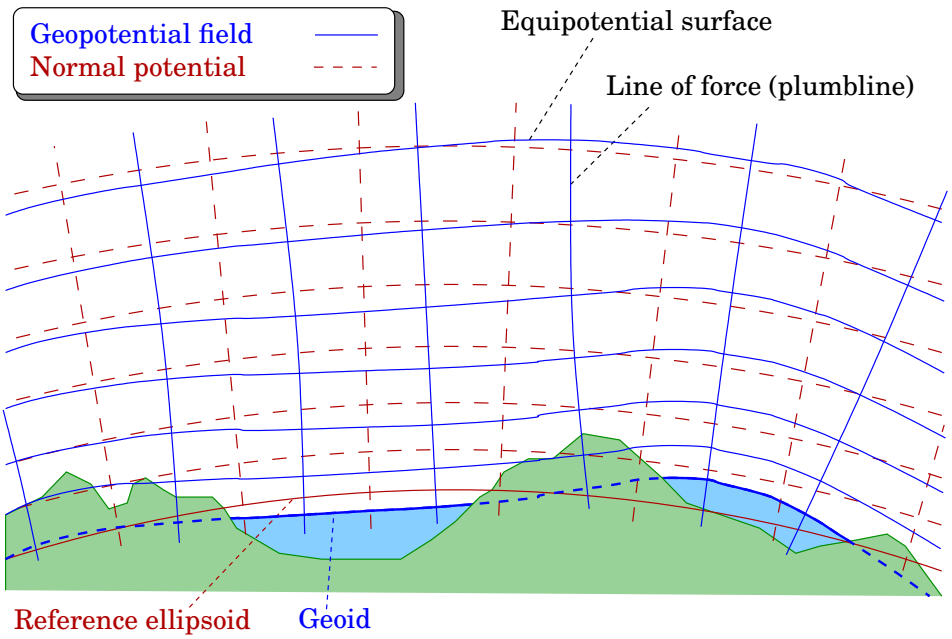


Figure 15.6. Level surfaces and lines of force of the geopotential and the normal potential.

and linearization with respect to height h yields

$$U_P \approx U_Q + \zeta \left.\frac{\partial U}{\partial h}\right|_P = U_Q + \zeta\gamma_P$$

in which ζ is the distance separating points P and Q.

Subtraction yields the *disturbing potential*

$$T_P \overset{\text{def}}{=} W_P - U_P = \zeta\gamma_P \implies \zeta = \frac{T_P}{\gamma_P}. \tag{15.1}$$

Equation 15.1 is the famous Bruns equation[6]. The quantity ζ is called the *height anomaly* (of point P); it is the distance between an equipotential surface of the Earth's gravity field and the *corresponding* surface of the normal field. The Bruns equation links this distance directly to the disturbing potential.

When point P is located on the geoid, we have $W_P = W_0$, and Q is located on the reference ellipsoid, i.e., $U_Q = U_0 = W_0$. In this case we use, instead of the notation ζ, the name N, the *geoid undulation*, the geoid height, i.e., the distance of the geoid from the reference ellipsoid. The Bruns equation is in this case

$$N = \frac{T_0}{\gamma_0},$$

in which the values of both T_0 and γ_0 are evaluated on the geoid. In practice $N \approx \zeta$, except in the mountains. At sea level, $\zeta = N$ exactly.

[6]Ernst Heinrich Bruns (1848–1919) was a gifted mathematician and astronomer, whose greatest achievements however were in the study of the Earth's gravity field.

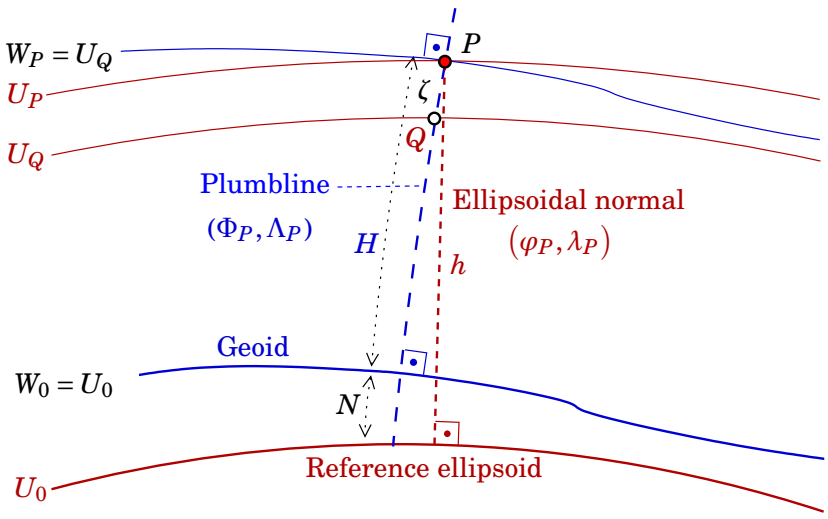


Figure 15.7. Equipotential surfaces of true and normal gravity field.

15.3 Gravity anomalies

In practice the height h of a point from the reference ellipsoid can be obtained empirically only using satellite positioning[7]. This is why one writes, using $h_P = H_P + N \approx \zeta + H_P$, see figures 15.7 and 15.8:

$$-\left.\frac{\partial T}{\partial h}\right|_P = g_P - \gamma_P = g_P - \gamma\left(\varphi_P, h_P\right),$$

in which (Taylor series expansion)

$$\gamma\left(\varphi_P, h_P\right) = \gamma\left(\varphi_P, H_P\right) + \left.\frac{\partial \gamma}{\partial h}\right|_P \cdot \zeta + \mathcal{O}\left(\zeta^2\right),$$

so, with the Bruns equation 15.1, we obtain

$$-\left.\frac{\partial T}{\partial h}\right|_P \approx g_P - \gamma\left(\varphi_P, H_P\right) - \left.\frac{\partial \gamma}{\partial h}\right|_P \cdot \zeta = g_P - \gamma\left(\varphi_P, H_P\right) - \left.\frac{\partial \gamma}{\partial h}\right|_P \cdot \frac{T_P}{\gamma_P}.$$

From this

$$g_P - \gamma\left(\varphi_P, H_P\right) = -\left.\frac{\partial T}{\partial h}\right|_P + \left(\left.\frac{\partial \gamma}{\partial h}\right|_P \frac{1}{\gamma_P}\right) T_P. \tag{15.2}$$

This expression is called the *gravity anomaly*, with the definition

$$\Delta g_P \overset{\text{def}}{=} g_P - \gamma\left(\varphi_P, H_P\right). \tag{15.3}$$

The gravity anomaly Δg_P can be computed if has been measured

[7]Nowadays, thanks to GNSS, it is easier to obtain the quantity

$$\delta g_P \overset{\text{def}}{=} g_P - \gamma\left(\varphi_P, h_P\right),$$

which is called the *gravity disturbance*.

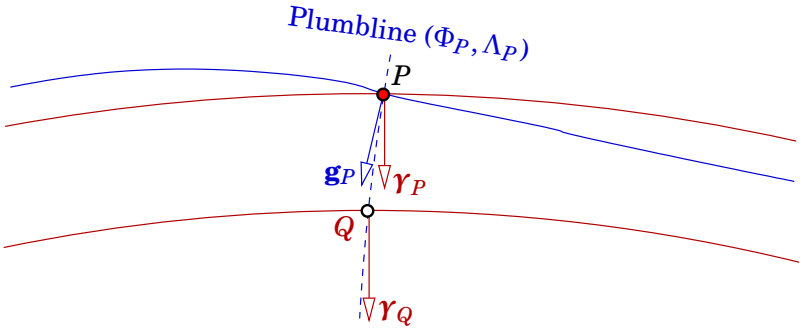


Figure 15.8. True and normal gravity vectors.

1. the gravity value g_P at point P by gravimetric measurements, and
2. the height H_P of the point from the geoid, i.e., "above sea level".

Most often — and before the satellite era, always —, the height of a gravity measurement point is determined by reading from a map, from a photogrammetric stereo model, by using a barometer, or from traverse levelling. In all cases, one obtains just the height H_P *above sea level*. The precision of measurement of the heights varies from several centimetres to around a metre; the error propagates straight into the anomaly values Δg, by the vertical gradient of normal gravity $\frac{\partial \gamma}{\partial h} \approx -0.3\,\mathrm{mGal/m}$.

The gravity anomaly Δg is an empirical quantity that can be calculated from measurements.

The most used gravity anomaly is the so-called *free-air anomaly* Δg, the definition of which was just given above as 15.3. Gravity anomaly values vary in the interval $\pm 100\,\mathrm{mGal}$, more rarely (mountains) $\pm 200\,\mathrm{mGal}$. Within Finland, the interval of variation is $\pm 60\,\mathrm{mGal}$. Gravity anomalies and their variations from place to place describe the irregularities of the interior mass distribution of the Earth, and are therefore of geophysical and geological interest.

As a curiosity may be mentioned that the Dutch researcher Vening Meinesz found, south of the island of Java, at the Java (today Sunda) deep-sea trench, a large deficiency in gravity. We know today that deep-sea trenches are those places on the Earth's surface where, as part of plate tectonics, the oceanic Earth's crust "dives" into the Earth's mantle in order to be geologically recycled: *subduction*, see figure 17.6. alityöntö

15.4 The gravimetric geoid

From the equation 15.2 above together with definition 15.3 is obtained

$$\Delta g = -\frac{\partial T}{\partial h} + \left(\frac{\partial \gamma}{\partial h}\frac{1}{\gamma}\right) T, \qquad (15.4)$$

fysiikaalisen geodesian perusyhtälö

which is called the *fundamental equation of physical geodesy*. So:

> *The gravity anomaly Δg is a linear combination of disturbing potential T and its vertical derivative of location $\frac{\partial T}{\partial h}$.*

The quantity $\frac{\partial \gamma}{\partial h} \approx -0.3\,\mathrm{mGal/m}$ is the, already mentioned, *vertical gradient of normal gravity*.

geoidi

The *geoid*, the equipotential surface of the Earth's gravity field that describes the figure of the whole field (the "mathematical figure of the Earth", Gauss), may be determined by gravimetric means, starting from the fundamental equation of physical geodesy 15.4. Let us assume the Earth to be a sphere. Then

$$\gamma = \frac{GM}{R^2}$$

and by differentiation

$$\frac{\partial \gamma}{\partial h} = \frac{\partial \gamma}{\partial R} = -\frac{2GM}{R^3},$$

from which follows

$$\Delta g = -\frac{\partial T}{\partial h} - \frac{2}{R}T,$$

an equation valid on the surface of a spherical Earth.

It is intuitively clear that there is some kind of link between variations of gravity, and variations of the geopotential, on the Earth's surface. Both are caused by the uneven distribution of masses inside the Earth. As figure 15.9 shows, the excess masses inside the Earth will cause both an excess in gravity (the level surfaces of the geopotential will be closer together) and a rising of the geoid above the surface of the reference ellipsoid, whereas, on the other hand, the mass deficiencies inside the Earth will lead to both a shortfall in gravity and to a depression of the geoid below the reference ellipsoid.

The *relationship between gravity anomalies Δg and geoid heights N* is however not simple. The quantities are connected by an integral equation, the so-called *Stokes equation*. George Gabriel Stokes[8] derived in 1849 the following classical integral equation[9]:

Stokesin yhtälö

$$N = \frac{R}{4\pi\gamma} \iint S(\psi)\, \Delta g\, d\sigma. \tag{15.5}$$

[8] George Gabriel Stokes (1819 – 1903) was a gifted English mathematician, physicist and geophysicist.

[9] Deriving the equation is difficult and uses the fundamental equation of physical geodesy 15.4 as a *boundary condition* for solving the Laplace field equation in the space exterior to the Earth. See Heiskanen and Moritz (1967, chapter 2).

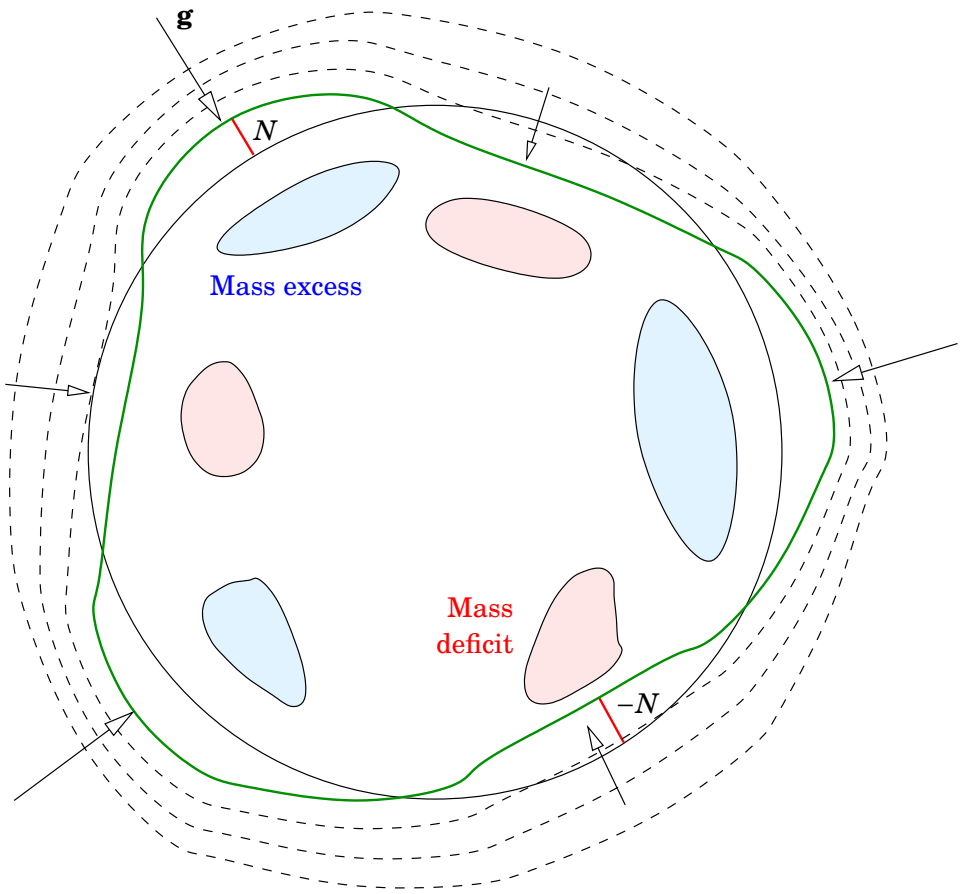


Figure 15.9. Relationship between variations in the Earth's gravity anomalies and those in geoid height.

With the equation, we may compute geoid heights from the global field of gravity anomalies. In the equation, R is the mean radius of the Earth, γ mean gravity on the Earth's surface, and $S(\psi)$ is the *Stokes function*, also called the "Stokes kernel". It depends only on the geocentric angular distance between the evaluation point for the quantity N and the measurement point for the anomaly Δg, i.e., ψ seen from the centre of the Earth:

$$S(\psi) = \frac{1}{\sin\psi/2} - 6\sin\frac{\psi}{2} + 1 - 5\cos\psi - 3\cos\psi\ln\left(\sin\frac{\psi}{2} + \sin^2\frac{\psi}{2}\right).$$

Look closer at equation 15.5. The quantity $d\sigma$ is the so-called *solid-angle element*, a surface element on a sphere of unit radius, in spherical co-ordinates $d\sigma = \cos\phi\, d\phi d\lambda$.

Open up the equation in the following way:

$$N(\phi,\lambda) = \frac{R}{4\pi\gamma}\int_0^{2\pi}\int_{-\pi/2}^{+\pi/2} S\left(\psi\left(\phi,\lambda,\phi',\lambda'\right)\right)\Delta g\left(\phi',\lambda'\right)\cos\phi'\, d\phi' d\lambda'.$$

In this, (ϕ,λ) are the co-ordinates — strictly speaking, the *geocentric* latitude and longitude — of the point in which the geoid height N is

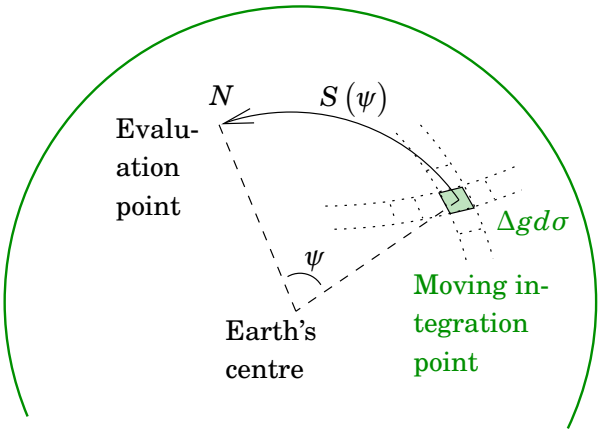


Figure 15.10. The geometry of the Stokes integral equation.

□

being computed. The co-ordinates (ϕ', λ') again are the co-ordinates of the point in which the gravity anomaly Δg is given, and which point traverses the whole Earth's surface with the computation of the double integral. The angle ψ is the angular distance between these two points as seen from the centre of the Earth[10].

We see here that computing even a single value N requires Δg values *from everywhere on the Earth's surface*, in order to evaluate the above integral completely. Divide the Earth's surface into cells, or *blocks*, of size $1° \times 1°$ — in total $360 \times 180 = 64{,}800$ of them — and compute the value of the integral numerically as a sum

$$N(\phi, \lambda) = \sum_{i=1}^{360} \sum_{j=-89}^{+90} \frac{R}{4\pi\gamma} \left(\frac{\pi}{180}\right)^2 \cdot S(\phi, \lambda, \phi', \lambda') \Delta g(\phi', \lambda') \cos\phi',$$

in which $\lambda' = i - 0°{,}5$ and $\phi' = j - 0°{,}5$.

Now one also understands why close international collaboration is so essential to studying the Earth's gravity field!

In practice, the greatest influence on the geoid height is by the *local* gravity anomalies. The function $S(\psi)$ is for small values of ψ approximately

$$S(\psi) \approx \frac{2}{\psi}.$$

---

[10]The formula for calculating the angular distance is

$$\cos\psi = \sin\phi \sin\phi' + \cos\phi \cos\phi' \cos(\lambda' - \lambda),$$

or more precisely for small angles ψ,

$$\sin^2\frac{\psi}{2} = \sin^2\frac{\phi' - \phi}{2} + \cos\phi \cos\phi' \sin^2\frac{\lambda' - \lambda}{2}.$$

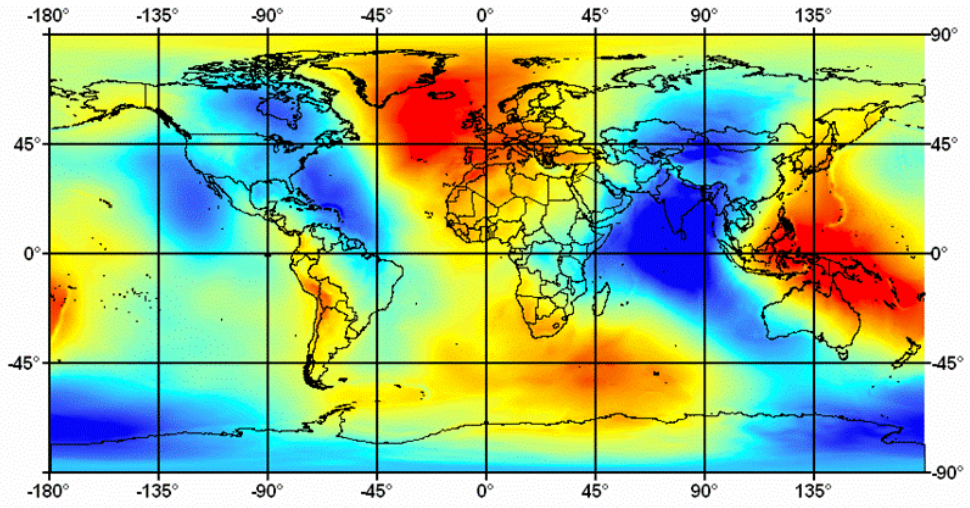


Figure 15.11. The global geoid model EGM2008. Geoid heights from the GRS80 reference ellipsoid −107 m (blue) . . . +86 m (red). © 2013 US National Geospatial-Intelligence Agency.

So, the values of anomalies in the immediate vicinity of the point of evaluation dominate the outcome of the computation. More remote areas have an effect also, but for taking them into account it suffices to use a more low-resolution, so-called *global gravity model* produced by satellite geodesy.

Globally, the geoid deviates from the reference ellipsoid by about ±100 m. The global mean sea surface in its turn follows the geoid, because it is an equipotential or equilibrium surface; the mean sea surface deviates from the geoid at most ±2 m. Of the deviation, the permanent part is called the *sea-surface topography*, see section 1.6. In addition, there are deviations varying in time, like, e.g., the phenomenon of the tides, and the deviations caused by winds and air-pressure variations.

15.5 The gravity field and heights

15.5.1 Geopotential and gradient

In figure 15.3 we see how one can depict the forms of the terrain on a map by means of *height contours*.

The most natural of all measures of height, the *geopotential*, is not a metric height. It is a measure of *energy*, it describes the level of potential energy of a test body (unit mass) in the gravity field of the Earth. For this reason it is a geophysically sensible quantity.

The geopotential is connected to *gravity*, in this way, that the gravity vector, **g**, is the derivative of place, or *gradient*, of the geopotential W,

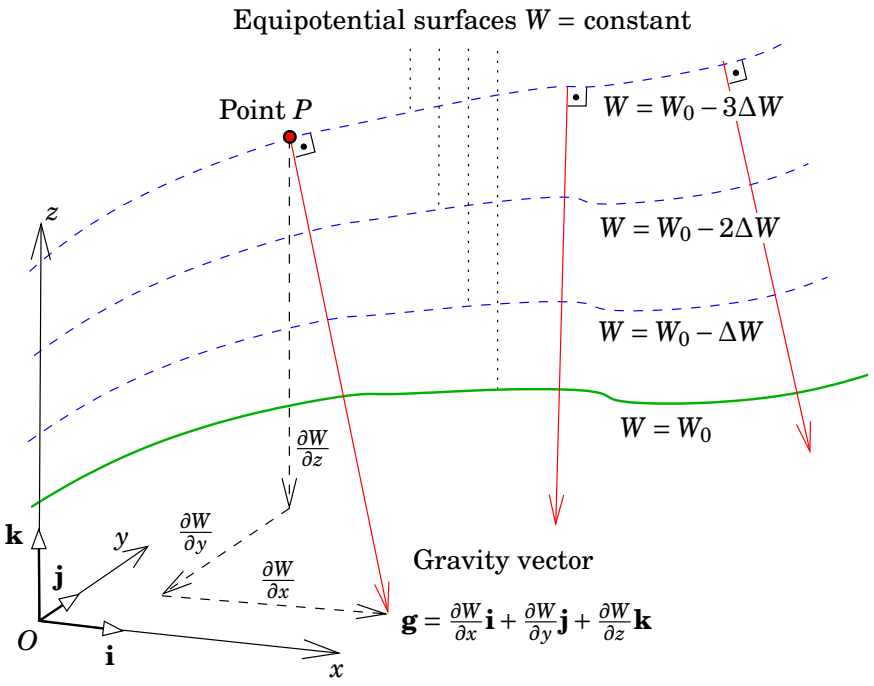


Figure 15.12. The gravity vector is the gradient of the geopotential, the derivative with respect to the three co-ordinates of location.

□

see figure 15.12:

$$\mathbf{g} = \nabla W = \text{grad} W = \frac{\partial W}{\partial x}\mathbf{i} + \frac{\partial W}{\partial y}\mathbf{j} + \frac{\partial W}{\partial z}\mathbf{k}, \tag{15.6}$$

in which again $\mathbf{i},\mathbf{j},\mathbf{k}$ are unit vectors in the directions of the axes of the (x,y,z) co-ordinate frame, forming an orthonormal basis.

Because of this, the local gravity vector is

1. always *perpendicular* to equipotential surfaces (also to sea level!)
2. the larger, the closer to each other the equipotential surfaces are.

The gravity field is a *conservative field*. This means that, when one transports a test mass around a closed path, no net work is done. In a conservative force field, the force vector can always be expressed as the gradient of a potential, in the way depicted in figure 15.12.

The potential difference between points A and B is now the same as the work to be done moving a unit mass from A to B, and the following integral applies (s is the path length along AB):

$$\begin{aligned} W_B - W_A &= \int_A^B dW = \int_A^B \frac{dW}{ds} ds = \\ &= \int_A^B \left(\frac{\partial W}{\partial x} dx + \frac{\partial W}{\partial y} dy + \frac{\partial W}{\partial z} dz \right) = \\ &= \int_A^B \langle \text{grad} W \cdot d\mathbf{x} \rangle = \int_A^B \langle \mathbf{g} \cdot d\mathbf{x} \rangle . \end{aligned} \tag{15.7}$$

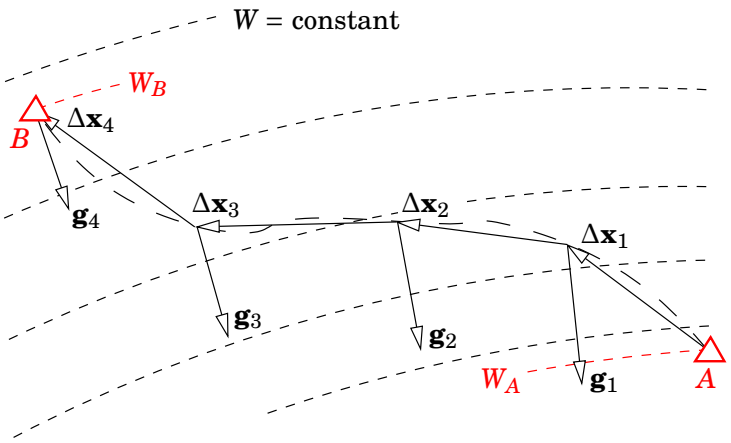


Figure 15.13. The path integral of work. The difference $W_A - W_B = \int_B^A \langle \mathbf{g} \cdot d\mathbf{x} \rangle \approx \sum_{i=1}^{4} \langle \mathbf{g}_i \cdot \Delta\mathbf{x}_i \rangle$.

From this can be seen that the work is the scalar product of the gravity vector $\mathbf{g}$ and the path vector element along the direction of the path $d\mathbf{x} = \begin{bmatrix} dx & dy & dx \end{bmatrix}^{\mathsf{T}}$.

In practice, often instead of the geopotential itself W, its difference $C \overset{\text{def}}{=} -(W - W_0)$ with the geopotential of mean sea level (or some other suitable reference surface) W_0 is used. This potential difference, which grows in the upward direction, is called *geopotential number*[11], and the above integral equation becomes

$$C_B - C_A = -\int_A^B \langle \mathbf{g} \cdot d\mathbf{x} \rangle .$$

In the case of a closed path, we have

$$\oint \langle \mathbf{g} \cdot d\mathbf{x} \rangle = 0.$$

Geopotential numbers are calculated from the measurement results of a *levelling* extending over the country. All metric heights of terrain points, like, e.g., orthometric height, are calculated from their geopotential numbers.

15.5.2 Geopotential unit, GPU

As we saw in section 15.1, gravity is expressed in the SI unit $\mathrm{m/s^2}$. As the measurement unit for the geopotential we use again the *geopotential unit*, or GPU. The SI unit of geopotential is $\mathrm{m^2/s^2}$: distance × $\mathrm{force/mass}$ = distance × acceleration = m × $\mathrm{m/s^2}$. In the gravity field of the Earth close to the surface, where the acceleration of gravity is $g \approx 9.8\,\mathrm{m/s^2}$, a height difference of one metre corresponds to a potential difference of about $9.8\,\mathrm{m^2/s^2}$.

[11] German *geopotentielle Koten*, French *cotes géopotentielles*.

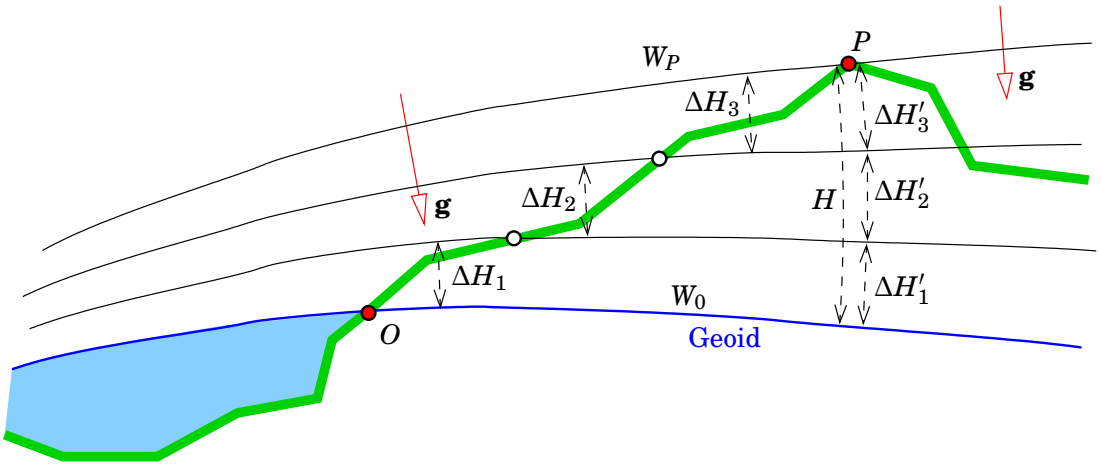


Figure 15.14. Heights and equipotential surfaces. Note that the stronger gravity g (always perpendicular to the equipotential surfaces), the closer together are the equipotential surfaces.

Define

$$1\,\mathrm{GPU} \stackrel{\mathrm{def}}{=} 10\,\mathrm{m^2/s^2},$$

then a height difference of one metre corresponds to a potential difference of 0.98 GPU; similarly a potential difference of 1 GPU corresponds to a height difference of about 1.02 m.

Thus one can, thanks to the fortuitous circumstance that g is close to $10\,\mathrm{m/s^2}$, express geopotential differences in a unit that is slightly more intuitive than the corresponding SI unit!

15.5.3 Orthometric heights

Orthometric heights H are known already. Let's take a closer look.

In figure 15.14 the orthometric height of point P is H. In this simple example it is the sum of three height differences:

$$H = \Delta H_1' + \Delta H_2' + \Delta H_3', \tag{15.8}$$

in which the $\Delta H_i'$ are the separations between the equipotential surfaces on the plumb line of the point.

Levelling, however, yields the height differences $\Delta H_1, \Delta H_2, \Delta H_3$ *on the terrain*, on the Earth's surface between point and coast. In this case a levelling has been carried out from coastal point O, the height of which is assumed zero. Now

$$H \neq \Delta H_1 + \Delta H_2 + \Delta H_3!$$

Height differences obtained from levelling may not just be added together to get the height of a point. This tells us that height, though a metric quantity, isn't a very nicely behaving quantity.

For this reason, in scientific work we always use, instead of metric heights, the already presented *geopotential numbers* $C = -(W - W_0)$.

We may simply write (note that also the units match):

$$\overset{[\mathrm{m^2/s^2}]}{\Delta C} = \overset{[\mathrm{m/s^2}]}{g} \cdot \overset{[\mathrm{m}]}{\Delta H}$$

in which ΔC is the *geopotential difference* (between two arbitrary points), the work that needs to be done in order to transport one unit mass over the height difference ΔH between the points.

If in point O it holds that $W = W_0$, it follows that $C_O = 0$. Then, in the example case of figure 15.14[12], the geopotential number of point P is

$$C = \Delta C_1 + \Delta C_2 + \Delta C_3 = g_1 \Delta H_1 + g_2 \Delta H_2 + g_3 \Delta H_3, \qquad (15.9)$$

which is *computable*, if in connection with the levelling, along the levelling line, also local gravity g is measured.

However, also the following holds:

$$C = g'_1 \Delta H'_1 + g'_2 \Delta H'_2 + g'_3 \Delta H'_3,$$

in which the g'_i are gravity values *inside the rock*, on the plumb line of point P.

Define the mean gravity along the plumb line by the following equation[13]:

$$\overline{g} \overset{\text{def}}{=} \frac{g'_1 \Delta H'_1 + g'_2 \Delta H'_2 + g'_3 \Delta H'_3}{\Delta H'_1 + \Delta H'_2 + \Delta H'_3} = \frac{g'_1 \Delta H'_1 + g'_2 \Delta H'_2 + g'_3 \Delta H'_3}{H}.$$

Now

$$C = \overline{g} H \Longrightarrow H = \frac{C}{\overline{g}},$$

the classical definition formula of orthometric heights. The formula tells that the amount of work needed to move a unit mass from the geoid to point P is force × distance:

$$\overset{\text{work}}{C} = \overset{\text{force}}{\overline{g}} \cdot \overset{\text{distance}}{H}.$$

---

[12] In the general case the equation is

$$C = \int_0^H g(z)\,dz.$$

[13] The general equation is again

$$\overline{g} = \frac{1}{H} \int_0^H g(z)\,dz,$$

where z is the arc length measured along the plumb line.

We are left with the problem of determining $\overline{g}$, the average of gravity along the plumb line. Measuring values g'_i inside the Earth's crust is usually impossible... therefore, in practice the determination is based on the value g_P measured on the Earth's surface, by *assuming* that gravity grows going downward, inside the Earth's crust, according to a certain formula[14]. In this way, an approximate value for the orthometric height is obtained, the accuracy of which at least for the Finnish territory is totally adequate.

Orthometric height is but one way of building a metric height system. There are other ways, like normal height and dynamic height. All are heights "above sea level", but the ways of definition and calculation are a bit different. And all three have their own pluses and minuses.

Orthometric heights are not without their problems. The tunnel network does not exist, and measuring gravity inside the rock — along the local plumb line — is normally not possible. In practice, orthometric heights are determined with the aid of *levelling*, starting from the coast, along the Earth's surface. If one wants to calculate orthometric heights from a levelling, we need for the calculations unfortunately very detailed data on

- the density of the rock below the height point
- the forms of the terrain around the height point, i.e., a *terrain model*.

So, even if orthometric heights are physically elegant, their precise determination may in practice be troublesome. Scientifically one says, that orthometric heights *are not hypothesis free*. The hypotheses required are precisely the density of the Earth's crust and the terrain's local forms. In practical computation often the effect of the terrain is omitted, and density values are taken from geological maps. The error thus made is usually small.

15.5.4 Normal heights

Normal heights H^* are, simply stated, orthometric heights computed from geopotential numbers C *as if* the true gravity field of the Earth *were* a regular, ellipsoid-of-revolution like mathematical model field, i.e., the *normal gravity field*. Therefore no information is needed related to the *true*, complicated gravity field. Normal heights are computed easily and precisely without any knowledge of local rock density or terrain models.

The equation for normal height is

$$H^* = \frac{C}{\overline{\gamma}},$$

[14] E.g., Poincaré – Prey reduction, see Heiskanen and Moritz (1967).

in which $\overline{\gamma}$ is the average of *normal gravity*, calculated again along the plumb line of the point[15].

However, unlike orthometric heights, normal heights have no direct physical interpretation.

On the Finnish territory the differences between orthometric and normal heights are of order millimetres. In the mountains they can easily be several decimetres.

In many countries — e.g., Russia, Sweden, nowadays also Finland — normal heights are used instead of orthometric heights. Their precise calculation is easier. Whereas orthometric heights are interpreted as heights from the geoid, normal heights are reckoned from a similar surface called the *quasi-geoid*. On the sea it coincides with the geoid and thus also with mean sea level, but under the land, and especially under the mountains, it differs from the geoid. The above proposed "tunnel-network metaphor" for letting in sea water doesn't work for the quasi-geoid.

15.5.5 Dynamic heights

Dynamic heights are rarely used. The are calculated simply by dividing the geopotential number C by the normal gravity at zero height and at latitude 45°, γ_{45}, which is a *constant*:

$$H^{\mathrm{dyn}} = \frac{C}{\gamma_{45}}.$$

15.5.6 Properties of different height types

Common among all height types is, that a metric height is obtained by *dividing* the geopotential number by some suitable *gravity value*: the unit of geopotential numbers is $\mathrm{m^2/s^2}$, the unit of acceleration of gravity is $\mathrm{m/s^2}$, so the unit of metric height is indeed $\frac{\mathrm{m^2/s^2}}{\mathrm{m/s^2}} = \mathrm{m}$, as it should be.

Independently of height type, all metric heights are computed from the *energy level* of the point, the forementioned *geopotential number* C. The only "heights" that can be measured and computed precisely, are the geopotential numbers $C = -(W - W_0)$. All other heights are *derived quantities*. In their computation are lost — besides precision — always some useful properties, just like when projecting a curved surface onto a flat one.

The user of height values desires from practical heights a number of good things, which are familiar from geometric heights within a small area:

[15] Actually the truth is more complicated: into the normal gravity equation $\gamma(h, \varphi)$ the height H above the *geoid* is substituted, although the equation asks for the height h from the reference ellipsoid. The equation values obtained for heights from zero to H^* are averaged.

Table 15.2. Properties of various height types.

Height type	Correctness		Hypothesis freeness	Equation
	Metric	Energetic		
Geopotential number	−−	+	+	C
Orthometric	+	−	−	$H = C/\overline{g}$
Normal	−	−	+	$H^* = C/\overline{\gamma}$
Dynamic	−−	+	+	$H^{\mathrm{dyn}} = C/\gamma_{45}$

- **Metric correctness**

  Metric correctness means that, if there are two points P and Q straight above each other, and the distance between them is 1 m, then also $H_P - H_Q$ is precisely 1 m. Only *orthometric heights* have this property. The metric correctness of dynamic heights is especially weak.

  Metric correctness is the better, the closer to the *true* mean gravity along the plumb line the expression in the denominator of the formula is.

- **Energetic correctness**

  This means that water always flows "down" in the sense of the height type in question. Of the three types mentioned, only *dynamic heights* are energetically correct — by virtue of their direct proportionality to geopotential numbers C.

- **Precise computability, non-dependence on uncertain hypotheses**

  Normal heights and dynamic heights may be precisely computed based on theory. With normal heights, it has however to be stated *which* normal field, or reference ellipsoid, has been chosen for the computations.

  Orthometric heights require knowledge both of the true gravity field and of the form and density of the topography. In practice, however, the uncertainty caused by these factors is fairly small.

15.6 Bouguer anomalies

Earlier we noted that free-air anomalies — equation 15.3 — tell us something about the interior mass distribution of the Earth. However, in the free-air anomaly Δg of a point is also along the effect of the whole topography underneath and around the point. The forms of the topography above sea level are visible and usually well known. Therefore it would seem logical to *remove* — computationally — the effect of the terrain forms from the free-air anomalies, in order to obtain a quantity that

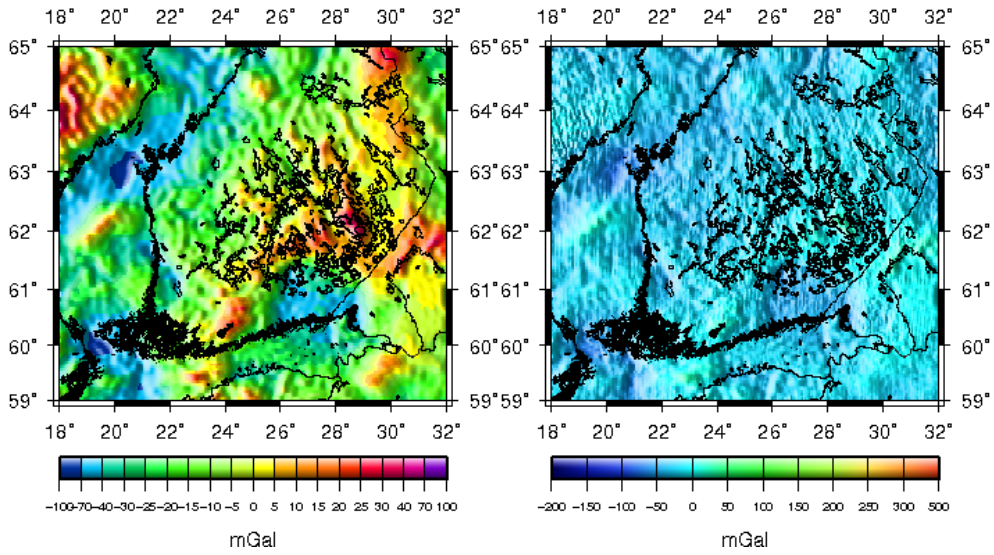


Figure 15.15. Free-air and Bouguer anomalies for Southern Finland computed from the EGM2008 geopotential model. Data © Bureau Gravimétrique International (BGI) / International Association of Geodesy.

tells us only about the mass distribution of the Earth below sea level.

This is how the *Bouguer anomaly* is obtained:

$$\Delta g_{\mathrm{B}} = \Delta g_{\mathrm{FA}} - g_{\mathrm{top}},$$

in which $\Delta g_{\mathrm{FA}} \overset{\mathrm{def}}{=} \Delta g$ is the free-air anomaly, Δg_{B} the Bouguer anomaly and g_{top} the vertical component of the attraction of the topography acting in the point.

Bouguer anomalies may be calculated precisely or approximately. In the first case we use a numerical model of the topography, a so-called digital terrain model (DTM). We also use a density model for the Earth's crust if one exists. In the approximate calculation, er only take into account the effect of the so-called *Bouguer plate*, as a simple closed formula:

$$g_{\mathrm{top}} = 2\pi G \rho d,$$

in which

g_{top} attraction of the plate, only in the vertical direction

G Newton's universal gravitational constant, see section 1.2

ρ density of the matter of the plate

d thickness of the plate.

If the density is $\rho = 2.67\,\mathrm{g/cm^3} = 2670\,\mathrm{kg/m^3}$, we obtain

$$g_{\mathrm{top}} = 0.1119 d,$$

in which d is in metres and g_{top} in milligals. The unit milligal was explained at the beginning of the chapter.

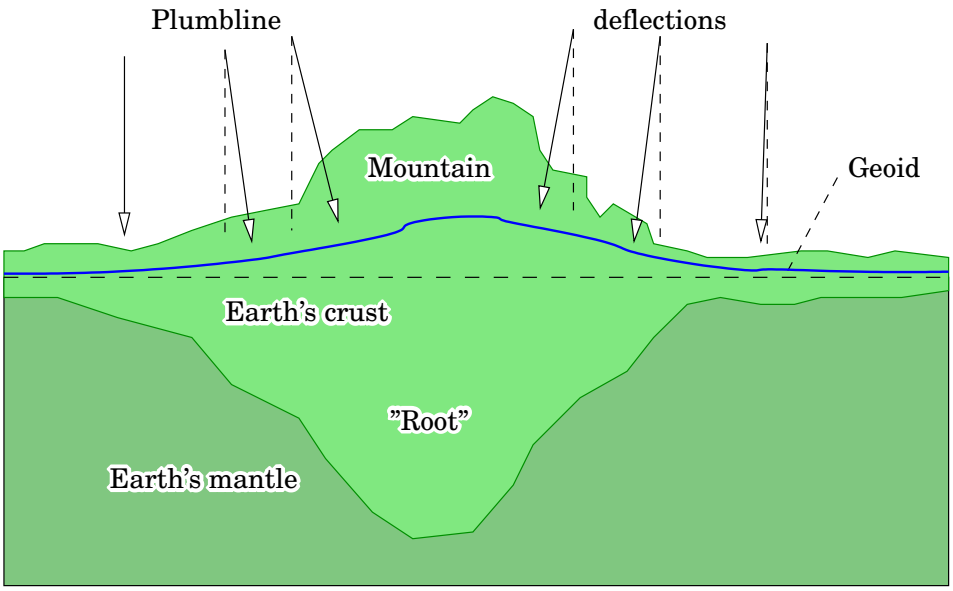


Figure 15.16. The root of a mountain range and its effect on the plumb line.

15.7 Astronomical position determination

The local plumb line, i.e., the direction of the local gravity vector, stands perpendicular upon the equipotential surface. Determining the direction of the plumb line in an absolute sense has been possible by traditional astronomical means.

One speaks of *astronomical position determination*, because the first practical application of the method was determining an unknown position, e.g., at sea. Later, the method was used to study the geophysically interesting variations (deviations) of the plumb-line direction on land. The already mentioned Pierre Bouguer noticed as the first in South America — like George Everest in India — that in the vicinities of mountain ranges the plumb line is deflected toward the mountain range, and interpreted this correctly as caused by the mountains' own gravitation.

The attempt to estimate the effect of the mass of the mountains, however, produced a result that was much larger than the actually observed plumb line deviations. The reason for this is today known to be *isostatic compensation*: under the mountains there is a *root* consisting of lighter rock, that keeps the mountain "afloat" on the plastically deforming Earth's mantle.

For astronomical position determination, various instruments are used like the *meridian circle*, the *astrolabe*, or the *zenith tube*.

The optical axis of an astrolabe points always upward by a fixed angle. The optical axis of a zenith tube again points always upward vertically, to the zenith, under an elevation angle $\eta = 90°$. Therefore, by observing the passage through the zenith of stars of which the declination δ is known, the astronomical latitude of the location is obtained by the for-

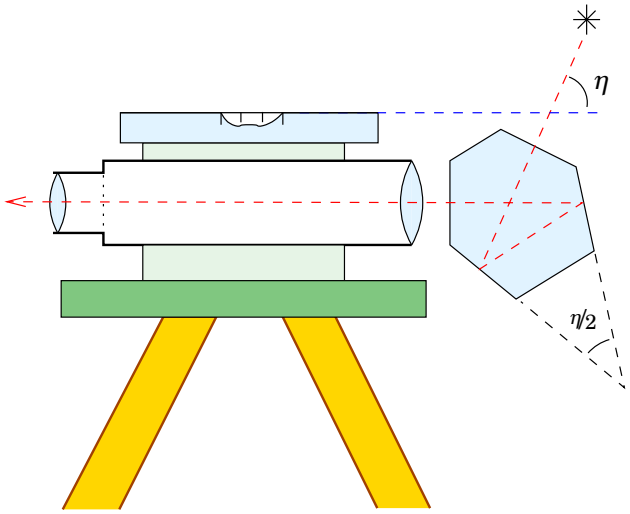


Figure 15.17. A levelling instrument converted to astrolabe.

mula $\Phi = \delta$. At the same time is measured the time of transit through the meridian, from which is obtained the longitude Λ — because the zenith lies in the plane of the meridian. A zenith tube is in a way both a meridian circle and an astrolabe.

In preparation for observations, a *star programme* is drafted, a list of stars that will transit the meridian, a certain elevation circle[16], or the zenith. In case of the astrolabe, one should take care that the stars are distributed evenly around the whole horizon. In this way, a precise determination of both Φ and Λ is achieved, and a minimization of the influence of atmospheric refraction.

An astrolabe may be built easily from a levelling instrument, by adding a sixty-degrees angle prism in front of the objective, see figure 15.17. The most precise (Danjon) astrolabes again use a mercury mirror together with a sixty-degree prism.

15.8 Measuring the gravity gradient

We have already a number of times spoken about the *potential* of gravity, which is a measure of the *potential energy* (energy content of location) of a test mass inside the gravity field. The acceleration vector of gravity, or free fall, is defined as the *gradient* of this geopotential W, its rate of change with place:

$$\mathbf{g} = g_x\mathbf{i} + g_y\mathbf{j} + g_z\mathbf{k} = \frac{\partial W}{\partial x}\mathbf{i} + \frac{\partial W}{\partial y}\mathbf{j} + \frac{\partial W}{\partial z}\mathbf{k},$$

i.e.,

$$\mathbf{g} = \operatorname{grad} W = \nabla W,$$

[16] In Arabic *almucantar*.

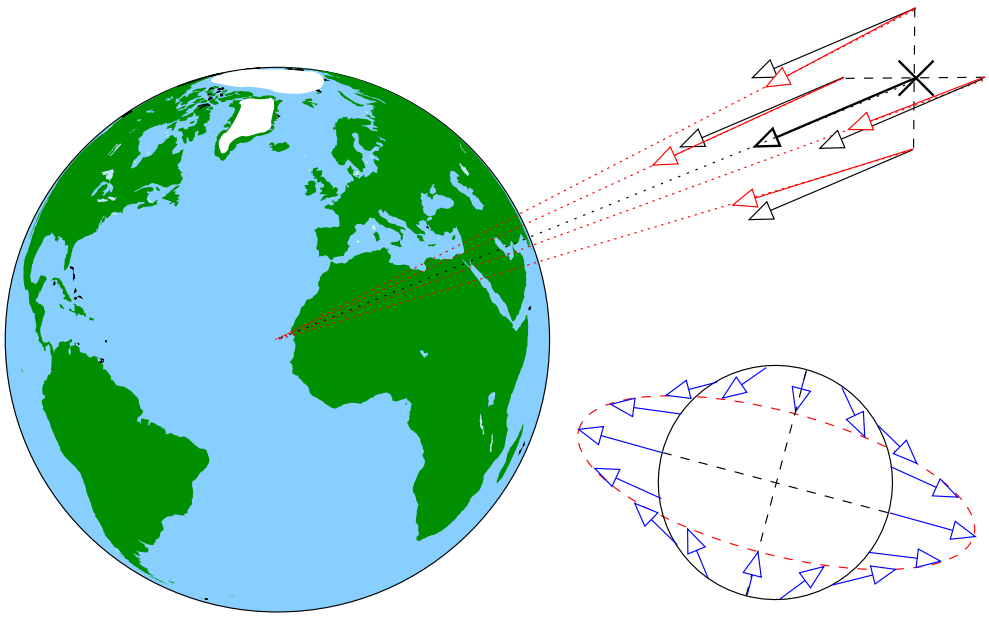


Figure 15.18. The gravity-gradient or tidal force field: gravity varies with place. The lower diagram gives the *visual ellipsoid* of the gravity-gradient tensor: it depicts the force field inside a freely falling object, e.g., a satellite.

where the grad *operator* is

$$\text{grad}\cdot = \nabla\cdot = \frac{\partial\cdot}{\partial x}\mathbf{i} + \frac{\partial\cdot}{\partial y}\mathbf{j} + \frac{\partial\cdot}{\partial z}\mathbf{k}.$$

Here $\{\mathbf{i},\mathbf{j},\mathbf{k}\}$ is an orthogonal triad of unit vectors, or orthonormal basis, oriented along the three axes of the (x,y,z) co-ordinate frame.

The acceleration vector of gravity is thus the *gradient of the geopotential*. This vectorial quantity is location dependent. We know that gravity grows going downward, at least in the free air. Going up, gravity diminishes, some 0.3 mGal for every metre in height.

Gravity as a vector varies in a more complicated way in the vicinity of masses, if one has the use of sufficiently accurate measurement devices. We talk of the *gravity-gradient tensor*, or Eötvös tensor:

$$M \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial^2}{\partial x^2} & \frac{\partial^2}{\partial x \partial y} & \frac{\partial^2}{\partial x \partial z} \\ \frac{\partial^2}{\partial y \partial x} & \frac{\partial^2}{\partial y^2} & \frac{\partial^2}{\partial y \partial z} \\ \frac{\partial^2}{\partial z \partial x} & \frac{\partial^2}{\partial z \partial y} & \frac{\partial^2}{\partial z^2} \end{bmatrix} W.$$

In a topocentric co-ordinate frame, where x points North, y East and z up, this matrix has the following approximate form:

$$M \approx \begin{bmatrix} -0.15 & 0 & 0 \\ 0 & -0.15 & 0 \\ 0 & 0 & 0.3 \end{bmatrix} \text{mGal/m}, \tag{15.10}$$

in which

$$\frac{\partial^2 W}{\partial z^2} = \frac{\partial}{\partial z} g_z = 0.3\,\mathrm{mGal/m}$$

is truly the so-called "free-air" standard gravity gradient, i.e., the vertical gradient of normal gravity: if we write according to Newton (note that the direction of $\mathbf{g}$ is down whereas the z co-ordinate grows going up, that's why the minus sign):

$$g_z = -\frac{GM}{(R+z)^2},$$

we obtain by taking the derivative

$$\frac{\partial}{\partial z} g_z = 2\frac{GM}{(R+z)^3}\frac{\partial (R+z)}{\partial z} = -\frac{2g_z}{(R+z)} \approx 3 \cdot 10^{-6}\,\mathrm{m/s^2/m} = 0.3\,\mathrm{mGal/m}.$$

The quantities $\frac{\partial^2 W}{\partial x^2}$ and $\frac{\partial^2 W}{\partial y^2}$ again describe the *curvatures* of the equipotential surfaces in the x and y directions, in the following geometrically intuitive way:

$$\frac{\partial^2 W}{\partial x^2} = -\frac{g}{r_x} \text{and} \frac{\partial^2 W}{\partial y^2} = -\frac{g}{r_y},$$

in which r_x and r_y are the radii of curvature in the x and y directions. Substituting $r_x = r_y = R \approx 6378\,\mathrm{km}$ (try!) gives

$$\frac{\partial^2 W}{\partial x^2} = \frac{\partial^2 W}{\partial y^2} = -1.5 \cdot 10^{-6}\,\mathrm{m/s^2/m} = -0.15\,\mathrm{mGal/m}.$$

The already mentioned researcher baron Loránd Eötvös did many ingenious experiments (Eötvös, 1998) in order to measure the components of the gravity-gradient tensor with the *torsion balance* built by him. The method continues in widespead use in geophysical research, because the gravity gradient as an observable is very sensitive to local variations in the density of the Earth's crust.

In the general case we can evaluate the gravity-gradient tensor by performing the partial differentiations. We do so for a central force field:

$$W = \frac{GM}{r},$$

where $r = \sqrt{X^2 + Y^2 + Z^2}$ is the distance from the geocentre. The co-

ordinates (X,Y,Z) are now geocentric. We obtain

$$\begin{aligned} M &= \begin{bmatrix} \frac{\partial}{\partial X} & \frac{\partial}{\partial Y} & \frac{\partial}{\partial Z} \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial X} \\ \frac{\partial}{\partial Y} \\ \frac{\partial}{\partial Z} \end{bmatrix} \frac{GM}{r} \\ &= \begin{bmatrix} \frac{\partial}{\partial X} & \frac{\partial}{\partial Y} & \frac{\partial}{\partial Z} \end{bmatrix} \begin{bmatrix} -\frac{GM}{r^3}X \\ -\frac{GM}{r^3}Y \\ -\frac{GM}{r^3}Z \end{bmatrix} \\ &= \frac{GM}{r^5} \begin{bmatrix} 3X^2-r^2 & 3XY & 3XZ \\ 3YX & 3Z^2-r^2 & 3YZ \\ 3ZX & 3ZY & 3Z^2-r^2 \end{bmatrix}. \end{aligned}$$

The Marussi tensor is the partial derivatives matrix of the gravity acceleration vector

$$\mathbf{g} = -\frac{GM}{r^3} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

with respect to place, as can be seen in the above equation:

$$M = \begin{bmatrix} \frac{\partial}{\partial X} & \frac{\partial}{\partial Y} & \frac{\partial}{\partial Z} \end{bmatrix} \begin{bmatrix} -\frac{GM}{r^3}X \\ -\frac{GM}{r^3}Y \\ -\frac{GM}{r^3}Z \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial X} & \frac{\partial}{\partial Y} & \frac{\partial}{\partial Z} \end{bmatrix} \mathbf{g}.$$

In honour of Eötvös we use as the unit of gravity gradient the Eötvös, symbol E:

$$1\,\mathrm{E} = 10^{-9}\,{}^{\mathrm{m/s^2}}\!/_{\mathrm{m}} = 10^{-4}\,{}^{\mathrm{mGal}}\!/_{\mathrm{m}}.$$

The tensor given above close to the Earth's surface 15.10 is now

$$M \approx \begin{bmatrix} -1500 & 0 & 0 \\ 0 & -1500 & 0 \\ 0 & 0 & 3000 \end{bmatrix} \mathrm{E}.$$

Note that in every case

$$\frac{\partial^2 W}{\partial x^2} + \frac{\partial^2 W}{\partial y^2} + \frac{\partial^2 W}{\partial z^2} = 0.$$

This condition, the so-called *Laplace field equation*[17], applies more generally for a gravity potential in vacuum[18].

The gravity gradients of Sun and Moon are known on the Earth's surface as the *tidal force field*, which causes the phenomenon of the ocean tides, with a dominant periodicity — due to the Earth's rotation — of twice a day.

[17] Pierre-Simon Laplace (1749 – 1827) was a French mathematician, physicist and astronomer, one of the 72 names on the Eiffel Tower, and surely the most gifted of them all.

[18] Here we have not considered, besides the atmosphere, also the centrifugal force of the Earth's rotation, which causes the Laplace field equation to not be valid in a co-rotating co-ordinate frame, not even in vacuum. The correction term can however be precisely calculated.

16. Space geodesy

16.1 Earth rotation, orbital motion, sidereal time

The Earth orbits the Sun in 365.25 days. She also rotates around her own axis in a day. See figure 16.1, where is shown both the physical situation, the orbit of planet Earth around the Sun, and the apparent situation, the annual path of the Sun along the zodiac or ecliptic.

During an exact civil day, 24 hours, the Earth turns once around her axis *with respect to the mean Sun*. With respect to the stars, however, the rotation period is a little shorter: when the Earth has, in a year, turned 365.25 times around her axis with respect to the Sun, she has, with respect to the stars, completed 366.25 rotations. Thus, the duration of one rotation is

$$T = \frac{365.25 \cdot 24^{\mathrm{h}}}{366.25} = 23^{\mathrm{h}}56^{\mathrm{m}}4^{\mathrm{s}}.$$

The name of this period is *sidereal day*.

The rotation of the Earth with respect to the stars (or, equivalently, the apparent rotation of the stars with respect to the Earth) is measured by an angle called *sidereal time*. It is calculated from clock time and calendar date using tables drafted for this purpose. In astronomical observatories, also *sidereal clocks* are used that show sidereal time, and run about 1/365.25 part or 0.27% faster than ordinary clocks.

Due to the annual motion of the Earth, the constellations that are visible in the evening hours shift slowly forward along with the season: every season has its own distinctive constellations visible in evening twilight, in winter, Orion, the stars Sirius and Procyon, in summer, the constellations Lyre, the Swan and the Eagle.

Seen from Earth, the Sun travels along a yearly path, the zodiac ("ring of beasts") or *ecliptic*[1]. At the beginning of spring, the Sun moves from the Southern hemisphere to the Northern one, in a point called the *vernal* or *spring equinox*. In the time of the ancient Greeks this point was

[1]The name "ecliptic" originates from the fact that this is where solar and lunar eclipses happen. Of course, because both types of eclipse require Sun, Moon and Earth to be on the same line.

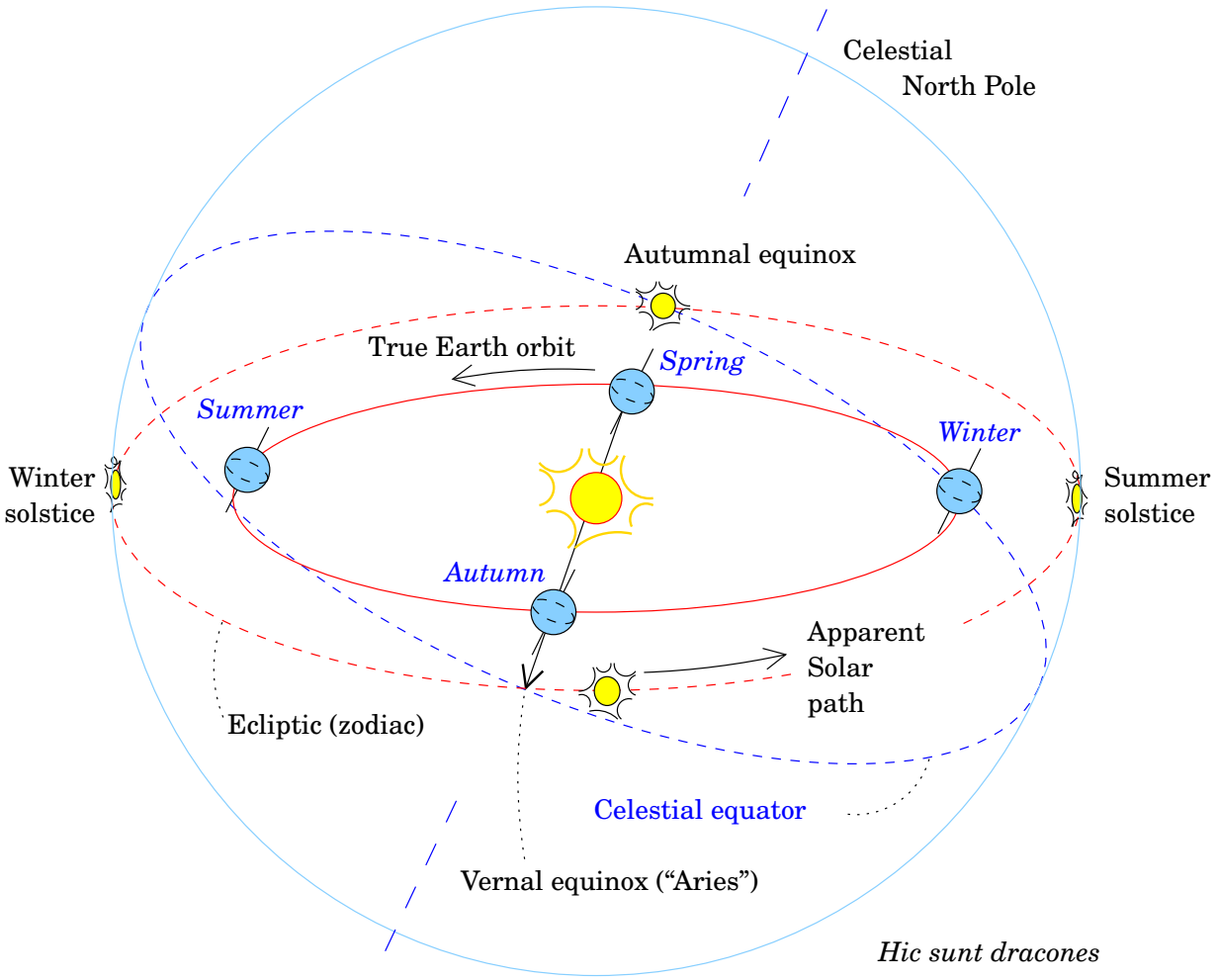


Figure 16.1. The orbit of the Earth around the Sun, and the apparent path of the Sun across the celestial sphere. Season names are boreal, i.e., refer to the Northern hemisphere.

in the constellation of the Ram, therefore the traditional name “First Point of Aries” — though, due to precession, it is nowadays located in the constellation of Pisces, the Fishes. In the same way, back then, the Sun was at the time of the summer solstice in the constellation of Cancer, the Crab, and at the time of the winter solstice in the constellation of Capricorn (a mythological goat-like creature), and the constellations gave their names to the *tropics*. Due to precession, today’s solstices happen in the constellations of Taurus (the Bull) and Sagittarius (the Archer) . . .

The absolute orientation of the whole globe with respect to the stars is described by Greenwich sidereal time (*GAST,* Greenwich apparent sidereal time). Local sidereal time (*LAST*, Local apparent sidereal time) is obtained using astronomical longitude:

$$LAST = GAST + \Lambda,$$

in which Λ is the longitude of the site reckoned *East*, and of course con-

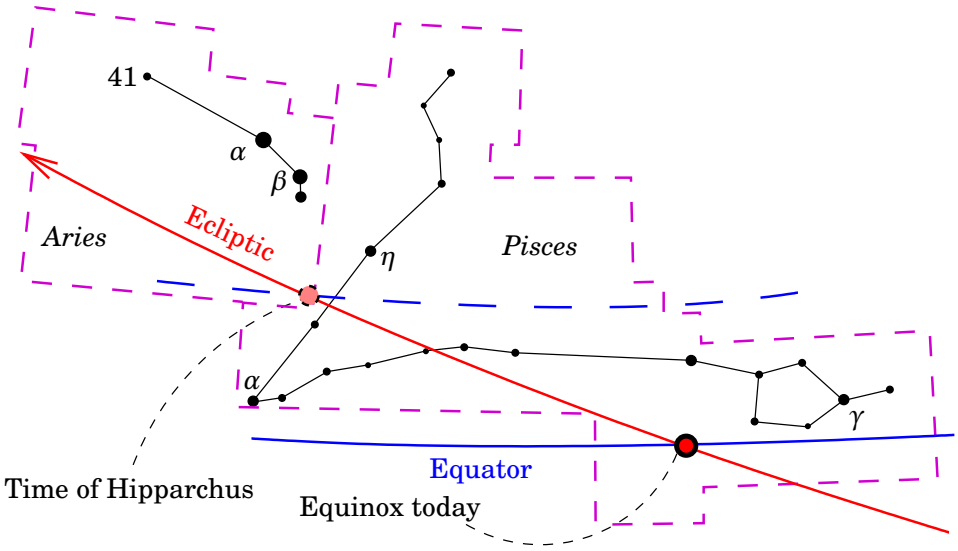


Figure 16.2. The vernal equinox and its movement, the *precession* of the equinoxes.

verted to time units ($1^h \triangleright 15°$, $1^m \triangleright 0°.25$, etc.).

16.2 Heavenly and Earthly co-ordinates

Celestial co-ordinates are *right ascension* α and *declination* δ. The right ascension is a longitude, measured however Eastward from the vernal equinox, the right ascension of which is thus 0.

In the local sky, however, we use the momentaneous co-ordinates *hour angle* h and declination δ. See figure 16.3.

The pair (h,δ) can be directly computed from azimuth and elevation (A,η), if the local astronomical latitude Φ is known.

During a sidereal day, a star moves apparently from East to West, and passes through the meridian plane two times: the so-called *upper and lower culmination*. The lower culmination remains unseen if the declination of the star is too small compared to the local latitude. Of the Southern constellations, even the upper culmination remains unseen by us!

The following relationship exists between sidereal time θ ($= LAST$), the hour angle h, and the right ascension α of a star:

$$h = \theta - \alpha.$$

If we write for Greenwich sidereal time ($GAST$), θ_0 and for the Eastern astronomical longitude of the observation site, Λ, we obtain

$$h = \theta_0 + \Lambda - \alpha.$$

If, of the four quantities, three are given, the fourth can be calculated. We speak of

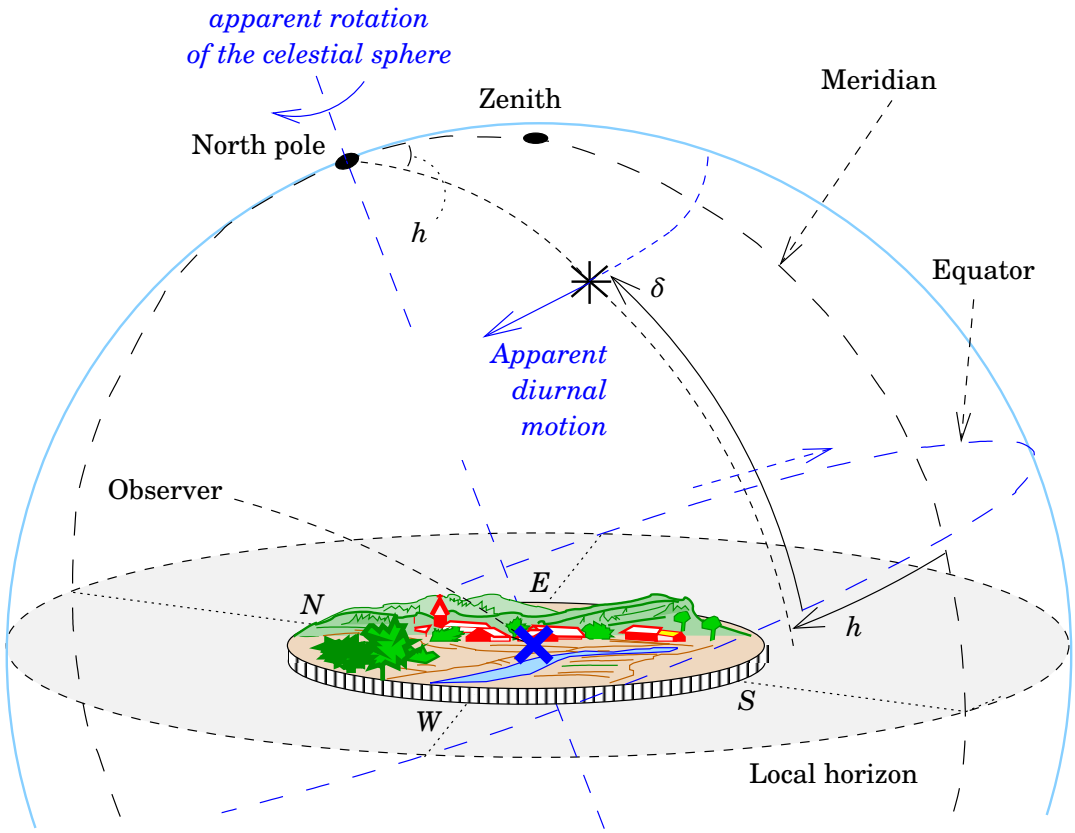


Figure 16.3. Hour angle h and declination δ on the celestial sphere.

1. *time determination*, if the unknown is θ_0. Back in time this was a service of astronomy to society, the maintenance of civil time. Today, with clocks being so much more precise, it is about monitoring Earth rotation.
2. *longitude determination* of the site, if the unknown is Λ. This was critical for navigation at sea (Sobel, 1998).
3. determination of the *right ascension* of a star, if the unknown is α. This is how star catalogues are constructed.

In these cases the measurement is generally done at the moment when $h = 0$: the *meridian transit*. Cases 2 and 3 require use of a precise clock; case 1 requires knowledge of one's own longitude and the right ascension of the star used.

meridiaanin läpikulku

An often used device is the *meridian circle*, used to precisely time the transit of a star through the meridian. A meridian circle is a telescope having only a horizontal axis, which has been built fixed into the East-West direction, in such a way that the sight axis of the telescope will always be in the local meridian plane. In the eyepiece one observes how the star moves underneath the crosshair. The precise moment is recorded electrically together with time signals.

A graduated circle is attached to the axis, which allows the reading of

the elevation or height angle η at the moment of transit. From this, the declination δ of the star can be calculated: between it, the latitude of the site Φ, and the elevation angle η exists a relationship (for the case of upper culmination)

$$\eta + \Delta\eta_{\text{refr}} = \left(90^\circ - \Phi\right) + \delta,$$

in which $\Delta\eta_{\text{refr}}$ is the correction for atmospheric refraction. The refraction correction must be made carefully based on local measurements of air pressure, temperature and humidity aimed at determining the index of refraction of the local air.

By combining the methods 1–3 above one can build catalogues of both the places of stars (α, δ) and of observation stations (Φ, Λ), and at the same time monitor the progress of sidereal time — i.e., the rotation of the Earth — θ_0. However, all this must be done in a consistent way.

Sobel (1998) is a fine book on the role of time in position determination at sea.

16.3 Väisälä's stellar triangulation

This method was invented by Yrjö Väisälä during the Second World War, while watching anti-aircraft shells exploding over Turku. In his 1946 article "*Maan toinen kuu*" ("*A second moon of the Earth*") (Väisälä, 1946) he described how targets high up in the sky could be photographed from different places on the Earth's surface, and thus a geodetic network built even connecting points between which there was no direct line of sight.

The method uses meteorological sounding balloons carrying powerful flashtubes. The flash train is photographed against the stellar background, the images are developed and the places of the flashes in among the stars are measured and calculated. Thus is obtained the momentaneous direction vector between photography site and balloon, in astronomical co-ordinates (Kakkuri, 1973). This balloon method has been tried in a production setting in Finland and Hungary (Czobor and Németh, 1981). The side lengths of the network were several hundreds of kilometres.

In short: from *two observation sites* A, B on the Earth's surface *two high signals* S, T visible in the sky are photographed against the stellar background. As cameras are used powerful so-called Schmidt-Väisälä reflector telescopes[2]. A glass image plate covered by light-sensitive emulsion is placed into the image plane of every camera.

After development, the image plates[3] are measured to extract, in an accurate measurement device, the places, in "plate co-ordinates", of both

[2] http://en.wikipedia.org/wiki/Schmidt_camera.

[3] Back then, glass plates were used instead of films for their geometric stability.

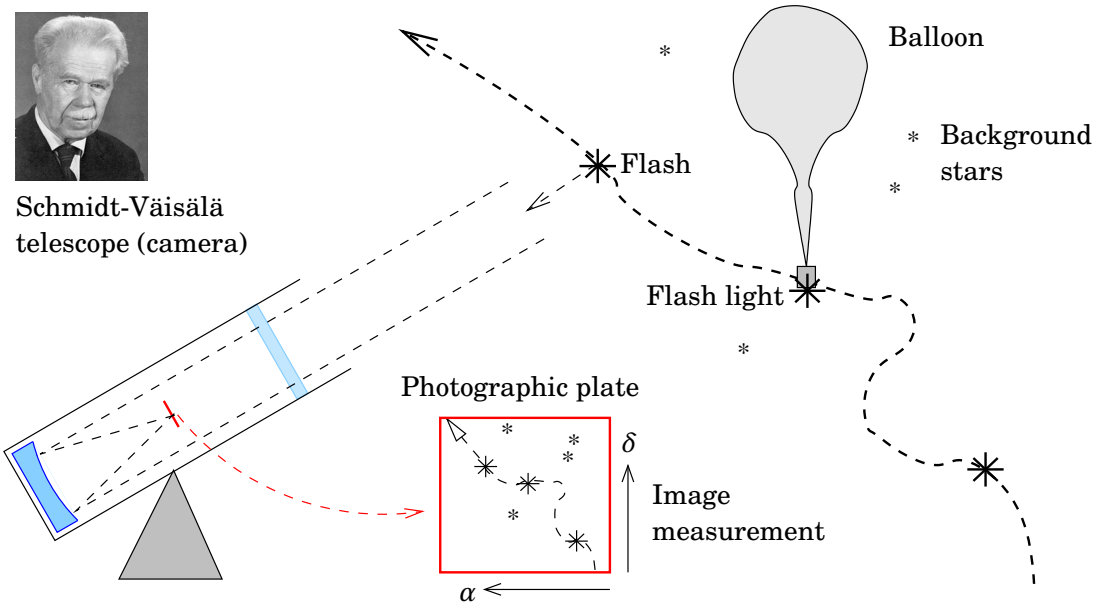


Figure 16.4. Yrjö Väisälä's stellar triangulation. Equipment used and principle of photography and direction determination. **(a)** Equipment used in the measurement.

the signals — the flashes — and chosen background stars. Because the places of fixed stars in celestial co-ordinates, right ascension (α) and declination (δ), are already known, we may derive for both plates a transformation formula, which will yield also the places of the *signals* or flashes in the sky (α, δ), seen from both observation sites.

These "places" on the celestial sphere are in reality directions in three-dimensional space. From the directions are formed planes $P = SAB$ and $Q = TAB$. The intersection line of the planes is AB. Thus has been obtained the direction vector from A to B in three-dimensional space.

In fact, the plane P is on the celestial sphere the same as the great circle though the image points of S taken in A and B, and Q is the great circle through the images of T. The intersection point of the great circles is the *direction* in space, with respect to the stars, of connecting line AB: stellar triangulation is *direction measurement* using heavenly auxiliary points.

When a sufficient number of directions between ground stations has been collected, a *network adjustment* can be carried out. Transforming the directions from the celestial co-ordinate frame to one co-rotating with the Earth, using a model of the Earth's rotation ("Greenwich sidereal time"!) requires that the flashes are accurately time tagged.

The places of the flashes are measured in plate co-ordinates, like also the places of the fixed stars captured on the plate. Using the known places of the fixed stars, the places of the flashes are solved in celestial co-ordinates (α, δ).

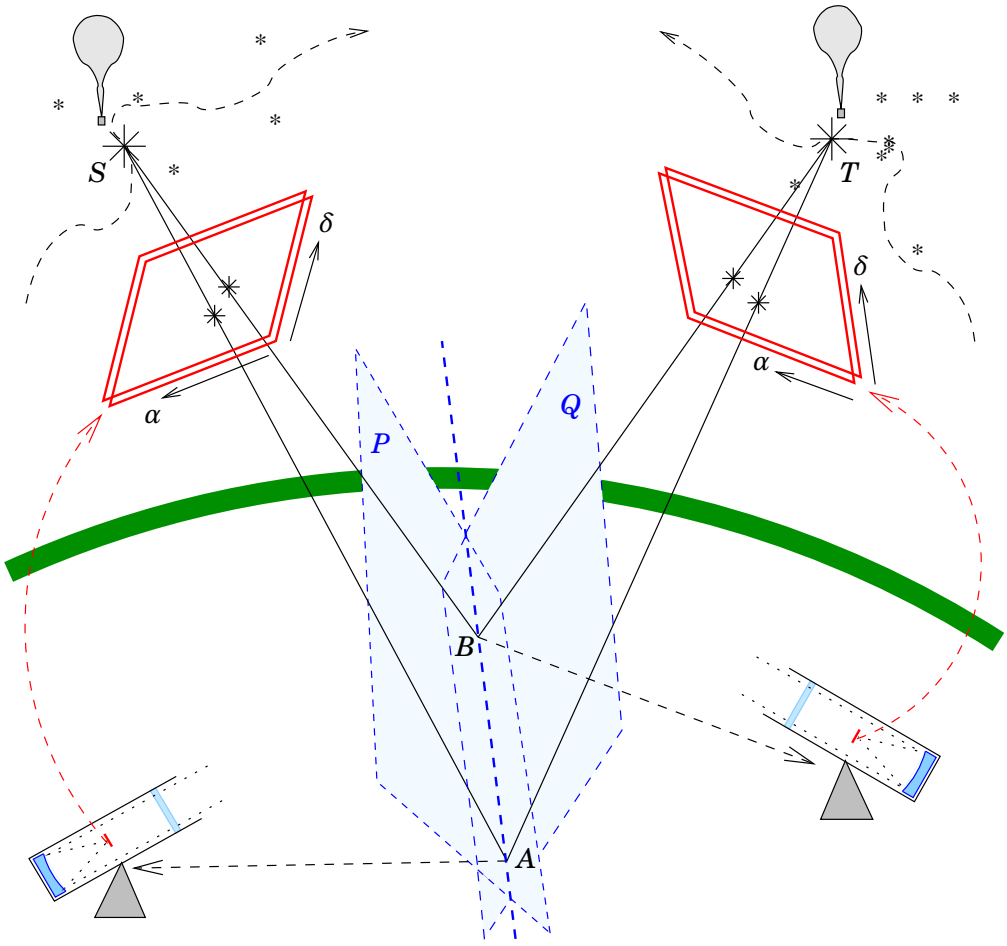


Figure 16.5. Yrjö Väisälä's stellar triangulation. Equipment used and principle of photography and direction determination. **(b)** Measurement geometry.

□

Väisälä's idea was, as may already be inferred from the name of his article, to use, instead of stratospheric balloons, artificial satellites orbiting the Earth. Also this method has been tried: the active satellite ANNA (Army-Navy-NASA geodetic satellite) from 1962 was equipped with powerful flashtubes. On the other hand, the passive satellite PAGEOS (Passive Geodetic Earth Orbiting Satellite) from 1966 was a balloon, over 30 metres in diameter, made of very thin, aluminized Mylar® film, in a 4000 km high orbit. She was clearly visible to the naked eye, until she disintegrated in the 1970s, possibly due to the corrosive effect of the Sun's ultraviolet rays[4].

Using the global network of twelve massive Baker-Nunn[5] cameras, it

[4]DuPont write in the document *Mylar® polyester film — Safe Handling*: "Mylar® is not recommended for applications requiring prolonged exposure to direct sunlight due to degradation when exposed to ultraviolet rays [...]" .

[5]The Baker-Nunn camera was a variant of the Schmidt camera, which was optimized for observing satellites. It used 55 mm broad Cinemascope motion-

(a) PAGEOS satellite, test inflation (NASA)

(b) Baker-Nunn camera.

Figure 16.6. Satellite geodesy from the photographic archives of NASA and Boller and Chivens. Right image © Boller and Chivens, http://bollerandchivens.com/?p=561, who built the cameras for the Smithsonian Institution.

became thus possible to build the first intercontinental geodetic triangulation networks, which however, with the advent of so much more precise methods (GNSS, VLBI, satellite laser ranging), have honourably faded into scientific history.

16.4 Variations in Earth rotation

The rotation of the Earth varies both in speed and in direction.

The momentaneous rotation axis of the Earth, the pole, moves with respect to the solid Earth's crust, a movement called *polar motion*.

- It comprises an annual motion, period 365 days. This is a forced motion, mainly caused by the atmosphere.

---

picture film, and it had a three-axis mount capable of fast tracking. The aperture was 50 cm, the weight of the whole instrument, 3.5 tonnes. The inventors were Harvard astronomer-optician James Baker (1914 – 2005) and mechanical engineer Joseph Nunn (1905 – 1968). The satellite tracking network consisted of 12 stations.

At least three Baker-Nunn cameras are or were spending their retirement as astronomical sky-survey cameras,
http://newt.phys.unsw.edu.au/~mcba/pubs/carter92.pdf,
https://arxiv.org/pdf/1211.5581v4, https://arxiv.org/pdf/0905.0361.

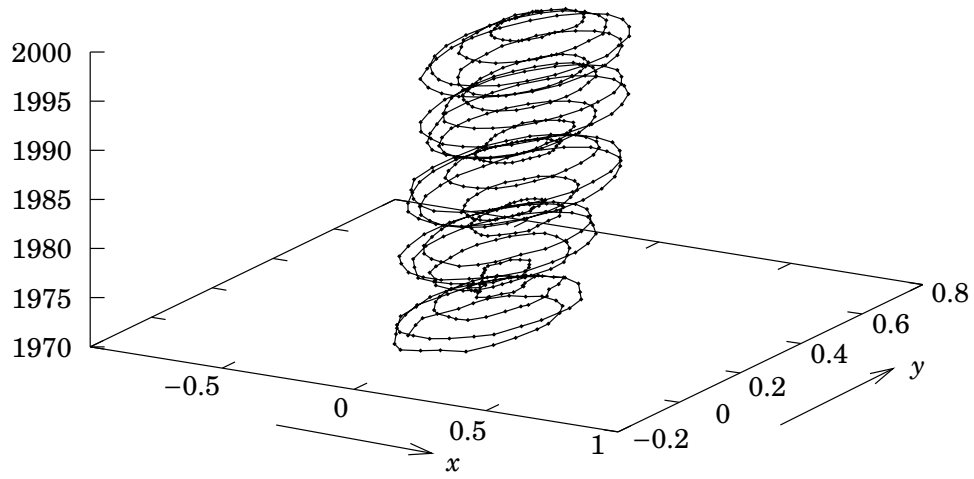


Figure 16.7. Polar motion for the period 1970 – 2000, unit second of arc. Source: International Earth Rotation and Reference Systems Service IERS.

- It also comprises the so-called "Chandler[6] wobble", period about 435 days.
- In addition to these, there has been observed a slow drift of the pole, which is related to changes in mass distribution of the solid Earth, e.g., the postglacial land uplift.

The amplitude (radius) of both circular motions is about $0'',1 - 0'',2$, on the Earth's surface about 3 – 6 m.

The Chandler wobble is theoretically understood: it is the so-called *free nutation*, strongly damped, of a flattened, elastic Earth, already predicted by Leonhard Euler, and variations in the pressure exerted by oceans and atmosphere are its driving force (Gross, 2000).

The *rate* of the Earth's rotation (Length of Day, LoD) is also varying. The phenomenon is being monitored in the same way as the polar motion. It is closely associated with variations in the *angular momentum* of the Earth's atmosphere, a quantity that numerical weather models (NWP, Numerical Weather Prediction) can nowadays calculate very well.

Monitoring of polar motion and length of day, historically and today:

1. Measuring the variations in latitude by astronomical means. Use of this method started in 1899 (Internal Latitude Service, from 1987 International Polar Motion Service), using six "International Latitude Observatories" (Misuzawa, Japan; Charjui, Turkestan, later Kitab, Uzbekistan; Carloforte, Italy; Gaithersburg, Ukiah, Cincinnati, USA) which are located on different continents on the same latitude, 39°08′ around the globe. As the instrument used at the observatories was a *zenith tube*, this allowed them to all use

[6] Seth Carlo Chandler, Jr. (1846 – 1913) was an American astronomer.

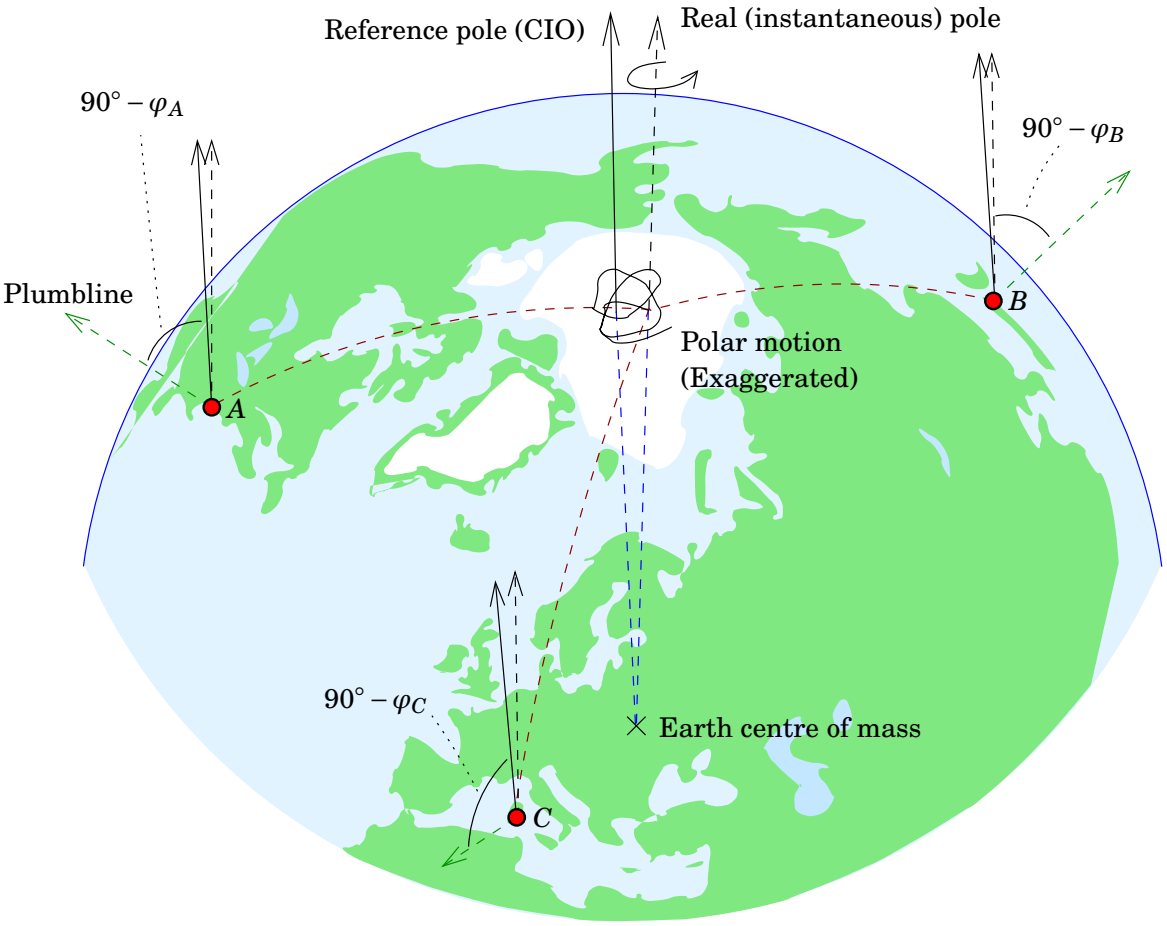


Figure 16.8. Polar motion causes variations in the latitudes of observation stations, with the help of which the phenomenon may be monitored.

the same selection of stars, enabling a uniform data analysis.

Also the monitoring of the rate of rotation of the Earth has been traditionally done by astronomical means. Civil time was originally defined by means of the Earth's rotation, i.e., the Earth was used as a clock. The instrument used was the meridian circle (e.g., in Greenwich, but also at the Helsinki observatory), with which the transit of a star through the meridian was observed. Back then, timekeeping was of vital importance to navigation at sea.

2. Using positioning satellites, first the Transit system, nowadays GNSS. The GNSS method is today more accurate than the astronomical method.
3. With VLBI, very long baseline interferometry, which provides the vectors between observation stations in an inertial or celestial system as a function of time. From this, one may compute the momentaneous direction of the Earth's rotation axis in the same system, and even the momentaneous orientation of the whole Earth. Accuracy is even better than for the GNSS method, order milliseconds of arc.

The *origin* of calculation of the polar motion, CIO, Conventional International Origin, was originally the mean place of the pole over the years 1900 – 1905. The currently used origin of calculation is close to this.

Polar motion and length-of-day variations are called, together with nutation and precession, *Earth orientation Parameters*, EOP. Its monitoring and publication is the task of the IERS (International Earth Rotation and Reference Systems Service). In precise geodetic work, the EOP must be taken into account! The correction information needed can be found on the Internet.

Not until the 1930s was it noticed that the rotation of the Earth is slightly irregular, and a more regular time scale was looked for. The first attempt was ET, *ephemeris time*, based on the orbital motion of the planets, especially the Moon. When sufficiently accurate atomic clocks appeared, they were taken into use, creating *atomic time* (TAI), suited also for demanding scientific use in which a time scale is needed that is strictly uniform.

Today's civil time is UTC, Universal Time Co-ordinated. One of its design objectives is to follow everywhere on Earth the cycle of daylight dominating daily life. Therefore it must follow with sufficient precision[7] the variations in the rotation of the Earth. UTC, which is based on TAI, follows the variation in Earth rotation with an error of at most 0.9 s. To this end, twice a year, at the end of December and at the end of June, UTC executes, if needed, a *leap second* or seconds[8]. The difference UTC – TAI, an integer number of seconds, is tabulated in the almanacs.

GPS time differs from both UTC and TAI. Like TAI, it is uniform and does not ever execute leap seconds. In 1980, when GPS timekeeping started, GPS time was identical with UTC. For this reason

$$\mathrm{GPS} = \mathrm{TAI} - 19\,\mathrm{s}$$

and the difference GPS – UTC is a *varying* number of seconds, included, e.g., in the GPS navigation message.

16.5 Precession and nutation of the Earth

The attraction of Sun and Moon, together with the flattening of the Earth, causes the slow motion of the Earth's rotation axis in space. This conical motion is called *precession*. The word "precession" denoted originally the earlier and earlier occurrence of the *equinoxes*, the times when

[7] In recent years there has been a public discussion in the time field on whether the leap seconds are worth the trouble they are causing, when the time of the time-zone system tied to UTC agrees with local Solar time anyway to no better than an hour or so.

[8] http://en.wikipedia.org/wiki/Leap_second.

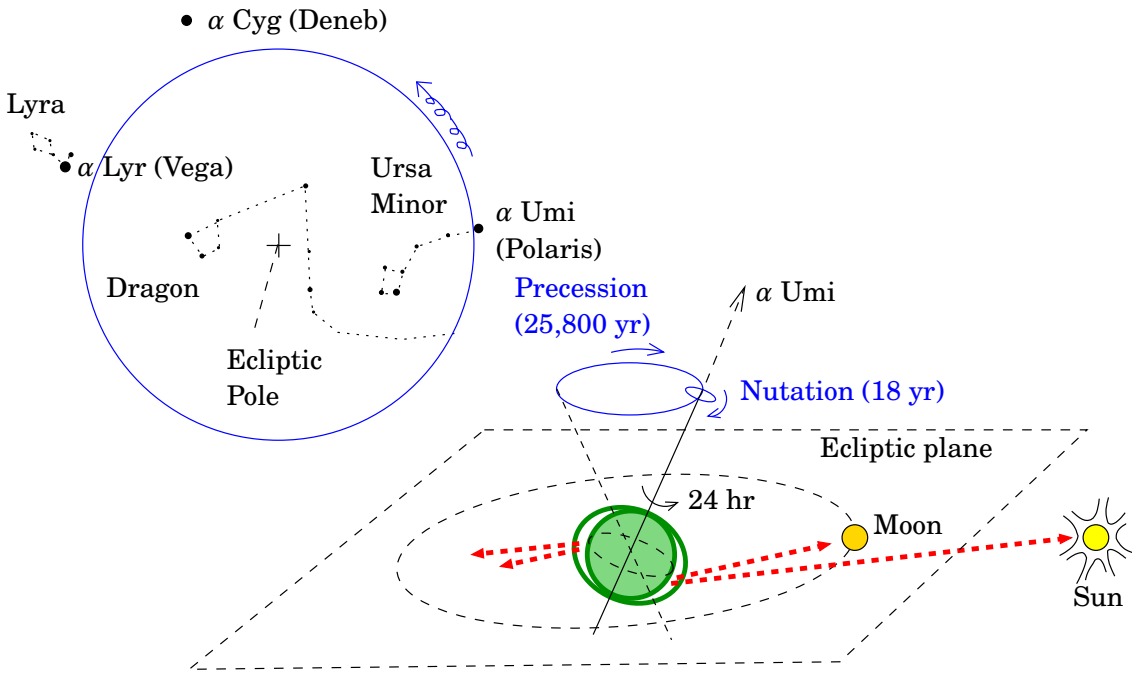


Figure 16.9. The Earth's precession. On the left is shown how, due to precession, the celestial pole describes a circle among the stars over a period of some 25,800 years.

day and night are equally long: this phenomenon, observed first by Hipparchus, brings on, over the centuries, the vernal and autumnal equinox earlier and earlier, as measured by the place of the Sun in the zodiac, amidst the stars[9].

In fact, the rotation axis of the Earth turns in some 25,800 years around an axis that stands perpendicular on the plane of the Earth's orbit. This plane, the apparent plane of the Sun's orbit as seen from Earth, is also called the zodiac or *ecliptic*, see above.

Precession is *not* the same as polar motion. Polar motion is the motion of the Earth's rotation axis with respect to the *solid Earth*. Its magnitude is under a second of arc, on the Earth's surface a few metres.

Besides precession, the Earth's rotation axis also goes through a small periodic motion called *nutation*. Its main period is 18 years and it is caused by periodic changes in the orbit of the Moon which affect through the Moon's attraction. The phenomenon can be precisely computed and is found in almanacs, just like the precession.

The precession makes the rotation axis of the Earth — and thus the place of the celestial pole among the stars — slowly describe a large

[9]Since then, the word has been used to denote the turning motion of the axes of other spinning objects, like spinning tops, gyroscopes, and atomic nuclei in a magnetic field. The latter phenomenon, Larmor precession, is fundamental to magnetic resonance imaging (MRI).

circle. The nutation again causes small "wobbles" on this regular motion. Whereas today, the North Star, Polaris or α Umi is close to the celestial pole, it will in the remote future be α Lyr or Vega, as it was also in the remote past.

Precession and nutation are, unlike the polar motion, motions of the Earth's rotation axis with respect to the *stars*.

16.6 Space weather

The Sun is a star, which produces in its interior nuclear power by "burning" hydrogen to helium. In the centre of the Sun, the temperature is about 15 million kelvins. The thermal energy in the solar core travels very slowly to the surface through many gas layers, first as radiation[10], closer to the surface carried by convection currents, and finally again as radiation, light, leaving the visible surface or *photosphere*, of which a vanishingly small fraction arrives also on Earth.

The outer layer of the Sun, of thickness about 27% of the solar radius, is in a permanent state of convective "bubbling". Because of this, continuously acoustic energy is being sent up to the Sun's highest layer made up of thin plasma, the *corona*, heating this layer to millions of degrees[11].

The corona, which is visible only during a total Solar eclipse, is so hot, that it "leaks" continuously to space. The plasma or particle flow can be observed near the Earth as the *solar wind*.

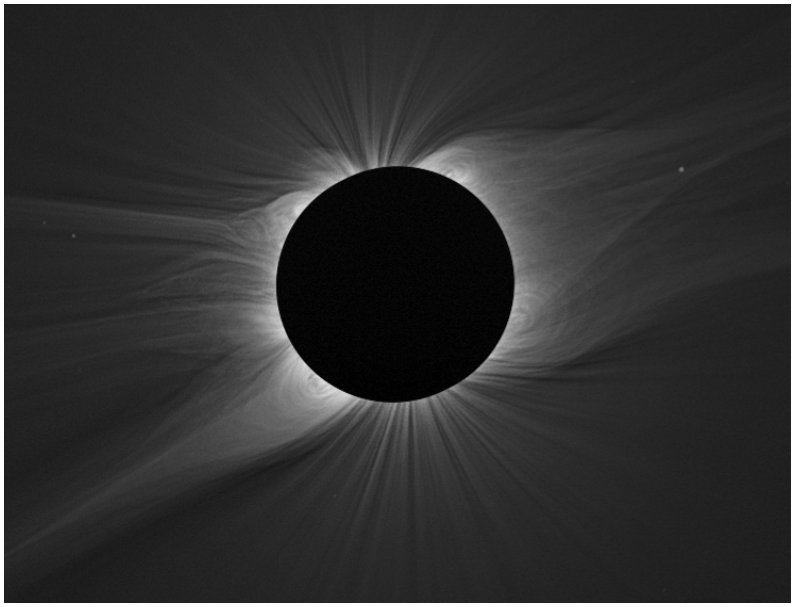
Inside the convection layer of the Sun "lives" a complicated magnetic field, a bit in the same way as in the liquid outer core of the Earth. The Sun does not rotate as a solid body: the higher-latitude zones lag clearly behind the equatorial zone. This, together with the magnetic field, acts as a complicated "natural dynamo". Like the Earth's magnetic field, also the Sun's magnetic field can change its direction, which happens approximately every 11 years.

The Sun's gas is fully ionized *plasma*, which conducts electric currents almost like a superconductor. Therefore the gas and the magnetic field are entangled in each other ("*frozen in*") in an inseparable way. The subject is studied by a discipline called *magnetohydrodynamics* (MHD), to which a short introduction is given in appendix B.

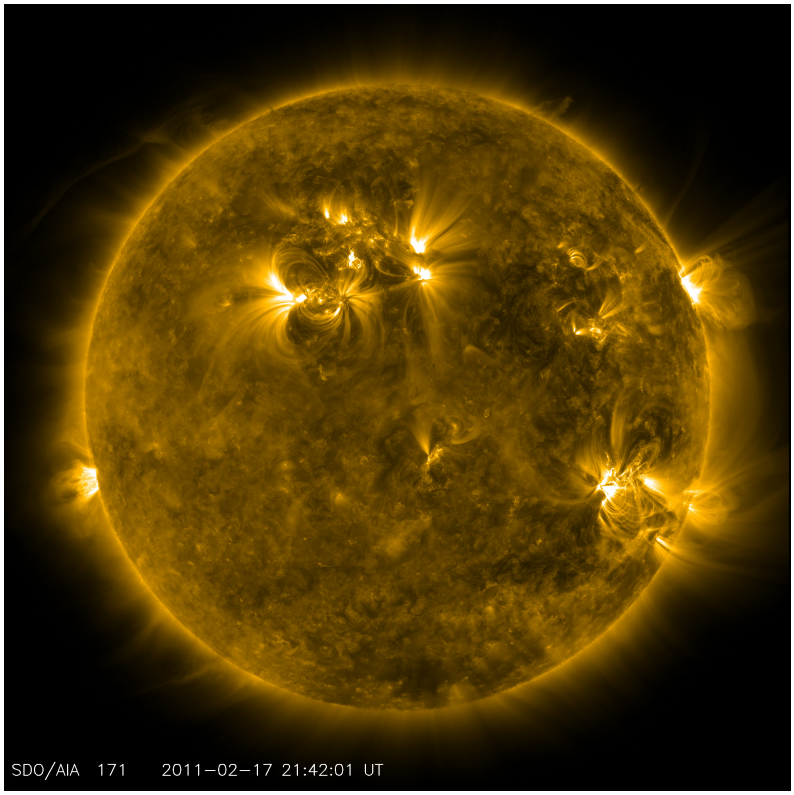
During the time of an active Sun, *sunspots* are seen on its surface, areas of a strong magnetic field where the field of the convection layer breaks out to the surface. They are born and show up always in pairs, magnetic North and South poles. The magnetism in the spots inhibits the

[10] "Thermal radiation", due to the high temperatures in the X-ray range.

[11] For comparison, the temperature of the Sun's surface, or photosphere, is "only" 6000 K.



(a) The corona of the Sun during a total eclipse (image processed) (© NASA).



(b) Sunspots and their magnetic field lines in UV light above the Solar surface (© NASA/SDO)

Figure 16.10. Solar plasmas and magnetic field. Note the patterns formed by the field lines in the corona and the solar wind escaping to space. Sunspots always occur in pairs, as a magnetic North and a magnetic South pole.

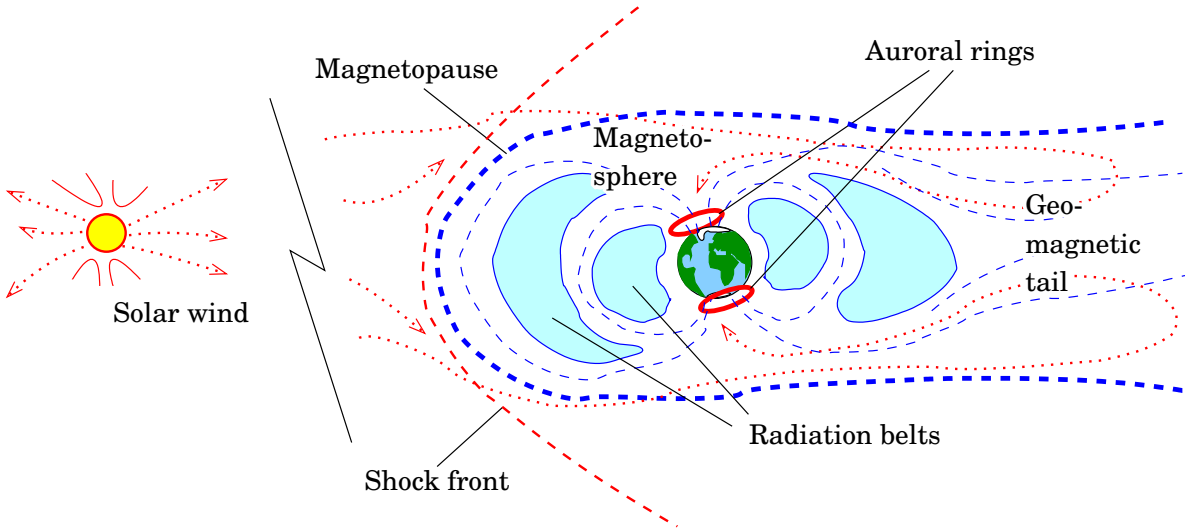


Figure 16.11. Space weather, the magnetosphere (http://en.wikipedia.org/wiki/Magnetosphere) and aurorae.

□

natural convection, and thus prevents energy from reaching the solar surface. In the centres of sunspots, temperatures are even a couple of thousand degrees lower than on the solar surface on average[12], about 3000 – 4000 K.

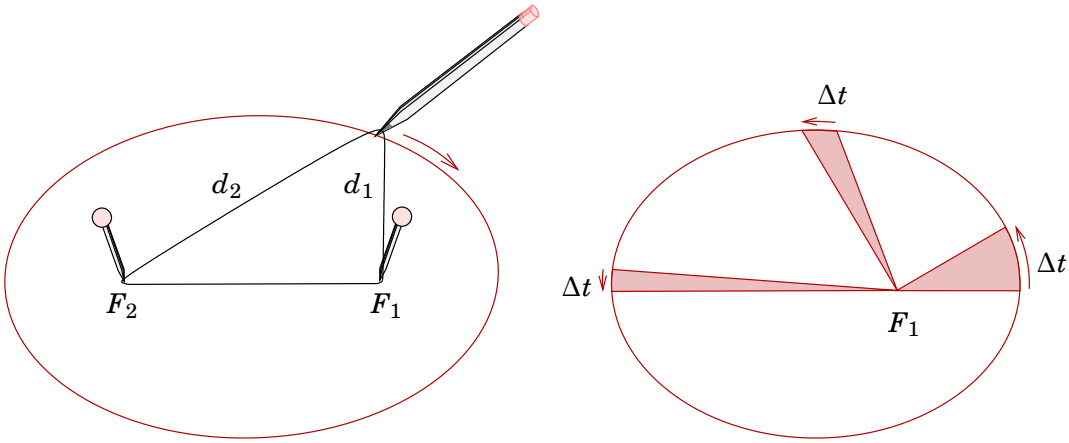
The magnetic field of the sunspots extends into the space above the spots, and affects the motion of the plasma there. With special imaging equipment[13] The field's lines of force can be observed as bright swirls of gas. The topology of the field may suddenly change (*reconnection*), causing the energy released from the field to throw the hot plasma into space. During the eruption, an excess of ultraviolet and X-ray radiation is generated, causing the lowest layers of the Earth's ionosphere to be excessively ionized. This adds to the absorption of short-wave radio, disrupting radio traffic.

About a day later, the plasma itself arrives in the neighbourhood of the Earth, and its interaction with the Earth's own magnetic field and ionosphere causes aurorae. These too affect radio traffic and the quality of GPS observations. During an eruption it may be impossible to collect useable GPS measurements, and generally during a solar maximum the quality of GPS observations is poorer than during a quiet Sun.

Also the Earth's radiation belts consist of hot plasma, fast, electrically charged particles, which the Earth's magnetic field keeps contained in a

[12] However even a single sunspot transferred to the night sky would still outshine the full Moon!

[13] Beautiful imagery and videos are found:
http://sohowww.nascom.nasa.gov/data/realtime-images.html,
http://sohowww.nascom.nasa.gov/data/realtime/mpeg/.



(a) Drawing an ellipse **(b)** Kepler's second law

Figure 16.12. An ellipse is the set of points for which the sum of the distances $d_1 + d_2$ from two focal points F_1 and F_2 is constant. This property, a consequence of the ellipse being a conic section —, is most easily proven using Dandelin spheres, https://en.wikipedia.org/wiki/Dandelin_spheres. According to Kepler's second law, the radius vector of a planet sweeps over an always same-sized area in the same time Δt.

"magnetic bottle".

16.7 Satellite orbital motion

Like the Earth orbits the Sun, also the orbital motion of satellites around the Earth follows the laws of Kepler[14]. We already touched upon the matter in the GPS part, section 11.9. Kepler's laws are:

1. The satellite moves around the Earth in an *elliptic* orbit in a *plane*. The centre of mass of the Earth is located in one of the focal points of the orbital ellipse.
2. The radius vector between satellite and centre of mass of the Earth sweeps in the same amount of time always over the same surface area (*law of surfaces*).
3. The *squares* of the periods of different satellites stand in the same ratios as the *cubes* of the semi-major axes of their orbital ellipses. See below (equation 16.1).

A satellite orbit is described by *six Kepler orbital elements*, figure 16.13. In appendix C the orbital elements are described in more detail. The angle θ_0 is Greenwich sidereal time, which describes the orientation of

[14]Johannes Kepler (1571 – 1630) was a German astronomer, mathematician and mystic.

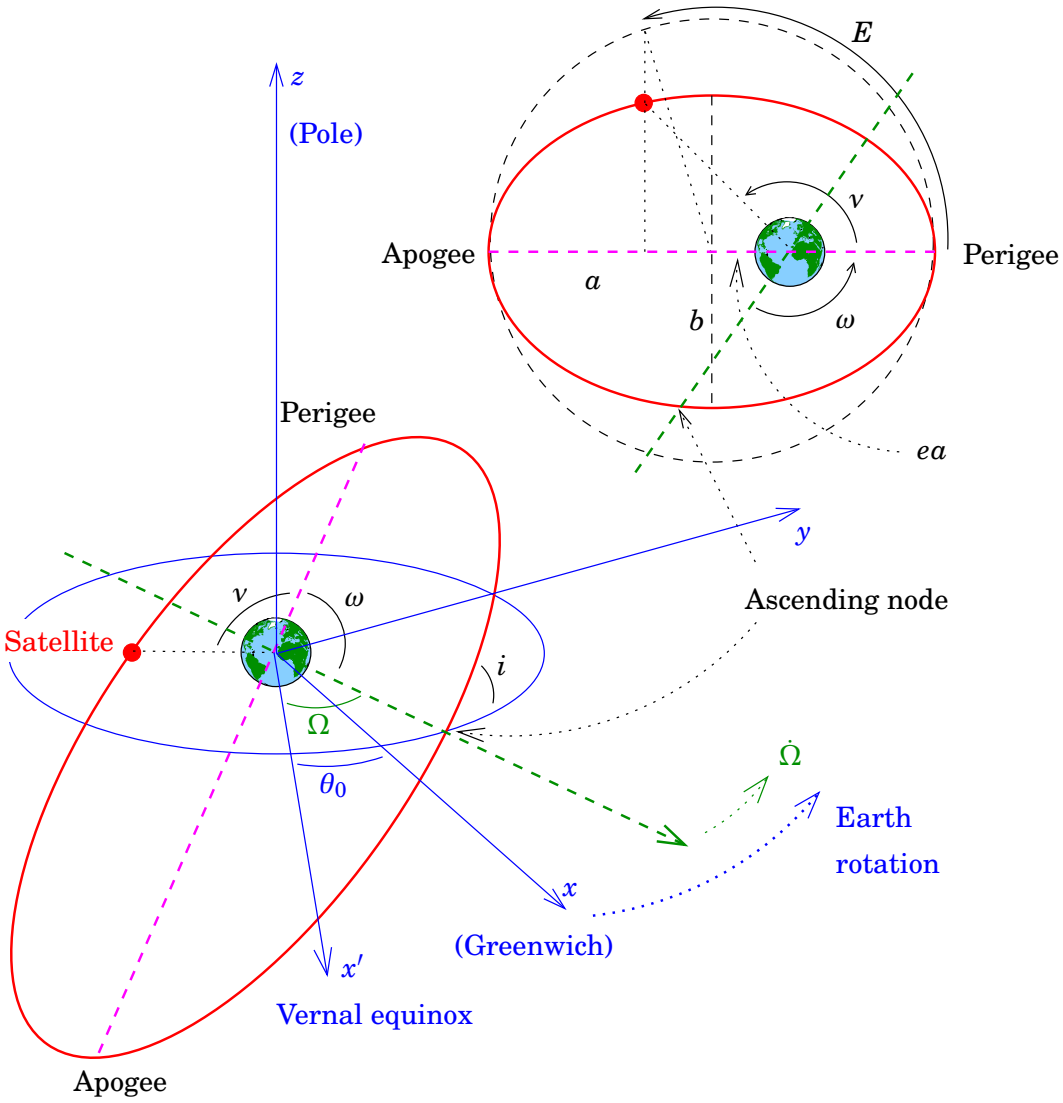


Figure 16.13. Kepler's orbital elements.

the Earth with repect to the stars.

Kepler's laws apply for low orbits *only approximately*. The uneven distribution of the Earth's masses, and especially her *flattening*, cause *orbit perturbations*. These are exploited for studying the internal mass distribution of the Earth. This is how satellite geodesy has become an important tool for studying the solid Earth.

16.8 Choosing a satellite orbit

The tilt of a satellite orbital plane with the equatorial plane, or its *inclination*, is an important parameter from the viewpoint of the intended use of the satellite. The inclination is in practice the same as the maximum Northern or Southern latitude over which the satellite can fly. So if, e.g., it is given that the inclination of some satellite is 55°, we may

Table 16.1. Kepler's third law for Earth satellites.

Height (km)	Period	Remark
0	$84^{\mathrm{m}}29^{\mathrm{s}}$	
400	$92^{\mathrm{m}}34^{\mathrm{s}}$	
800	$100^{\mathrm{m}}52^{\mathrm{s}}$	
20,183	$11^{\mathrm{h}}58^{\mathrm{m}}$	GPS
35,785	$23^{\mathrm{h}}56^{\mathrm{m}}$	Geostationary
376,603	$27^{\mathrm{d}}07^{\mathrm{h}}$	Moon

conclude that the satellite[15] will never come down over Finland.

From the viewpoint of low-flying weather and remote-sensing satellites, the importance of the inclination is, that the area that a satellite can properly map lies approximately between these maximum latitudes.

For high-flying satellites, this limitation does not apply. E.g., geostationary satellites may perfectly well map the Nordic area. The imaging angle would not, however, be good.

The choice of the height of a satellite orbit is made using Kepler's laws of orbital motion. Kepler's third law says:

$$GMP^2 = 4\pi^2 a^3, \tag{16.1}$$

in which $a = a_{\mathrm{e}} + h$ is the semi-major axis of the satellite orbit, the mean distance from the centre of the Earth. The quantity h again is called the mean height of the satellite. P is the orbital period, the time to go around the Earth.

The perigee and apogee *heights* from the Earth's surface are formally calculated as follows:

$$\begin{aligned} h_{\mathrm{P}} &= (1-e)a - a_{\mathrm{e}}, \\ h_{\mathrm{A}} &= (1+e)a - a_{\mathrm{e}}, \\ \text{siis } h_{\mathrm{A}} - h_{\mathrm{P}} &= 2ea. \end{aligned}$$

In this, a_{e} is the *equatorial radius* of the Earth, according to the GRS80 reference ellipsoid 6378,137 m.

16.9 Satellite orbital precession, Sun-synchronous orbit

The figure of the Earth affects satellite orbital motion. E.g., the quantity J_2, the *dynamic flattening*, the value of which is $J_2 = 1082.6267 \cdot 10^{-6}$, and which is just one of many so-called spherical-harmonic coefficients

[15] Assuming it is not capable of aerodynamic flight upon atmospheric entry.

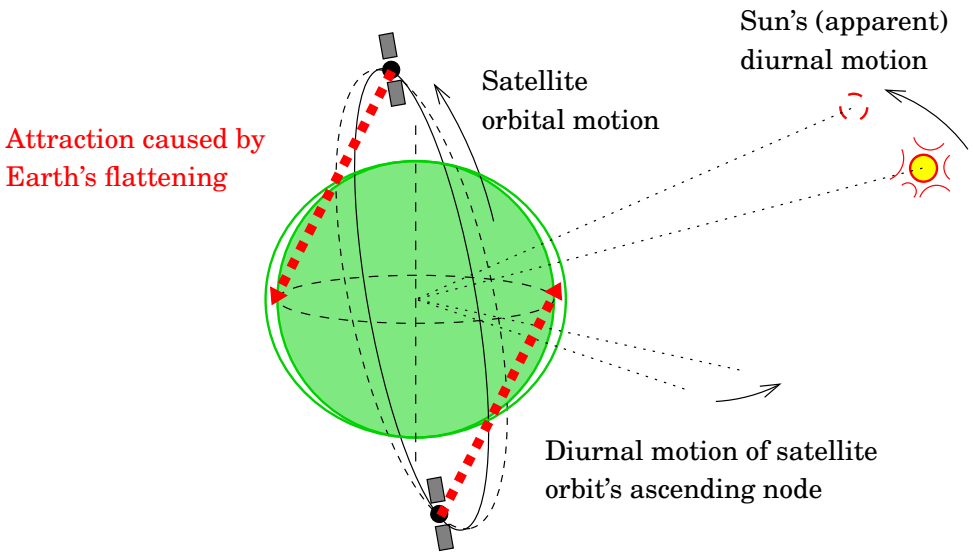


Figure 16.14. Sun-synchronous orbit.

□

describing the figure of the Earth and affecting satellite motion. In the case of J_2, one important effect is, that the satellite orbital plane *turns* at a certain rate around the rotation axis of the Earth, i.e., *orbital precession*. This leads to the circumstance that the satellite will fly over the same place every day a number of minutes earlier. The rate of precession, for a circular orbit of radius a, is described by the equation

$$\dot{\Omega} = \frac{d\Omega}{dt} = -\frac{3}{2}\sqrt{\frac{GM}{a^3}}\left(\frac{a_\mathrm{e}}{a}\right)^2 J_2 \cos i,$$

in which a_e is the equatorial radius of the Earth and i the inclination angle of the orbit with respect to the equator. If we substitute numerical values into this, we obtain

$$\frac{d\Omega}{dt} = -1.318{,}95 \cdot 10^{18} \frac{\cos i}{a^{3{,}5}} \left[\mathrm{m}^{3.5}\,\mathrm{s}^{-1}\right].$$

If, as a calculational example, we substitute into this as the satellite height

$$h = 800\,\mathrm{km} \Longrightarrow a = 6378{,}137\,\mathrm{m} + 800{,}000\,\mathrm{m} = 7178{,}137\,\mathrm{m},$$

we obtain

$$\begin{aligned}\frac{d\Omega}{dt} &= -1.331{,}02 \cdot 10^{-6} \cos i\,[\mathrm{rad\,s^{-1}}] = \\ &= -6.589^\circ/\mathrm{day} \cdot \cos i. \end{aligned} \tag{16.2}$$

For practical reasons (solar panels!) we often choose the satellite orbit so, that the orbital plane turns with the apparent annual motion of the Sun:

$$\frac{360^\circ}{365.25\,\mathrm{days}} = 0^\circ.9856/\mathrm{day}.$$

If the inclination of the orbital plane is chosen in the range $96° - 102°$, depending on the height, the flattening J_2 of the Earth will cause just the suitable rotation of the orbital plane ("no-shadow orbit, Sun-synchronous orbit").

Height (km)	500	750	1000	1500
Critical inclination	97°.4	98°.4	99°.5	102°.0

Self-test questions

1. A sidereal day is $23^{\mathrm{h}}56^{\mathrm{m}}4^{\mathrm{s}}$. Why is is shorter than a civil day, 24^{h}?
2. What is the ecliptic? How did it get its name?
3. What is the vernal equinox?
4. What mechanism causes the four seasons?
5. What is precession, and what is nutation? What causes them?
6. How did the Tropics of Cancer and of Capricorn get their names?
7. What geometrical quantity does Greenwich Apparent Sidereal Time (*GAST*) represent?
8. Describe the co-ordinates on the celestial sphere, i.e., right ascension and declination.
9. What is the hour angle of a celestial object? How is it related to its right ascension?
10. What means upper and lower culmination of a star? Can they both always be observed (assuming clear skies)?
11. What is a meridian transit, and how is it observed?
12. Explain Väisälä's stellar triangulation.
13. Describe the components of polar motion and variations in length of day. How are they observed, historically and today?
14. In what way are precession and nutation different from polar motion?
15. What is a plasma?
16. What is the solar wind? Is it a sub- or supersonic flow? Why?
17. What is a *Carrington Event*? Ask Google.
18. Could the wreck of the International Space Station (assuming it isn't brought down in a controlled way) ever impact Finnish territory? Why / why not?
19. How does one engineer a Sun-synchronous orbit? Why is it useful?

17. Geodesy and geophysics

17.1 Geodynamics

Geodynamics is the field of study within geophysics that studies the motions taking place in the Earth, like plate tectonics, post-glacial land uplift, as well as other motions of the Earth's crust, local, global, natural or caused by man. Normally also the study of the Earth's rotation is included with geodynamics.

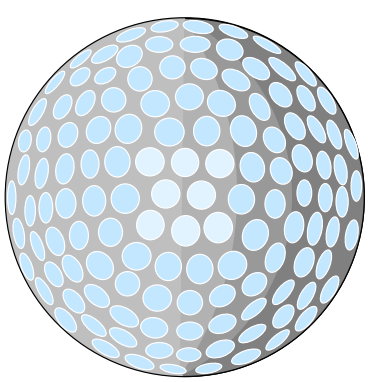
17.1.1 Geodetic observation techniques

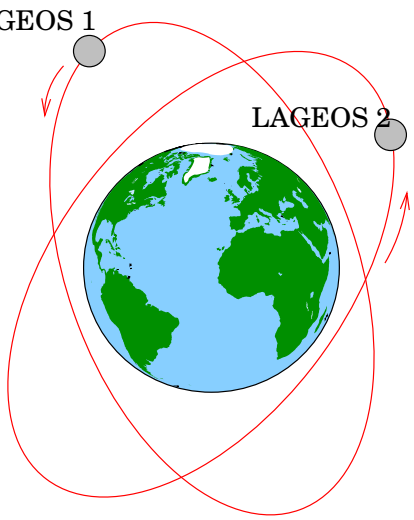
Of the observation techniques suitable for geodynamic research, we already discussed satellite positioning. All GNSS methods, like GPS, GLONASS, BeiDou and Galileo, are suitable for precise geodynamics measurements.

However, in scientific research we always want to use as many as possible, as independent as possible techniques for the study of the same phenomenon, in as versatile a fashion as at all possible. Therefore we use, in geodynamics research, also some very precise, but also very expensive, research methods:

- satellite laser ranging
- laser ranging to the Moon
- VLBI, very long baseline interferometry with radio telescopes
- DORIS, *Doppler Orbitography and Radiopositioning Integrated by Satellite* (The IAG's DORIS service http://ids-doris.org/), a French satellite orbit determination system that serves geodetic and geodynamic research.

The satellite laser and lunar laser ranging techniques are very similar. Laser satellites are very heavy, massive spheres manufactured from high-density materials like bronze, depleted uranium or some other metal, studded with reflective corner-cube prisms. On the Moon, the Apollo astronauts left several reflective panels covered in prisms serving the same purpose. Also the Russian vehicles deposited on the Lunar surface in unmanned flights, Lunokhod 1 and 2, carry reflective prisms.





	a	e	i
LAGEOS 1	12273 km	0.00444	109°.84
LAGEOS 2	12273 km	0.01377	52°.68

Figure 17.1. A LAGEOS satellite. The diameter of the satellite is 60 cm, it consists of an exterior aluminium sphere to which are attached 426 corner-cube prism reflectors (see subsection 6.4.3), and an interior sphere made of bronze. The mass of the satellite is about 400 kg. The LAGEOS 1 satellite also carries a map designed by Carl Sagan of, i.a., the current locations of the continents, as a message to possible alien finders (https://en.wikipedia.org/wiki/LAGEOS#Time_capsule). On the right, the orbits of the LAGEOS 1 and 2 satellites.

□

The distance to satellites or to the Moon is measured by measuring the two-way travel time of a light pulse, and dividing by two. The influence of the atmosphere on signal propagation must be taken into account. Unlike GNSS, this technique measures *real distances* (without clock unknowns), not pseudo-ranges. Therefore the global laser station network is somewhat stronger geometrically than the global GNSS network. For this reason it gives a valuable boost to the global monitoring network's geometric strength.

VLBI (*very long baseline Interferometry*) uses the radio signals of remote radio sources, *quasars*. Quasars are so far way and compact[1], that in practice they are point sources. Therefore the wave form of the noise-like radio signal that they transmit is precisely the same, wherever on Earth it is received.

VLBI observations are carefully planned. Many radio telescopes around the globe participate in the campaigns. All participants execute an agreed programme, in which the same objects are observed simultaneously within

[1]The dominant theory is that quasars are supermassive black holes in the cores of remote, i.e., young, galaxies. The radiation is generated when the hole sucks up matter from its surroundings. Also our own Milky Way galaxy has a black hole at its centre (Sagittarius A*), which however radiates much more weakly.

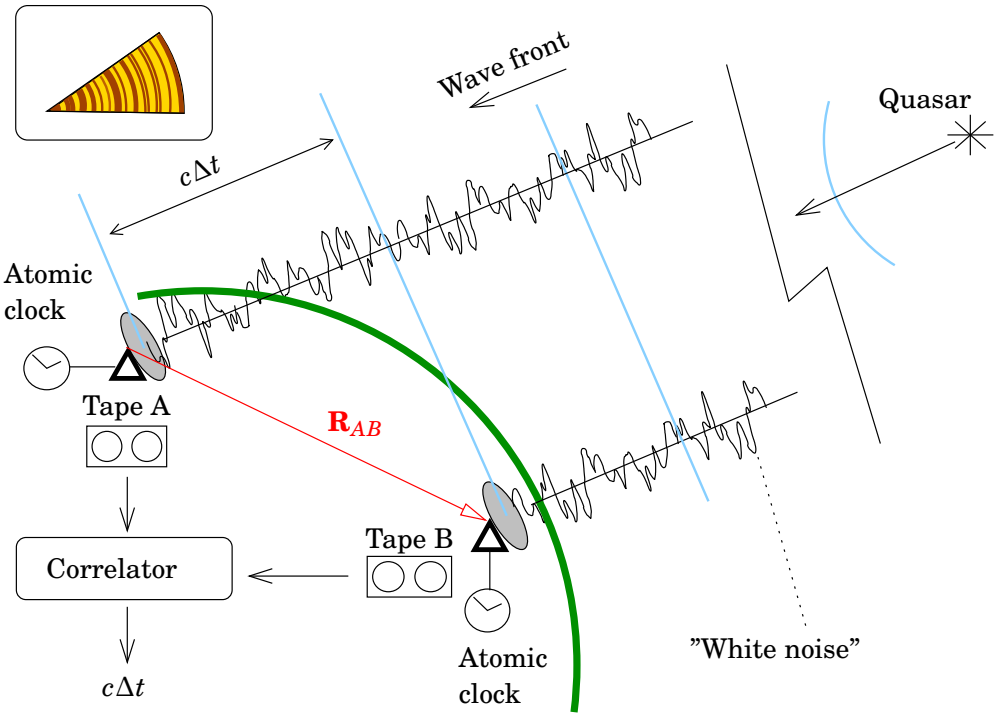


Figure 17.2. Principle of operation of very long baseline interferometry.

a certain frequency band.

For storing the signal there are various solutions and generations of instrumentation. ”Mark II” used common-or-garden VHS video tapes. The signal was converted by special equipment to a video signal and stored to tape. Today, the recording is often done digitally to a hard-drive pack, and also direct transfer over the Internet using the UDP/IP protocol is becoming popular.

The recordings are read and *correlated* with each other: corresponding spots are found, just like in comparing the tree rings of two pieces of wood (see figure 6.7). This correlation is a compute intensive process which earlier required specialized equipment. As a result of the correlation process, we obtain the *difference in reception times* Δt at the two radio telescopes, which is thus the observable of this technology. It may also be interpreted as the *projection* $c\Delta t$ of the vector $\mathbf{R}_{AB}$ connecting the radio telescopes A and B on the direction vector to the quasar, see figure 17.2.

The method requires, in order to work, like the famous Väisälä interferometry, that the signal is random, i.e., *white noise*: the coherence length has to be short. It's the same with GPS code measurement, where the noise is generated artificially, as so-called pseudo-random code.

By carrying out measurements to many different quasars, one may solve for the vector $\mathbf{R}_{AB}$ between radio telescopes A and B in an absolute — i.e., inertial or celestial, tied to the stars — co-ordinate reference frame. Just like satellite laser observations, also VLBI observations strengthen

the global geodynamic monitoring networks. Also in Finland, geodetic VLBI observations are carried out, with the Metsähovi radio telescope, see figure 3.15.

One essential difference between, on the one hand, VLBI observations, and on the other, both laser and GNSS observations, is that the centre of mass of the Earth is not along in the VLBI observation equations. Quasars are at such huge distances, that any small change in the location of the centre of mass of the Earth would affect in precisely the same way the observations by the radio telescopes at both ends of the VLBI vector, and the effect on the end result would be zero. On the other hand the momentaneous direction of the rotation axis of the Earth is along in the measurement geometry, and VLBI has become, besides GNSS, a favourite means of monitoring the Earth's rotation, both polar motion and variations in length of day.

Close to the Metsähovi research station, at three kilometres distance in Sjökulla, shielded by the landscape from the radio telescope, is a DORIS beacon containing an active radio transmitter. The French DORIS (Doppler Orbitography by Radiopositioning Integrated on Satellite) is an unusual system: *the observation stations are active* and the satellites passively collect their signals. The advantage of this solution is the centralization of data collection. Around the world there are 60 DORIS stations, there are stations on all major continental plates.

17.1.2 Global geodynamics

The German Alfred Wegener[2] proposed already in 1912 that the continents, fragments of the Earth's crust, were moving slowly over the soft interior of the Earth, like ice floes on the sea. As evidence he offered the shapes of the continental coast lines, which fit together remarkably well (figure 17.3) and rock types and fossils, which are very similar on corresponding coastal locations.

Nobody believed him at the time. It didn't help matters that seismology showed the Earth's mantle to be as hard as steel: transversal — sideways oscillating — waves, *S waves*, only travel in solids, and they travel very well through the Earth's mantle.

Not until the 1960's did the theory receive more support, especially based on research into the Earth's magnetic field. Back then, it was already possible to measure the local magnetic field from an aircraft, military technology developed for detecting submarines. The minute variations in the field were mapped, and everywhere on the maps appeared *parallel magnetization stripes*, see figure 17.5. The stripes run in the direction of

[2]Alfred Lothar Wegener (1880 – 1930) was a German meteorologist, geophysicist and Greenland explorer, where he also died at the age of 50, and into the continental ice sheet of which he was buried.

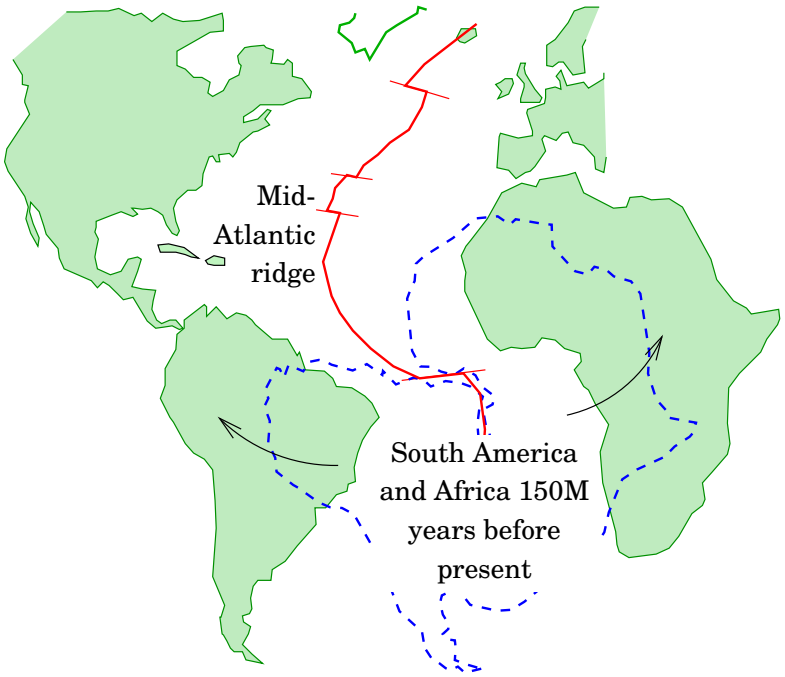


Figure 17.3. Alfred Wegener's continental drift theory and the Mid-Atlantic Ridge.

the Mid-Atlantic Ridge, or similar ridges in other oceans. Also their pattern was everywhere the same, even if sometimes broader, sometimes narrower — like the tree rings from the same period in different trees, see tableau 11.2.

The theoretical explanation is, that new Earth crust — sea floor — is being formed all the time at the Mid-Atlantic Ridge: hot, liquid magma cools down and solidifies, and the iron-ore particles in the magma are permanently oriented along the direction of the magnetic field at that point in time.

At this moment, the Earth's North magnetic pole — in Canada — is physics-wise a South pole, S; the Southern magnetic pole, near Tasmania, is an N type pole. These roles, however, swap at irregular intervals, of order a million years. These alternating directions are recorded into the sea floor in the same way that sound is recorded on the magnetic tape of a tape recorder.

The Earth's internal structure according to our best current knowledge is given in figure 17.4.

At the same time, the British Arthur Holmes[3] developed *radiometric dating*, which exploits the uniform decay rate of radioactive isotopes of various elements occurring in the Earth's crust as a geological clock (Lewis, 2000). Among the useful decay processes are potassium ▷ argon,

[3]Arthur Holmes (1890 – 1965) was a British geologist, a pioneer of radiometric dating and understanding the mechanisms of plate tectonics.

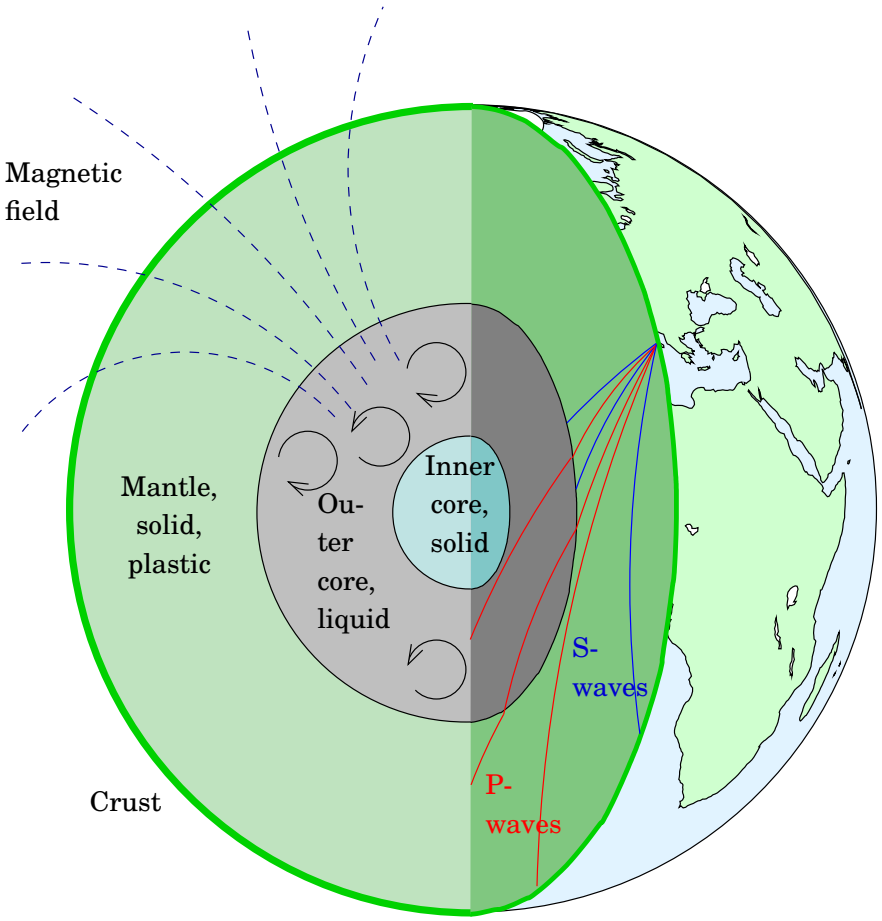


Figure 17.4. The internal structure of the Earth. Convection currents in the outer core generate the Earth's magnetic field (dynamo theory). The core consists fairly certainly of an iron-nickel mixture with lighter-element impurities.

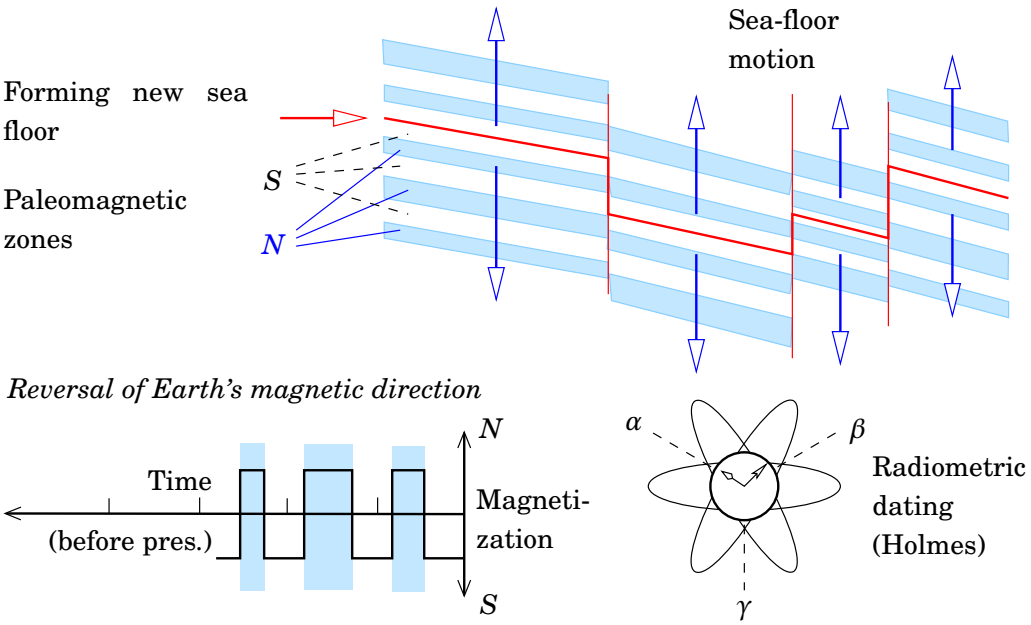


Figure 17.5. Paleomagnetism and sea-floor spreading.

half-life 1.25 billion years, and uranium 238 ▷ lead, 4.47 billion years.

By measuring the concentrations of the decay product and the original isotope, one may infer how much time has elapsed since the solidification of the rock. The ages determined in this way for different places on the sea floor agree very well with the pattern of the magnetic stripes: the longer the distance from the central ridge, the older the sea floor.

We say that the *continental* plates move, but the plates include just as well sea floor. We speak of the Eurasian plate, although it contains also the whole North Atlantic sea floor North of the Azores and East of the Mid-Atlantic Ridge. What is happening is a very slow *convective motion*, which carries the plates of the Earth's crust along with it. The rates of motion vary from a few centrimetres per year to as much as ten centimetres. The precise form of the convective pattern is still unclear: does it take place throughout the mantle, or in two layers?

The thermal energy that maintains the convection originates from two sources: radioactive decay, and "primordial heat", the ongoing cooling down and statification after formation of the Earth. Both parts are believed to be of similar magnitude[4]. The heat produced by the Earth's core is entirely primordial: the solid inner core continues to grow at the expence of the liquid outer core.

The convection theory is presented in figure 17.6, and more details on the mechanism of continental motion in figure 17.7. At the Mid-Atlantic Ridge (and the Mid-Pacific, Mid-Indian etc. ridges) new sea floor crust is being formed out of the magma rising from the mantle (however, as already said, the Earth's mantle itself *is not liquid*). At the edges of the oceans again are found *deep-sea trenches*, under which the oceanic crust dives down into the mantle (*subduction*).

In precise geodetic work, the continental plate motion must be taken into account. For computing the motion, the so-called NUVEL models may be used, which tabulate to motion of each plate as a *rotation* around a given *pole* at a given rotation rate. The model NUVEL-1A (DeMets et al., 1994) has been computed from geological research material. The results agree, within their measurement uncertainties, with the results obtained by GPS and other methods.

17.1.3 Regional geodynamics

17.1.3.1 Example: the Fennoscandian land uplift

The land is rising in Finland. Like it is in Sweden, Norway, Denmark, Scotland and Canada. All these vertical movements are caused by the melting, after the last ice age, of continental glaciers or *ice sheets*, some 11,000 years ago. In Fennoscandia a continental ice sheet, thickness

[4] https://en.wikipedia.org/wiki/Earth's_internal_heat_budget.

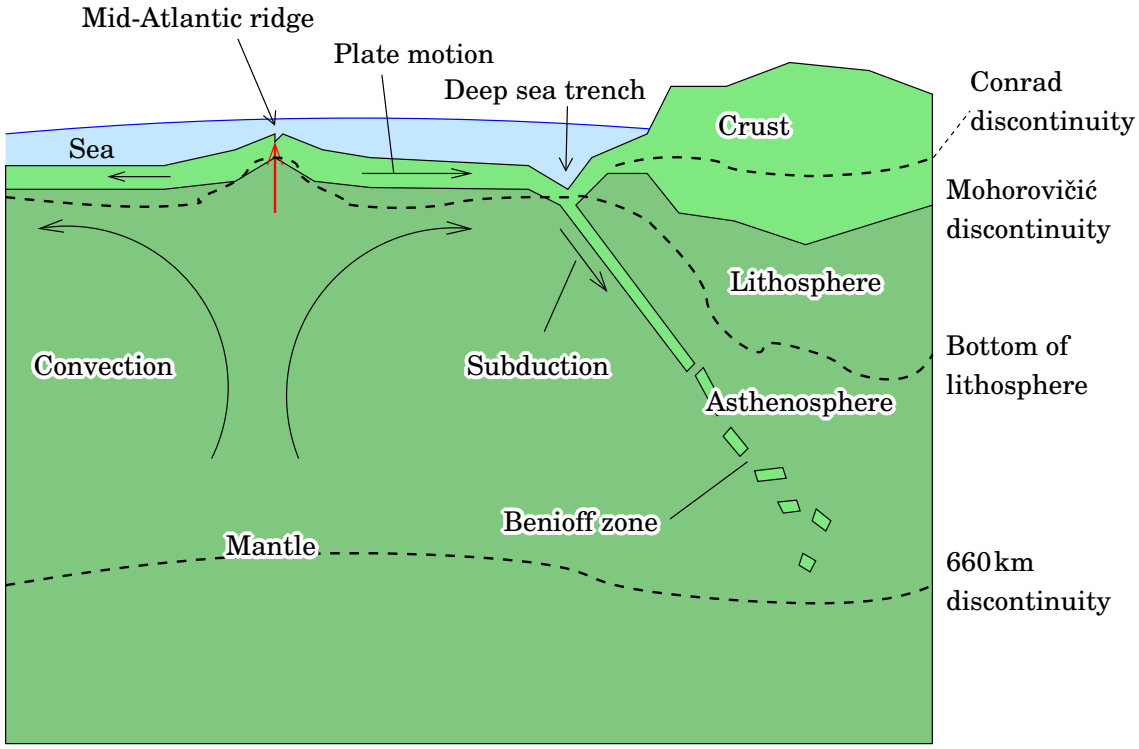


Figure 17.6. Global plate tectonics.

up to two-three kilometres, pressed the Earth's crust down. The phenomenon is called *isostatic adjustment*; although both the Earth's crust and the mantle consist of solid matter, they nevertheless give slowly and *plastically* way under great loads.

Elastic: The object responds immediately, and returns to its original shape immediately when the load vanishes. E.g., a spring, a tennis ball.

Plastic: The object responds slowly and continuously. It does not return to its original shape when the load vanishes. E.g., syrup, modelling clay. The resistance offered to deformation by a plastic substance is characterized by a concept called *viscosity*.

Viscosity: The more solid a substance, the higher the value of its *viscosity*. Unit: Pa s (Pascal second) or N s m^{-2}.

The physical character of post-glacial land uplift is one of *plastic* rebound. The phenomenon is studied because it offers a possibility to *determine viscosity values for the Earth's mantle*. Results obtained point to the following structure:

1. the "lid" or *lithosphere*, thickness 50 – 100 km, which has a high (in practice, infinite) viscosity and responds purely elastically[5]
2. under the lithosphere, the *asthenosphere*, a layer with a relatively

[5] . . . until the loading becomes too heavy and the rock fractures: it is *frangible*.

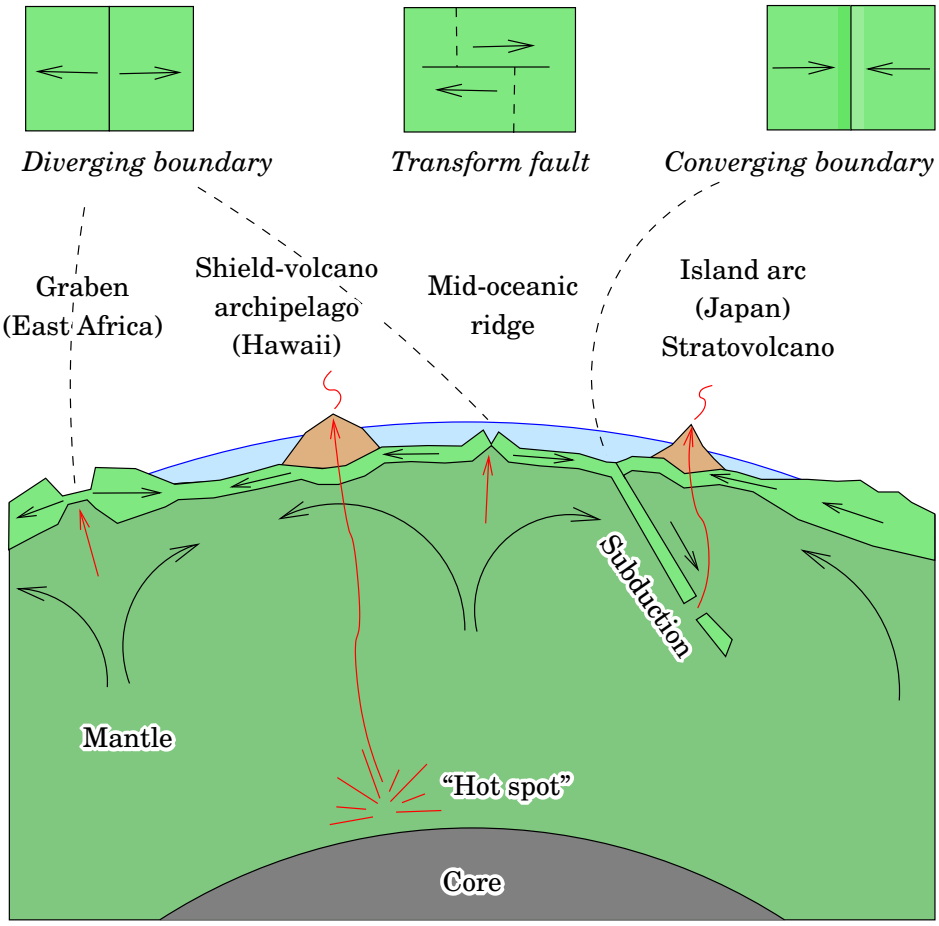


Figure 17.7. Mechanisms of plate tectonics, plate boundaries and volcanism. The thickness of the Earth's crust is exaggerated. Note that the deep "hot spot" theory is not generally accepted among geophysicists.

□

low viscosity, $10^{20} - 10^{21}$ Pa s, thickness several hundred km

3. under the asthenosphere, the lower mantle, which has a relatively high viscosity, order 10^{22} Pa s or even higher.

The numerical values given above are highly uncertain and may well change as a result of ongoing research.

For comparison, the viscosity of running water at a temperature of 20° C is 0.001 Pa s, that of liquid sodium (used as coolant in breeding reactors) 0.0007 Pa s at its melting point 98° C.

Fennoscandia isn't the only area in the world where the Earth's crust is rising as a result of the termination of the last ice age. We speak of GIA, *glacial isostatic adjustment*. Other similar extended areas are:

- the Laurentide (Northern Canada) land uplift area. Much more extensive than the Fennoscandian one
- the West Siberian land uplift area. Little studied (Ehlers et al., 2015).

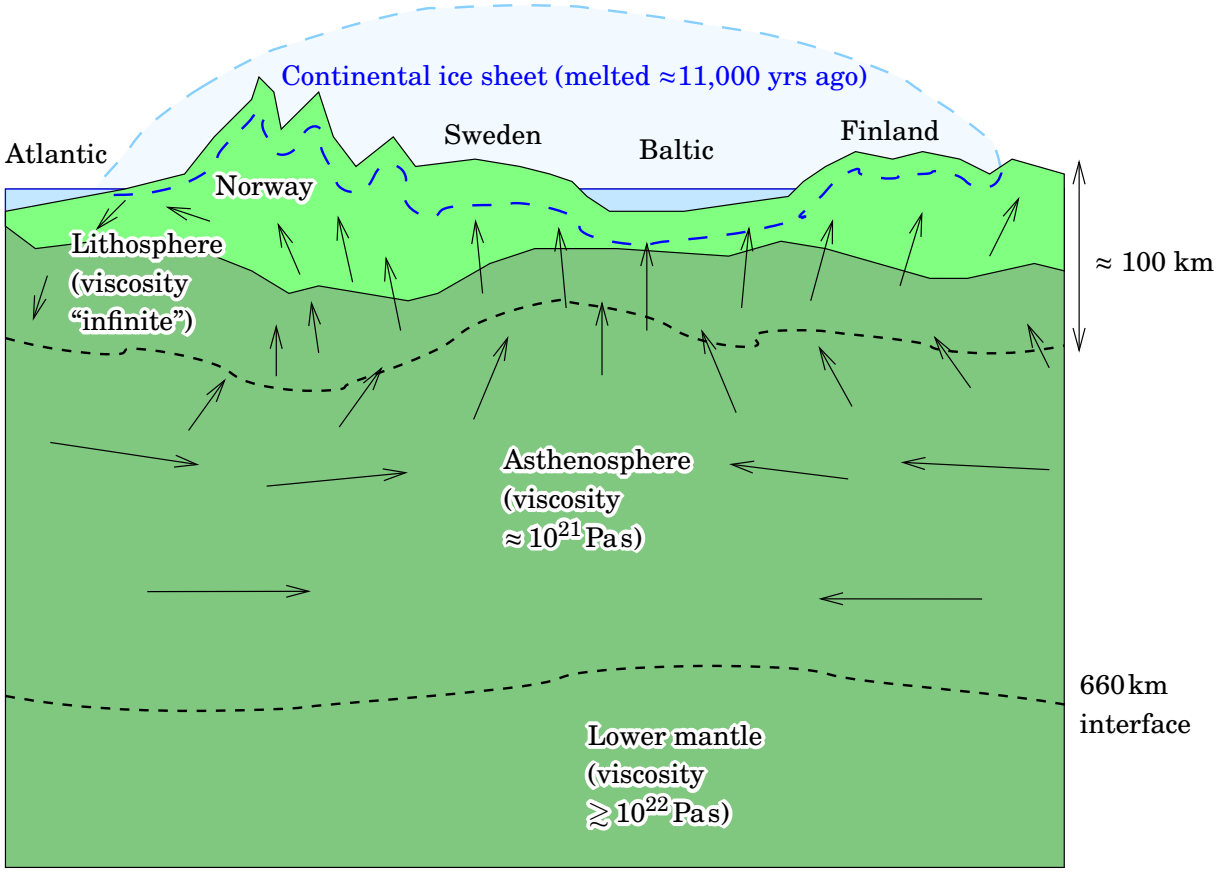


Figure 17.8. Post-glacial land uplift in Fennoscandia. The figure gives a vertical cross-section of the area; in fact, the phenomenon is three-dimensional. Topography and ice-sheet thickness are exaggerated.

The post-glacial land uplift in Fennoscandia takes place in an area in the shape of an ellipse, the centre of which, the land-uplift maximum, is located at the narrow ("Kvarken") of the Gulf of Bothnia, on the Swedish side, where it amounts to over 9 mm/a. Going outward from here, the rate diminishes in all directions. At the Finnish South-Eastern border, only 3 mm/a remains. The zero land-uplift line runs through Denmark and Northern Poland, curving on the Lithuanian and Russian territories to the North-East. Outside, i.e., to the South of the line, the land is slowly subsiding.

When the Earth's crust rises, material flows slowly inward in the asthenosphere, toward the uplift centre, and fills the void that was created under the crust. The land uplift is not only a vertical motion: also horizontal motions occur, as GNSS monitoring over extended periods has shown.

Techniques for studying and measuring the post-glacial land uplift are:

- regularly repeated (at intervals of decades) precise levelling
- monitoring sea level with respect to the Earth's crust by using

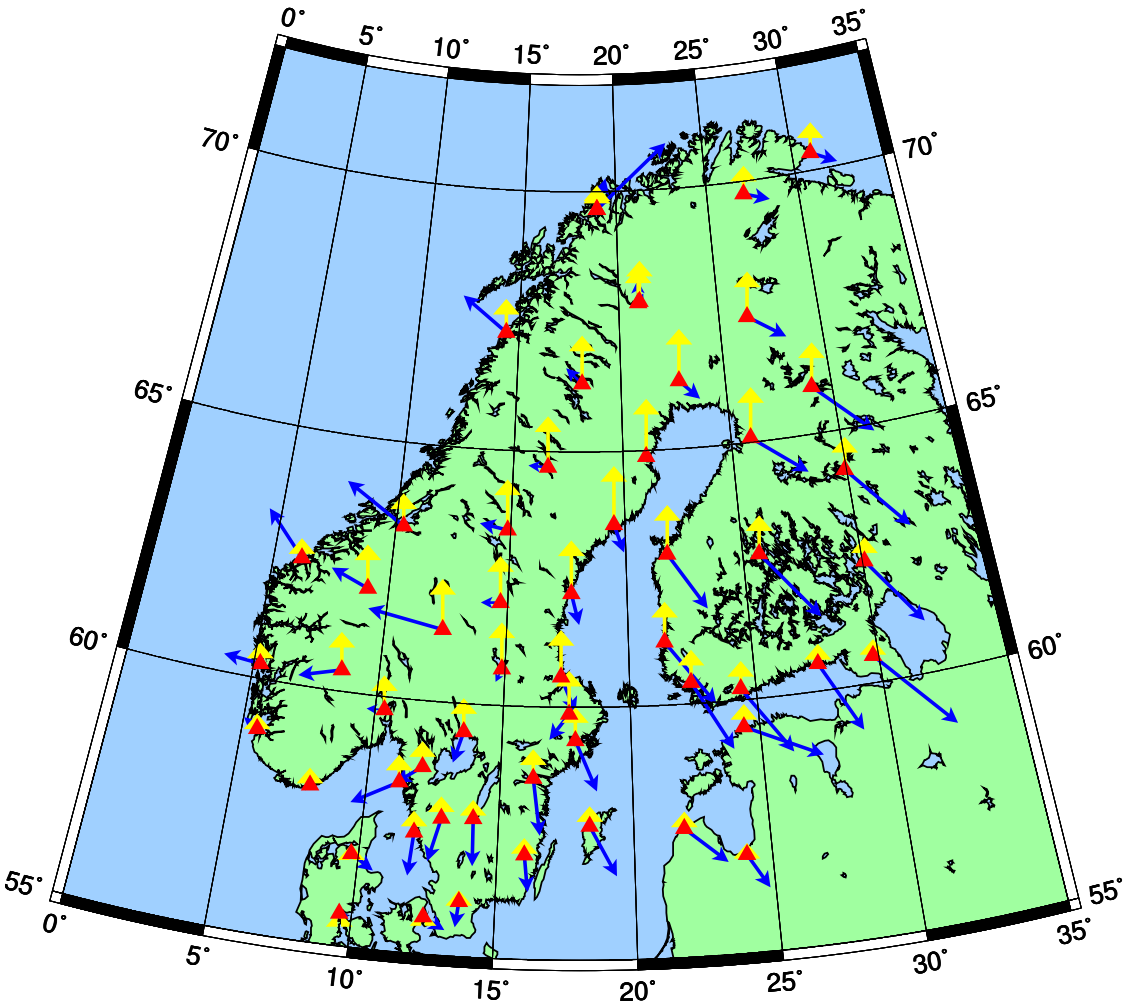


Figure 17.9. Horizontal and vertical motions in Fennoscandia as determined by the BIFROST project. The horizontal and up arrows are not on the same scale. Source Lidberg et al. (2009).

mareographs, also called tide gauges, at the coast

- monitoring changes in gravity, e.g., the Nordic gravity profiles (Mäkinen et al., 2010). The changes in gravity have *two causes*:
  1. The point of measurement shifts, because of the land uplift, away from the Earth's centre, and gravity diminishes.
  2. Due to the flow of mass under the Earth's crust, the amount of mass under the point increases and gravity grows stronger.

  The observed change in gravity is the resultant of both effects.
- GNSS monitoring in three dimensions. This is a fairly young technique and the time series are still short, because GNSS receivers with the necessary precision, and sufficiently precise processing methods, have existed only since the 1990s.

17.1.3.2 Ocean and atmospheric loading

Variations in sea level, e.g., the tides, may be reflected in the level of the Earth's crust in coastal areas. The motion may even be a couple of centimetres, but peters out quickly going in-land. Only in recent years has it been possible to measure this motion using GNSS, but the uncertainties are large. Also in the long time series of gravimetric monitoring, this tidal loading is visible. It is one way of studying the local elastic properties of the solid Earth.

Also the effect of the atmosphere, mostly variations in air pressure, should be visible in this way. The phenomenon is however very weak and hard to observe with confidence. The problem with gravimetric techniques again is the difficulty of separating the effect of loading from the atmosphere's own varying attraction on the measurement device.

17.1.4 Local geodynamics

17.1.4.1 Anthropogenic motions of the Earth's crust

Mankind, by its activities, often causes motions of the Earth's crust, so-called *anthropogenic* motions. E.g., building a reservoir dam and the filling of its reservoir often causes an additional loading of the local Earth's crust, and may even cause tiny earthquakes.

In Venezuela, the pumping of oil in the Caracas region has caused a very noticeable local subsidence of the land. There are many other, similar areas in the world. In the Netherlands, in Groningen, the pumping of natural gas has caused a subsidence, for which farmers are financially compensated. For this reason, the motions are measured with geodetic precision, at regular intervals.

Also the pumping of drinking and irrigation water from porous layers (*aquifers*) can cause a subsidence of the land by as much as metres. This is everywhere, but especially in developing countries, a very common problem[6].

17.1.4.2 Natural motions of the Earth's crust

Local movements of the Earth's crust happen often close to geological faults, like the San Andreas fault in California, the boundary between two tectonic plates, where the plates move slowly with respect to each other. When the movement of the plates gets stuck, stresses inside the Earth's crust build up, and may after years be released destructively in the form of an earthquake.

For monitoring motions of the Earth's crust, nowadays various geodetic methods are used. Also at the mid-ocean ridges, like in Iceland, all kinds of geodetic measurement activity take place in the service of geophysical research.

[6] http://en.wikipedia.org/wiki/Groundwater-related_subsidence.

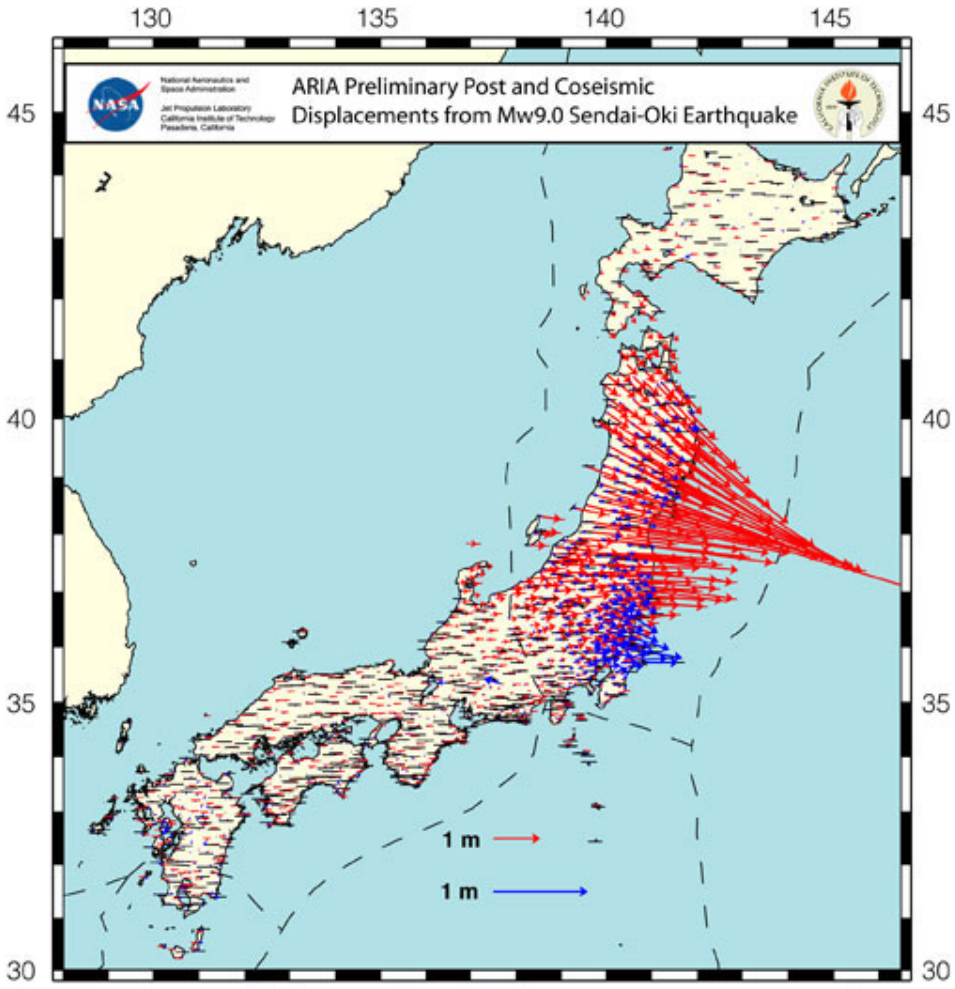


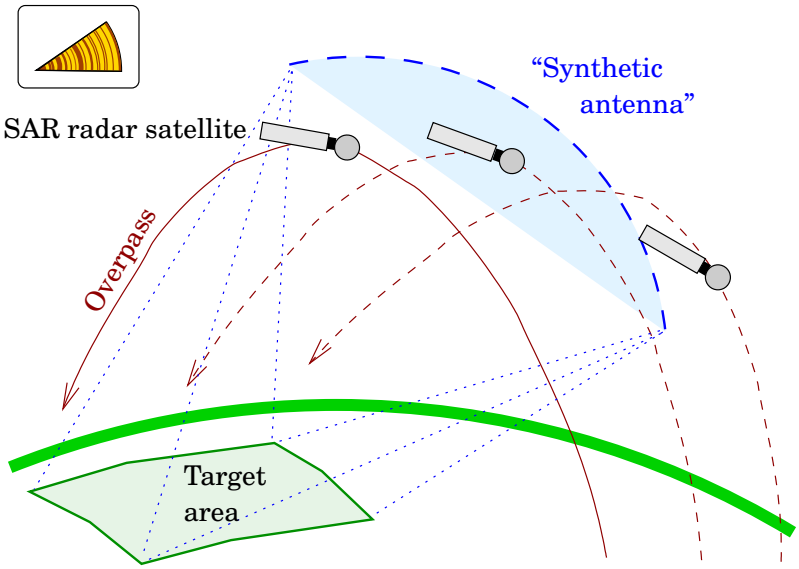
Figure 17.10. Horizontal motions in connection with the Japanese Sendai earthquake of 2011, as measured with the Japanese national continuously operating GNSS network. The original motion of the earthquake in red, aftershocks in blue. © and computation: Advanced Rapid Imaging and Analysis (ARIA) team (JPL and CalTech). Data: Japan's Geospatial Information Authority (GSI). At the address goo.gl/wfrJ6T an animation is found. The data is collected by 1200 receivers a rate of one per second.

□

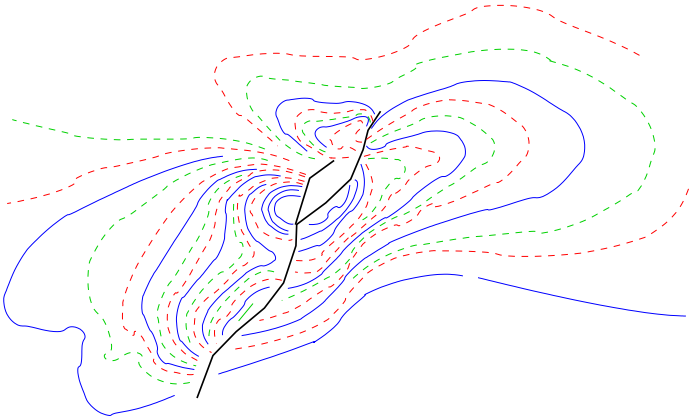
The continental plates are not necessarily completely rigid and of one piece, rather also inside them there are all kinds of faults along which have taken place — and surely still take place — tectonic movements. In addition to observing microseismicity, one could try to observe these with geodetic monitoring techniques (Ahola, 2001).

Deformation monitoring For monitoring local deformations in the Earth's crust, one may use both traditional (total-station measurement, precise levelling) and modern techniques like GNSS monitoring.

In earthquake prone areas (e.g., Japan, California) GNSS monitoring networks comprising many hundreds of continuously operating receivers



(a) The geometry of SAR imaging



(b) Example of an InSAR image

Figure 17.11. InSAR image. The deformation interval between interference fringes of the same colour is some 10 cm. The black line represents the fault.

□

have been built. However, also more traditional techniques, like monitoring with automatic total stations, is used. Collecting and processing the materials is done in real time.

Detecting deformations from the observational material is a similar task as detecting gross errors: the same kind of *statistical testing* (section 14.3) may be used, with a suitably chosen alternative hypothesis H_a.

Also the SAR (synthetic-aperture radar) technique from satellites[7] is used for deformation monitoring. "Synthetic aperture" means that the radar images taken during different satellite overpasses are computationally combined in such a way, that a "virtual objective" the size of the distance between overpasses is created. The method is interferomet-

[7] http://www-radar.jpl.nasa.gov/insar4crust/.

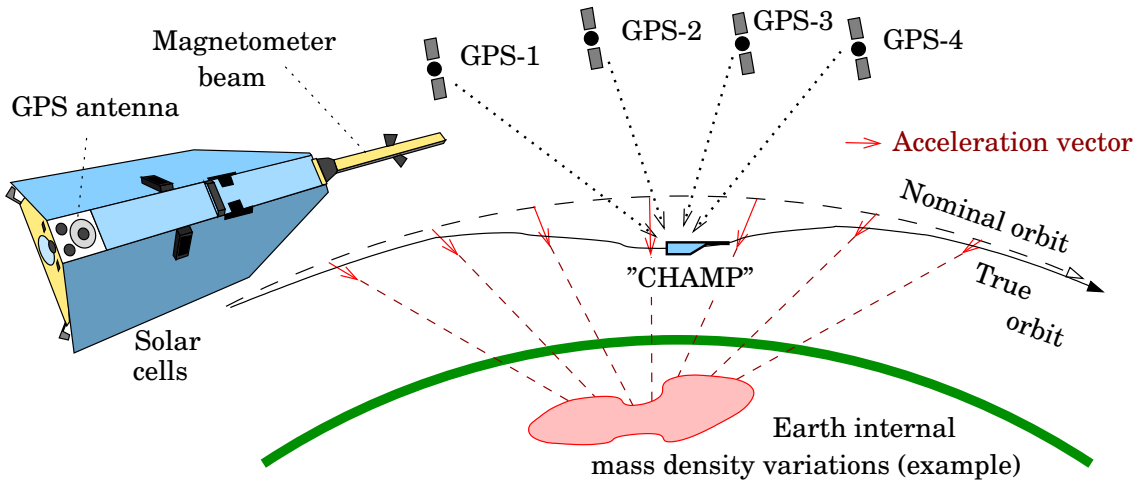


Figure 17.12. Determining the Earth's gravitational field by tracking the orbit of a low-flying satellite using a GPS receiver.

ric (InSAR), after suitable processing, the deformations show up in the images as interference fringes.

17.2 Studying the Earth's gravity field from orbit

During the decade 2000 – 2010, three satellite missions were launched to study the fine structure of the Earth's gravity field or geopotential, in other words, to draft a global geoid map.

CHAMP (Challenging Minisatellite Payload for Geophysical Research and Applications, http://op.gfz-potsdam.de/champ) was launched into orbit from the Plesetsk cosmodrome in Russia on July 15, 2000. The orbital height of CHAMP was initially only 450 km, which during the flight diminished to 350 km. (For comparison: the orbital height of the GPS satellites is 20,000 km.)

CHAMP carried a GPS receiver for precise orbit determination of the satellite. From GPS data, one may compute the precise location $\mathbf{x}(t)$ of the satellite as a function of time. From this, one may calculate the geometric acceleration $\mathbf{a}(t)$ by differentiation:

$$\mathbf{a}(t) = \frac{d^2}{dt^2}\mathbf{x}(t).$$

The satellite also carried an accelerometer, which served to measure the satellite accelerations caused by the atmosphere's aerodynamic forces acting on the satellite. After eliminating these, what is left are only the accelerations caused by the Earth's gravitational field, from which a precise geopotential or *geoid model* may be computed.

The data collected by CHAMP has been used to compute global geopotential models. The geographical resolution of the models

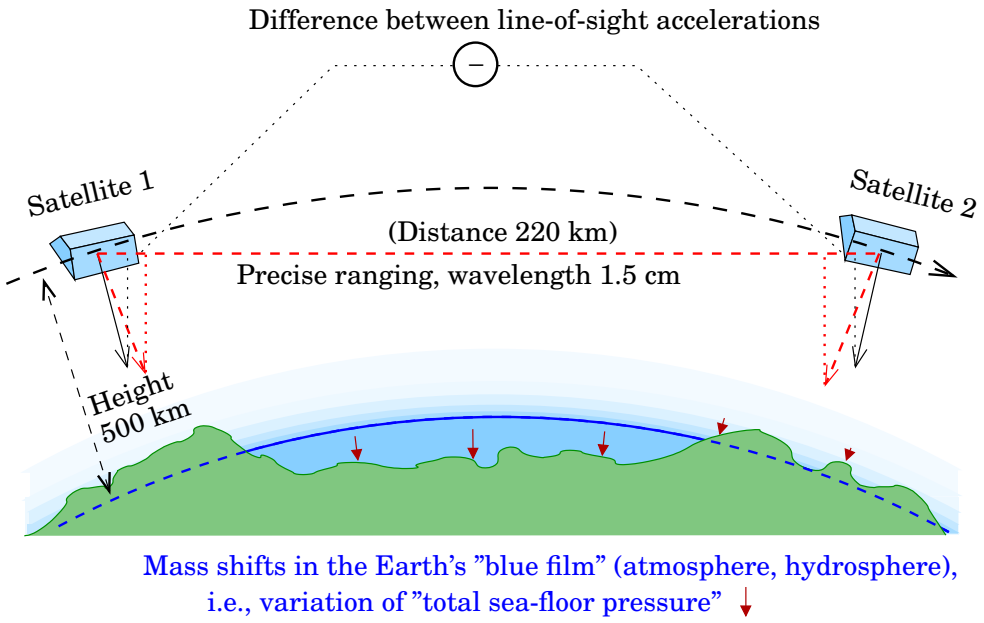


Figure 17.13. Basic idea of the GRACE satellite pair: measuring the tiny temporal changes in the Earth's gravity field using SST (Satellite-to-Satellite Tracking).

has been modest, order 1000 km. The satellite returned into the atmosphere on September 19, 2010 and burned up after 58,277 orbits.

GRACE (Gravity Recovery And Climate Experiment Mission[8]) measured the *temporal changes* in the Earth's gravity field, very precisely, at a time resolution of about a month, at a fairly crude spatial resolution. These temporal changes are mostly caused by movements of the "blue film", the atmosphere and hydrosphere. The quantity measured is also called the *sea-floor pressure*, perhaps a bit surprisingly. The explanation is, that the quantity is proportional to all of the mass contained in a column of air and water.

GRACE was a *satellite pair*: the satellites ("Tom and Jerry") flew in a tandem formation at some 450 km height, at an average separation of 220 km. A microwave link measured changes in the distance between the satellite at an accuracy of 1 μm/s. Both satellites also carried sensitive accelerometers to measure and eliminate atmospheric drag.

The measurement system was so sensitive, that changes in a water layer of even a millimetre thickness could be noticed, if they extended over an area the size of a continent (some 500 km).

[8] http://www.gfz-potsdam.de/portal/gfz/Struktur/Departments/Department+1/sec12/projects/grace.

The successful launch took place in 2001. The data has been a treasure for hydrologists[9]. It wasn't until 2017 when one of the satellites developed a malfunction and measurements came to a close. In December 2017 the satellite re-entered the atmosphere.

A GRACE Follow-On mission (GRACE-FO) is being planned for launch in 2017[10].

GOCE (Gravity Field and Steady State Ocean Circulation Explorer) was the most ambitious of the satellite missions. The satellite was launched from Plesetsk, Russia, on March 17, 2009. The orbital height was only 250 km, and the satellite carried a rocket engine (an ionic engine) and a stock of propellant (xenon) for orbit maintenance against atmospheric drag. The GOCE payload contained a *gravity gradiometer*, an instrument for precisely measuring components of the *gradient* of the Earth's attraction, the dependence of the attraction on the three co-ordinates of place. The gradiometer consisted of several extremely sensitive accelerometers mounted on a frame.

GOCE has worked well. However, in July 2010 a serious malfunction occurred in the telemetry link, which was repaired in August. The mission ended October 21, 2013, and on November 11, the satellite returned into the atmosphere and was seen burning up over the Falkland islands.

It has been theoretically established, that gradiometry is the best way to measure the very local features of the gravity field, better than by orbital tracking with GNSS. The smallest details in the geoid map that can be seen in the GOCE material are 100 km in diameter, their precision being as good as ±2 cm.

One important application for a global geoid map this precise is mapping the deviations of the mean sea surface from the geoid, an equipotential surface, with similar precision. The true location in space of mean sea level is obtained by satellite radar altimetry, also at a precision level of several centimetres. This difference in level between the sea surface and an equipotential surface, the *sea-surface topography*, may be *inverted* into a map of ocean currents. The theory behind this is explained in subsection 17.6.2. See also figure 17.15.

The name of the GOCE mission was inspired by this possibility.

[9] https://commons.wikimedia.org/wiki/File:Global_Gravity_Anomaly_Animation_over_LAND.gif.

[10] https://gracefo.jpl.nasa.gov/mission/overview/.

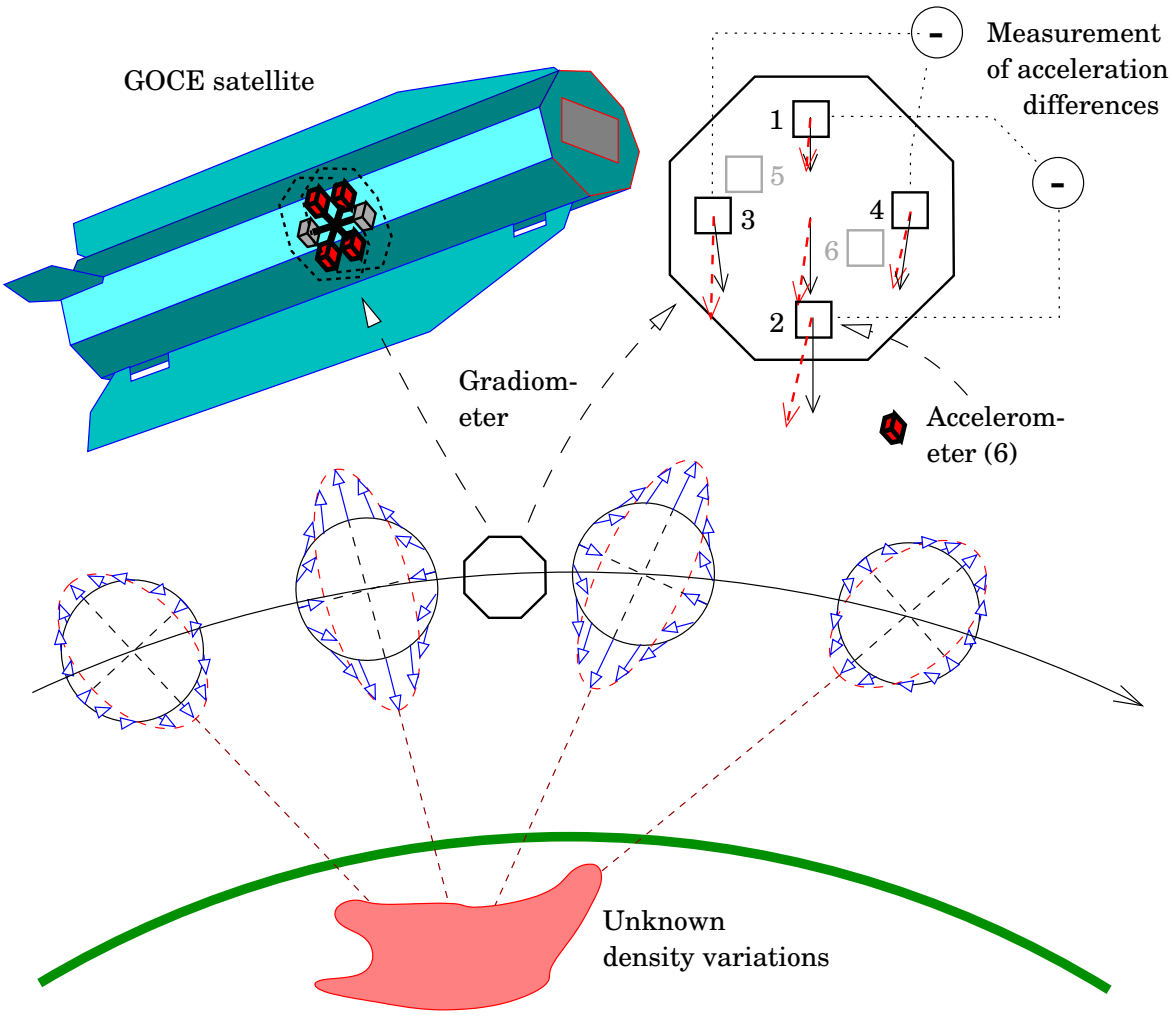


Figure 17.14. Determining the Earth's gravitational field with the gravity gradiometer on-board the GOCE satellite.

17.3 Atmospheric research and GNSS

17.3.1 Water-vapour values from a GNSS network

The atmosphere has an effect on the propagation of radio waves, also of the GNSS signal. A can be seen from refractive-index equation for microwaves 11.4:

$$N_{\mathrm{M}} = 10^6 \cdot (n_{\mathrm{M}} - 1) = \frac{77.624\,\mathrm{K/hPa}}{T}(p-e) + \frac{64.70\,\mathrm{K/hPa}}{T}\left(1 + \frac{5748\,\mathrm{K}}{T}\right)e$$

both the total air pressure p and the partial pressure of water vapour, or *absolute humidity*, e affect propagation, but in different ways. If we substitute $T \approx 285\,\mathrm{K}$ (approximately 12 °C), we obtain

$$N_{\mathrm{M}} = 0.27\,(p-e) + 4.8e.$$

Write this into the form

$$N_{\mathrm{M}} = a\,p_{\mathrm{dry}} + b\,p_{\mathrm{wet}},$$

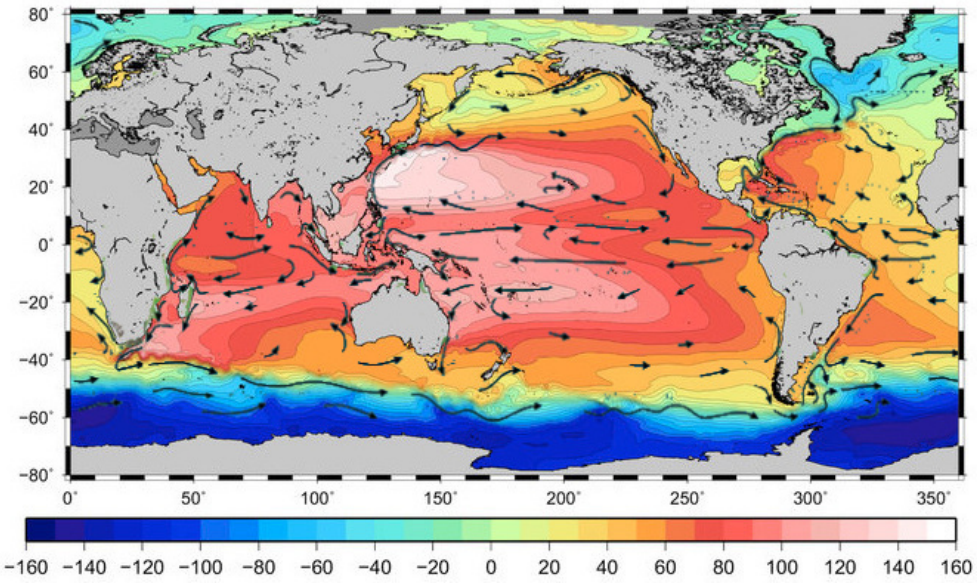


Figure 17.15. A sea-surface topography map produced by the GOCE mission © European Space Agency. Unit cm. Ocean surface currents drawn on top © NOAA / Rick Lumpkin (Lumpkin and Garraffo, 2005). Compare with figure 17.21!

in which $p_{\text{dry}} = p - e$ is the total of the partial pressures of the dry constituents of the atmosphere, and $p_{\text{wet}} = e$ is the partial pressure of water vapour. The coefficients a and b describe the influences of dry air (mostly nitrogen, oxygen, argon, and carbon dioxide) and water vapour. The values $a \approx 0.27\,\text{hPa}^{-1}$ and $b \approx 4.8\,\text{hPa}^{-1}$ depend somewhat on the atmosphere's temperature profile with height. However, b is some 18 times a: water vapour affects the propagation of the GNSS signal some 18 *times stronger than dry air*. Behind this phenomenon is the *chemical polarity* of the water molecule, see section 6.5.

The partial pressure of water vapour is a less used way to describe how much water vapour there is in the atmosphere. The more used way is *relative humidity* in percent, the amount of water vapour compared to the amount at *saturation*, when there is as much water vapour in the air as there can be at a certain temperature, before it starts condensing out as liquid. The saturation partial pressure e_{sat} is given by the equation of *Clausius*[11] and *Clapeyron*[12] :

$$e_{\text{sat}}(T) = e_{\text{sat}}(T_0) \exp\left(-4895\,\text{K} \cdot \left(\frac{1}{T} - \frac{1}{T_0}\right)\right).$$

When we know that the temperature $T_0 = 100\,^\circ\text{C} = 373.15\,\text{K}$ is the boil-

[11] Rudolf Julius Emanuel Clausius (1822 – 1888) was a German physicist and methematician, one of the founders of thermodynamics.

[12] Benoît Paul Émile Clapeyron (1799 – 1864) was a French engineer and physicist, one of the 72 names on the Eiffel Tower.

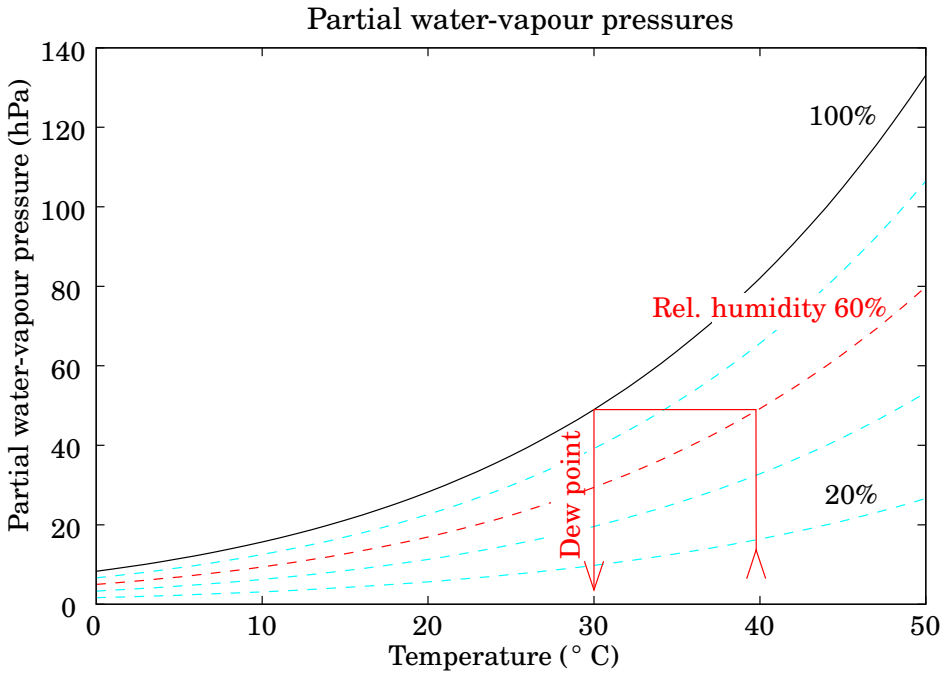


Figure 17.16. Saturation partial pressure e_{sat} and partial pressures e of water vapour at various temperatures and relative humidities. The red path tells how from temperature (40° C) and "wet-bulb temperature" (30° C) follows relative humidity (60%). Many hygrometers — devices for measuring the humidity of air — are based on this principle.

ing point of water, i.e., $e_{\mathrm{sat}}(T_0) = 1\mathrm{atmosphere} = 1013.25\,\mathrm{hPa}$, it follows that

$$e_{\mathrm{sat}}(T) = 1013.25\,\mathrm{hPa} \cdot \exp\left(-4895\,\mathrm{K} \cdot \left(\frac{1}{T} - \frac{1}{373.15\,\mathrm{K}}\right)\right).$$

The path through the atmosphere is the longer, the lower the satellite is in the local sky. If the zenith angle is ζ, we may describe the signal propagation delay in the atmosphere by the following equation ($\widetilde{a}$ and $\widetilde{b}$ are the effective averages of the above mentioned coefficients a and b over the full height of the atmosphere):

$$d_{\mathrm{trop}} = \frac{\left(\widetilde{a}\,p_{\mathrm{dry}} + \widetilde{b}\,p_{\mathrm{wet}}\right)}{\cos\zeta} = \frac{\left(\widetilde{a}\,(p-e) + \widetilde{b}\,e\right)}{\cos\zeta} = \frac{d_{\mathrm{zenith}}}{\cos\zeta}. \qquad (17.1)$$

In a GNSS network adjustment, we may solve the station values d_{zenith}zenith propagation delays, as unknowns, by adding equation 17.1 to the GNSS observation equations. (In a small area, one common constant term will remain unresolved.)

If we measure, at the same station, with a barometer also the total air pressure

$$p = p_{\mathrm{dry}} + p_{\mathrm{wet}},$$

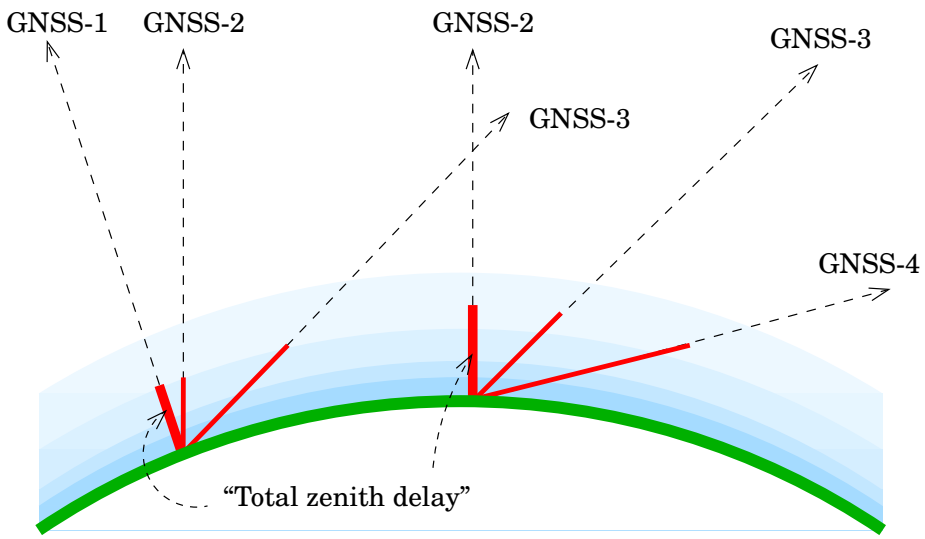


Figure 17.17. Use of GNSS for studying the troposphere.

we may solve separately for p_{dry} and p_{wet}.

The latter quantity, converted to amount of matter in an air column, is also called *integrated water-vapour content* or *total precipitable water-vapour content* (converted to liquid water, unit mm). It is a quantity that weather and climate researchers are very much interested in!

17.3.2 GNSS radio occultation or "limbsounding"

Using low-flying satellites it is possible to measure how the radio signal from a GNSS satellite is slowed down as the lowest point of the ray dives deeper and deeper into the atmosphere. The technique is called GNSS limbsounding or radio occultation. The first satellite exploiting this method was GPS/MET; also the Danish Ørsted satellite, and the aforementioned CHAMP, used this method.

The method requires having a GNSS receiver on-board the satellite. As there are, only in the GPS system, already some 30 satellites, and the low-orbit satellite goes around the Earth once every 1.5 hours, a quite substantial amount of information is collected every 24 hours.

The method is important because it allows determination of temperatures high up in the atmosphere, above the tropopause, where there isn't much water vapour left. This is the layer in which climatic global warming would become most visible[13]. The ozone layer is located at the top of the measurement range, and processes taking place there also affect the temperature.

[13]Climate models predict surprisingly, and observations confirm, that while temperatures at the Earth's surface and in the troposphere go up, a compensating lowering of temperatures is expected for the stratosphere. The cause is enhanced radiative cooling by increasing CO_2 concentrations.

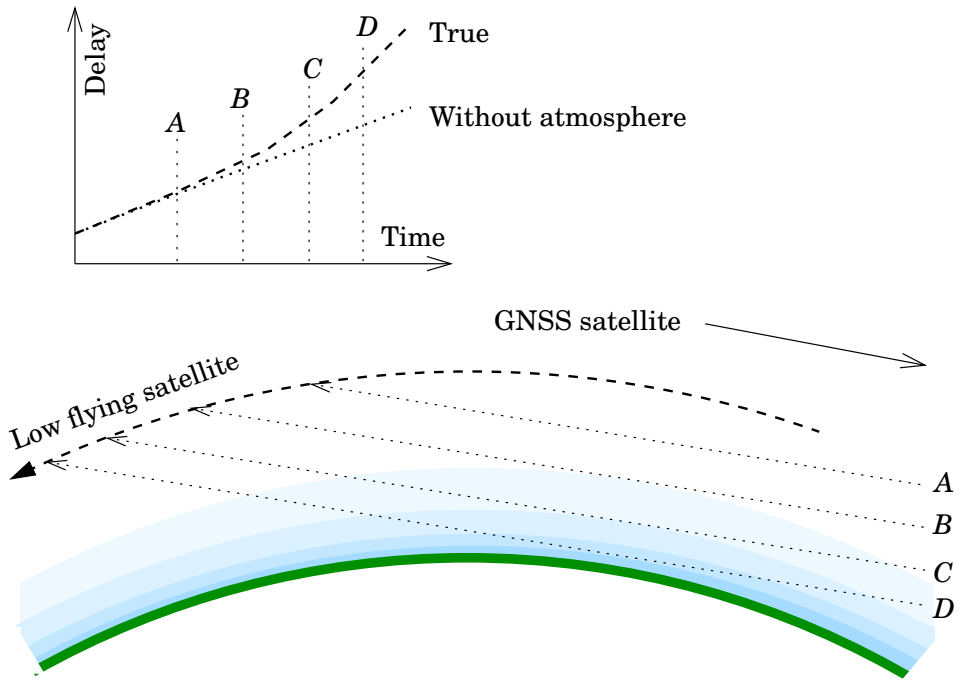


Figure 17.18. GNSS radio occultation.

The method is the following: the delay caused by the atmosphere is proportional to air density, which is an exponential function of height[14]:

$$\rho(H) = \rho(H_0) \cdot \exp\left(-\frac{H - H_0}{S(t,k,g)}\right),$$

in which $S(t,k,g)$ is the *scale height*, a function of air temperature t, air composition k, and gravity g. If we assume k[15] and g known, we may calculate from S the temperature t.

17.3.3 Ionosphere sounding

From GPS measurements at two frequencies, L_1 and L_2, we may compute *ionospheric models*, in which is mapped, as a function of place (φ, λ, h) and time, the ionospheric *electron density* (TEC, total electron content). This is being done, nearly in real time, using the global GNSS network, since 1998 by the IGS, the International GNSS Service. A data format, IONEX, Ionosphere Map Exchange Format, has been developed in support of this activity.

[14] It is a bit more complicated than this, as the GNSS signal travels through many atmospheric layers, not only the layer of closest approach to the Earth's surface.

[15] The water vapour content at these heights is practically zero due to the low temperature.

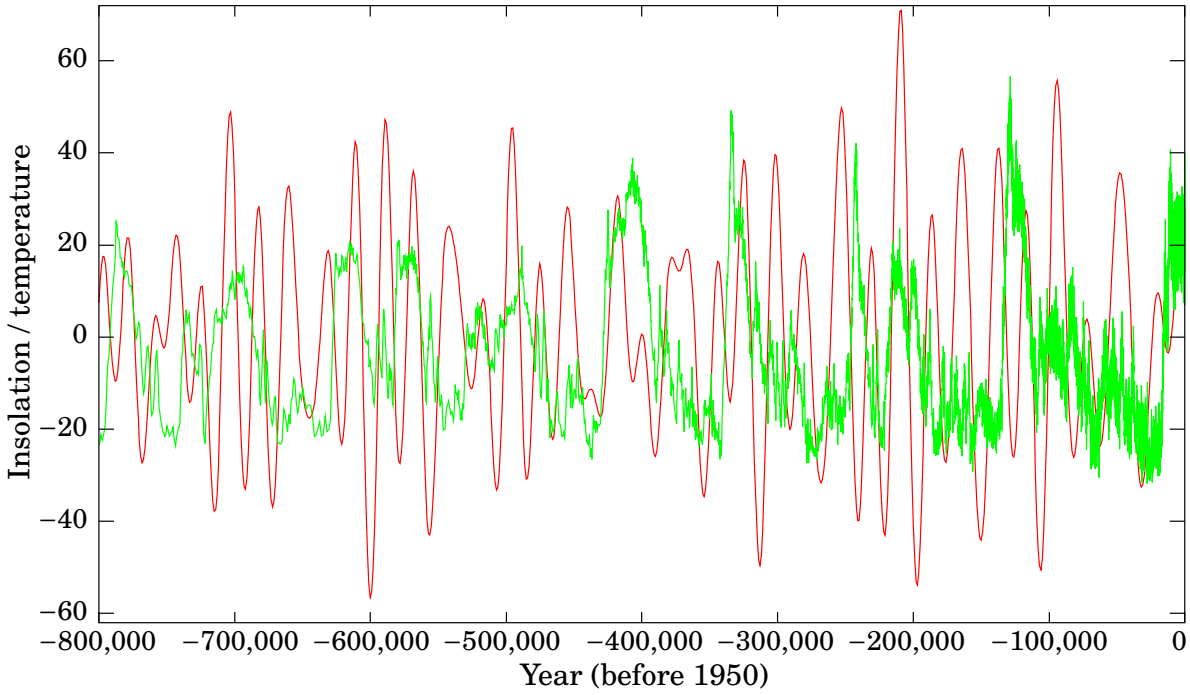


Figure 17.19. Milanković cycles over the past 800,000 years on both hemispheres.
Red, theoretically computed summer-solstice insolation at 65° Northern latitude (outside the atmosphere).
Green, Antarctic temperature estimate, measured using the deuterium isotope.
Arbitrary scales. Sources: Laskar et al. (2004); Jouzel ym. (2007). See also the *Science* article Jouzel et al., 2007.

17.4 Long-term variations in the Earth's rotation axis and orbit

Milanković[16] proposed already in 1941 a hypothesis, that variations in the *insolation* — the radiative energy coming from the Sun — at the edges of the large continental ice sheets cause their growth and retreat, and that these variations would be caused by astronomical factors. The work of Milanković is based on earlier work, i.a., by James Croll[17], to explain the ice ages and interglacials.

17.4.1 Tilt of the rotation axis

Besides the precessional motion, also the obliquity, or *tilt* of the Earth's axis of rotation with respect to the ecliptic varies slowly. Currently

[16] Milutin Milanković (1879 – 1958) was a Serbian polymath, engineer and climatologist.

[17] James Croll (1821 – 1890) was a Scottish autodidact physicist, astronomer and climatologist, as well as a fellow of the Royal Society.

(epoch 2000.0) the tilt angle is $23°26'21''$ and it is slowly diminishing, by about $0''.47$ per year.

Actually the variation is periodic, between $22°.1$ and $24°.5$, with a period of some 41,000 years. This variation is related to orbital perturbations caused by the planets on the Earth's orbital plane.

17.4.2 Climatological precession

Astronomical precession is the above mentioned phenomenon, that the rotation axis of the Earth turns around in a conical figure in a period of about 25,800 years. Additionally also the semi-major axis of the Earth's orbital ellipse rotates in the orbital plane, the period being 112,000 years (apsidal precession). This motion is caused by orbit perturbations by the other planets.

The resultant of both processes is *climatological precession*, the period of which is

$$\left(\frac{1}{112{,}000}+\frac{1}{25{,}800}\right)^{-1}\text{years}=21{,}000\,\text{years}.$$

This is the time period in which the season when the Earth is closest to the Sun cycles through the calendar. Currently the Earth is closest to the Sun in the beginning of January. Still 11,000 years ago, at the termination of the last ice age, the Earth was closest to the Sun in July. The result was 7% more solar energy for the Arctic summers.

Note that the direct impacts of climatological precession are opposite in the Northern and Southern polar areas, because the boreal (Northern) summer is the austral (Southern) winter, and the austral summer is the boreal winter.

17.4.3 Variation of the orbital eccentricity

The eccentricity of the Earth's orbit is also variable, the main periods being 108,000 and 412,000 years, between the values 0 and 0.05. This variation, caused by orbital perturbations from the other planets, modulates the effect of climatological precession.

17.5 Land-ice research, climate change

Much geophysical research is aimed at continental ice sheets, where, thanks to *polar amplification*, the change in our planet's climate is first becoming visible. A continental ice sheet is a *glacier*: it grows above the snow line from falling snow that slowly compacts to ice, flows slowly and plastically to the coasts, where it loses mass through melting and formation of icebergs ("calving").

The thicknesses of both the Greenland and the Antarctic ice sheet have been monitored by satellite techniques, e.g., radar altimetry. Also radar mapping from orbit (SAR, synthetic-aperture radar) has been used for this. Observing the small changes of course requires long time series, and this is still a deficiency of the satellite methods.

Also GRACE (see section 17.2) has been used for studying the change in mass of these continental ice sheets. Contrary to the change in thickness, this change represents directly the contribution of these ice sheets to sea-level rise.

Also the GNSS technique has been used for studying both crustal and ice motions, both in Greenland and in Antarctica, e.g., in the surroundings of the Finnish Antarctic base Aboa in Dronning Maud Land (Koivula and Mäkinen, 2003; Khan et al., 2010).

17.6 Geodetic oceanography

17.6.1 Satellite altimetry and the geoid

It is possible to measure from a satellite the location of the momentaneous sea surface in a geocentric system. The instrument used is a *radar altimeter* Rummel and Sansò (1993) and its precision of measurement is a few cm.

A satellite altimetric radar works in this way, that it sends a short microwave pulse straight downward, where it is reflected back from the sea surface. The reflection is received and the shape of the return pulse is analyzed. The reflection doesn't just come from the point straight underneath the satellite, but from a whole area, the so-called *footprint*. This footprint may have a diameter of several kilometres, more if the sea state includes high waves.

Based on the analysis of the pulse shape, the distance between satellite and sea surface is inferred. If the satellite orbit is known — nowadays altimetric satellites always carry a GNSS receiver — the momentaneous, geocentric location of the sea surface may be computed. The measurement points together form *arcs* running either from South to North or from North to South.

17.6.2 Ocean currents and sea-surface topography

Mean sea level aligns pretty nicely with the geoid, i.e., an equipotential surface of the Earth's gravity field. If sea water were in a "state of rest", i.e., an equilibrium state like in communicating vessels, the mean sea surface would be exactly the same as the geoid. This is, however, not the case. The causes of this, so-called, *sea-surface topography* are:

1. local variations in *temperature* and *salinity*

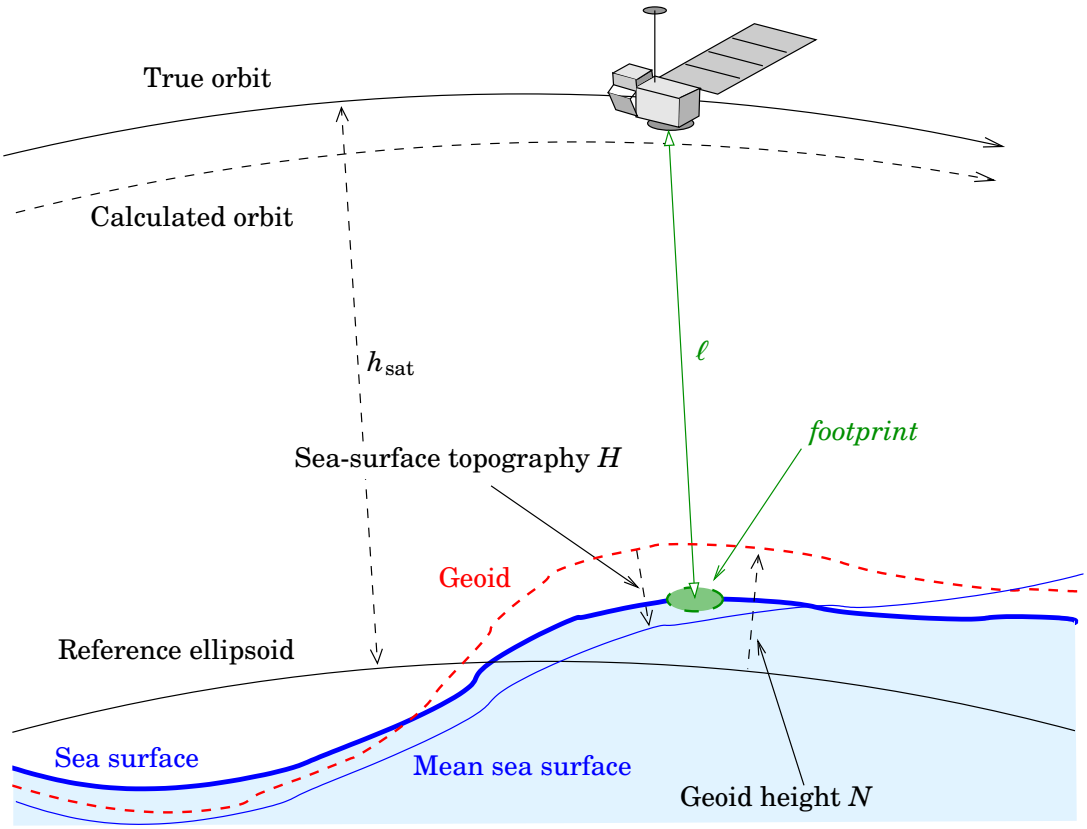


Figure 17.20. The measurement geometry of satellite radar altimetry.

2. variations in *air pressure* from place to place. The sea surface responds like an "inverted barometer" (IB) to these
3. winds cause *drag* with the sea surface, pushing water masses in certain directions
4. ocean currents respond to the rotation of the Earth though the

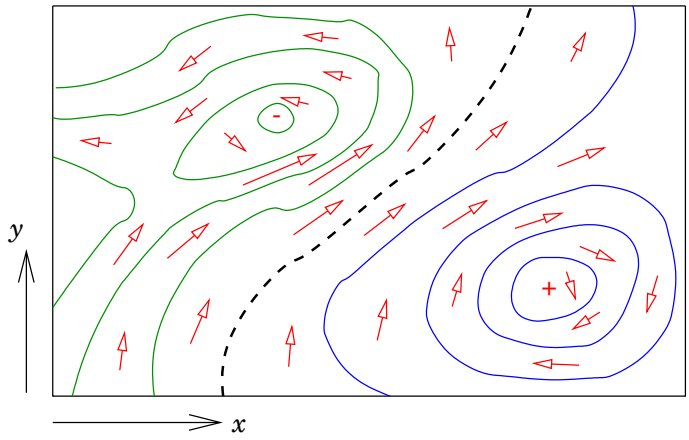


Figure 17.21. Theoretical connection between sea-surface topography and ocean currents.

Coriolis[18] force, in the same way as air currents (Buys-Ballot[19]'s law: currents try to turn to the right on the Northern hemisphere, to the left on the Southern one. This causes a *transversal tilt* of ocean currents, which is proportional to the flow velocity and the sine of latitude, $\sin\varphi$. We speak of *geostrophic flow*.

The magnitude of the sea-surface topography is globally $\pm 1\,\mathrm{m}$. Also the Baltic Sea has a sea-surface topography, a tilt of the sea surface, as a result of which the bottoms of the Gulfs of Finland and Bothnia are about $20-30\,\mathrm{cm}$ higher than the Western parts of the Baltic Sea. There is also a height difference across the Danish straits. The reasons are mentioned above: (1) and (3).

Earlier on, some satellite missions were mentioned that were intended to determine the precise location in space of the *geoid*. The sea-surface topography, i.e., the difference between the levels of mean sea level and geoid, is caused, on the oceans, mostly by *ocean currents*. The rotation of the Earth causes currents to tilt in the transversal direction. Using this observable phenomenon, it is possible to theoretically compute the current volume, i.e., the amount of water transported, and from this in turn the thermal energy carried along with the water. For climate research, this is an essential matter. This was one objective of the ambitious GOCE project (section 17.2).

In figure 17.21 we see what is the relationship between ocean currents and sea-surface topography. We may also describe this as equations, the so-called *geostrophic equations*:

$$\frac{\partial H}{\partial x} = +2v_y \frac{\omega}{\gamma}\sin\varphi,$$
$$\frac{\partial H}{\partial y} = -2v_x \frac{\omega}{\gamma}\sin\varphi,$$

in which $H(x,y)$ is the sea-surface topography, ω the rotation rate of the Earth, γ gravity, x and y plane co-ordinates in the East and North direction, and v_x and v_y are the East and North components of the flow velocity. Because the equations contain $\sin\varphi$, the effect is opposite on the Northern and Southern hemispheres[20].

By simple partial differentiation, we may infer from the observed sea-surface topography $H(x,y)$ to the flow pattern $\mathbf{v}(x,y)$, in which $\mathbf{v} = v_x\mathbf{i} + v_y\mathbf{j}$ is the current velocity vector in the (x,y) plane.

---

[18] Gaspard-Gustave de Coriolis (1793 – 1843) was a French mathematician, one of the 72 names on the Eiffel Tower.

[19] Christoph Hendrik Diederik Buys Ballot (1817 – 1890) was a Dutch chemist and meteorologist.

[20] . . . and in the immediate vicinity of the equator, it vanishes.

17.6.3 Mareographs, satellite altimetry, sea-level rise

Starting in 1992, the French-American satellite TOPEX/Poseidon and its follow-up satellites Jason-1 (http://topex-www.jpl.nasa.gov/), Jason-2 and Jason-3 have been doing pioneering work monitoring the rise of global sea level. Traditionally, sea level has been monitored using *mareographs* or *tide gauges*; the rise of sea level over the whole 20th century has been estimated to have been of order $1.3-2\,\mathrm{mm/a}$, i.e., $13-20\,\mathrm{cm}$ over the whole century[21].

A mareograph measures the rise of sea level with respect to the Earth's crust on which it has been built. A long tube connects the instrument with the open sea, in order to dampen wave motion.

This method has two problems:

1. The Earth's crust itself may move. Important here is mostly the so-called post-glacial isostatic land uplift, more commonly GIA (Glacial Isostatic Adjustment) which affects most strongly in Fennoscandia and Canada, but the effects of which extend to all of Europe and North America. Outside the land-uplift area there is a broad zone in which the land subsides slowly, the so-called "periglacial bulge". This phenomenon must be carefully modelled and removed from the mareograph data, if one wishes to compute global sea-level rise precisely. Nowadays there are continuously operating GNSS stations co-located with many mareographs, with the help of which the vertical land motion may be empirically determined. Unfortunately these time series are still short.
2. There are only mareographs on the coasts, and there, only in individual points. There are large ocean areas where there is not a single mareograph nearby. And the further one goes back in time, the weaker the situation becomes. On the Southern hemisphere, the situation has been really weak before around 1950.

Altimetric satellites can measure the location of the sea surface *geocentrically* — using an on-board GNSS positioning device — and *everywhere* underneath their orbits. E.g., the inclination angle of the orbit of the satellite TOPEX/Poseidon is 66°, and the whole ocean surface between latitudes 66°S and 66°N is being mapped at an interval of some ten days. When the plane of the satellite orbit precesses, with respect to the Sun, once in 60 days, we obtain the long-term trend in global mean sea level with this temporal resolution. During the last 17 years, mean sea level has risen some $3.2\pm0.4\,\mathrm{mm/a}$, clearly more than during the 20th century on average[22]. The reason for the acceleration of the rise is undoubtedly

[21] A new analysis (Dangendorf et al., 2017) proposes that sea-level rise before 1990 was even substantially slower, only some $1.1\pm0.3\,\mathrm{mm/a}$.

[22] http://sealevel.colorado.edu/.

Figure 17.22. Tide gauges (mareographs) of the Baltic Sea, and some ground tracks of the Seasat satellite back in 1978.

the warming of the climate.

17.6.4 Height systems, mareographs and sea-surface topography

Also *mareographs* may be used to observe the sea-surface topography, but only on the coast. This presupposes, that the mareographs are connected to the same height system. In Finland, the matter has been organized so, that the 13 mareographs on the Finnish coast operated by the Finnish Institute of Marine Research, nowadays a part of the Finnish Meteorological Institute, are connected to the national precise levelling network, which the Finnish Geodetic Institute, nowadays the Finnish National Land Survey's Geospatial Research Institute FGI, maintains. The necessary connecting levellings are carried out by the geodesists at

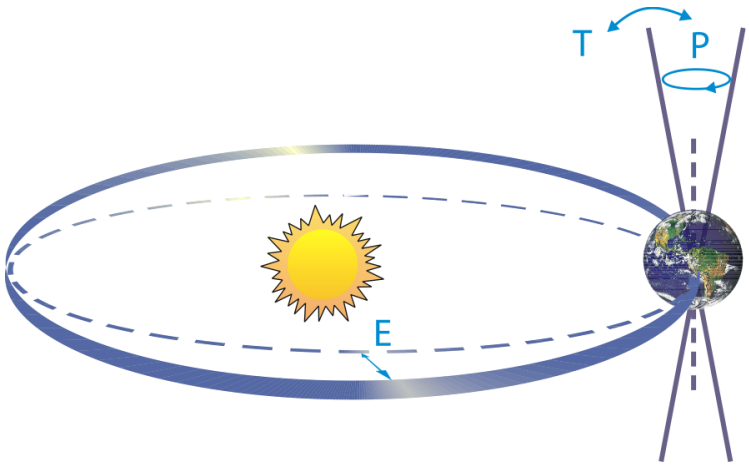


Figure 17.23. *"Schematic of the Earth's orbital changes (Milankovitch cycles) that drive the ice age cycles. 'T' denotes changes in the tilt (or obliquity) of the Earth's axis, 'E' denotes changes in the eccentricity of the orbit (due to variations in the minor axis of the ellipse), and 'P' denotes precession, that is, changes in the direction of the axis tilt at a given point of the orbit."*
From IPCC (2007), Frequently Asked Question 6.1.

a few years' intervals.

The Finnish mareographs measure the level of the local sea surface with respect to the local Earth's crust once every half hour. The measurement values are transferred electronically to the headquarters of the Finnish Meteorological Institute in Helsinki, where they are processed and archived.

Near five mareographs, *pillars* have been erected by the Finnish Geodetic Institute in order to carry out regular GNSS measurements. In this way one may obtain the local sea surface also in a geocentric reference frame by means of satellite positioning.

Also, use of GNSS buoys is becoming more common: with these, one obtains the geocentric location of the momentaneous sea surface, also far away from the coast. Also, sea state may be monitored.

Self-test questions

1. Which geodetic observation techniques useful for geodynamics do you know?
2. What is the evidence showing that the Earth's outer core is liquid?
3. Describe the polar motion and its monitoring.
4. In what way were the mission objectives of the GRACE and GOCE satellite missions different?

5. How can a network of GNSS stations be used to determine the water-vapour content of the atmosphere above it?
6. What is the atmospheric scale height, and how can GNSS radio occultation ("limbsounding") be used to determine it?
7. What are the three changes in the Earth's rotation axis and orbital motion causing long-term climatic variations — the glacial cycle — according to the theory of Milanković?
8. See figure 17.23. After what you have learned about Kepler's laws of orbital motion, what is wrong with this picture (and caption text)?
9. According to a newspaper article, https://goo.gl/1MFPGi, military aircraft that crash-landed in Greenland during the Second World War were found 46 years later 260 feet deep in the ice. Based on your knowledge of how glaciers work, explain how they ended up this deep. What is the annual snowfall rate at the site of the crash?
10. The geostrophic equations are

$$\frac{\partial H}{\partial x} = +2v_y \frac{\omega}{\gamma} \sin\varphi,$$
$$\frac{\partial H}{\partial y} = -2v_x \frac{\omega}{\gamma} \sin\varphi.$$

Explain each occurring symbol.

Bibliography

Joel Ahola. Nuottavaaran siirroksen deformaatiotutkimus GPS-mittausten avulla. Master's thesis, Teknillinen korkeakoulu, Maanmittausosasto, Espoo, February 2001.

Anon. *Elements of surveying*. Headquarters, Department of the Army, 1971. URL: https://books.google.fi/books?id=564XAAAAYAAJ&printsec=frontcover#v=onepage&q&f=false.

Anon. IAG Resolutions adopted at the XXIIth General Assembly in Birmingham, 1999. URL: https://iag.dgfi.tum.de/fileadmin/IAG-docs/IAG_Resolutions_1999.pdf, 1999. Resolution 3.

Willem Baarda. *A testing procedure for use in geodetic networks*, volume 2 no. 5 of *Publications on geodesy, new series*. Netherlands Geodetic Commission, Delft, 1968.

Willem Baarda. *S-transformations and criterion matrices*, volume 5 nr. 1 of *Publications on geodesy, new series*. Netherlands Geodetic Commission, Delft, second revised edition, 1981. URL: http://www.ncg.knaw.nl/Publicaties/Geodesy/pdf/18Baarda.pdf.

Silvio A. Bedini. Along Came a Spider – Spinning Silk For Cross-Hairs. *The American Surveyor*, March/April, May 2005. URL: http://www.amerisurv.com/PDF/TheAmericanSurveyor_BediniAlongCameASpider-SpinningSilkForCross-HairsPart1_March-April2005.pdf, http://www.amerisurv.com/PDF/TheAmericanSurveyor_BediniAlongCameASpiderPart2_May2005.pdf.

Mirjam Bilker-Koivula and Matti Ollikainen. Suomen geoidimallit ja niiden käyttäminen korkeuden muunnoksissa. Tiedote 29, Geodeettinen laitos, 2009.

Claude Boucher and Zuheir Altamimi. Memo : Specifications for reference frame fixing in the analysis of a EUREF GPS campaign. URL: http://users.auth.gr/kvek/20070327-MEMO-ver6.pdf, 2007.

M. A. R. Cooper. *Modern Theodolites and Levels*. Granada Publishing Ltd, London, second edition, 1982.

M. A. R. Cooper. *Control surveys in civil engineering*. Collins, Department of Civil Engineering, The City University, London, 1987.

Nicholas Crane. *Mercator – the man who mapped the planet*. Weidenfeld and Nicholson, London, 2002.

Á. Czobor and Zs. Németh. Balloon-borne beacons for stellar triangulation. *Advances in Space Research*, 1(11):29–34, 1981.

Sönke Dangendorf, Marta Marcos, Guy Wöppelmann, Clinton P. Conrad, Thomas Frederikse, and Riccardo Riva. Reassessment of 20th century global mean sea level rise. *Proceedings of the National Academy of Sciences*, 2017. doi: 10.1073/pnas.1616007114. URL http://www.pnas.org/content/early/2017/05/16/1616007114.abstract.

Charles DeMets, Richard G. Gordon, Donald F. Argus, and Seth Stein. Effect of recent revisions to the geomagnetic reversal time scale on estimates of current plate motions. *Geophysical Research Letters*, 21(20):2191–2194, 1994. http://www.earth.northwestern.edu/people/seth/Texts/nuvel1a.pdf.

J.L.E. Dreyer. The Well of Eratosthenes. *The Observatory*, 37:352–353, 1914. http://articles.adsabs.harvard.edu//full/1914Obs....37..352D/0000352.000.html, http://articles.adsabs.harvard.edu//full/1914Obs....37..352D/0000353.000.html.

Jürgen Ehlers, Philip Hughes, and Philip L. Gibbard. *The Ice Age*. Wiley-Blackwell, 2015. ISBN 978-1-118-50781-0.

Loránd Eötvös. *Three Fundamental Papers of Loránd Eötvös, ed. Zoltán Szabó*. Loránd Eötvös Geophysical Institute of Hungary, 1998. ISBN 963-7135-02-2. 299 pp.

FIG Commission 6. Engineering Surveys. Technical report, FIG, 1998.

Erik W. Grafarend, Horst Kremers, Juhani Kakkuri, and Martin Vermeer. Modelling vertical refraction in the SW Finland Triangular Network TAGNET adjustment. Report 87:2, Finnish Geodetic Institute, Helsinki, 1987.

Richard S. Gross. The excitation of the Chandler wobble. *Geophys. Res. Lett.*, 27:2329–2332, 2000.

Thomas Gruber, Reiner Rummel, Johannes Ihde, Gunter Liebsch, Axel Rülke, Uwe Schäfer, Michael Sideris, Elena Rangelova, and Philip Woodworth. Height System Unification with GOCE – Summary and Final Report. Technical report, European Space Agency, 2014. http://www.goceplushsu.eu/ext/doc/fin-doc/GO-HSU-RP-0021_1.0_Final_Summary_Report.pdf.

Werner Gurtner and Lou Estey. RINEX, The Receiver Independent Exchange Format, Version 3.00. Technical report, International GNSS Service, 2007. http://igscb.jpl.nasa.gov/igscb/data/format/rinex300.pdf.

Pasi Häkli, Jyrki Puupponen, Hannu Koivula, and Markku Poutanen. Suomen geodeettiset koordinaatistot ja niiden väliset muunnokset. Tiedote 30, Geodeettinen laitos, Masala, 2009. URL: http://www.fgi.fi/julkaisut/pdf/GLtiedote30.pdf.

Markku Heikkinen. Solving the shape of the Earth by using digital density models. Report 81:2, Finnish Geodetic Institute, Helsinki, 1981.

V. A. Heiskanen. *Kenttämittaus ja kartoitus*. Werner Söderström Oy, 1943. Toinen, lisätty painos.

V. A. Heiskanen and Seppo Härmälä. *Maastomittaus ja kartoitus*. Werner Söderström Oy, 1963. Viides, muuttamaton painos.

Weikko A. Heiskanen and Helmut Moritz. *Physical Geodesy*. W.H. Freeman and Company, San Francisco, 1967.

B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collins. *GPS Theory and Practice*. Springer-Verlag, Vienna, fifth corrected edition, 2001.

IPCC. Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, editors, *Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge Univ. Press, Cambridge UK, 2007.

J. Jouzel, V. Masson-Delmotte, O. Cattani, G. Dreyfus, S. Falourd, G. Hoffmann, B. Minster, J. Nouet, J.M. Barnola, J. Chappellaz, H. Fischer, J.C. Gallet, S. Johnsen, M. Leuenberger, L. Loulergue, D. Luethi, F. Parrenin H. Oerter, G. Raisbeck, D. Raynaud, A. Schilt, J. Schwander, E. Selmo, R. Souchez, R. Spahni, B. Stauffer, J.P. Steffensen, B. Stenni, T.F. Stocker, J.L. Tison, M. Werner, and E.W. Wolff. Orbital and Millennial Antarctic Climate Variability over the Past 800,000 Years. *Science*, 317 (5839):793–797, 2007.

Jouzel ym. EPICA Dome C Ice Core 800KYr Deuterium Data and Temperature Estimates. IGBP PAGES/World Data Center for Paleoclimatology Data Contribution Series # 2007-091. NOAA/NCDC Paleoclimatology Program, Boulder CO, USA., 2007.

JUHTA. JHS163. Suomen korkeusjärjestelmä N2000. WWW page, Julkisen hallinnon tietohallinnon neuvottelukunta, 2010. URL: http://www.jhs-suositukset.fi/web/guest/jhs/recommendations/163/full, accessed March 20, 2015.

JUHTA. JHS178. Kunnan paikkatietopalvelurajapinta. WWW page, Julkisen hallinnon tietohallinnon neuvottelukunta, 2012a. URL: http://www.jhs-suositukset.fi/web/guest/jhs/recommendations/178/full, accessed March 20, 2015.

JUHTA. JHS184. Kiintopistemittaus EUREF-FIN-koordinaattijärjestelmässä. WWW page, Julkisen hallinnon tietohallinnon neuvottelukunta, 2012b. URL: http://www.jhs-suositukset.fi/web/guest/jhs/recommendations/184/full, accessed March 20, 2015.

JUHTA. JHS185. Asemakaavan pohjakartan laatiminen. WWW page, Julkisen hallinnon tietohallinnon neuvottelukunta, 2014. URL: http://www.jhs-suositukset.fi/web/guest/jhs/recommendations/185/full, accessed March 20, 2015.

JUHTA. JHS196. ETRS89-järjestelmän mukaiset koordinaatit Suomessa. WWW page, Julkisen hallinnon tietohallinnon neuvottelukunta, 2016a. URL: http://www.jhs-suositukset.fi/suomi/jhs196, accessed August 2, 2016.

JUHTA. JHS197. ETRS89 -koordinaattijärjestelmät, niihin liittyvät muunnokset ja karttalehtijako. WWW-sivu, Julkisen hallinnon tietohallinnon neuvottelukunta, 2016b. URL: http://www.jhs-suositukset.fi/suomi/jhs197, accessed August 2, 2016.

Heribert Kahmen and Wolfgang Faig. *Surveying*. Walter de Gruyter, Berlin - New York, 1988.

Juhani Kakkuri. *Stellar triangulation with balloon-borne beacons and satellites*. PhD thesis, Helsinki University, 1973.

Ulla Kallio. *Tasoituslasku*. Number 587 in Otatieto. 1998.

Kari-Pekka Karlsson. Maastotietojärjestelmä 2030 -selvitys – Loppuraportti. Technical report, Ministry of Agriculture and Forestry, 2015. http://mmm.fi/documents/1410837/1801204/Maastotietojarjestelma_2030_-selvitys_Loppuraportti_28.2.2015.pdf.

Shfaqat Abbas Khan, Lin Liu, John Wahr, Ian Howat, Ian Joughin, Tonie van Dam, and Kevin Fleming. GPS measurements of crustal uplift near Jakobshavn Isbræ due to glacial ice mass loss. *Journal of Geophysical Research: Solid Earth*, 115(B9):B09405, 2010. ISSN 2156-2202. doi: 10.1029/2010JB007490. URL http://dx.doi.org/10.1029/2010JB007490.

Hannu Koivula and Jaakko Mäkinen. Geodetic activities at Finnish Antarctic research station Aboa. In *Fifth International Antarctic Geodesy Symposium AGS'03, Lviv, Ukraine, Sept. 15 – 17*, 2003. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.518.4145&rep=rep1&type=pdf.

Yoshihide Kozai. The Motion of a Close Earth Satellite. *Astronomical Journal*, 64(1274):367–377, 1959. http://adsabs.harvard.edu/full/1959AJ.....64..367K.

T.J. Kukkamäki and Pekka Lehmuskoski. Influence of the earth magnetic field on Zeiss Ni 002 levels. Report 84:1, Finnish Geodetic Institute, Helsinki, 1984.

Erkki Kääriäinen. The Second Levelling of Finland in 1935-1955. Publication 61, Finnish Geodetic Institute, Helsinki, 1966.

J. Laskar, P. Robutel, F. Joutel, M. Gastineau, A.C.M. Correia, and B. Levrard. A long term numerical solution for the insolation quantities of the Earth. *Astronomy and Astrophysics*, 428:261–285, 2004. https://www.imcce.fr/fr/presentation/equipes/ASD/preprints/prep.2004/La_2004_prep.pdf.

Leica. *TPS-System 1000 Version 2.2 quick start*. Leica Geosystems AG, Heerbrugg, Switzerland, 1997. https://issuu.com/landsurveyor/docs/tps1000_quick_start_en.

Cherry L. E. Lewis. *The Dating Game*. Cambridge University Press, 2000.

Martin Lidberg, Jan M. Johansson, Hans-Georg Scherneck, and Glenn A. Milne. Recent results based on continuous GPS observations of the GIA process in Fennoscandia from BIFROST. *Journal of Geodynamics*, 50(1): 8–18, July 2009.

Rick Lumpkin and Zulema Garraffo. Evaluating the Decomposition of Tropical Atlantic Drifter Observations. *Journal of Atmospheric and Oceanic Technology*, pages 1403–1415, 2005. http://journals.ametsoc.org/doi/pdf/10.1175/JTECH1793.1.

Javier Arribas Lázaro. *GNSS Array-based Acquisition: Theory and Implementation*. PhD thesis, Universitat Politécnica de Catalunya, 2012. http://theses.eurasip.org/theses/449/gnss-array-based-acquisition-theory-and/download/.

Maanmittauslaitos. Kaavoitusmittausohjeet 2003. Julkaisu 94, Maanmittauslaitos, 31. tammikuuta 2003. URL: http://www.maanmittauslaitos.fi/sites/default/files/kaavoitusmittausohjeet_2003_0.pdf.

Stephen Malys, John H. Seago, Nikolaos K. Pavlis, P. Kenneth Seidelmann, and George H. Kaplan. Why the Greenwich meridian moved. *Journal of Geodesy*, 89(12):1263–1272, 2015. doi: 10.1007/s00190-015-0844-y. https://link.springer.com/article/10.1007/s00190-015-0844-y.

Patrap Misra and Per Enge. *Global Positioning System – Signals, Measurements, and Performance*. Ganga-Jamuna Press, 2001. ISBN 0-9709544-0-9.

Oliver Montenbruck, Peter Steigenberger, and André Hauschild. Broadcast versus precise ephemerides: a multi-GNSS perspective. *GPS Solutions*, 19 (2):321–333, 2015. http://link.springer.com/article/10.1007%2Fs10291-014-0390-8.

J. Mäkinen, A. Engfeldt, L. Engman, B.G. Harsson, T. Oja, S. Rekkedal, K. Røthing, P. Rouhiainen, H. Ruotsalainen, H. Skatt, G. Strykowski, H. Virtanen, K. Wieczerkowski, and D. Wolf. The Fennoscandian Land Uplift Gravity Lines: comparison of observed gravity change with observed vertical motion and with GIA models. Technical report, Nordiska Kommissionen för Geodesi, 2010. http://www.nkg.fi/nkg/sites/default/files/Makinen_et_al_land_uplift_gravity_lines.pdf.

V. R. Ölander. Trigonometriset korkeusmittaukset pitkillä välimatkoilla ja niiden tarkkuus. *Maanmittaus*, 1932.

Marko Ollikainen. Kaavoitusmittausohjeiden uudistaminen poiki uuden JHS-suosituksen. *Positio*, 2013. http://www.paikkatietoikkuna.fi/c/document_library/get_file?uuid=b562d56d-b26e-45ca-94f0-117101c1f541&groupId=108478.

Matti Ollikainen. GPS-koordinaattien muuntaminen Kartastokoordinaateiksi. Tiedote 8, Geodeettinen laitos, Helsinki, 1993.

B. Overgaauw, B.A.C. Ambrosius, and K.F. Wakker. Analysis of the EUREF-89 GPS data from the SLR/VLBI sites. *Bulletin Géodésique*, 68:19–28, 1994. http://link.springer.com/article/10.1007%2FBF00806749.

Teuvo Parm. Kansallisen koordinaattijärjestelmän luominen Suomessa. *Maanmittaus*, 63(1), 1988.

Jerry Penry and David Lee Ingram. The American Theodolite. *The American Surveyor*, 10(9), 2013. http://www.amerisurv.com/PDF/TheAmericanSurveyor_PenryIngram-TheAmericanTheodolite_Vol10No9.pdf.

Mika Pitkä. Sataman tehokkuuden osatekijät – tarkastelussa lastauksen ja purun automatisointi. Insinöörityö, Kymenlaakson ammattikorkeakoulu, 2009. http://www.merikotka.fi/safgof/mika_pitka_OPINNAYTETYO.pdf.

Markku Poutanen. *GPS-paikanmääritys*. Ursa, Helsinki, 1998. Ursan julkaisuja 64.

Markku Poutanen. *GPS-paikanmääritys*. Ursa, Helsinki, 2016.

RAKLI ry. Kiinteistöalan yhteiskunnallinen ja kansantaloudellinen merkitys. Technical report, Asunto-, toimitila- ja rakennuttajaliitto RAKLI, 2014. http://www.rakli.fi/media/tietoa-kiinteistoalasta/faktaa-alasta/2014_kiinteistoalan-yhteiskunnallinen-ja-kansantaloudellinen-merkitys_nettires.pdf.

J. Reijnoudt. Het hele IJsselmeer waterpas, 1996. URL: http://www.digibron.nl/search/detail/012de334d1e3bd5cea566f97/het-hele-ijsselmeer-waterpas.

Jean M. Rüeger. *Electronic Distance Measurement*. Springer-Verlag, Berlin Heidelberg, kolmas, uusittu edition, 1990.

Jean M. Rüeger. Refractive indices of light, infrared and radio waves in the atmosphere. Technical Report S-68, School of Surveying and Spatial Information Systems, University of New South Wales, 2002. http://www.sage.unsw.edu.au/about/school_pubs/pdfrep/s68.pdf.

Reiner Rummel and Fernando Sansò, editors. *Satellite Altimetry in Geodesy and Oceanography*, Lecture Notes in Earth Sciences, 1993. Springer-Verlag.

Hannu Salmenperä. Runko- ja kartoitusmittaukset. Technical Report 4/1998, Tampereen teknillinen korkeakoulu, Rakennustekniikan osasto, Geoinformatiikka, Tampere, 1998.

Gavin Schrock. CERN. *xyHt*, July 2, 2014. http://www.xyht.com/civiltransportation/cern-2/.

Günter Seeber. *Satellitengeodäsie. Grundlagen, Methoden und Anwendungen*. De Gruyter, Berlin New York, 1989.

Juha Simonen. Maalaserkeilaus infrarakentamisen mittauksissa. Master's thesis, Aalto-yliopisto, 2012. http://personal.inet.fi/koti/jsimonen/dippa/dippa_jsimonen.pdf.

Dava Sobel. *Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time*. Walker and Company, New York, 1998. Paperback.

G. Strang van Hees. Stokes' Formula Using Fast Fourier Techniques. *Manuscripta geodaetica*, 15:235–239, 1990.

M. Takalo, P. Rouhiainen, P. Lehmuskoski, and V. Saaranen. On Calibration of Zeiss DiNi12. In *FIG Working Week 2001 6-11 May 2001*, Seoul, Korea, 2001. FIG Commission 5.

Mikko Takalo. Computer supported Finnish EDM height traversing. Part I: technical description. *Surveying Science in Finland*, 13(2):56–99, 1995.

Mikko Takalo and Paavo Rouhiainen. On system calibration of digital level. In *Ingenieurvermessung 2004, 14th International Conference on Engineering Surveying*, Zürich, 2004. http://www.researchgate.net/profile/Mikko_Takalo/publication/254398461_Development_of_a_System_Calibration_Comparator_for_Digital_Levels_in_Finland/links/00b49528a004fbfee1000000.pdf.

Martti Tikka. *Käytännön Geodesia III. Mittausmenetelmät*. Otakustantamo, Espoo, 1985.

Martti Tikka. *Käytännön Geodesia I. Mittaustekniikan perusteet ja rakennusteknilliset sovellutukset*. Otatieto, Helsinki, 1991. Kolmas korjattu painos.

Wolfgang Torge. *Geodesy. Third completely revised and extended edition*. Walter de Gruyter, Berlin – New York, 2001.

Yrjö Väisälä. Maan toinen kuu. *Tähtitaivas*, (6), 1946.

Peter Vaníček and Edward Krakiwsky. *Geodesy – The Concepts*. Elsevier Science Publishers, Amsterdam, 1986.

WebTechTix. Civil Engineering Dictionary, 2014. http://www.aboutcivil.org/compound-curves-in-engineering-surveying.html, accessed April 6, 2017.

Index

Q

R

A. Properties of matrices

Matrices are used very widely in science to describe multidimensional phenomena, like co-ordinate transformations and mappings from one vector space to another. The matrix is written out as a two-dimensional table of numbers between parentheses or square brackets. The *symbolic* notation for a matrix is an italic[1] capital letter, like A, B, I etc.

Formally, matrices are a *type of numbers* (like real numbers, complex numbers, etc.. . .). Many operations on numbers, like addition and subtraction, exist also for matrices: same sized matrices may be added simply by adding together the corresponding elements. Other operations, like multiplication, don't always work between matrices: it requires that the *sizes* of matrices, the numbers of their rows and columns, are suitable. More precisely, the number of columns in the left matrix must be the same as the number or rows in the right matrix, for multiplication to be even possible.

A.1 Adding matrices

Same sized matrices are added by adding together the corresponding elements. I.e.,

$$C = A + B$$

means

$$c_{ij} = a_{ij} + b_{ij}$$

for every pair i, j, $i = 1 \ldots n, j = 1 \ldots m$, if the matrices A, B, C are of size $n \times m$. E.g., in the following way ($n = m = 2$):

$$\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix},$$

with numeric example

$$\begin{bmatrix} 5 & 2 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} + \begin{bmatrix} 2 & 2 \\ 1 & 2 \end{bmatrix},$$

[1] Alternatively, many authors use bold capital letters.

in which $c_{11} = a_{11} + b_{11}$, $C = \begin{bmatrix} 5 & 2 \\ 1 & 4 \end{bmatrix}$ etc.

Multiplying a matrix with a constant

This amounts to multiplying every element with this constant. E.g.,

$$k \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} ka_{11} & ka_{12} \\ ka_{21} & ka_{22} \end{bmatrix}.$$

A.2 Matrices and vectors

A matrix with only one row is called a *row vector*; a matrix with only one column, a *column vector*. E.g.,

$$\mathrm{r} = \begin{bmatrix} r_1 & r_2 & r_3 & r_4 \end{bmatrix}, \mathrm{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}.$$

Vectors are commonly denoted by small letters.

A.3 The unit matrix

In the same way that for numbers there exists a *zero element* 0, for which holds $a + 0 = 0 + a = a \, \forall a \in \mathbb{R}$, and a *unit element* 1 with the property $a \cdot 1 = 1 \cdot a = a \, \forall a \in \mathbb{R}$, these exist also for matrices. The unit matrix is simply[2]

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

A.4 Matrix multiplication

Multiplication of matrices is done following the simple scheme “row times column”. If

$$C = A \cdot B,$$

this means that for every i, k:

$$c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk},$$

[2] So, it is always square!

i.e., the matrix

$$\overbrace{\begin{bmatrix} c_{11} & \cdots & c_{1k} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots & \cdot & \vdots \\ c_{i1} & \cdots & [c_{ik}] & \cdots & c_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nk} & \cdots & c_{nn} \end{bmatrix}}^{C}$$

is obtained by multiplying every row of matrix A

$$\overbrace{\begin{bmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots & \cdot & \vdots \\ [a_{i1}] & \rightarrow & \left[a_{ij}\right] & \rightarrow & [a_{in}] \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nj} & \cdots & a_{nn} \end{bmatrix}}^{A}$$

with every column of matrix B:

$$\overbrace{\begin{bmatrix} b_{11} & \cdots & \lceil b_{1k} \rceil & \cdots & b_{1n} \\ \vdots & \ddots & \downarrow & \ddots & \vdots \\ b_{j1} & \cdots & \left|b_{jk}\right| & \cdots & b_{jn} \\ \vdots & \ddots & \downarrow & \ddots & \vdots \\ b_{n1} & \cdots & \lfloor b_{nk} \rfloor & \cdots & b_{nn} \end{bmatrix}}^{B};$$

so, one sums $c_{ik} = a_{i1}b_{1k} + a_{i2}b_{2k} + \ldots + a_{in}b_{nk}$, "row i times column k", for all rows and columns of matrix C. In the above formula we have marked with square brackets $[\cdot]$ only one row, row i, in the matrix A, and one column, column k, in the B matrix, multiplication of which thus yields the element of C called c_{ik}.

Here it is assumed, for the sake of simplicity, that all matrices are square, $n \times n$ rows and columns. This is not inevitable, however the number of columns of A and the number of rows in B must be the same for multiplication to be possible.

Exercise: Verify that for an arbitrary square matrix A it holds that $A \cdot I = I \cdot A = A$.

A.5 The transpose

The transpose of matrix A , the matrix A^{T}, is defined as

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}^{\mathsf{T}}}_{n \times m} = \underbrace{\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{bmatrix}}_{m \times n},$$

or more compactly,

$$\left(A^{\mathsf{T}}\right)_{ij} = (A)_{ji}, i = 1,\ldots,n; j = 1,\ldots,m.$$

The transpose A^{T} of matrix A is thus obtained by interchanging rows and columns, i.e., by "mirroring" all elements though the main diagonal. Trivially, if the transposition is done twice, one obtains back the original matrix:

$$\left(A^{\mathsf{T}}\right)^{\mathsf{T}} = A.$$

The transpose of a *vector* changes a row vector into a column vector and vice versa:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}$$

and

$$\begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}.$$

The latter notation is often used in running text to save paper in presenting a column vector.

A.6 The inverse matrix

The inverse matrix N^{-1} of a square matrix N is defined as the matrix K, for which $K \cdot N = N \cdot K = I$, in which I is the already defined *unit matrix*. A property of the unit matrix again is $I \cdot N = N \cdot I = N$ for all matrices N. The role of the unit matrix is the same as for real numbers the role of the number 1. For a real number a again, the concept corresponding to the inverse matrix is the number $1/a$.

There exist good numerical algorithms for inverting a matrix. Inversion is however heavy for large matrices: the computational work required is proportional to the matrix size n to the *third* power! So, inverting a size

20×20 matrix takes about 8 times the computation time as inverting a size 10×10 matrix.

Manual inversion is only readily possible for a matrix of size up to 3×3 — and is not worth the bother[3].

E.g., if

$$A = \begin{bmatrix} 3 & 1 \\ 2 & 4 \end{bmatrix},$$

its inverse matrix is

$$A^{-1} = \begin{bmatrix} 0.4 & -0.1 \\ -0.2 & 0.3 \end{bmatrix},$$

which is easily verified:

$$A \cdot A^{-1} = \begin{bmatrix} 3 \cdot 0.4 - 1 \cdot 0.2 & -3 \cdot 0.1 + 1 \cdot 0.3 \\ 2 \cdot 0.4 - 4 \cdot 0.2 & -2 \cdot 0.1 + 4 \cdot 0.3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I,$$

and similarly $A^{-1} \cdot A = I$.

A valuable tool in matrix computations is the MatLab™ software which is in widespread use within the science community. For home users there are similar open-source offerings like octave or scilab.

In practice, the computation is based on solving systems of linear equations: if $X = A^{-1}$ is the inverse of the matrix A, it holds, for the case of a 2×2 matrix, that

$$AX = I,$$

i.e.,

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

i.e.,

$$\left.\begin{aligned} a_{11}x_{11} + a_{12}x_{21} &= 1 \\ a_{21}x_{11} + a_{22}x_{21} &= 0 \end{aligned}\right\} \Rightarrow x_{11}, x_{21},$$

$$\left.\begin{aligned} a_{11}x_{12} + a_{12}x_{22} &= 0 \\ a_{21}x_{12} + a_{22}x_{22} &= 1 \end{aligned}\right\} \Rightarrow x_{12}, x_{22}.$$

If the matrix is of size $n \times n$, similarly n systems of equations are created, with in each, n equations in n unknowns.

Singularity: Just like division $1/a$ doesn't work for all numbers a — i.e., specifically not if $a = 0$ — also the computation of the inverse matrix doesn't work for all matrices. A simple example of a non-invertible matrix is

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

[3] "One matrix, inverted".

and the solution equations above become conflicting:

$$x_{11} + x_{21} = \text{both 1 and 0},$$
$$x_{12} + x_{22} = \text{the same}.$$

Such a matrix is called *singular*.

A.7 Vectorial products

The scalar product of two vectors $\mathbf{a}$ and $\mathbf{b}$ can be defined using the matrix product: if

$$\begin{aligned} \mathbf{a} &= a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + \ldots + a_n\mathbf{e}_n, \\ \mathbf{b} &= b_1\mathbf{e}_1 + b_2\mathbf{e}_2 + \ldots + b_n\mathbf{e}_n, \end{aligned}$$

in which $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ is an orthonormal basis[4], and

$$\overline{\mathbf{a}} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \text{ and } \overline{\mathbf{b}} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

the scalar product *of the vectors is*

$$\langle \mathbf{a} \cdot \mathbf{b} \rangle \stackrel{\text{def}}{=} \overline{\mathbf{a}}^{\mathsf{T}} \overline{\mathbf{b}} = \overline{\mathbf{b}}^{\mathsf{T}} \overline{\mathbf{a}} = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n.$$

The *vectorial product* of vectors is more complicated: it is (only in three dimensions, $n = 3$)

$$\langle \mathbf{a} \times \mathbf{b} \rangle \stackrel{\text{def}}{=} (a_2 b_3 - a_3 b_2)\mathbf{e}_1 + (a_3 b_1 - a_1 b_3)\mathbf{e}_2 + (a_1 b_2 - a_2 b_1)\mathbf{e}_3 =$$
$$= \det \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{bmatrix}.$$

[4] An *orthonormal basis* is a set of vectors $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ spanning a vector space, for which

$$\langle \mathbf{e}_i \cdot \mathbf{e}_j \rangle = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

B. A short introduction to magnetohydrodynamics

B.1 Plasma

A *plasma* is a gas of which the atoms have been stripped of part of their electrons — so-called ionization — which are moving around freely. In space, plasmas are commonly hot, with temperatures varying from thousands to billions of degrees.

Plasmas in space are so extended and rarefied, that they behave like *superconductors*: electric currents run in them for long times without noticeable weakening. For this reason, also a magnetic field cannot move though a plasma: matter and field lines move hand in hand, so-called *frozen-in* magnetism or Alfvén′s theorem. If a plasma is compressed, the magnetic field contained in it will grow stronger. Similarly if the plasma is stretched and rolled up like puff pastry, as is happening inside the Sun's convection layer. This is how natural dynamos work.

B.2 Maxwell's equations

Maxwell's field equations are[1]

$$\begin{aligned} \langle \nabla \times \mathbf{B} \rangle &= \mu_0 \mathbf{j} + \epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t}, \\ \langle \nabla \times \mathbf{E} \rangle &= -\frac{\partial \mathbf{B}}{\partial t}, \\ \langle \nabla \cdot \mathbf{E} \rangle &= \frac{\rho}{\epsilon_0}, \\ \langle \nabla \cdot \mathbf{B} \rangle &= 0. \end{aligned} \tag{B.1}$$

(**E** electric field vector, **B** magnetic field vector, **j** vector of electric current density, ρ density of electric charge, ϵ_0, μ_0 constants of nature.)

[1]This compact form of the equations we owe to Oliver Heaviside (1850 – 1925). Heaviside was a telegraph researcher who developed the theory of electric circuits. He also proposed the existence of the *ionosphere* as an explanation for the long range — beyond the horizon — of radio waves.

The partial differential operators $\nabla\times$ and $\nabla\cdot$ are known in vector calculus as *curl* and *divergence* operators, curl and div. They are defined as follows:

$$\begin{aligned}\operatorname{curl}\mathbf{E} &= \langle\nabla\times\mathbf{E}\rangle \stackrel{\text{def}}{=} \left(\frac{\partial E_z}{\partial y}-\frac{\partial E_y}{\partial z}\right)\mathbf{i}+\left(\frac{\partial E_z}{\partial x}-\frac{\partial E_x}{\partial z}\right)\mathbf{j}+\left(\frac{\partial E_x}{\partial y}-\frac{\partial E_y}{\partial x}\right)\mathbf{k},\\ \operatorname{div}\mathbf{E} &= \langle\nabla\cdot\mathbf{E}\rangle \stackrel{\text{def}}{=} \frac{\partial E_x}{\partial x}+\frac{\partial E_y}{\partial y}+\frac{\partial E_z}{\partial z},\end{aligned}$$

with $\{\mathbf{i},\mathbf{j},\mathbf{k}\}$ being the orthonormal basis of the (x,y,z) co-ordinate frame, and

$$\mathbf{E}=E_x\mathbf{i}+E_y\mathbf{j}+E_z\mathbf{k}.$$

B.3 "Frozen-in" magnetic field

In a superconductor there can be no net electric field integrated around a loop — as it would cause an infinite electric current! So

$$\oint_{\partial S}\langle\mathbf{E}\cdot\mathbf{t}\rangle\,ds=0,$$

in which $\mathbf{t}$ is the tangent vector of loop ∂S. The *Stokes loop integral theorem* (http://en.wikipedia.org/wiki/Stokes%27_theorem) — also called the Kelvin-Stokes theorem because lord Kelvin[2] apparently discovered it — says that this integral is the same as the surface integral of the *curl* $\nabla\times\mathbf{E}$ of $\mathbf{E}$ over the surface S, the edge of which is ∂S. Also that integral must thus vanish:

$$\iint_S\langle\langle\nabla\times\mathbf{E}\rangle\cdot\mathbf{n}\rangle\,dS=\oint_{\partial S}\langle\mathbf{E}\cdot\mathbf{t}\rangle\,ds=0.$$

In this, $\mathbf{n}$ is the normal on the surface S. From the *Maxwell-Faraday* equation B.1 follows now

$$\iint_S\left\langle\frac{\partial\mathbf{B}}{\partial t}\cdot\mathbf{n}\right\rangle dS=0\quad\Longrightarrow\quad\frac{\partial}{\partial t}\iint_S\langle\mathbf{B}\cdot\mathbf{n}\rangle\,dS=\iint_S\left\langle\mathbf{B}\cdot\frac{\partial\mathbf{n}}{\partial t}\right\rangle dS.$$

Here, the integrand on the right, $\left\langle\mathbf{B}\cdot\frac{\partial\mathbf{n}}{\partial t}\right\rangle$, vanishes if the loop surface S is chosen perpendicular to the field vector $\mathbf{B}$. Then

$$\frac{\partial}{\partial t}\iint_S\langle\mathbf{B}\cdot\mathbf{n}\rangle\,dS=0.$$

The expression $\iint_S\langle\mathbf{B}\cdot\mathbf{n}\rangle\,dS$ is the magnetic *flux* through the loop surface S, the total number of field lines passing though the loop, which

[2] Sir William Thomson (1824 – 1907) FRS FRSE (lord Kelvin) was a British physicist, engineer and inventor. He was ennobled in 1866 for his work on the trans-Atlantic telegraph cable.

thus must be constant in time and cannot change. And this applies for an arbitrary (cross-field) loop inside the plasma: the plasma and the field are moving hand in hand — though the plasma can freely flow along the field lines. If the plasma is compressed, the magnetic field strength increases[3].

B.4 History of the field

The name "plasma" was invented by Nobel laureate Irving Langmuir[4] (1927). The scientific discipline studying plasmas is called *magnetohydrodynamics* (MHD). It has been estimated that, in the universe, over 99% of normal — not "dark" — matter is plasma.

A well-known magnetohydrodynamicist was the Swedish Nobel laureate Hannes Alfvén, 1908 – 1995. He described in 1942 how in a superconductor the magnetic field lines move along with the matter.

[3] A spectacular example of this is the magnetic field of a neutron star or pulsar, the collapsed core of a star.

[4] Irving Langmuir (1881 – 1957) was an American chemist and physicist known for his research on matter.

C. The Kepler orbital elements for satellites

C.1 Angular elements describing the orbit's *orientation* in space

Ω The *right ascension*, or astronomical longitude, of the ascending node. The zero point of this longitude is the place among the stars where the plane of the ecliptic and the plane of the equator intersect and the Sun corsses the equator at the beginning of spring: the *vernal equinox point*.

i The inclination, i.e., the *tilt angle* of the orbital plane with respect to the equatorial plane. For GPS satellites $i = 55°$. This is also the highest latitude (North and South) where the satellite moves through the zenith.

ω The argument of perigee. The angular distance, seen from the Earth's centre, between the ascending node and the perigee of the satellite orbit.

C.2 Elements describing the orbit's *size and shape*

a The semi-major axis of the satellite orbit, i.e., the "mean radius" of the orbit. The mean height of the satellite is obtained from this by subtracting the Earth radius.

e The eccentricity of the satellite orbit. $e = \sqrt{\frac{a^2 - b^2}{a^2}}$, in which b is the semi-minor axis. This describes how much the Earth's centre is "to the side" from the centre of the orbital ellipse, i.e., how large is the height difference between perigee (the lowest point of the orbit) and apogee (the highest point).

C.3 Elements describing the satellite's place in its orbit, its "time table"

There are three alternatives:

$\nu(t)$ true anomaly. The direction angle of the satellite in the orbital plane as seen from the Earth's centre, reckoned from the perigee in the direction of motion.

$E(t)$ eccentric anomaly. The direction angle of the satellite in the orbital plane as seen from the *centre of the ellipse*, after the orbital ellipse has been inflated into a circle.

$M(t)$ mean anomaly. The time that has elapsed since the last passage through the perigee, in units of the satellite's orbital period P.

Between these exist the following relationships, which may be used to convert them into each other:

$$\frac{\tan\frac{1}{2}\nu(t)}{\tan\frac{1}{2}E(t)} = \sqrt{\frac{1+e}{1-e}}$$

$$E(t) = M(t) + e\sin E(t)$$

See figure 16.13. E and ν are geometric quantities, the mean anomaly M on the other hand is a mere measure of time, scaled to the period P of the satellite, and referenced to its time of passage τ through the perigee:

$$M(t) \stackrel{\text{def}}{=} 2\pi\frac{t-\tau}{P}.$$