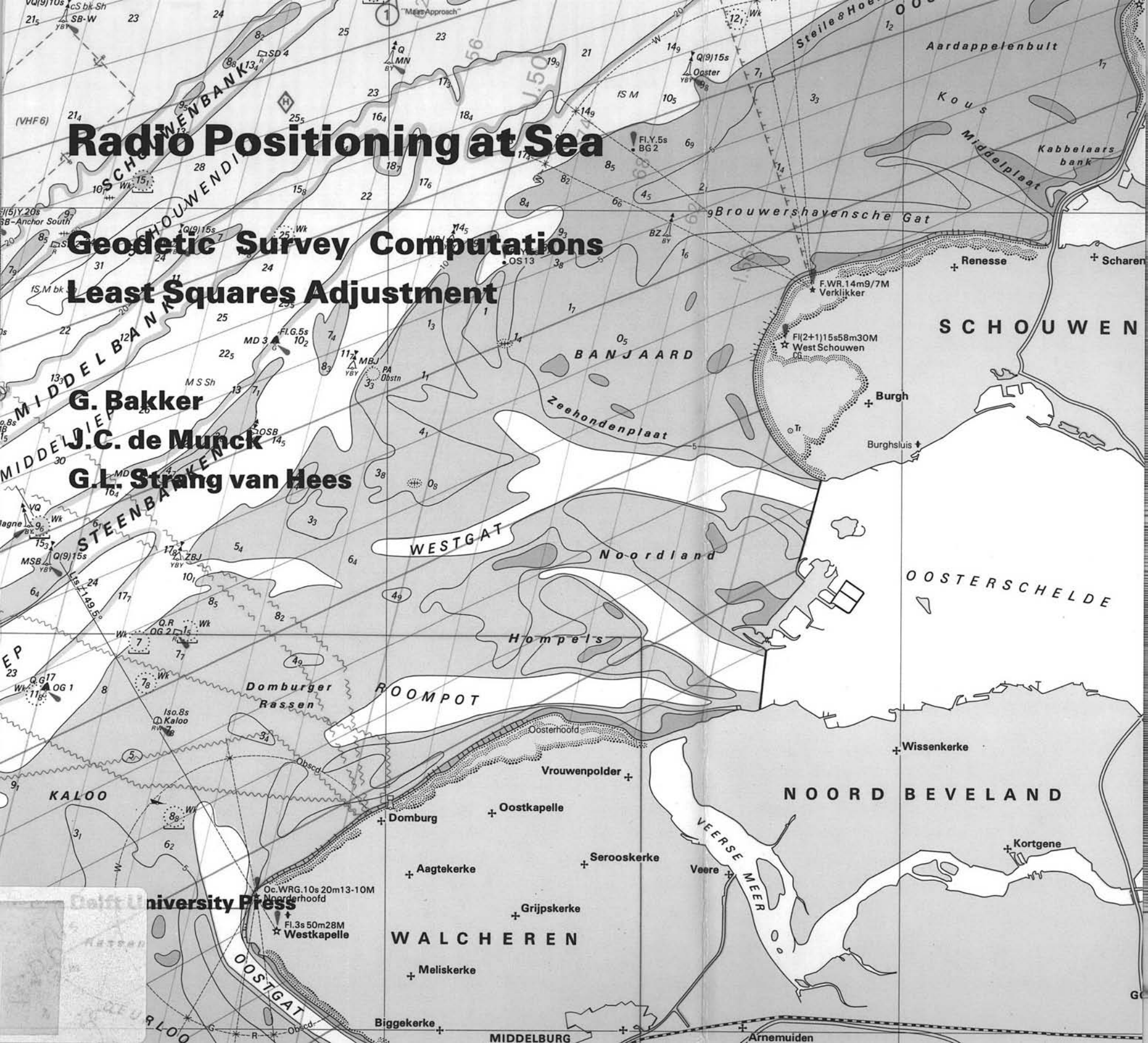
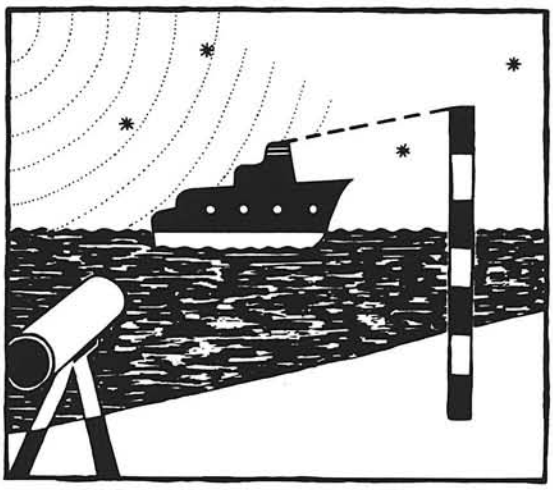

Radio Positioning at Sea
Geodetic Survey Computations
Least Squares Adjustment
G. Bakker
J.C. de Munck
G.L. Strang van Hees
niversity Press
SCHOUWENBANK
MIDDELBANK
STEENBANKEN
WESTGAT
OOSTGAT
ROOMPOT
Hompels
Domburger Rassen
Zeehondenplaat
BANJAARD
Brouwershavensche Gat
Aardappelenbult
Kous
Middelplaat
Kabbelaars bank
Steile Hoek
SCHOUWEN
NOORD BEVELAND
WALCHEREN
KALOO
OOSTERSCHELDE
Noordland
VEERSE MEER
Renesse
Scharen
Burgh
Burghsluis
Wissenkerke
Vrouwenpolder
Oosterhoofd
Oostkapelle
Domburg
Aagtekerke
Serooskerke
Veere
Kortgene
Grijpskerke
Meliskerke
Westkapelle
Biggekerke
MIDDELBURG
Arnemuiden
F.WR.14m9/7M Verklikker
Fl(2+1)15s58m30M West Schouwen
Fl.Y.5s BG 2
Q(9)15s Ooster
Oc.WRG.10s 20m13-10M Noorderhoofd
Fl.3s 50m28M Westkapelle
Iso.8s Kaloo
Obscd
"Maas Approach"

Radio Positioning at Sea

MARINE



GEODESY

Radio Positioning at Sea

Geodetic Survey Computations
Least Squares Adjustment

G. Bakker
J.C. de Munck
G.L. Strang van Hees

Student edition

Delft University Press / 1989

Published and distributed by:

Delft University Press
Stevinweg 1
2628 CN Delft
The Netherlands
Telephone (0)15 - 78 32 54

Address of the authors:

Faculty of Geodesy
Delft University of Technology
Thijsseweg 11
2629 JA Delft
The Netherlands
Telephone (0)15 - 78 25 83

Cover, in three versions, shows part of the hydrographic map of the North Sea.
Kindly supplied by: The Hydrographic Service of the Royal Netherlands Navy.
The Hague.

Printed in the Netherlands.

CONTENTS

PREFACE

In summer 1985 a course on radio positioning at sea was held at the faculty of geodesy of the Delft University of Technology.
The course took one week and its main purpose was to provide the participants with a synopsis of the more theoretical aspects of the subject. After the course it soon appeared that there was a fair demand for the lecture notes. This encouraged the authors to improve their contributions, what resulted in this book.
In **chapter 1** the systems of radio positioning are introduced and classified on the basis of wave length and propagation properties and their effect on accuracy, range and applicability.
Chapter 2 deals with the geometrical aspects. More specifically the geodetic computations on the ellipsoid and the conformal map projections are explained.
Chapter 3 deals with the least squares adjustment of the observations and testing methods to find gross-errors. The formulas for the precision and reliability of the parameters are derived. These methods are applied to the radio positioning systems.
Finally attention is given to the Kalman filtering method.
The last two chapters have a more general scope and their use is not restricted to radio positioning only.
The authors are very grateful to Mrs. W. Coops who was the key-figure in both the organization of the course and the edition of this book.

G. Bakker
J.C. de Munck
G.L. Strang van Hees

1. RADIOPOSITIONING SIGNALS AND SYSTEMS

J.C. de Munck

Contents.

Systems for radiopositioning will be classified on the basis of wavelength, discussing the propagation properties and their consequences for accuracy, range and application.

The different methods will also be arranged based on geometry and on the used signals.

A number of examples will be discussed. In addition, the set-up, the possibilities for checks and the calibration will be reviewed.

1.1 Effects of the wave propagation.

Radiowaves are travelling electromagnetic waves: a distortion of the electric and magnetic field strength, being at any fixed point some function of time, is travelling with a velocity of about 300,000 km/s through the space. It is often useful to consider sine waves of the form:

$$U(t,x) = U_o \sin \{2\pi(\frac{t}{T} - \frac{x}{\lambda} + \varphi)\}$$

where V_o (amplitude), T (period), λ (wavelength) and φ (phase) are considered as constants.

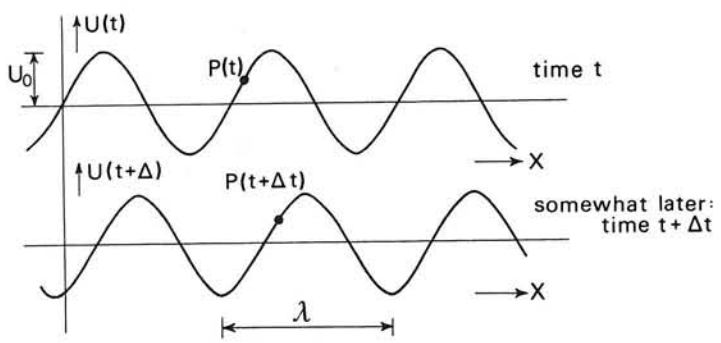


Figure 1.

If one follows some characteristic point of the wave, for instance P, the argument $2\pi(\frac{t}{T} - \frac{x}{\lambda} + \varphi)$ will not change in time; see figure 1. In the short time

interval Δt point P, and in fact the whole wave, has travelled over a distance Δx so that

$$2\pi\left(\frac{t}{T} - \frac{x}{\lambda} + \varphi\right) = 2\pi\left(\frac{t+\Delta t}{T} - \frac{x+\Delta x}{\lambda} + \varphi\right)$$

or $$\frac{\Delta x}{\Delta t} = \frac{\lambda}{T} \quad .$$

Consequently, the propagation velocity v of the wave is given by:

$$v = f \cdot \lambda \qquad (1)$$

where $f = \frac{1}{T}$ is the frequency.

The properties of electromagnetic waves depend largely on their wavelength. In table 1 a common classification of radiowaves is presented. For comparison, also some other electromagnetic waves are mentioned. The free space wavelength λ_0 used in the table is defined by

$$\lambda_0 = c/f$$

where c is the free space light velocity, a constant of nature (c = 299,793,458 m/s by definition).

	Radiowaves				Radarbands		
	Names		f	λ_0	Names	frequency	wavelength
	V.L.F.	very low frequency	< 30 k Hz	> 10 km	P-band	0.23-1 G Hz	130-30 cm
	L.F.	low "	30-300 k Hz	10-1 km	L-band	1-2 "	30-15 cm
	M.F.	medium "	300-3000 k Hz	1000-100 m	S-band	2-4 "	15-7,5 cm
	H.F.	high "	3 M Hz-30 M Hz	100-10 m	C-band	4-8 "	7,5-3,75 cm
	V.H.F.	very high "	30 M Hz-300 M Hz	10-1 m	X-band	8-12,5 "	3,75-2,4 cm
micro-	U.H.F.	ultra high "	300-3000 M Hz	100-10 cm	K_u-band	12.5-18 "	2,4-1,67 cm
waves	S.H.F.	super high "	3 G Hz-30 G Hz	10-1 cm	K-band	18-26,5 "	1,67-1,13 cm
	E.H.F.	extremely high "	30-300 G Hz	10-1 mm	K_a-band	28,5-40 "	1,13-0,75 cm

Table 1: Classes of radiowaves.

Depending on the classes of radiowaves (table 1) different descriptions of the propagation are useful and other aspects are relevant. In figure 2 the different types of propagation through the atmosphere are shown. Indicated are the

earth's surface (including the sea surface), the ionosphere and the troposphere, all described in the sections 1.1.1.1.-1.1.1.3. The properties of the different classes of waves and the corresponding systems are discussed in section 1.1.2.

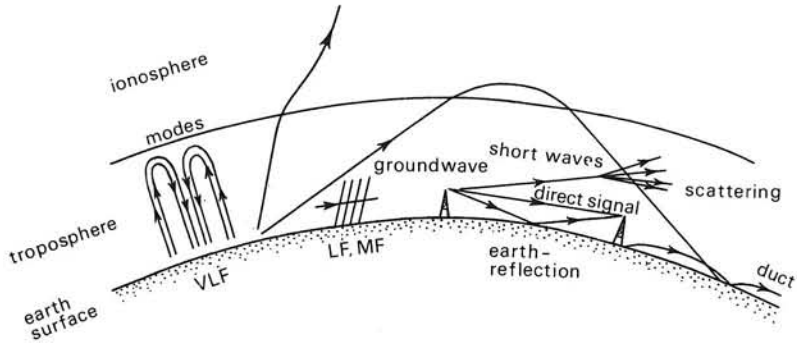


Figure 2.

1.1.1. The different media.

1.1.1.1. The troposphere.

The troposphere is the part of the atmosphere where free electrons (see 1.1.3) hardly exist. In the troposphere the propagation velocity depends slightly on temperature T, pressure P and humidity e, the partial pressure of the water vapour, according to the formula

$$N = (n-1).10^6 = 77.62 \frac{P}{T} - 12.92 \frac{e}{T} + 37.19 \cdot 10^4 \cdot \frac{e}{T^2} \quad (2)$$

where

n = c/v = the refraction index
c = the propagation velocity in free space
v = the propagation velocity in the air
P = the total air pressure in Pascal (1 Pascal = 1 Newton/m^2 = $= 10^{-5}$ bar = 0.01 mbar ≈0.0075 Torr (mm mercuri))
T = 273 + the temperature in degrees centigrade ≈ temperature in Kelvin.
e = the partial pressure of the water vapour in Pascal

and N is called the refractivity.

From (1) it follows that, under normal circumstances (P $\approx$ 1000 mbar, T = 288 K ($\approx$ 15 ^{o}C), e $\approx$ 10 mbar) an increase of 10^{-5} in n can be caused by an increase in P of 37 mbar, by a decrease in T of 10.6 ^{o}C, or by an increase in e of 2.3 mbar.

It follows also from (1) that for sea level conditions the refraction index of the air changes between 1.00023 and 1.00054, with 1.0003 as an average value.

Because distances or distance differences derived from measurements are proportional to the adopted refractionindex, registration of temperature, pressure and humidity is required for measurement with very high proportional precision (some $1/10^{4}$ or better).

The refraction index of the air varies with space and time. These variations in space cause a curvature of the radiopaths, the so-called refraction. This curvature makes a ray, going from two points A to B, follow the quickest route. In figure 3 the radiowaves prefer therefore the upper path with the lower refraction index (higher velocity) rather than for instance the straight line which would give a longer travelling time because of the lower velocity. This idea of the quickest path is called the principle of Fermat.

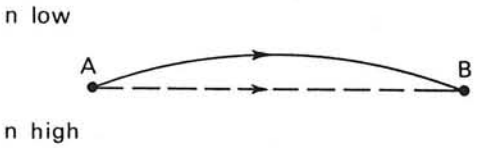


Figure 3.

Above the earth, and particularly above the sea, the refraction index often decreases so sharply with the height that radiopaths are more curved than the earth, so that nearly horizontal rays are "trapped" between the surface and a layer in the air. This is called a duct; see figure 4a. Another form of duct is the case of figure 4b.

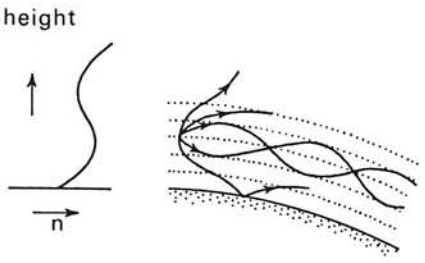


Figure 4a.

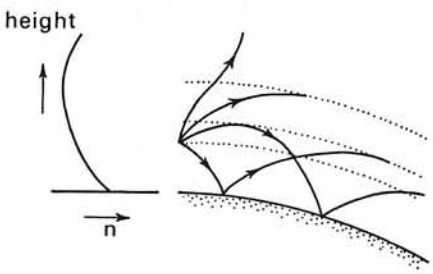


Figure 4b.

A necessary condition for radiowaves to be "trapped" in a duct is that the wavelength is much shorter than the thickness of the duct. In practice, ducts are therefore important for short waves ($\lambda < 1$ meter). Propagation over long distances is then possible with low losses of signal strength.

Another important property of the troposphere is the extinction of electromagnetic waves. As everyone knows, this extinction is very troublesome for visible light. Eventually it is one of the main advantages of radiowaves that they pass quite well through haze, fogg and clouds. Only the very short wavelengths which are comparable with the diameter of waterdropplets are considerably attenuated.

1.1.1.2. The earth surface.

The earth and the sea have electrical properties which are quite different from the tropospheric air. The relevant properties are the (relative) permittivity ε_r and the conductibility σ. See table 2 for some examples.

terrain	LF, MF, HF		VHF and higher	
	ε_r	σ siemens/m	ε_r	σ siemens/m
dry sand; urban and industrial regio	2-5	0.0001 - 0.001	2	0.03
dry, sandy grounds	5-10	0.002	-	-
humid bottom (meadows, woods, clay).	10-15	0.002 - 0.01	-	-
wet bottom (heavy clay, certain meadows, moor)	15-20	0.01 - 0.02	24	0.6
fresh water	80	0.001 - 0.002	80	2
salt water	81	4.6	80	6

Tabel 2.

Radiowaves in the air are more or less reflected on the surface depending on these quantities ε_r and σ. The not reflected part of the power is propagating into the ground, where it will be absorbed.

However, at distances within about one wavelength from the surface the apparent properties of radiowaves in the air depend on the properties of the soil material in a complicated way: the apparent velocity and the attenuation of these surface waves depend on the ε_r and the σ of the soil between transmitter and receiver, and on the refraction index of the air. This last influence is however of less importance because of the large differences and uncertainty of the properties of the soil.

1.1.1.3. The ionosphere.

In the high layers of the atmosphere a significant number of molecules in the air are ionized by radiation from the sun, in particular X-rays and ultra violet. Below heights of 100 to 50 km the radiation is so much attenuated by these processes that ionization hardly occurs.

Radiowaves in these ionized layers, in the ionosphere, are influenced in a very complicated way by the free electrons, the earth magnetic field and the collisions of the electrons with other particles.

The ionosphere is a quite variable medium and the number of free electrons depends on time (the day, season, solar activity) and place. In the lower layers, below some 120 km, the ionospheric characteristics are directly related to the intensity of radiation (solar elevation and solar intensity). The higher layers react in a more complicated way and with delays of many hours.

Of prime importance for wave propagation is the refraction index, being a measure for the velocity ($n = c/v$). If the frequency of the waves is not too low, the refraction index may be written as

$$n^2 = 1 - (f_p/f)^2 \qquad (3)$$

where f_p is the so-called plasma frequency. This is the resonance frequency of the local free electrons depending only on the (free) electron density N:

$$f_p^2 = 80.5\ N \qquad (4)$$

if N is the number of free electrons per m^3

and f_p is the plasma frequency in Hertz.

From (3) one can infer that:

1. The refraction index in the ionosphere is normally smaller than unity, i.e. the (so-called phase-) velocity is greater than in the non-ionized atmosphere. So the radiowaves will tend to refract downwards (see figure 1, the skywave).
2. If the frequency of the waves becomes much higher than the plasma frequency, the influence of the ionosphere tends to disappear. (Visible light is not influenced by the ionosphere, microwaves hardly are).
3. Radiowaves can only pass through the ionosphere if their frequency is higher than the highest plasma frequency or critical frequency f_c.

In the lower layers of the ionosphere, below some 100 km, where the density of the particles is relatively high and where collisions occur frequently, the radiowaves are highly absorbed, particularly if the ionization degree is high (i.e. if N is large, which occurs when the sun is well above the horizon.

1.1.2. Positioning systems in view of wave propagation.

In this section some typical examples of radiopositioning systems will be discussed, arranged according to the wavelength and their related propagation characteristics.

1.1.2.1. The Omega system, covering the whole earth with eight transmitters on frequencies near 10 kHz (VLF). At the unknown point (ship, aeroplane) signals are received from at least three of the eight transmitters. Because all transmitters are synchronized with each other, it is possible to find the position of the ship from the measured time intervals.

The wavelength of 30 km is comparable to the height of the lower boundary of the ionosphere. So ray tracing is not very realistic and the waves are considered as wavepatterns (modes) travelling horizontally in the area between the ionosphere and the earth (or sea) surface. The propagation is quite variable, particularly because of variations of the ionosphere. So the predictions of signal strength and time of arrival are not very precise. The Omega system is worldwide usable, but its accuracy is limited (standard deviation: a few kilometres).

1.1.2.2. L.F. waves are influenced by the conductivity of the earth (sea) surface, so that the surface waves reach far beyond the horizon, particularly over sea (high conductivity). However, at low elevation angles, i.e. over long distances, these

waves are effectively refracted by the ionosphere. If such skywaves interfere with the surface waves the required phase measurements become severely deteriorated. Only if the lowest layer of the ionosphere is highly ionized - at (summer) days - the skywave is absorbed and cannot disturb the measurements.

Another difficulty with these waves is the influence of the ground on the time of arrival and on the signal strength. Particularly poor conducting land (dry sand) yields a considerable reduction of signal strength and a change in travelling time.

Important deviations of the expected behaviour of these long radiowaves occur near (i.e. within a few wavelengths of) discontinuities like great iron constructions, coast lines, etc.

Interesting but quite troublesome is the phase jump of waves crossing a coast line. See figure 5. At considerable heights (several wavelengths above the ground) the wave fronts will be undisturbed equidistant flat vertical surfaces. Near a bad conducting surface (land surface) the waves tend to enter into the ground. So the wavefront will be inclined towards the earth. Above "well-conducting" seawater the wave fronts are nearly vertical. So a jump must exist near the coast line. See figure 5. Indeed such effects are found up to several wavelengths from the coast. This makes the calibration of long wave systems quite difficult.

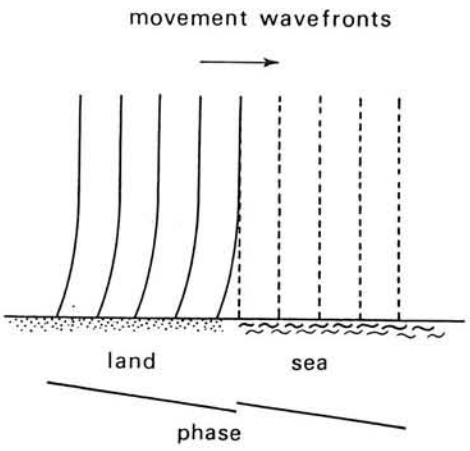


Figure 5.

Examples of systems on these wavelengths are: Loran C and Decca Navigator. With such systems a precision of some tens of metres can be obtained over distances of several hundreds of kilometres if there is no skywave or if the skywave effect is eliminated like for Loran C. Systematic errors may be much larger.

1.1.2.3. Systems on M.F. like Argo and Hyperfix. These waves have about the same propagation characteristics as the LF waves: a surface wave up to far beyond the horizon, influence of the underground, skywave. The difference is that the waves are not only affected by the conductivity but also by the permittivity of the soil.

Corresponding to the wavelength all lineair measures are smaller than for LF: the antennae, the standard deviation, the range. The accuracy is normally between 10 m and 20 m standard deviation under good conditions.

1.1.2.4. Short waves i.e. waves shorter than about one metre HF and VHF (wavelengths between 100 m and 1 m are hardly used for positioning because of the strong ionospheric influence and because of need of these frequencies for communication purposes).

For these short waves the surface wave-bending along the earth surface is insignificant because a great part of the radiopath is many wavelengths away from the earth. Just like a light-beam, the radiopath is about a straight line, but it is influenced by refraction, reflection and scattering. So, in general these waves may be used up to the radio horizon which is somewhat further away than the optical horizon owing to refraction by humidity gradients. In case of duct, the range can be significantly larger.

Reflection on the surface can be quite troublesome if the phase difference between the direct signal and reflected signal is about half a wavelength or any other odd number of half wavelengths, for in that case the signals extinct each other. So if the radiopath is over water (good reflection) one often finds zones of no signal (dead zones) for specific distances and heights of the antennae.

1.2 Geometry and signals.

The consequences of different geometries and of different forms of the signals on radiopositioning will be dealt with.

1.2.1. The geometry.

In general position fixing on the surface (in two-dimensions) can be decomposed in the determination of two or more lines of position (L.O.P.'s). A L.O.P. is the line on which the unknown point (e.g. the ship) can lay, given a certain measurement. Although a LOP is a curved line it may locally be approximated by a straight line. The most important radio measurements are:

a. A distance measurement to a known point (circular LOP).
b. The difference of the distances to two known points (so called hyperbolic LOP).
c. The azimuth to the unknown point from a fixed station (LOP approximately a straight line or a great circle).
d. Other measurements like resection (with radiodirectioning) are less important for accurate radiopositioning.

Although any combination of these measurements is possible most methods may be reduced to:

A. A combination of two or more distance measurements, the circular method, also called "range-range method".
B. A combination of two or more distance differences, the hyperbolic method.
C. A combination of distance and azimuth (b. and c.), the polar method.
D. A mixture of A and B is the measurement of "pseudo ranges". Here distances are measured with a range bias which is supposed to be constant, at least over a short time. A precise clock on board is needed to maintain this bias over some time. This method is called the "rho-rho-method" or method with independent clocks.

1.2.1.1. The circular system.

An "interrogator" (a transmitter-receiver) is placed on the unknown point P (the ship). On at least two known fixed points T_i a "transponder" (also a transmitter-receiver) is placed (figure 6).

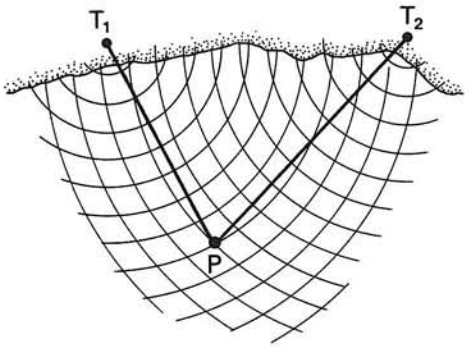


Figure 6. A circular chain.

The interrogator transmits a signal that is received by one of the transponders which transmits an answer backwards. At the interrogator the travelling time is measured and with some known propagation velocity (about $3 \cdot 10^8$ m/s) the distance is calculated. This is done for each of the transponders, simultaneously or in time-sharing. The intersection of at least two L.O.P.'s gives the position of P.
If two or more ships are to be fixed, time-sharing between that different users is necessary because the transmitting moments of the transponders are necessarily controlled by the unknown point(s).

1.2.1.2. The hyperbolic system.

A transmitter is placed at a number of fixed positions, Z_i. Each Z_i transmits a radiosignal and these signals are synchronized, i.e. they have known time intervals or phase intervals. The signals are received at the unknown point P (the ship), where the time (or phase) intervals of the received signals are measured, giving the range differences from P to Z_i. The LOP for the two transmitters (say Z_0 and Z_1) is, at least, in a flat plane, a hyperbola with Z_0 and Z_1 as foci. (On the spherical (or ellipsoidal) earth the LOP's cannot be hyperbolae). With three or more transmitters a position fix of P is possible. In figure 7 a pattern of two transmitters is sketched.

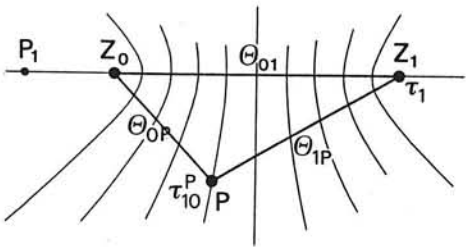


Figure 7. A hyperbolic pattern.

The synchronization of the transmitters is mostly done by receiving the signal of one transmitter, Z_0 (sometimes, but not always, named "master"), at the other transmitter(s), here Z_1 (slave). With some known delay τ_1 this slave transmits its own signal. So the measured time interval at P, τ_{10}^P, can be expressed as:

$$\tau_{10}^P = \theta_{01} + \tau_1 + \theta_{1P} - \theta_{0P} \qquad (5)$$

if θ_{01} is the propagation time from Z_0 to Z_1,
θ_{1P} is the propagation time from Z_1 to P and
θ_{0P} is the propagation time from Z_0 to P.

Out of (5) the difference in travelling time from the transmitters Z_i to the unknown point P can be solved:

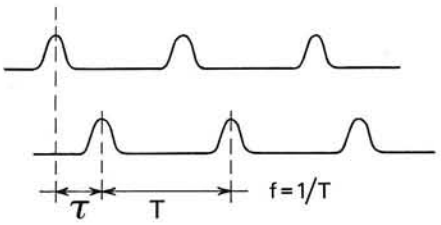
$$\theta_{1P} - \theta_{0P} = \tau_{10}^P - \theta_{01} - \tau_1$$

Assuming a constant propagation velocity v one finds for the difference of the distances:

$$\overline{Z_1P} - \overline{Z_0P} = (\tau_{10}^P - \tau_1)v - \overline{Z_0Z_1} \qquad (6)$$

The adjusted delay τ_1 and the length of the base Z_0Z_1 may assumed to be constant, besides some noise. τ_{10}^P is the measured quantity.

In practice the transmitted signals do not consist of one single pulse, but rather of a periodical repetition of pulses, or even of a continuous wave. Instead of time intervals τ, phase differences $\tau/T = f \cdot \tau$ are often introduced in the calculations. T is the repetition interval of the pulses or the period of the continuous waves, and $f = 1/T$ the frequency, see figure 8.



phase difference $\varphi = \frac{\tau}{T} = f\tau$ in periods

or $2\pi \frac{\tau}{T} = 2\pi f\tau$ in radians

or $360 \frac{\tau}{T} = 360\ f\tau$ in degrees

Figure 8. Phase measurements.

So, however, an ambiguity of an integer number of periods N is introduced. In terms of phase differences φ our equation may than be written as:

$$\overline{Z_1P} - \overline{Z_0P} = (\varphi^P_{10} - \varphi_1 + N_{10})\frac{v}{f} - \overline{Z_0Z_1} \qquad (7)$$

This formula may be used to calculate the position, or more correctly the LOP, from the measured phase difference φ^P_{10}. Formula (7) may also be used to find the influence of deviations (errors) in the measurement and in the parameters φ, v, f and Z_0Z_1.

NOTE Many hydrographers have the rather confusing practice to adjust the phase shift φ_1 in the slave and to choose N_{10} so that the reading $(\varphi_{10} + N_{10})$ becomes zero if P is situated in one line with the base Z_1Z_0 at the side of the master ($P \rightarrow P_1$ in figure 7). With this adjustment (7) becomes for P_1:

$$\overline{Z_1P_1} - \overline{Z_0P_1} = -\varphi_1 \frac{v}{f} - \overline{Z_0Z_1},$$

or, because $Z_1P_1 - Z_0P_1 = Z_0Z_1$ (see figure 7):

$$\varphi_1 = \frac{f}{v} \cdot 2\overline{Z_0Z_1}$$

Substitution in (3) gives for this adjustment for any point P:

$$\overline{Z_1P} - \overline{Z_0P} = (\varphi^P_{10} + N_{10})\frac{v}{f} + \overline{Z_0Z_1} \qquad (7a)$$

Although this formula is adequate to calculate the position in a pattern with this adjustment, the form is not so useful to find disturbing influences, because here Z_0Z_1 does not only contain the time delay over the base but also the phase adjustment of Z_1.

To fix a position with a hyperbolic system, at least two patterns are needed. See figure 9. The accuracy of a fix in such a system depends on the propagation of the radiowaves, on the geometry, and in a less extent, on the electronics and on the accuracy of the transmitter coordinates. The geometrical accuracy in a point P depends on the intersection angle of two hyperbola and on the lane width, which is the distance (e.g.) in meters between two LOP's (hyperbolae) with a

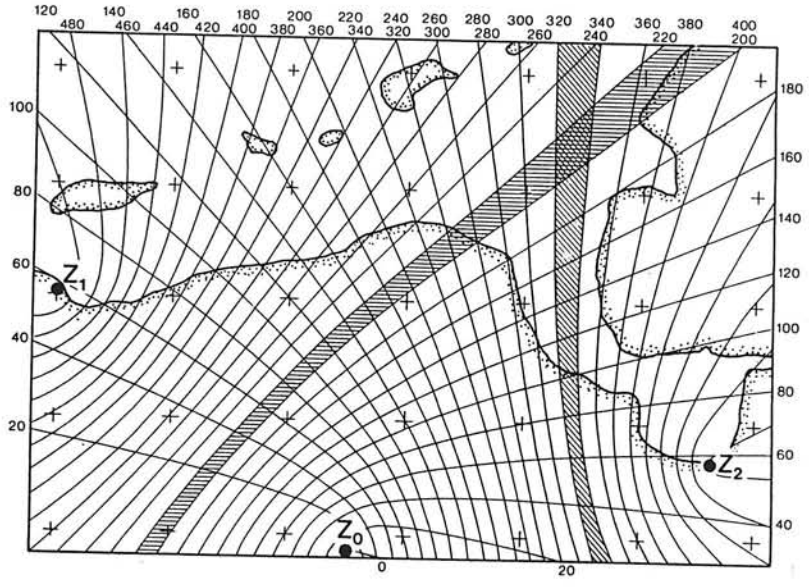


Figure 9. A hyperbolic chain.

phase difference of one period (often called a lane). The lane width is minimal on a base (a halve wavelength). See figure 2.4, where zones of 20 lanes are shaded.

It is also possible to measure in the pattern Z_1Z_2, which has not been drawn in figure 9. Besides possible receiver noise this measurement will however give in combination with for instance Z_0Z_1 exactly the same position as the pair Z_0Z_1, Z_0Z_2.

It is also possible to measure in two patterns with no common transmitter, even if they belong to different chains (i.e. if the pairs are not synchronized to each other).

1.2.1.3. The independent clocks.

If one has a clock on board which keeps the time accurate enough during a certain time, it is possible to find distances by reading on this clock the times of arrival of signals transmitted from known stations at known epochs. Because the clocks of the transmitters and the ship born clock are not synchronized they will after some time deviate too much from each other. So more or less frequent updates are necessary, for instance by position fixing with (doppler-) satellites or by working from time to time in hyperbolic mode. So continuous reception of only two stations can suffice.

1.2.1.4. The polar method.

It is possible to combine distance measurements (circular method) with radio angle measurements by using very short radiowaves ($\lambda \approx 3$ cm) and broad antennas which direct themselves mutually. If one antenna is placed on a ship and the other one on a fixed platform, the direction of the last one determines, together with the travelling time of the radiowaves, the position of the ship. See figure 10.

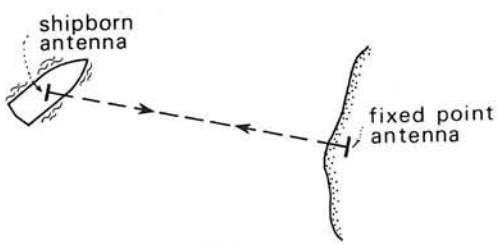


Figure 10. Polar system.

1.2.1.5. A comparison between the different methods.

The hyperbolic method and the method with independent clocks have the advantage that an unlimited number of ships can fix their position simultaneously (multi-use system).

Distance measurements (circular method or independent clocks) have the advantage that the lane width is a half wavelength anywhere in the pattern. It is also an advantage that the propagation effects are easier to understand for distance measurements than for hyperbolic or angle observations.
The polar system has the advantage that one fixed station suffices for a complete, though not redundant, fix.

1.2.2. Signals.

For position fixing one can use light, radiowaves (both electro-magnetic waves) or acoustic waves. Radiowaves propagate quite well through the atmosphere (nearly straight path, constant velocity, few absorption). So does light, if clouds or fog do not obstruct the path. Under water however electro-magnetic waves are hardly practicable because of high absorption; here one uses acoustic waves with fair propagation characteristics.
Radiowaves may be longer or shorter. They may be unmodulated sine waves or they may exist of sine modulated waves: continuous waves, see figure 11.

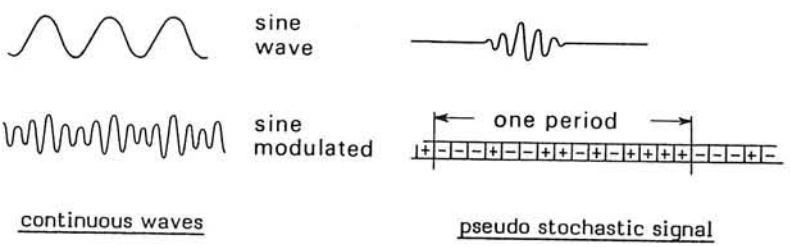


Figure 11. Signal forms.

Another simple sort of signal is the short puls. A more complicated signal, but increasingly more in use, is the pseudo random code. Table 3 gives a survey of the different signals for different carriers, with a number of applications and the relevant characteristics.

1.2.2.1. The continuous wave (c.w.).

The c.w., and more particularly the sine wave, has the advantage of a very narrow bandwidth, so that the interference of the transmitters to other people is quite limited (at least in the frequency domain) and so that in the own receivers noise and interference can easily be filtered out. A disadvantage is the ambi-

Signal	light (or i.r.)	radiowaves		acoustic	power max. mean	band-width	separ. of reflect.
		short	long				
Puls	Laser dist. meter	Radar	Loran C	Echo sounding Acoust. position.	high	large	+
PS:rand.code	x	Syledis G.P.S.	x	x	low	large	+
Cont.wave	El.optic distance meter for surv.	.NNSS satel.	Decca Navig. Hyperfix	Doppler sonar	low	small	-
Directivity	very good	bad---fair	no directivity	depending on wavelength	Legenda + good - bad x not applicable P pressure T temperature		
through fog	-	+	++	x			
velocity dependent on	P,T	P,T,H_2O vapour ground		P,T,salinity			

Table 3. Carrierwaves and signals, examples and characteristics.

guity of an integer number of periods. This ambiguity may be improved by using two or more frequencies. Then the period of ambiguity is the smallest period divisible or nearly divisible by each of the used periods. See figure 12.
The use of more frequencies reduces however the advantage of narrow bandwidth.

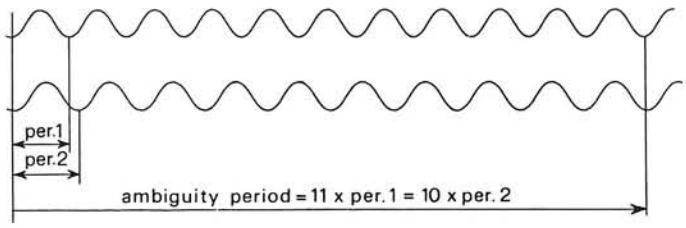


Figure 12. 11 periods "2" = 10 periods "1".

1.2.2.2. The short pulse.

For a high resolution, i.e. for a precise measurement one needs sharp pulses. But a pulse is physically and mathematically a composition of sines of different frequencies (and phases) within a certain frequency band (the spectre), and the sharper the pulse the broader this band. For more mathematical treatment see textbooks about Fourier transforms. Figure 13 gives an illustration.

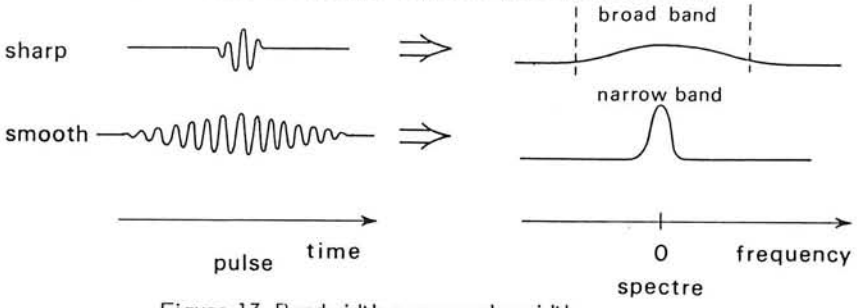


Figure 13. Bandwidth versus pulsewidth.

The ambiguity can become irrelevant if the repetition period of the pulses is long enough. Another advantage of the short pulse is that in case of reflections one can often choose the correct one of all the received pulses e.g. the first arriving one.
The most important disadvantage of the pulse is its high maximum power. To recognize and to use the signal, the total energy over the measuring time is of importance. Consequently some minimum average power is needed at the reception, and a narrow pulse must be very high (high maximum power) to give sufficient mean power.

1.2.2.3. The pseudo random code.

The pseudo random code or maximal-length sequence is a sequency of binary numbers which has a structure similar to a random sequence, but which is periodical.

A pseudo random code can simply be made with a shift register and a modulo 2 arithmic. See figure 14.

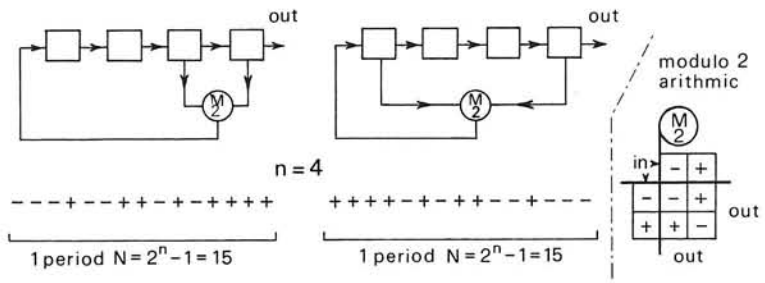


Figure 14. An example of a pseudo random code.

With one shift register one can make a number of codes, increasing sharply with the number N of elements of the register. The autocorrelation function consists of a sharp pulse with a repetition period of N times the duration of one element. So by correlation techniques the pseudo random signals can be transformed into short pulses with a long repetition period. See figure 15.

The code is often used so that only the phase of the carrier wave is shifted by e.g. 120^{o} or 180^{o} for a "-" with respect to a "+". See figure 16.

Such a signal has a broad bandwidth, but a low maximum power, a long ambiguity period, and reflections may be filtered out as with short pulses. Another advantage is that different codes can be used to discriminate between different transmitters (to use the wanted one) and between users (to oblige them to pay or to provide unauthorized use).

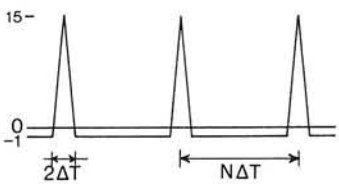


Figure 15.
Autocorrelation function of pseudo random code with N = 15.

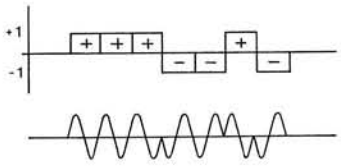


Figure 16.
Binary phase-coded signal.

1.3 Systems for radiopositioning.

In this chapter a survey will be given of most of the existing systems. Further a number of typical systems will be treated in some more details.
Description and features of systems can be found in literature and in information from manufacturers. Some critical attention is however justified because not only manufacturers may tend to be optimistic, but also the users may be inclined to accentuate the more positive features.

1.3.1. A survey of systems.

More detailed information may be found in:
R.C. Munson, "Positioning systems", Proc. Féd. Int. Géom. (FIG) Commission 4, Stockholm 1977.
Féd. Int. Géom (FIG), "Hydrographic survey equipment", catalogue and supplement, Hydrogr. Soc. Spec. Publ. 11.

1.3.1.1. A survey of satellite systems.

Only two systems are more or less available for precise navigation:
N.N.S.S. (Navi Navigation Satellite System) doppler satellites.
The satellites and their tracking are controlled by the U.S. government. Receivers are available from many manufacturers. The system works world wide. Navigation and (also precise) relative positioning is possible without the use of restricted data or codes. The accuracy for moving ships is a few hundred meters. By integration with other systems an (absolute) accuracy of some 30 meters is possible (standard deviation). A frequency of 400 MHz and an additional frequency of 150 MHz for accuracies better than a few hundred meters, are used.

G.P.S. (Global Positioning System) = Navstar.
The satellites and their tracking are controlled by the US government. Receivers are made by different companies. For using the system in normal modes ("Precise" or "Coarse Acquisition") special codes are required the use of which will probably be quite restricted. Certain alternative modes without such restrictions may be used for fixed stations and for moving ships. When in full operation (early 90's) continuous positioning of ships will be possible with accuracies of a few meters or even better; for fixed points probably subcentimeter accuracies may become possible, also without the knowledge of the codes.

1.3.1.2. A survey of terrestrial systems on short waves.

These systems are in use for precision navigation. The ranges of most of these systems are up to somewhat beyond the (radio) horizon, say twenty or thirty kilometers. At favourable conditions some of the systems reach up to hundreds of kilometers (Maxiran, Syledis). Most systems work in range-range mode, some (also) in hyperbolic mode; with two of the systems bearings can be measured. Sources of errors: reflections, and beyond the horizon also variations in the turbulence of the higher atmosphere. No land-sea effects (except for reflections), no sky wave. A typical accuracy is a few meters. For a survey see table 4.

1.3.1.3. A survey of terrestrial systems on MF (Medium Frequency).

Most of these systems work on frequencies near 2 MHz. The maximum ranges are a few hundreds of kilometers at daytime. In the night the effective ranges may be much shorter. The maximum effective range is highly dependent on over-land paths, on ionospheric reflections and on the possibility of filtering (depending on the movements of the ship).

The accuracy is quite variable and often not very well defined. Often the systems are used in hyperbolic mode for an unlimited number of users. Sometimes the rho-rho mode is used, also for an unlimited number of users, or the range-range mode (active mode) for a small number of users in time sharing. The most important sources of errors are the propagation characteristics over land, land-sea crossings, and the sky waves. Large errors may occur within a few wavelengths from constructions and from the coast. The systems are used for hydrographic work. A survey is given in table 5.

1.3.1.4. A survey of systems on LF and on VLF (Low Frequency, Very Low Frequency).

These systems are in the first place used for general navigation; the LF-systems are also used for ocean-wide hydrographic work. These long wave systems are almost exclusively used in passive modes (hyperbolic or sometimes rho-rho).

The accuracy of the LF-systems may be a few tens of meters under good conditions and with careful use.

The main sources of error are: propagation over land and ice, and particularly for unmodulated systems: sky waves. One may also find large errors within a wavelength from constructions and from coasts.

Name of the system	Manufacturer Address	Mode(s) of users	Number of users	Frequency (roughly)
Artemis	Christiaan Huygens Lab. Noordwijk, The Netherlands	range and bearing	1	9 GHz
Audister	S.P.C. Electronic Co. Tokyo, Japan	ranges	1	3 GHz
Autotape	Cubic Western Data San Diego, California	ranges	1	3 GHz
Aztrac	Odom Offshore Surveys Inc. Baton Route, Louisiana	bearing	1	10,4 GHz
Hydroflex	Tellurometer Chessington, U.K.	ranges	6	3 GHz
Maxiran	Navigation Management Inc. Ocala, Florida	ranges	6	0,45 GHz
Microfix	Racal Positioning System Leatherhead, U.K.	ranges	?	5 GHz
Miniranger	Motorola Inc. Temple, Arizona	ranges	> 10	5,5 GHz
Syledis	Sercel Carquefou, France	hyperbolic ranges	∞ 4	0,45 GHz
Trident	Thomson C.S.F. Malakoff, France	ranges	≤ 50	0,5 GHz or 1,2 GHz
Trisponder	Del Norte Technology Inc. Euless, Texas	ranges	8	9 GHz or 0,4 GHz

Table 4. A survey of terrestrial positioning systems on short waves.

Name	Manufacturer	Mode(s)
Argo	Cubic Western Data San Diego, California	range-range (option) hyperbolic
Geoloc	Sercel Carquefou, France	hyperbolic ranges (integr. w. Syledis and GPS)
Hifix	Racal Survey Ltd. Leatherhead, U.K.	hyperbolic ranges
Hydrotac	Odom Offshore Surveys Inc. Baton Rouge, Louisana	hyperbolic ranges
Hytrac	Gardline Surveys Grest Yarmouth, U.K.	hyperbolic ranges
Lorac	Lorac Service Co. Houston, Texas	hyperbolic rho-rho integr. w. satellites
Microphase	Offshore Navigation Harahan, Louisiana	rho-rho
Omi	Ocean Measurements Inc. W. Palmbeach, Florida	ranges rho-rho
Raydist	Teledyne Hastings-Raydist Hampton, Virginia	hyperbolic ranges
Toran	Sercel Carquefou, France	hyperbolic

Table 5. A survey of positioning systems on M.F.

On VLF only one system is of importance:

Omega.

This system works with unmodulated wave trains on frequencies near 10 kHz. It has a world wide coverage with 8 transmitters in hyperbolic mode. The accuracy is a few kilometers.

Sources of error: the variations of the ionosphere. The transmitters are controlled by the U.S. government. Receivers obtainable from different manufacturers.

On LF three systems will be mentioned:

Loran C.

The frequency is exactly 100 kHz, transmitted in pulses. These pulses are used for lane identification and for elimination of the sky wave. Ranges up to 1000 or 2000 km day and night. Accuracy under good conditions some 30 m standard deviation. Mostly used in hyperbolic mode. Also the rho-rho mode is often used. Range-range mode is possible.

Many transmitters are controlled by the U.S. government, other ones by local authorities. User apparature available from different manufacturers.

Pulse 8.

Almost identical with Loran C, but with less power, so that the ranges are shorter. Transmitters and receivers manufactured by Racal Survey.

Decca Navigator.

Frequencies between 70 kHz and 130 kHz, unmodulated. Ranges up to 500 km at day and 100 km at night and at twilight. Mostly used in hyperbolic mode.

1.3.2. Satellite systems.

Only one or two satellite navigation systems are at the moment (1988) operational for general use. Some other systems are operational for restricted use or are in study. We mention the Argos system for meteorological and oceanographic bouys and ships (see D.E. Wells, Hydr. J., no. 18, Aug. 1980, p. 49), Glonass of the USSR, and Geostar with geostationary satellites. See Proc. Global Civil Satellite Navigation Systems, London, May 1984.

1.3.2.1. N.N.S.S.

This Navy Navigation Satellite System, often mentioned as "Doppler satellites", also called "Transit", the only system fully operational for general use, works

with a number (4 to 6) satellites in polar orbits on a height of 1100 km. The period of orbit is about 100 minutes. For any point on earth a satellite is 20 minutes or less above the horizon.

The satellites transmit modulated radiowaves on 400 MHz and on 150 MHz. The modulation contains information about orbit, time and position (ephemeris). Special receivers are used to receive and to process the signals, and often to compute their position.

Essentially the number of periods of the carrier wave(s) are counted in intervals of about 4,6 sec., 30 sec. or 2 minutes. These counts are measures for range differences from the receiver to two satellite positions. Out of the counts of a reasonable passage (maximum elevation between say 30^{0} and 70^{0}) a position fix is possible. The accuracy is some 40 m for a fixed platform. For a moving ship the accuracy degrades by an order of magnitude. If the movements of the ship are measured with an other system an accuracy of 50 m or better is possible.

If some 20 passages are used on stationary points, an accuracy better than a meter is possible in relative accuracy. In this case "translocation" may be used, i.e. from 2 or more terrestrial points simultaneous measurements are executed, so that the errors in the ephemeris and also the atmospheric errors are more or less eliminated. Also on a moving vessel translocation can be useful if the satellite receiver works integrated with an other system.

The system is world wide, but it is an disadvantage that only 10 to 20 passages a day are available.

Note: the NNSS system can be considered as an hyperbolic system (range differences) where the transmitter (one satellite on different epoques) is synchronized by the satellite born oscillator which works as a phase memory. Table 6 gives a summary of obtainable accuracies.

Fixed station	. One passage 30 m . A few days (> 20 passages) 1 m
Moving ship	. Unknown velocity: a few hundreds of meters . Velocity vector measured: 50 m . Velocity vector measured + translocation } better than 50 m

Table 6. Typical accuracies (standard deviation) with NNSS doppler satellites.

1.3.2.2. G.P.S.

The Global Positioning System, also called Navstar, works with a minimum number of 18 satellites arranged in 6 orbits on about 20.000 km height. The orbital period of each satellite is 12 hours relative to the earth. This configuration is so designed that always at least 4 satellites will be sufficiently high above the horizon in any place on earth. The satellites transmit signals in pseudo random code with long repetition periods, so that ambiguities are not annoying. The transmitted sigals are synchronized, i.e. they have known relations in time. In the most direct mode differences of the time of arrival are measured for at least 4 satellites (quasi) simultaneously. So this may be considered as an hyperbolic system with 4 synchronized transmitters. This number of 4 is the minimum number to fix a point in the three-dimensional space (compare the plane, where three transmitters are needed).

The measurements may also be done in doppler mode, where sequential time (= range) intervals are measured from one satellite, as with the NNSS satellites. Another mode is the correlation mode or interferometer mode. Here the knowledge of the codes is not strictly needed. Signals are compared in time by correlation.

These two alternative modes may also become adequate for navigation, but in this case some communication with the shore is required for on-line processing.

1.3.3. Terrestrial systems on short waves.

1.3.3.1. Artemis.

With this system bearing and range is measured between two points, of which usually one is fixed. So with this system it is possible to position a vessel (without redundancy) if only one fixed point is available.

The antennae of both stations are automatically pointed towards each other (see figure 17). At the fixed station the bearing of the antenne is sensed and transmitted to the mobile. The distance is measured by transmitting signals from the mobile to the fixed station and back from the fixed station to the mobile. All these signals are modulated on the carrier waves of about 9 Hz. Because the system works with pointing narrow beam antennes the radio connection is very efficient in signal to noise ratio.

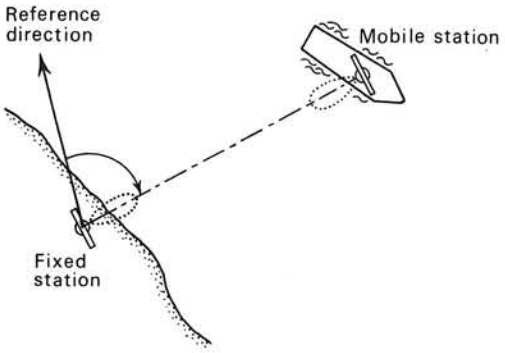


Figure 17. Artemis.

At the beginning of an operation and in case the connection is lost, a search program is executed. When the connection has been found the servo mechanics take care of the preservation.

The angle measurement.

The most particular parts of the Artemis are the antennea: so-called slotted waveguide radiators. An antenna exists of an horizontal tube of more than one meter with a pattern of slots. The antenna has very small sides lobes and it comes up to high requirements for phase symmetry.

The phase difference is measured in the carrier wave received in the left and the right half of the antenna. The error signal for the servo mechanisme is obtained from this phase difference to correct the bearing of the antenna. In order to find a reference for the bearing at the fixed station at least once the antenna has to be pointed to some fixed direction (for instance to the sun) with the help of an optical telescope.

1.3.3.2. Syledis.

This system, working on a wavelength of about 70 cm is a range-range system with the possibility to work also in the hyperbolic mode. The precision within line of sight (LOS) is given by a standard deviation of 0.5 m to a few meters. Beyond the horizon up to about 1.5 x LOS this precision may grow to 5 to 10 meters. By the use of directional antennae and high power (if permitted) ranges of much more than 100 km may be possible, but the errors may increase to tens of meters.

The working principle of Syledis.

The signals have a repeating frequency of at least 10 Hz. Such a maximum period of 100 milliseconds contains 30 time slots. See figure 18.

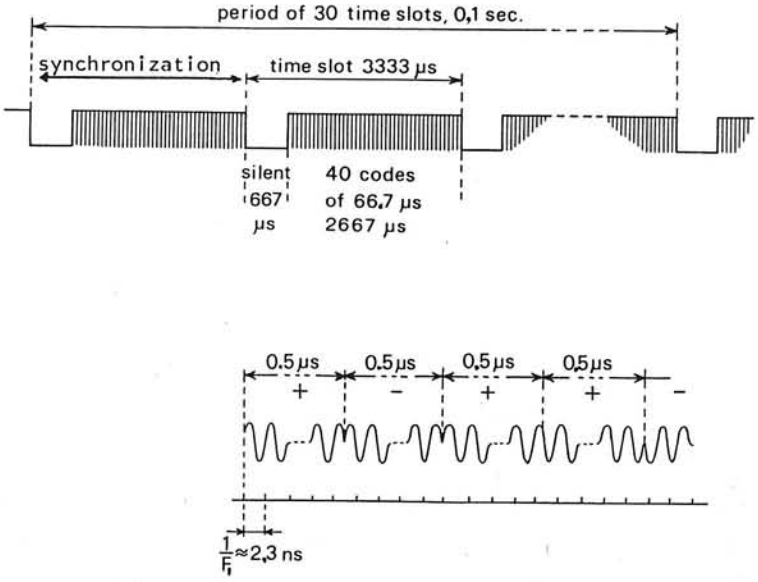


Figure 18. The Syledis signals.

A time slot of 3.33 millisec. starts with a silent period of 667 μs, followed by an active period of 2.667 ms. The first slot of each period provides the synchronization for the choice of the right time slot in each of the participating instruments. This synchronization slot exists of an unmodulated signal of frequency F2 differing by 5 kHz, 10 kHz or 20 kHz from the carrier wave F1 in the other time slots. That signal is transmitted by one of the stations and received by all other stations. During the active time of each of the following slots a code signal of 66.67 μs is repeated 40 times.

Such a code signal is a pseudo random signal of $2^7 - 1 = 127$ elements (+ or -) each of about 0.5 μs. The + and - signals differ in phase by 180^0. Because with the 127 elements a number (128) of codes may be constructed, different transmitter to receiver connections can be recognized. So each of the fixed and mobile stations has a number of code generators (normally 4).

A distance measurement is schematically indicated in figure 19. During the first time slot of a period all programs of time slots are synchronized by the frequency F2 from one of the stations. In the next slot a mobile station (interrogator) transmits a certain pseudo random code on frequency F1. This code is made by a code generator of the interrogator. During the same slot an identical code is generated in the wanted fixed responder. In the correlator the crosscorrelation is measured between the signal from the interrogator and the delayed signal of the responder itself.

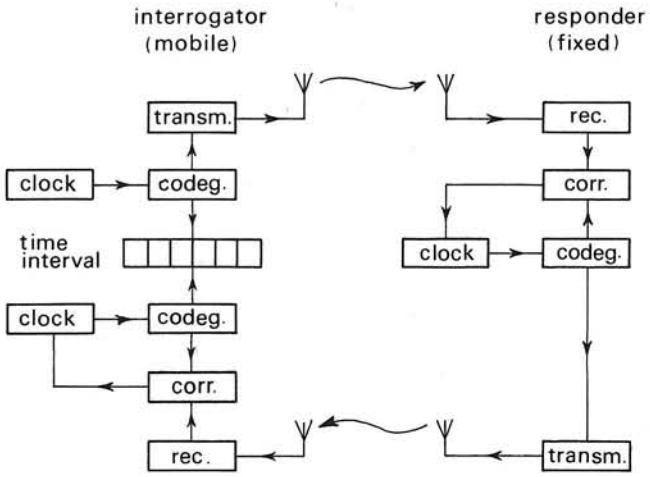


Figure 19. Diagram of distance measurement with Syledis.

The delay is increased in small steps until the correlation is high, i.e. until both signals are almost identical. This process lasts about one or two minutes at the start of a measurement, but when locked the process goes on very fast. In the interrogator the delay for maximum cross correlation is measured and so the travelling time and consequently the range can be found.

The other time slots can be used for other ranges and one of the slots may be used for the hyperbolic mode. After a whole period (maximum 100 ms) all the measurements are repeated.

The ambiguity has here a period of 66.7 μs corresponding with a distance of $\frac{1}{2}$ c . 66.7 . 10^{-6} = 10 km which is quite acceptable.
The resolution is prescribed by the length of one element. This time interval of 0,5 μs corresponds to $\frac{1}{2}$. c . 0.5 . 10^{-6} = 75 m distance. Owing to an integration over at least 10 groups of 40 elements the resolution may be less than a meter. In this integration the velocity of the vessel is included. Changes in the magnitude and in the direction of the velocity and also velocities higher than 100 knot may cause errors.

Remark: A compromis has been made in the Syledis between the ambiguity period and the integration time. Doubling the ambiguity period would result in doubling the integration time for the same number of time slots, i.e. for the same resolution. In that case also the search time when sailing into the radio pattern will be doubled.
During this initial search all possible delays with steps of 0,5 μs have to be tried. That are 127 steps per time slot. For each of these 127 steps the integration time T_i is needed. On an average half of this time will do for a search, which will than last $\frac{1}{2}$. 127 . T_i which is about one minute.

Some possibilities of Syledis.

With the standard equipment, where each station contains 4 code generators, at most 4 mobiles can measure distances to 3 fixed stations quasi simultaneously. With the same equipment it is also possible to form two hyperbolic patters with 3 fixed stations, but then only 3 mobiles can each measure 3 distances to the 3 fixed stations.
For these modes not all 30 time slots are needed so that the period can become shorter than 100 msec. It is however also possible to use some of the time slots for the possibility to switch from one fixed station to another one; for instance if it gives a better reception.

1.3.4. Terrestrial systems on longer waves.

Two typical examples will be treated: Loran C and Argo.

1.3.4.1. Loran C.

This general navigation system works on a carrier wave of exactly 100 kHz (wavelength about 3 km). Most frequently it is used in the hyperbolic mode. The master transmits periodically a group of 9 pulses. See figure 20.

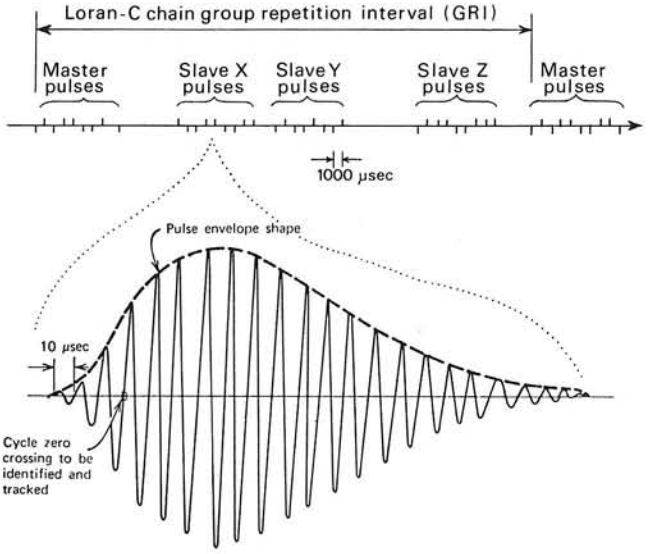


Figure 20. The format of Loran C.

The period (GRI, group repetition interval) can have different durations between 5 and 10 msec. for different chains in order to make discrimination between different chains possible. The GRI is so long that within the maximum range of the transmitters the signals of the old period do not interfere in the new period: 5 msec correspond to a distance of 0.005 c = 1500 km.

The slaves X, Y and Z transmit groups of 8 pulses somewhat later within the GRI. The pulses differ by 180^{0} in phase according to a certain key. See figure 20 where the spikes are up or down. These phase differences are introduced to recognize the different transmitters and to eliminate heavy delayed sky waves.

In the receiver phase differences of the carrier waves are measured. For the lane identification (the integral number of carrier wave periods) the enveloping curve of the pulse is used to select the beginning of the third sine of the pulse.

Note: Nowadays one often prefers the point 25 μsec from the start instead of 30 μsec.

The advantage of this pulsed system is that the sky wave can be eliminated because this wave arrives at least more than 30 μs (= 3 periodes) later than the ground wave. A disadvantage is the broader bandwidth and the higher peak power compared with an unmodulated signal.
The Loran C signals are also suitable for time transmission owing to its discrete time signals. So it is important that Loran C transmissions are controlled by good atom clocks and kept by international time services.

1.3.4.2. The Argo system.

This system, working on Medium Frequency, is intended for hydrography and for other precision navigational work. The mark DM54 will be treated here.
The Argo system is in the first place a range-range system but it can also be used in the hyperbolic mode. Argo works on one pair of frequencies of which one frequency is only used for lane identification. Both frequencies can be chosen by the user between 1.6 and 2.2 MHz. So the lane width is about 85 m (in the hyperbolic mode on the base). Dependent on the number of users and the number of required ranges and/or range differences a certain timing program can be entered for the fixed and mobile stations. Figure 21 gives an example of such a timing program for 6 mobiles each measuring 4 ranges plus a hyperbolic mode for an unlimited number of users.

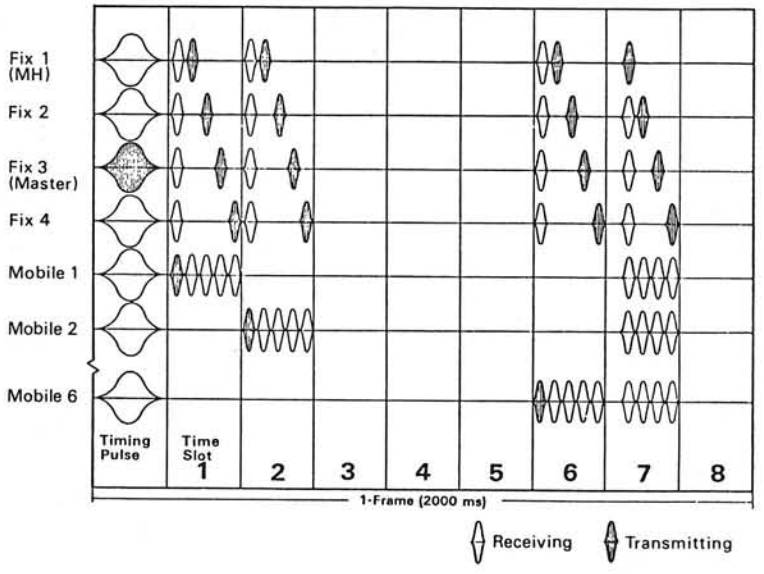


Figure 21. Example of an Argo timing program (from Cubic Western Data).

If wanted it is also possible to make a timing program for 9 mobiles measuring 3 ranges each, or other timing programs.
The outlined program has a fixed period of 2 seconds. The first interval of 144 msec (timing pulse) is used to synchronize the program for all mobile and fixed stations. The timing pulse is transmitted by one of the fixed or mobile stations (here fix 3) and received by all other stations. In the Argo system this transmitter is called the "master". If the distances are unfavourable it is also possible to use one of the stations as a "relay". In this case the master and the relay transmit alternately a timing pulse so that the relay is synchronized by the received master signal.
After the timing pulse there are a number of time slots (here 8 ones). Within each of the time slots 1-6 one of the mobiles is serviced. Time slot 7 is used here for the hyperbolic mode; time slot 8 can be used by any of the interrogators for lane identification.
Time slot 1 starts with a 56 msec period for the transmission of an unmodulated signal from mobile 1 which is received by the 4 fixed stations. In the next four intervals of 44 msec the fixed stations respond by transmitting one by one the carrier wave synchronized by the just received signals. As a result in time slot 1 mobile 1 has measured the ranges to the 4 fixed stations. So the times slots 2-6 are used for the other mobiles. In time slot 7 the fixed station 1 works as phase synchronizer for the other fixed stations (station 1 could be called hyperbolic phase master). In this time slot no mobile is transmitting and an unlimited number of mobiles may use the hyperbolic mode with their receivers.
In the last time slot a carrier frequency of about 10% above the normal frequency is used for lane identification.

Checking and smoothing.

As a first check for a valuable received signal, the signal strength is detected. Further the measured ranges are continuously compared with predictions from the last few measurements. If the difference is too large a warning is given. If the difference is acceptable the new measurement is used together with preceding ones to find a more or less smoothed value. The amount of smooting is in table 7 indicated by an instrument code and by the "time constant". This time constant is the number of seconds over which the measurements are included in the smoothing with a weight of more than about $1/e = 0{,}37$. A strong smoothing gives less influence of noise, interference and small skywaves, but the results can not follow quick changes in course and velocity.

SMOOTHING CODE	TIME CONSTANT	APPLICATION
0	N/A	Raw data computer processing
1	2.8	High speed — survey (launch)
2	3.0	High speed — survey
3	3.9	Moderate speed — survey
4	5.6	Moderate speed — seismic
5	9.0	Low dynamics — seismic
6	13	Very low dynamics — seismic
7	19	Large ship — no maneuvers
8	26	Very large ship — no maneuvers
9	39	Pipe lay barge — at anchor

Table 5. Data smoothing for Argo (from Cubic Western Data).

1.4 Literature on radiopositioning.

1.4.1 Books.

(1) A.E. Ingham: "Sea Surveying" (2 volumes) Wiley 1975.

(2) A.E. Ingham: "Hydrography for the surveyor and engineer", Granada, London, 1984.

(3) C.D. Burnside: "Electromagnetic Distance Measurement", Granada, London, second ed. 1982.

(4) S.H. Laurila: "Electronic surveying in practice", Wiley 1983.

(5) S.H. Laurila: "Electronic surveying and navigation", Wiley 1976.

(6) G.J. Sonnenberg: "Radar and Electronic Navigation" Newness Butterworths, London, 1978.

(7) FIG "Hydrographic Survey Equipment Catalogue", Publ. SP11 of the Hydrographic Society.

(8) D.B. Thompson, D.E. Wells, W.H. Falkenburg: "An introduction to hydrographic surveying", Dept. of Surveying Engineerig,Univ. New Brunswick, Fredericton N.B., Canada, Lecture Notes 53 , 1981.

(9) M.I. Skolnik: "Introduction to radar systems", McGraw Hill, 1980.

(10) P.G. Sluiter: "Positioning for marine seismic surveys", Thesis Delft University of Technology, Geodetic Department, May 1988.

1.4.2 Periodicals with papers on radiopositioning.

- Navigation (3 monthly) Journal of the Institute of Navigation, Washington.
- The Hydrographic Journal (Quarterly), Hydrographic Society, London.
- The International Hydrographic Review (6-monthly), IHB, Monaco.
- Lighthouse (6-monthly), Canadian Hydrographers' Association.
- Marine Geodesy (6-monthly), Crane Russah, New York.
- IEEE: the series: Transactions on antennae and propagation.
  Transactions on geoscience electronics.
  Transactions on aerospace and electronic systems.

1.5 Exercises.

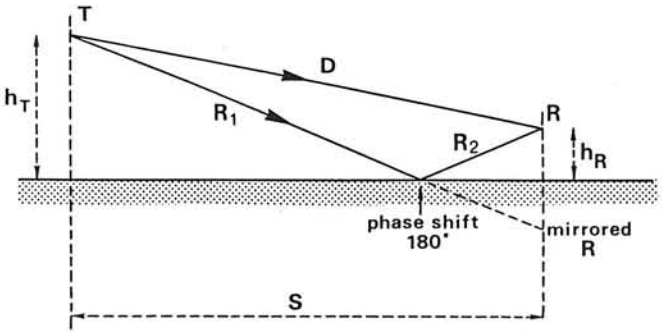
1. Dead zone.

A transmitter T, with as antenna a vertical dipole at a height h_T, transmits a carrier wave of frequency f.

The receiving antenna R (also a vertical dipole) is placed at a height h_R on some mobile.

What is the longest distance S, for a dead zone will to occur if the phase changes by 180^o at the reflection point?

f = 400 MHz
h_T = 30 m
h_R = 10 m



Solution

The first minimum will occur when the difference between the reflecting path $R_1 + R_2$ and the direct path D, $R_1 + R_2 - D$, equals the wavelength $\lambda = c/f. = 3.10^8/4.10^8 = 0.75$ m.

$$R_1 + R_2 = \sqrt{S^2 + (h_T + h_R)^2} \approx S\{ 1 + \tfrac{1}{2}\frac{(h_T + h_R)^2}{S^2} \} \qquad (h_T + h_R << S)$$

$$D = \sqrt{S^2 + (h_T - h_R)^2} \approx S\{ 1 + \tfrac{1}{2}\frac{(h_T - h_R)^2}{S^2} \} \qquad (h_T - h_R << S)$$

$$R_1 + R_2 - D \approx \frac{4h_Th_R}{2S} = \frac{2h_Th_R}{S} = \lambda$$

For the above numbers: $S = \frac{2 \cdot 30 \cdot 10}{0.75} = \underline{\underline{800}}$ m.

Exercises signals and systems.

2. The influence of meteorological assumptions.

Find with formula (2) the influence of an error of ΔP = 10 mbar in the assumed air pressure on a measured distance.

[A distance S is calculated from the measured travelling time t with S = V.t = (c/n)t where V = the propagation velocity, n the refraction index and $c \simeq 3.10^8$ m/s the free space propagation velocity.]

Do the same for an error in temperature ΔT of 10^0 C and for an error Δe of 10 mbar in the pressure of the water vapour.

Solution

$$S = ct/n$$

$$\Delta S = \frac{\partial S}{\partial P}\Delta P + \frac{\partial S}{\partial T}\Delta T + \frac{\partial S}{\partial e}\Delta e$$

$$= \frac{dS}{dn}\frac{\partial n}{\partial P}\Delta P + \frac{dS}{dn}\frac{\partial n}{\partial T}\Delta T + \frac{dS}{dn}\frac{\partial n}{\partial e}\Delta e$$

$$= \{-\frac{ct}{n^2}\frac{77.62}{T}\Delta P - \frac{ct}{n^2}(-77.62\frac{P}{T^2} + 12.92\frac{e}{T^2} - 37.19\cdot 10^4\frac{2e}{T^3})\Delta T -$$

$$- \frac{ct}{n^2}(-12.92\frac{e}{T^2} - 37.19\cdot 10^4\frac{1}{T^2})\Delta e\}\ 10^{-6}$$

$$\frac{\Delta S}{S} = -\frac{10^{-6}}{n}\{-\frac{77.62}{T}\Delta P + (-77.62\frac{P}{T^2} + 12.79\frac{e}{T^2} - 37.19\cdot 10^4\frac{2e}{T^3})\Delta T +$$

$$+ (-12.92\frac{1}{T} + 37.19\cdot 10^4\frac{1}{T^2})\Delta e\}$$

For T = 300 K, P = 1000 mbar, e = 10 mbar, $n \simeq 1$.

$$10^6\frac{\Delta S}{S} = \frac{77.62}{300}\Delta P + (77.62\frac{1000}{300^2} - 12.92\frac{10}{300^2} + 37.19\cdot 10^4\frac{20}{300^3})\Delta T$$

$$+ (12.92\frac{1}{300} - \frac{37.19}{300^2}\cdot 10^4)\Delta e$$

Now substitute ΔP = 10 mbar, ΔT = 0 and Δe = 0 to find the influence of ΔP = 10 mbar: $\Delta S/S \simeq 2.6.10^{-6}$.

In the same manner one finds the influence of temperature:

$\Delta T = 10^0 C \rightarrow \Delta S/S \simeq 12.10^{-6}$, and for humidity Δe = 10 mbar

$\Delta S/S \simeq 40.10^{-6}$.

Exercises signals and systems.

3. Independent clocks.

For distance measurement one uses on board a good quartz clock which differs in frequency not more than $1:10^{10}$ from the transmitter clock, i.e. $|f_{receiver} - f_{transmitter}| / f_{transmitter} = |\Delta f|/f \leq 10^{-10}$.

During which time interval one can maintain an accuracy of 30 m?

Solution.

After 1 sec. sailing one looses not more than 10^{-10} sec, which corresponds with $10^{-10}/c = 0.03$ m.

So 30 m will be lost after not less than $30/0.03 = 1000$ sec.

Exercises signals and systems.

4. Radiopositioning in different situations.

Which sort of methods would you choose in the following situation? (which wavelength, geometry, terrestrial, satellites, etc.). Motivate your choice.

a. - A position fix at some 10 km from the coast. Required accuracy 5 m.
b. - Navigation of many ships at some 100 km from the coast. Required accuracy 100 m.
c. - Positioning at the ocean, say 2000 km from any coast. Required accuracy 2 km.
d. - Measurement of the position of a fixed rigg at 200 km from any coast. Accuracy one m or a few m.

Solution.

a. Short waves ($\lambda < 1$ m). For one position fix the circular method is often the most useful. Good accuracy. Well within the maximum range.
b. In this case a multi-use system is necessary, so a hyperbolic system is a good solution.
The NNSS doppler system is not sufficiently accurate. GPS is possible, but expensive and monopolistic. A frequency of some 2 MHz is reasonable.

c. Terrestrial systems on wavelengths of at least 3 km. These systems are hardly used as circular systems. Hence "hyperbolic" or "rho-rho". NNSS and GPS also very useful.

d. Only satellite methods will satisfy the required accuracy. A measuring time of several hours is allowed.

Exercises signals and systems.

5. Lane slips.

A hyperbolic system works on a continuous wave with a frequency f_1 = 2 MHz. To find the integer lane number one uses a second the frequency f_2. Which would be a useful frequency for f_2:
3 MHz, 2.2 MHz, 2.02 MHz or 2.002 MHz? Why? Depending on what?

Solution.

2.2 MHz is quite reasonable because the accuracy of the phase measurement is normally much better than 10% (0.1 lane).
2.02 MHz would be reasonable if the phase measurement is a lot better than 0.01 lane.

2. SURVEY COMPUTATIONS.

G. Bakker

2.0 Introduction.

One of the main purposes of surveying is the determination of the relative position of points on the earth surface. That purpose is attained once the coordinates of these points are known in a three-dimensional (3d.) coordinatesystem that is firmly anchored to the earth. Such a coordinatesystem is often denoted by the word datum.
In our 3d. so-called Euclidean space E^3 it is possible to introduce the most simple coordinatesystem, the rectilinear cartesian coordinates $\{x,y,z\}$, but curvilinear coordinatesystems, for instance the geodetic coordinates $\{\phi,\lambda,h\}$, are also feasible.
To find the coordinates of a point set $\{P,Q,...\}$, one has to formulate the relations between these coordinates $\{x^j \mid j = 1,2,3\}$ and the geodetic observables l^i, $\{l^i \mid i = 1,...,n\}$ symbolically denoted by $l^i = f^i\{(x^j)P,Q,...)\}$.

These relations are obtained with the tools of a branch of mathematics, called analytical geometry.

Gradually, 3d. computations have come more and more into use, especially for large networks and when large computers are available. At the same time this latter constraint clearly indicates the main drawback of 3d. methods: although the mathematics is rather simple, the number of relations is such that only large computers can handle these methods. Happily however most surveying work is of limited extent and occurs at or near the earth surface that may be conceived as predominantly two-dimensional.
Three-dimensional methods may therefore safely be replaced by computation on a curved surface, viz. a sphere or an ellipsoid. Such a surface is mathematically called a Riemann space and indicated by R^2. The computations in such a space are more difficult than in an E^3 since it is impossible to use cartesian coordinates.

Choosing the ellipsoid as computational surface, this surface is supposed to fit as close as possible the equipotential surface of the earth gravity field that coincides with mean sealevel.
Although in hydrographical survey, where the third coordinate, the height h, plays hardly any role and therefore 2d. methods are explicitly in use, this chapter dealing with survey computations starts with a section on coordinatesystems in 3d. space.

This has two reasons,

first because ellipsoidal computations are better understood if the coordinates $\{\phi, \lambda\}$ on the ellipsoid are looked upon as a subset of the 3d. geodetic coordinates $\{\phi, \lambda, h\}$ and

second because the important datumtransformations can only be explained in the 3d. context.

Section 1 ends with the reductions that should be applied to the actual observations in order to fit them into the ellipsoidal model. The reduction formulas are given without any derivation. They may serve only as an easy reference. For most hydrographic work they will rarely be applied.

When planar coordinates in a given map projection are the ultimate purpose, one can once again proceed in two ways:

either the ellipsoidal observations are transferred to ellipsoidal coordinates $\{\phi, \lambda\}$ by means of ellipsoidal computations, where after the mapping formulas are applied to obtain the gridcoordinates $\{Y,X\}$,

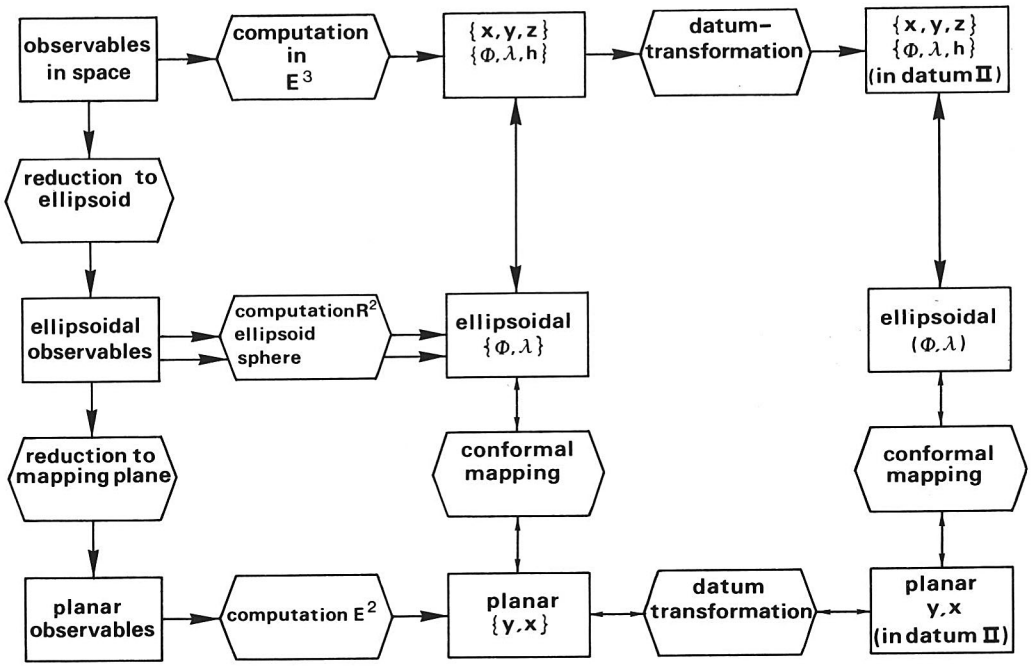
or the ellipsoidal observations are provided with reductions induced by the map projection in question, where after the computation of the $\{Y,X\}$ is carried out with the simple formulas of the plane.

For networks of limited extent, the latter strategy is often pursued.

In section 2 the conformal map projections are briefly outlined with emphasis upon the conformal projections of the ellipsoid onto the sphere. The reductions to the observations, induced by this projection can be neglected for nearly all practical problems to the effect that difficult computations on the ellipsoid may be replaced by computations on the conformal sphere using the simple rules of spherical trigonometry.

As an example both Bowrings method for the geodesic on the ellipsoid and Ballarins famous method for the computation of a hyperbolic pattern are dealt with, using spherical trigonometry.

In section 3 finally, the computations on the ellipsoid are dealt with. Basic to these computations are the so-called direct and inverse problem for the geodesic. Algorithms are given for small and large scale computers. In combination with the linearized equations for the observables, derived in section 2, any geometric problem on the ellipsoid can be tackled. In the diagram on the next page it has finally been tried to indicate the frame in which this chapter has been set.


observables in space
computation in E^3
{x,y,z} {Φ,λ,h}
datum-transformation
{x,y,z} {Φ,λ,h} (in datum II)
reduction to ellipsoid
ellipsoidal observables
computation R^2 ellipsoid sphere
ellipsoidal {Φ,λ}
ellipsoidal (Φ,λ)
reduction to mapping plane
conformal mapping
conformal mapping
planar observables
computation E^2
planar {y,x}
datum transformation
planar y,x (in datum II)

2.1 Geodetic coordinates and reductions.

2.1.1 Introduction.

To locate points in 3-dimensional space, the most obvious way to do so is to introduce a rectangular (orthogonal) cartesian coordinate system (c.s.) {x,y,z}.
In the origin O of this c.s. three unit vectors along the coordinate-lines constitute its base. The coordinates (c.n.) themselves should be considered as numbers, quantities without dimension.
Coordinate-surfaces are defined as surfaces on which one of the coordinates assumes a constant value, whereas the remaining ones are varying. A surface parallel to the plane Oyz is indicated by {x = constant}, a.s.o.
In consequence there are three different families of parallel coordinate-surfaces in 3d. geometrical space.
Two surfaces of different families intersect in a straight line on which the remaining coordinate varies. These lines are generally called coordinate-lines and indicated by {x = variable},a.s.o.
Starting from the cartesian coordinates {x,y,z}, other c.s. can be derived at will by applying a so-called coordinate transformation. These transformations can be of linear or non-linear type, resulting in rectilinear or curvilinear c.s. respectively.
In this chapter attention is focussed on the special group of geodetic coordinates.
In figure 1 it is shown that a point is also determined if its distance r to the origin O is given together with the angles subtended in O by the coordinate surfaces {z = 0} and {y = 0}. These angles are denoted by ϕ and λ.
Closely related to the set $\{\phi,\lambda,r\}$ is the set $\{\phi,\lambda,h\}$. The variable polar radius is split up in a constant part of length a, the radius of a sphere, and a variable part of length h, the height above the sphere.

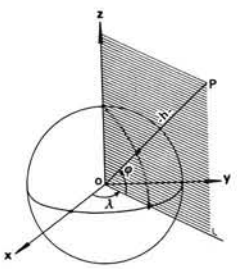


Figure 1

The transformation formulas from the set $\{\phi,\lambda,h\}$ to the set $\{x,y,z\}$ are:

$$\begin{aligned} x &= (a + h) \cos \phi \cos \lambda \\ y &= (a + h) \cos \phi \sin \lambda \\ z &= (a + h) \sin \phi \end{aligned} \qquad (1)$$

The inverse formulas are:

$$\begin{aligned} \phi &= \arccos(x^2 + y^2)^{\frac{1}{2}} / (x^2 + y^2 + z^2)^{\frac{1}{2}} \\ \lambda &= \arctan y/x \\ h &= (x^2 + y^2 + z^2)^{\frac{1}{2}} - a \end{aligned} \qquad (2)$$

The coordinate-surfaces $\{h = \text{constant}\}$ are concentric spheres. The coordinate-surface $\{h = 0\}$ is the sphere with radius a, that serves as reference surface for the height coordinate h. This latter surface intersects:

- on the one hand: the plane coordinate-surfaces $\{\lambda = \text{constant}\}$ in the coordinate-curves $\{\phi = \text{variable}\}$. These curves are called the meridians;
- on the other hand: the coordinate-surfaces $\{\phi = \text{constant}\}$ in the coordinate-curves $\{\lambda = \text{variable}\}$. These curves are called the parallels or latitude-circles.

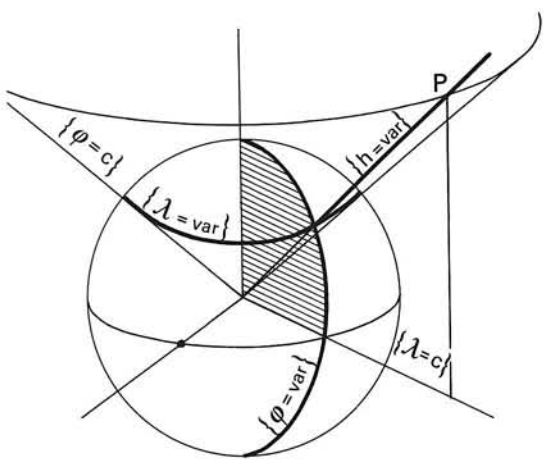


Figure 2.

The set $\{\phi,\lambda,h\}$ is part of a much larger group, the so-called geodetic coordinates. In this group the surface $\{h = \text{constant}\}$ is no longer a sphere but in general an ellipsoid of rotation, generated by an ellipse rotating around its minor axis.

Prior to the derivation of the transformation formulas from $\{\phi,\lambda,h\}$ to $\{x,y,z\}$ and vice versa, some important properties of the ellipse will be derived in the following section. The formulas of this section are also applied in section 2.3, dealing with ellipsoidal computations.

2.1.2. The geometry of an ellipse.

The equation of a circle in implicit form is:

circle: $x^2 + y^2 - a^2 = 0$.

The equation in the so-called parameter form is (see fig. 3):

circle: $\begin{cases} x = a \cos \beta \\ y = a \sin \beta \end{cases}$

The parameter is β with $0 \leq \beta \leq \pi$

An ellipse can be conceived to be generated from a circle by reducing the y-coordinate with a factor b/a, thus:

ellipse: $\boxed{\begin{cases} x = a \cos \beta \\ y = b \sin \beta \end{cases}}$ (3)

The implicit equation is found by eliminating β:

ellipse: $x^2/a^2 + y^2/b^2 - 1 = 0$. (4)

The parameters a and b respresent half the length of the major and minor axis of the ellipse. There are many parameters to characterize the flattening of an ellipse:

$e^2 = (a^2 - b^2)/a^2$ (e is called the first eccentricity)
$e'^2 = (a^2 - b^2)/b^2$ (e' is called the second eccentricity)
$f = (a - b)/a$ (f is called the flattening).

For other parameters and their relations see table 1 on page 51

The parameter β is called the reduced latitude. Another parameter which is sometimes used for the parametrization of an ellipse is the geocentric latitude ψ. In most cases however the parameter ϕ is employed for its parametrization. This parameter can be represented geometrically by the angle between the perpendicular to the ellipse and the x-axis, and is generally called the geographic latitude. The parameter equations of the ellipse in this parameter will now be derived.

For an ellipse the following properties hold:

- the tangents UV and QV intersect in a point of its major axis;
- UR : QR = a : b.

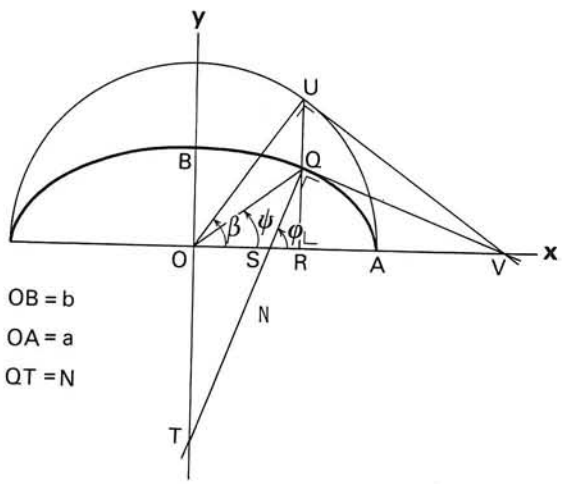


Figure 3.

With the aid of figure 3 the following formulas can be derived:

$$\cot g\,(\pi/2 - \beta) = \tan\beta = VR/UR$$
$$\cot g\,(\pi/2 - \phi) = \tan\phi = VR/QR$$

and thus: $\boxed{\tan\beta = \frac{b}{a}\tan\phi}$ (5)

and also $\boxed{\tan\psi = \frac{b}{a}\tan\beta}$

After defining two auxiliary latitude functions

$$\boxed{\begin{aligned} W &= (1 - e^2 \sin^2 \phi)^{\frac{1}{2}} \\ V &= (1 + e'^2 \cos^2 \phi)^{\frac{1}{2}}, \end{aligned}}$$

it follows from (5):

$$\boxed{\begin{aligned} \sin \beta &= V^{-1} \sin \phi = \frac{b}{a} W^{-1} \sin \phi \\ \cos \beta &= W^{-1} \cos \phi = \frac{a}{b} V^{-1} \cos \phi \end{aligned}} \qquad \begin{matrix} (6) \\ (7) \end{matrix}$$

After denoting the distance QT by N the following formulas can be derived:

$$\begin{aligned} QT &= a \cos \beta / \cos \phi = aW^{-1} \equiv N \\ OR &= a \cos \beta = N \cos \phi \\ QR &= b \sin \beta \\ QS &= b \sin \beta / \sin \phi = (1 - e^2)N \\ QR &= N(1-e^2) \sin \phi \end{aligned} \qquad (8)$$

Thus for point Q the following coordinates are found:

$$\boxed{\begin{aligned} x &= N \cos \phi \\ y &= N(1 - e^2) \sin \phi \end{aligned}} \qquad (9)$$

These formulas are the parameter-equations of the ellipse in the parameter ϕ, with $0 \leq \phi \leq \pi$, and N is a function of ϕ.

The meridian arc length.

To end this section an expression will be derived for the length s along the ellipse between two points with parameter value $\phi = 0$ and $\phi = \phi$. (This length is equivalent with the meridian arc length between these points).

To that purpose it is necessary to know the radius of curvature at an arbitrary point of the ellipse. This radius is usually denoted by M and may be conceived as the radius of the circle that is osculating the ellipse (a three-point contact). From differential calculus the following general formula is known.

Let $x = x(t)$, $y = y(t)$ be the parameter equations of a plane curve, then the radius of curvature is

$$r = \frac{(\dot{x}^2 + \dot{y}^2)^{3/2}}{\dot{x}\ddot{y} - \dot{y}\ddot{x}}$$

The dots mean one- or twofold differentiation with respect to t.
Applying this formula to (9) one finds

$$M = a(1 - e^2)W^{-3} \tag{10a}$$
$$N = aW^{-1} \tag{10b}$$

In the geometry of an ellipsoid, M and N is the standard notation for the so-called first and second principal radius of curvature, being defined as the radii of the oscillating circle of the normal section respectively in the direction of and perpendicular to the meridional plane.
For the meridian arc length s one gets:

$$s = \int_0^\phi M \, d\phi = a(1 - e^2) \int_0^\phi W^{-3} \, d\phi \tag{11}$$

The integral is of the elliptic type, which can not be solved analytically. Therefore the integrand is developed in a Taylor series.
As $W^{-3} = (1 - e^2 \sin^2 \phi)^{-3/2}$ is of the binomial type, reference is made to the following general expressions:

$$(1 \pm x)^n = 1 \pm nx + \frac{n(n-1)}{2!} x^2 \pm \frac{n(n-1)(n-2)}{3!} x^3 + \ldots$$

$$(1 \pm x)^{-n} = 1 \mp nx + \frac{n(n+1)}{2!} x^2 \mp \frac{n(n+1)(n+2)}{3!} x^3 + \ldots$$

The first expansion can be written as $\Sigma(\pm 1)^n \binom{n}{k} x^n$ (k = 0,1,...).
So one finds

$$W^{-3} = 1 + \frac{3}{2} e^2 \sin^2 \phi + \frac{3.5}{2.4} e^4 \sin^4 \phi + \frac{3.5.7}{2.4.6} e^6 \sin^6 \phi + \ldots.$$

By introducing the following notation for the so-called Wallis integrals (attention: both the Wallis integral and the latitude function are denoted by W!):

$$W_{2p} = \int_0^\phi \sin^{2p} \phi \, d\phi \quad (p = 0,1,2,\ldots) \tag{12}$$

one gets

$$\int_0^\phi W^{-3} d\phi = W_0 + \frac{3}{2} e^2 W_2 + \frac{3.5}{2.4} e^4 W_4 + \frac{3.5.7}{2.4.6} e^6 W_6 + \ldots \tag{13}$$

There are two main methods to arrive at a numerical value for this integral.

1. The n^{th} powers of the sinus function of the angle are replaced by equivalent cosinus functions of the multiple angle which can be easily integrated.
2. By applying the method of partial integration the following recurrent relation for the Wallis integrals is obtained:

$$\boxed{W_{2p} = \frac{2p-1}{2p} W_{2p-2} - \frac{1}{2p} \sin^{2p-1}\phi \cos\phi} \qquad (14)$$

Proof of (14) : By partial integration one finds:

$$\int \sin^n\phi \, d\phi = -\sin^{n-1}\phi \cos\phi + (n-1) \int \cos^2\phi \sin^{n-2}\phi \, d\phi =$$
$$-\sin^{n-1}\phi \cos\phi - (n-1) \int \sin^n\phi \, d\phi + (n-1) \int \sin^{n-2}\phi \, d\phi$$

Hence : $$\int_o^\phi \sin^n\phi \, d\phi = \frac{-\sin^{n-1}\phi \cos\phi}{n} + \frac{n-1}{n} \int_o^\phi \sin^{n-2}\phi \, d\phi$$

By substituting n = 2p one finds (14)

By repeated application one arrives at :

for n = even : $\int_o^\phi \sin^0\phi \, d\phi = \phi$

for n = odd : $\int_o^\phi \sin^1\phi \, d\phi = 1 - \cos\phi$.

The recurrent relation (14) enables the developement of a consize computerprogram for the evaluation of (11) (see also page 114)

The inverse problem, being the computation of ϕ when the meridian arc length s is given, can be solved by an iterative process.

Eqs. (11) and (13) are written as follows:

$$s/(a(1-e^2)) = W_0 + u \qquad \text{and thus:}$$
$$\phi = s/(a(1-e^2)) - u \quad \text{with } W_0 = \phi \text{ and}$$
$$u \ll W_0$$

$$u = \frac{3}{2} e^2 W_2 + \frac{3.5}{2.4} e^4 W_4 + \frac{3.5.7}{2.4.6} e^6 W_6 + \ldots.$$

Using as first approximation $u^{(1)} = 0$ we find $\phi^{(1)}$.
This value $\phi^{(1)}$ is used to evaluate $u^{(2)}$ and then a second, more close approximation $\phi^{(2)}$ is found a.s.o..

Many different parameters and latitude functions, most of them of equal benefit, are being employed for the description of the geometry of the ellipse and the ellipsoid. For easy reference they are summarized in the tables 1 and 2 on the next page. In the pages to follow, these tables are frequently referred to.

a and	b	c	e^2	e'^2	f
b =	b	$\frac{a^2}{c}$	$a(1-e^2)^{\frac{1}{2}}$	$a(1+e'^2)^{-\frac{1}{2}}$	$a(1-f)$
c =	$\frac{a^2}{b}$	c	$a(1-e^2)^{-\frac{1}{2}}$	$a(1+e'^2)^{\frac{1}{2}}$	$\frac{a}{1-f}$
e^2 =	$\frac{a^2-b^2}{a^2}$	$\frac{c^2-a^2}{c^2}$	e^2	$\frac{e'^2}{1+e'^2}$	$f(2-f)$
e'^2 =	$\frac{a^2-b^2}{b^2}$	$\frac{c^2-a^2}{a^2}$	$\frac{e^2}{1-e^2}$	e'^2	$\frac{f(2-f)}{(1-f)^2}$
f =	$\frac{a-b}{a}$	$\frac{c-a}{c}$	$1-(1-e^2)^{\frac{1}{2}}$	$1-(1+e'^2)^{-\frac{1}{2}}$	f

Table 1 DIMENSION PARAMETERS

	W	V	w	v
W =	$(1-e^2\sin^2\phi)^{\frac{1}{2}}$	$\frac{b}{a}$ V	$\frac{b}{a}$ w^{-1}	v^{-1}
V =	$\frac{a}{b}$ W	$(1+e'^2\cos^2\phi)^{\frac{1}{2}}$	w^{-1}	$\frac{a}{b}$ v^{-1}
w =	$\frac{b}{a}$ W^{-1}	V^{-1}	$(1-e^2\cos^2\beta)^{\frac{1}{2}}$	$\frac{b}{a}$ v
v =	W^{-1}	$\frac{a}{b}$ V^{-1}	$\frac{a}{b}$ w	$(1+e'^2\sin^2\beta)^{\frac{1}{2}}$
$\frac{d\phi}{d\beta}$ =	$\frac{a}{b}$ W^2	$\frac{b}{a}$ V^2	$\frac{b}{a}$ w^{-2}	$\frac{a}{b}$ v^{-2}
M =	$\frac{b^2}{a}$ W^{-3}	c V^{-3}	c w^3	$\frac{b^2}{a}$ v^3
N =	a W^{-1}	c V^{-1}	c w	a v

Table 2 LATITUDE FUNCTIONS

2.1.3 Coordinatetransformation.

Using (9) one can easily find the relation between cartesian c.n. { x,y,z } and geodetic c.n. {ϕ,λ,h }. Let in (9) y be replaced by z and N by (N+h) and let the angle of rotation around the z-axis be denoted by λ, then one finds successively (see figure 4):

For point Q: $x = OR\cos\lambda = N\cos\phi\cos\lambda$

$y = OR\sin\lambda = N\cos\phi\sin\lambda$

$z = QS\sin\phi = N(1-e^2)\sin\phi$

For point P:

$$x = (N+h)\cos\phi\cos\lambda \qquad (15a)$$

$$y = (N+h)\cos\phi\sin\lambda \qquad (15b)$$

$$z = (N(1-e^2)+h)\sin\phi \qquad (15c)$$

For N see table 2 on page 51

By substituting $e^2 = 0$, it can easily be demonstrated that the formulas (1) are a special case of (15).

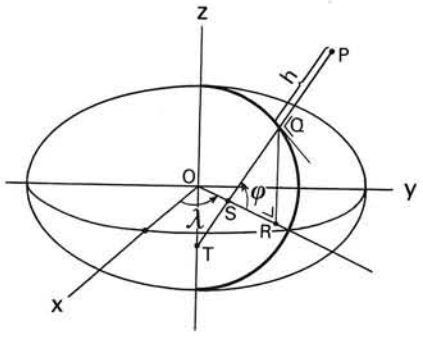


Figure 4.

Although they are by far not as simple as the cartesian c.n., the geodetic c.n. $\{\phi,\lambda,h\}$ are almost exclusively used in geodesy. This stems from the fact that geodetic c.n. better meet the distinction people are used to make between 'situation' and 'height or 'horizontal'and 'vertical'. For the geodesist this means that geodetic c.n. better allow for a splitting up of 3-dimensional geodetic problems, at or near the earth's surface, into a 2-dimensional problem on a slightly curved ellipsoidal surface and a separate 1-dimensional height problem.
The parameters of the ellipsoid are chosen in such a way that the ellipsoid is a good fit to the local, regional or global geoid, the equipotential-surface of the earth's gravity field that coincides with mean sea level.

In table 3 the dimensions of some, relatively well-known ellipsoids are given.

Ellipsoid Name (Year computed)	Semi-Major Axis (a) (m)	Inverse Flattening 1/f
Airy (1830)	6378563.396	299.324964
Bessel (1841)	6377397.155	299.152813
Clarke 1866	6378206.4	294.978698
Clarke 1880 (modified)	6378249.145	293.4663
Clarke 1880	6378249.145	293.465
Everest (1830)	6377276.345	300.8017
International (1909)	6378388	297.00
Hayford	6378388	297.00
Krassovski (1940)	6378378	297.00
Geodetic Reference System 1967	6378160	298.2471674273
IAG (1975)	6378140	298.257222101
Geodetic Reference System 1980	6378137	298.257222

Table 3. Ellipsoid parameters.

2.1.3.1. The inverse relations.

The equations (15) are of the non-linear type and direct inverse formulas for the computation of $\{\phi,\lambda,h\}$ from $\{x,y,z\}$ can only be obtained for the spherical case cf. (2). Although there are general algorithms for the solution of n non-linear equations in n unknowns, we shall give an algorithm that is specially taylored for the equations (15).

The coordinate λ can be found in a direct way by dividing (15a) and (15b):

$$\lambda = \text{atan}(y/x) \qquad (16)$$

For the coordinate ϕ a non-linear equation can be formulated.
From (15a) and (15b):

$$r = (N + h)\cos\phi \qquad (17a)$$

with

$$r^2 = x^2 + y^2 \quad .$$

From (15c):

$$z + Ne^2 \sin\phi = (N + h)\sin\phi \quad . \qquad (17b)$$

Eliminating h from (17a) and (17b) gives:

$$\phi = \text{atan}\left(\frac{z}{r} + \frac{Ne^2 \sin\phi}{r}\right) \quad . \qquad (18)$$

As the second term on the right hand side is much smaller than the first one, the following iterative solution for ϕ is possible.
As a first approximation take for instance $\phi^{(1)} = 0$. From the general iterative procedure

$$\varphi^{(k+1)} = \text{atan}\left(\frac{z}{r} + \frac{N^{(k)} e^2 \sin \phi^{(k)}}{r}\right)$$

the second approximation $\phi^{(2)}$ can be found and so on. The iteration is terminated when the absolute value of the difference between two consequtive approximations is less than a prescribed tolerance ε
Customary values for ε are:

on sea $\varepsilon = 10^{-7}$ rad (corresponding with 60 cm in position)

on land $\varepsilon = 10^{-9}$ rad (corresponding with 6 mm in position).

Usually 3 or 4 iterations are sufficient.
After ϕ is found, the height h can be obtained with (17a) or (17b):

$$\text{if } \phi < 45^0: \qquad h = \frac{r}{\cos \phi} - N$$

$$\text{if } \phi \geq 45^0: \qquad h = \frac{z}{\sin \phi} - N(1 - e^2) \quad . \tag{19}$$

2.1.4 Datum transformation.

In the preceding sections geodetic c.n. have been introduced in a formal mathematical way, without paying any attention to its application in surveying and navigational problems. Which problems can arise in daily practice? First, the hydrographic surveyor is confronted with geodetic c.n., based on different ellipsoidal parameters. As a matter of fact these differences cause only minor inconveniences that can easily be overcome. Second, what is worse and less easy to evaluate, is that various sets of geodetic c.n. also differ with respect to the location and scale of the bases of the corresponding cartesian c.n.

The determination of such a base with regard to fundamental earth's points and directions (and their underlying observations) is called the determination of the datum. In surveying practice the relative position of the various datums is of utmost importance. National and international scientific geodetic institutes, concerned with the connections of datums, regularly publish new updatings based on recent additional observations and adjustments.

Although the determination of the datum transformation parameters is not his immediate concern, it seems worthwhile that a hydrographic surveyor has some knowledge of how these parameters are obtained.

How to obtain the transformation parameters.

Let for both datums the two sets of cartesian c.n. be denoted by $\mathbf{x}_I$ and $\mathbf{x}_{II}$, with $\mathbf{x} = \{x,y,z\}$. Let the c.n. of a common point k be denoted by $\underset{(k)}{\mathbf{x}}$ $(k = 1,..,n)$ and the centre of mass (c) of the point set by $\underset{(c)}{\mathbf{x}}$.

For vectors letters in bold face type are used.

The relation between both c.n. sets is given by a 3d. similarity transformation

$$\mathbf{x}_{II} = \lambda A \mathbf{x}_I + \mathbf{t}$$

With	$\mathbf{t}$	is translation vector	(3 parameters)
	A	is rotation matrix	(3 parameters)
	λ	is scale factor	(1 parameter)

The estimation of the unknown 7 parameters narrows down to an adjustment problem with non-linear equations. There are a number of methods of solutions. The following one lends itself best to a geometric interpretation.

The set of n points is considered as a distribution of mass points with unit mass. In the theory of the dynamics of rigid bodies, moving under the influence of central and non-central forces, the following characteristic quantities of the body are introduced:

- the centre of mass;
- the set of body-axes, three mutual orthogonal unit vectors in the centre of mass in the direction of the axes of main inertia;
- the radius of gyration, the root mean square distance from the points of the body to its centre of mass.

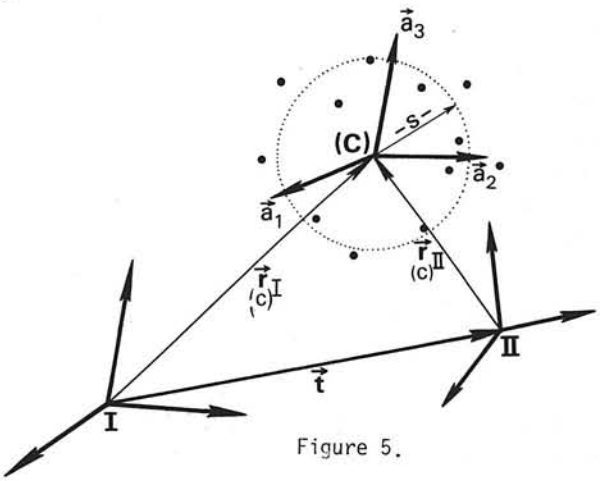


Figure 5.

The underlined quantities are often called intrinsic, since they belong to the object itself and bear no relation to whatever c.s.

If we can now dispose of two sets of c.n. for the point set, two estimations of the above quantities can be evaluated.

A comparison between both sets of estimations gives the relation between the bases of the two c.s.

The computation proceeds along the following steps.

1. The centre of mass is computed by

$$x_{(c)} = \frac{1}{n} \sum_{k=1}^{n} x_{(k)} \quad , \quad y_{(c)} = \frac{1}{n} \sum_{k=1}^{n} y_{(k)} \quad , \quad z_{(c)} = \frac{1}{n} \sum_{k=1}^{n} z_{(k)}$$

In shorthand vector notation:

$$\mathbf{x}_{(c)} = \{ x_{(c)}, y_{(c)}, z_{(c)} \}.$$

These c.n. are subtracted from the original c.n.

Without changing the notation: $\mathbf{x}_{(k)} - \mathbf{x}_{(c)} \longrightarrow \mathbf{x}_{(k)}$

2. The symmetric inertia tensor I can be formed by computing the sum of products and cross products of the c.n. of the points:

$$I = \begin{pmatrix} \Sigma xx & \Sigma xy & \Sigma xz \\ \Sigma yx & \Sigma yy & \Sigma yz \\ \Sigma zx & \Sigma zy & \Sigma zz \end{pmatrix}$$

The normalized eigenvectors of I represent the body-axes $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$. They can be found by applying any simple eigenvalue routine.

In short hand matrix notation: $R = \{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$.

remark : In the 2-dimensional case is, if α is the angle of $\mathbf{a}_1$ with the x-axis :

$$\tan 2\alpha = \frac{2\,\Sigma xy}{\Sigma xx - \Sigma yy} \quad \text{and} \quad \mathbf{a}_1 \text{ orthogonal to } \mathbf{a}_2 .$$

3. The radius of gyration is obtained by

$$s = \frac{1}{n} \sqrt{\sum_{k=1}^{n} \left(x_{(k)}^2 + y_{(k)}^2 + z_{(k)}^2 \right)}$$

The summations in the above formulas are from k = 1 until n.

The above computations can be made for both c.s. $\mathbf{x}_I$ and $\mathbf{x}_{II}$ and the relation between both c.s. is determined by

$$\lambda = s_{II}/s_I ,$$

$$A = R_{II} \cdot R_I^{-1}$$

$$\mathbf{t} = \mathbf{x}_{(c)II} - \lambda A\, \mathbf{x}_{(c)I}$$

In most cases only the translational vector **t** assumes a significant value, whereas scale and orientation differences are usually small and may safely be neglected. Hence only step 1 has to be carried out.

Example.

We like to compute the transformation parameters between two different coordinate systems : xa and xb. The coordinates of four points are given in both systems:

number of points : 4

coordinates in the two systems :

xa	ya	za	xb	yb	zb
585.435	755.475	102.520	929.590	422.800	-0.210
553.175	988.105	104.190	575.360	480.900	2.370
424.045	785.635	106.125	812.370	200.820	-0.240
394.950	1061.700	106.070	396.280	283.240	0.410

In both systems we compute successively the coordinates of the center of mass, the relative coordinates of the four points with respect to the center of mass, the radius of gyration, the inertia tensor, the matrix of eigenvectors and the eigenvalues:

coordinates of the center of mass :

a-system	489.401	897.729	104.726
b-system	678.398	346.940	0.583

coordinates relative to center of mass :

xa	ya	za	xb	yb	zb
96.034	-142.254	-2.206	251.182	75.860	-0.793
63.774	90.376	-0.536	-103.038	133.960	1.788
-65.356	-112.094	1.399	133.972	-146.120	-0.823
-94.451	163.971	1.344	-282.118	-63.700	-0.173

elements of the inertia tensor :

a-system			o-system		
0.6786E+05	0.1606E+05	0.4644E+03	0.4911E+05	-.3647E+04	0.4448E+03
0.1606E+05	0.2649E+05	-.3289E+03	-.3647E+04	0.1713E+06	-.3105E+03
0.4644E+03	-.3289E+03	0.9434E+05	0.4448E+03	-.3105E+03	0.2204E+06

matrix of eigen vectors :

a-system			b-system		
-.1704E-01	0.9459E+00	0.3241E+00	0.2736E-02	-.2979E-01	0.9996E+00
0.8138E-03	0.3242E+00	-.9460E+00	-.6526E-02	0.9995E+00	0.2981E-01
-.9999E+00	-.1586E-01	-.6293E-02	0.1000E+01	0.6597E-02	-.2543E-02

eigenvalues :

a-system : 0.9435E+05 0.7336E+05 0.2099E+05
b-system : 0.2204E+06 0.1714E+06 0.4900E+05

Now we are able to compute the transformation parameters :

scale factor : 0.15282338772767E+01

elements rotation matrix :

0.29584933298135E+00	-.95522964261692E+00	-.30819423370025E-02
0.95497048760243E+00	0.29584193204196E+00	-.22560923510770E-01
0.22456381700247E-01	0.37309410489324E-02	0.99974082562774E+00

translation vector:

.17681365144645E+04 -.76956755502304E+03 -.18133641591091E+03

After world war II it was decided, especially as a result of bad military experiences, to come to an integration of the various national c.s. A combined readjustment of several networks in Europe led to European Datum 1950, briefly ED50, and nowadays ED80 is on its way.
Other continental datums are the North American Datum (NAD '27), the South American, the Asian and the African Datum.
In the Netherlands only two datums are involved in coordinate computation: the national datum, to which the R.D. coordinates refer and ED '50 on which U.T.M. coordinates are based.

Geodesists have always strived after the ideal of one common datum for the whole earth. Such a world datum became reality when satellite-methods entered the domain of geodesy.
A well-known datum is WGS'72 (World Geodetic System 1972). Almost every year new world datums are published, based on the ever-increasing stream of new data. Recent updatings, however, change the coordinates hardly anymore.
A characteristic feature of a world datum is that its origin is supposed to be located in the centre of mass of the earth, where as its z-axis is supposed to coincide with the earth's rotational axis. In table 4 some datums are compared with ED'50.

DATUM	a	$e^2.10^9$	ED50 - D				
			Δx	Δy	Δz	Δa	$\Delta e^2.10^9$
HOLLAND Bessel	6377397.155	6674372	+676.52	+121.57	+595.27	+990.845	+48298
BELGIUM Hayford	6378388	6722670	+165	+215	+246	0	0
ENGLAND OSGB	6377563.396	6670540	+452	-17	+552	+824.604	+52130
WGS '72	6378135	6694318	+84	+103	+127	+253	+28352
NWGL 10	6378135	6694318	+83	+110	+118	+253	+28352
ED50 Hayford	6378388	6722670	0	0	0	0	0

Table 4 SHIFTPARAMETERS

This table should be handled with great care since updating is continuously taking place. It is advised to contact the national institutes concerned with these updatings.

For the transformation from ED '50 to WGS '84 see Appendix 4 on page 133.

In table 4 the translation from a specific datum D to ED'50 is specified by the set of translation (or shift) parameters: $\mathbf{t} \equiv \Delta\mathbf{x} = \{\Delta x, \Delta y, \Delta x\}$.
These parameters may be considered as the coordinates of the origin of datum D in the c.s. ED'50: $x_{ED} = x_D + \Delta x$.

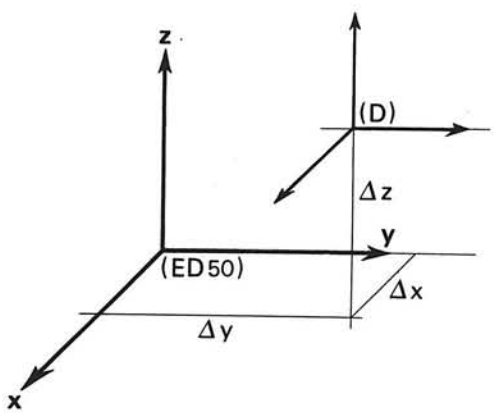
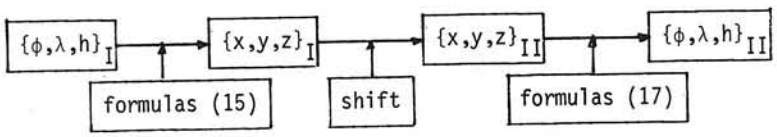


Figure 6.

How to perform a datum transformation.

1. The transformation of geodetic c.n. $\{\phi,\lambda,h\}$ from datum I to datum II can be represented by the following scheme:



2. As a matter of fact the differences between both sets are small and therefore the transformation can also be carried out by the linearized equations, obtained after differentiation of the above chain of relations:

$$\begin{pmatrix} a.\Delta\phi \\ a\cos\phi.\Delta\lambda \\ \Delta h \end{pmatrix} = \begin{pmatrix} -\sin\phi\cos\lambda & -\sin\phi\sin\lambda & \cos\phi \\ -\sin\lambda & \cos\lambda & 0 \\ \cos\phi\cos\lambda & \cos\phi\sin\lambda & \sin\phi \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix} +$$

$$+ \begin{pmatrix} e^2\sin\phi\cos\phi & a\sin\phi\cos\phi \\ 0 & 0 \\ \frac{1}{2}e^2\sin^2\phi - 1 & \frac{1}{2}a\sin^2\phi \end{pmatrix} \begin{pmatrix} \Delta a \\ \Delta e^2 \end{pmatrix}$$

Remark:

Some institutes issue additional small-scale contourmaps representing $\Delta\phi$- and $\Delta\lambda$-isocorrectioncurves. When they are available and when the utmost accuracy is required, these maps can be used to assess the final additional corrections to the coordinates.

2.1.5. Reduction of observations onto the ellipsoid.

Horizontal directions and angles.

Assume that directions have been observed on the surface of the earth with a theodolite. In order to use these directions in computations on the ellipsoid they have to be reduced first. Since the vertical axis of the theodolite is aligned with the direction of the plumb line at the station the directions are measured in the local horizon, i.e. the plane orthogonal to the plumb line. The first correction, therefore, has to be such that the corrected values are those which would have been obtained if the vertical axis had been aligned with the ellipsoidal surface normal.

Assume that you are given a set of deflections of the vertical $\{\xi , \eta\}$ which refer to the ellipsoid to which you should reduce the observations. Let A be the astronomical azimut of the target point, referred to C.I.O. (Conventional International Origin), α the corresponding azimuth of the ellipsoidal normal section, then $\alpha = A + \delta_1$ with

$$\delta_1 = -\eta \tan \phi - (\xi \sin \alpha - \eta \cos \alpha)\cot z$$

where z is the zenith distance of the observed direction.
The correction term $\eta \tan \phi$ is common to all directions of the same station whereas the other correction term is a function of the azimuth. In fact, for the reduction of angles the second correction term will be the only term necessary to be considered since angles are the difference of directions and the $\eta \tan \phi$ -term thus cancels.

There are two reductions left: the reduction δ_2 taking into account the projection of the target point onto the ellipsoid, and the reduction δ_3, representing the transfer from normal section to geodesic on the ellipsoid

$$\delta_2 = 0.''108 \cos^2 \phi_1 \sin 2\alpha_1 \cdot h_{km}$$

$$\delta_3 = -0.''028 \cos^2 \phi_1 \sin 2\alpha_1 \frac{S^2_{km}}{100}$$

Summarizing:

$$\alpha_{geodesic} = A_{observed} + \delta_1 + \delta_2 + \delta_3$$

Distances.

Most of the modern geodetic measurement techniques primarily measure the spatial distance between points. Figure 7 shows the plane containing the stations P_1 and P_2, and the geocenter 0.

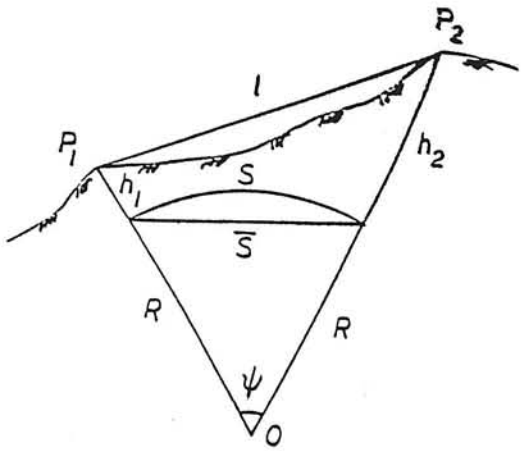


figure 7

The ellipsoidal arc is approximated by a circle whose radius R is equal to the radius of curvature of the ellipsoid at the midpoint and in the direction of pointing. The law of cosines applied to the triangle OP_1P_2 gives

$$l^2 = (R + h_1)^2 + (R + h_2)^2 - 2(R + h_1)(R + h_2)\cos\,\psi$$

Since

$$\cos\,\psi = 1 - 2\,\sin^2\,\frac{\psi}{2}$$

this expression becomes:

$$l^2 = (h_2 - h_1)^2 + 4R^2(1 + \frac{h_1}{R})(1 + \frac{h_2}{R})\sin^2\,\frac{\psi}{2}$$

With the chord length

$$\overline{S} = 2R\,\sin\,\frac{\psi}{2}$$

and the ellipsoidal height difference

$$\Delta h = h_2 - h_1$$

we obtain

$$l^2 = \Delta h^2 + (1 + \frac{h_1}{R})(1 + \frac{h_2}{R})\bar{S}^2$$

and

$$\bar{S} = \left(\frac{l^2 - \Delta h^2}{(1 + \frac{h_1}{R})(1 + \frac{h_2}{R})} \right)^{\frac{1}{2}}$$

The ellipsoidal distance is

$$S = R\psi = 2R \arcsin \frac{\bar{S}}{2R}$$

A more accurate expression for the $S - \bar{S}$ correction is

$$\bar{S} = S[1 - \frac{1}{24}(1 + 2\eta_1^2 \cos^2\alpha_1)\frac{S^2}{N_1^2} + \frac{1}{8} t_1 \eta_1^2 \cos\alpha_1 \frac{S^3}{N_1^3} + \frac{1}{1920}\frac{S^4}{N_1^4}]$$

The above correction formulae are accurate enough to be useful in ordinary triangulation/trilateration network computations.

2.2 Conformal mapping and spherical computations.

2.2.1. Introduction.

Map projections may be looked upon either from a cartographic or a geodetic point of view.

In cartography a small scale map is, in most cases, the ultimate output.

As a consequence the required accuracy of the mapping formulae is low and a sphere may safely be introduced as the surface to be mapped.

In geodesy a map projection is conceived as a mathematical device that transfers the set of geographical c.n. $\{\phi, \lambda\}$ into a set of cartesian planar coordinates $\{x,y\}$ without loss of information.

That latter requirement implies that an ellipsoid should be introduced as the surface to be mapped and that the set of geographical c.n. should refer to this surface.

With respect to the inevitable distortion of the geometric elements, distance and angle, map projections are usually subdivided into conformal, equivalent (or equal area) and conventional projections.

Conformal projections preserve the angle of intersection of any two curves. In fact it means that a conformal projection may be considered as a similarity mapping in an infinitesimal small region. It differs from a similarity mapping in the plane in that its scale is not constant but varying over the mapping area.

It is because of this property that only conformal projections are employed for geodetic purposes.

The family of conformal projections has an infinite number of members, however, only four of them, are explicitely in use: the mercator, the transverse mercator, the lambert conical projection and the stereographic projection.

Although in geodesy a map projection is merely looked upon as a numerical transformation of c.n., it is sometimes useful to have in mind the geometric concept of 'projection'.

With regard to this aspect, projections are divided into cylindrical, conical and azimutal projections, depending on whether the 'plane of projection' is a cylinder, cone or a plane. After the projection the cylinder and the cone can be spread out in a plane.

Loosely speaking, the mercator projection can geometrically be looked upon as a projection onto a cylinder tangent to the equator.

The Lambert conformal conical p. may be considered as a p. onto a conic tangent to one standard parallel or secant to two standard parallels.

In the same way the stereographic projection can be seen as a projection on a plane tangent to the pole.

Speaking from a geometric standpoint, any projection can be applied in the so-called normal, oblique and transverse position of the cylinder, cone or plane.

In the normal case, the axis of the cylinder and the cone or the normal to the plane, usually called the axis of projection, coincides with the minor axis of the ellipsoid.

In that case the mapping formulae $\{x,y\} \rightleftarrows \{\phi,\lambda\}$ are very simple, since the set $\{\phi, \lambda\}$ also refers to this minor axis.

In the oblique and transverse case however the minor axis and the axis of projection have different position and the transformation formulae become rather complicated. In the Netherlands for instance, the centre of the oblique stereographic projection is the O.L.V. church in Amersfoort.

In order to obtain the transformation formulae from ellipsoid to plane for the non-normal cases, a sphere is introduced as a intermediate surface.

The computational program is then divided into three modules:

1. the ellipsoid is conformally mapped upon this sphere, thus: $\{\phi\ \lambda\} \rightleftarrows \{\Phi,\Lambda\}$
2. the 'polar' geographical c.n. are transformed to 'local' geographical c.n., with the centre of projection as local pole, thus $\{\Phi,\Lambda\} \rightleftarrows \{\bar{\Phi},\bar{\Lambda}\}$. This transformation boils down to the solution of the 'inverse geodetic problem' for the sphere, cf. section 2.3 and 2.2.5.3;
3. finally the simple mapping formulae from sphere onto plane for the normal case are applied, thus $\{\bar{\Phi},\bar{\Lambda}\} \rightleftarrows \{x,y\}$

It is now obvious that the conformal projection of the ellipsoid onto the sphere, named after Gauss, plays an important role in the theory of map projections. The combination of the Gauss projection and the subsequent projection from sphere onto plane is called 'double projection'.

A not less important property of this projection is that the so-called 'conformal sphere' may beneficially serve as a substitute surface for computations on the ellipsoid. The induced distortion of geometric elements (lengths and angles) by the projection is small to such an extent that in many problems it may be neglected. That is why in many cases difficult elaboration of differential equations and elliptic integrals on the ellipsoid can be replaced by simple spherical computations. Bowring's method for the direct and inverse problem on the ellipsoid and Ballarin's method for the computation of ellipsoidal hyperbolic patterns are well-known examples.

All these things considered, this chapter is framed into the following sections.
In 2.2.2 the preliminary mathematics of conformal mapping is given.
In 2.2.3 the Gauss conformal p. from ellipsoid onto sphere is treated.
In 2.2.4 those computations on the conformal sphere that are relevant in surveying are dealt with. They are divided into direct algorithms (Bowring, Ballarin) using closed trigonometrical formulaes, and indirect algorithms using the linearized trigonometrical formulas.
In 2.2.5 the essential formulae for the 4 main conformal projections are summarized, without derivation. They only may serve as an easy reference to the surveyor.
In 2.2.6 the problem of transformation of map coordinates in between neighbouring countries is briefly discussed.

2.2.2 Preliminary mathematics.

In 2.1 the geographic, geocentric and the reduced latitude have been introduced as coordinates along the meridian. Now a fourth latitude will be defined: the isometric latitude, denoted by ω for the sphere and by q for the ellipsoid.

Let us start with the sphere with radius R. Let P and Q be two infinitesimal close points with c.n. $\{\phi, \Lambda\}$ and $\{\phi + d\phi, \Lambda + d\Lambda\}$.
The distance ds between both points is expressed by the so-called formula of the line element:

$$(ds)^2 = R^2(d\phi)^2 + R^2 \cos^2 \phi (d\Lambda)^2 \qquad (20)$$

For equal increments $d\phi$ and $d\Lambda$, the corresponding increments in length, $Rd\phi$ and $R \cos \phi \, d\Lambda$, are not equal: the c.n. $\{\phi,\Lambda\}$ are not isometric, in contrast to the cartesian planar c.n. $\{x,y\}$.

The isometry can be restored by introducing the 'isometric latitude' ω. This is done by the following expedient. In the following the squares of the differentials are written without ( ).

$$ds^2 = R^2 d\phi^2 + R^2 \cos^2 \phi \, d\Lambda^2$$

$$ds^2 = R^2 \cos^2 \phi \left(\frac{d\phi^2}{\cos^2 \phi} + d\Lambda^2\right)$$

$$\text{Let } d\omega = \frac{d\phi}{\cos \phi}$$

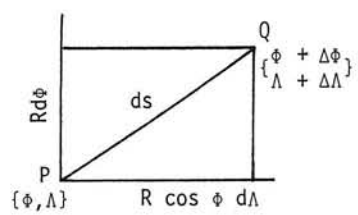


figure 8.

By integration we find:

$$\boxed{\omega = \ln \tan \left(\frac{\pi}{4} + \frac{\phi}{2}\right) = \operatorname{atanh} (\sin \phi)} \qquad (21)$$

which shows the relation between both latitudes.

We can now write:

$$ds^2 = e^2 \{d\omega^2 + d\Lambda^2\} \qquad \text{with } e = R \cos \phi$$

or $$ds^2 = d^{-2} \{d\omega^2 + d\Lambda^2\} \qquad \text{with } d = R^{-1} \cos^{-1} \phi$$

As the factor e before the brackets still depends on ϕ, the spacing of the coordinate curves (meridians and parallels) is varying over the surface. It can be said that a set of isometric c.n. on a surface covers this surface with a grid of infinitesimal small squares with varying mesh size.
e is called the (equal) unit of measure along the coordinate curves and d is called the density of the coordinate system: the larger d, the larger the number of squares per surface-unit.

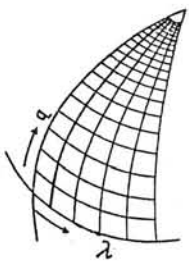


figure 9.

Warning: Do not mix up the unit of measure, the first eccentricity and the base of the natural logarithms which are all denoted by e.

For the cartesian c.n. $\{x,y\}$ in the plane: $d = 1$. For the c.n. $\{\omega, \Lambda\}$ on the sphere: $d = (R \cos \phi)^{-1}$.
For the ellipsoid an analogous reasoning can be pursued.
The formula of the line element is

$$ds^2 = M^2 d\phi^2 + N^2 \cos^2 \phi \, d\lambda^2 \quad \text{(see 2.4.2)}$$

$$ds^2 = N^2 \cos^2 \phi \left(\frac{M^2}{N^2 \cos^2 \phi} d\phi^2 + d\lambda^2\right)$$

Let $$dq = \frac{M}{N \cos \phi} d\phi = \frac{1 - e^2}{(1 - e^2 \sin^2 \phi)\cos \phi} d\phi \quad \text{(cf. table 1)}$$

In partial fractions:

$$dq = \frac{d\phi}{\cos \phi} - \frac{e(e \cos \phi)}{2(1 + e \sin \phi)} d\phi - \frac{e(e \cos \phi)}{2(1 - e \sin \phi)} d\phi$$

Integration gives

$$q = \ln \tan\left(\frac{\pi}{4} + \frac{\phi}{2}\right) - \tfrac{1}{2}e \ln(1 + e \sin\phi) + \tfrac{1}{2}e \ln(1 - e \sin\phi) \qquad (22)$$

or $\quad q = \operatorname{atanh}(\sin\phi) - e.\operatorname{atanh}(e.\sin\phi)$

The set $\{q, \lambda\}$ constitutes a set of isometric c.n. on the ellipsoid with density $d = (N \cos\phi)^{-1}$.

The use of isometric coordinates is fundamental for the design and the numerical elaboration of a conformal mapping. That is why the transfer from geographic to isometric latitude and vice versa is an essential subroutine in any software packet dealing with conformal mapping. Locally, in the small, both $\{\omega, \Lambda\}$ and $\{q, \lambda\}$ may be conceived as cartesian coordinates on the sphere resp. on the ellipsoid.

Let the general mapping formulas from the ellipsoid onto the plane be written as functions between isometric coordinates:

$$\begin{aligned} y &= y(q, \lambda) \\ x &= x(q, \lambda) \end{aligned} \qquad (23)$$

The question now arises how one should specify these functions in order to get the category of conformal mappings.

By differentiating one finds locally the general relation

$$\begin{pmatrix} dy \\ dx \end{pmatrix} = \begin{pmatrix} \frac{\partial y}{\partial q} & \frac{\partial y}{\partial \lambda} \\ \frac{\partial x}{\partial q} & \frac{\partial x}{\partial \lambda} \end{pmatrix} \begin{pmatrix} dq \\ d\lambda \end{pmatrix} \qquad (24)$$

From planar geometry we know that in cartesian c.n. a similarity mapping from plane onto plane can be written (omitting the irrelevant shift parameters):

$$\begin{pmatrix} Y \\ X \end{pmatrix} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} p\cos\omega & -p\sin\omega \\ p\sin\omega & p\cos\omega \end{pmatrix} \begin{pmatrix} y \\ x \end{pmatrix} \qquad (25)$$

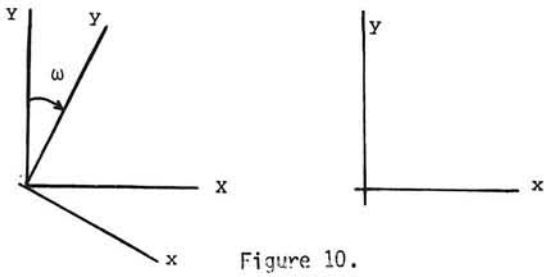


Figure 10.

with p = scale factor

ω = angle of rotation.

By comparing (24) and (25), it can be seen that, since $\{q,\lambda\}$ may be considered as 'local cartesian' c.n., (24) represents locally a similarity transformation if

$$\boxed{\begin{aligned}\frac{\partial y}{\partial q} &= \frac{\partial x}{\partial \lambda}\\ \frac{\partial x}{\partial q} &= -\frac{\partial y}{\partial \lambda}\end{aligned}} \qquad (26)$$

Since 'locally similar' is synonymous with 'conformal', (23) and (24) represent a conformal projection if the conditions (26) are satisfied.

The equations (26) are called the Cauchy Riemann equations. They constitute a set of differential equations and its general solution can be given in the form:

$$\boxed{y + i\,x = f(q + i\,\lambda) \quad (i^2 = -1)} \qquad (27)$$

The rigorous proof of it is omitted, however the correctness of the reverse statement will be illustrated by an example.
Take $w = z^2$ and verify that (26) is satisfied.

Proof:

$$\begin{aligned} y + ix &= (q + i\lambda)^2\\ y &= q^2 - \lambda^2\\ x &= 2q\lambda\\ \frac{\partial y}{\partial q} &= \frac{\partial x}{\partial \lambda} = 2q\\ \frac{\partial x}{\partial q} &= -\frac{\partial y}{\partial \lambda} = -2\lambda \end{aligned} \quad \rightarrow \text{Cauchy Riemann is satisfied}$$

The results, as derived above, will now be generalised for a conformal mapping between any two arbitrary surfaces Σ and $\bar{\Sigma}$.

Let the isometric coordinates and their unit of measure on both surfaces be denoted by $\{u, v, e\}$ and $\{\bar{u}, \bar{v}, \bar{e}\}$ respectively, then any function $\bar{u} + i\bar{v} = f(u + iv)$ represents a conformal mapping from Σ onto $\bar{\Sigma}$ with (in analogy with (6)):

the (point) scale factor $$m = \frac{\bar{e}}{e}\left\{\left(\frac{\partial \bar{u}}{\partial u}\right)^2 + \left(\frac{\partial \bar{v}}{\partial u}\right)^2\right\}^{\frac{1}{2}} = \frac{\bar{e}}{e}\left\{\left(\frac{\partial \bar{u}}{\partial v}\right)^2 + \left(\frac{\partial \bar{v}}{\partial v}\right)^2\right\}^{\frac{1}{2}} \qquad (28)$$

the (point) meridian convergence $$\gamma = \arctan\left(\frac{\partial \bar{v}}{\partial u}\Big/\frac{\partial \bar{v}}{\partial v}\right) = \arctan\left(-\frac{\partial \bar{u}}{\partial v}\Big/\frac{\partial \bar{u}}{\partial u}\right)$$

The scale factor is defined by the ratio of two corresponding line elements: $m = \frac{d\bar{s}}{ds}$; the meridian convergence by the angle between the y-axis in the plane and the representation of the meridian in the plane (see fig. 11).

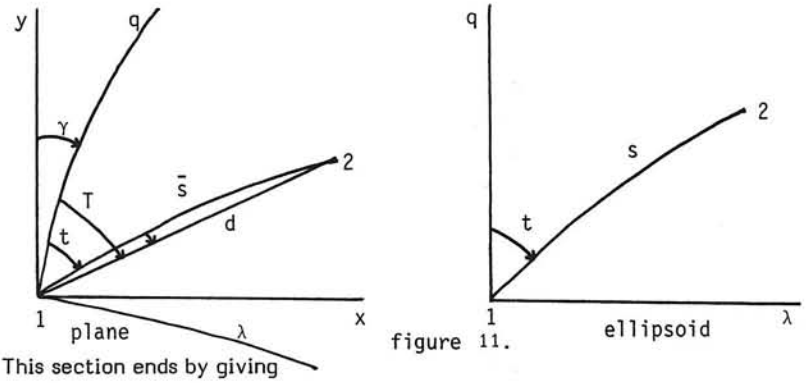


figure 11.

This section ends by giving

1) the general formulae for the reduction of distances and directions, induced by the mapping,
2) the administrative conventions for the numbering of c.n.

ad 1. Reduction formulas - The reduction of a distance consists of two parts:

1. Let $\bar{s}$ and s be the distances along corresponding curves then

$$\bar{s} = \int_0^s m \, ds$$

In most cases it suffices to compute $\Delta s = \bar{s} - s$ by applying Simpson's rule for the approximation of an integral:

$$\Delta s = \{ \frac{1}{6}(m_0 + 4m_{\frac{1}{2}} + m_1) - 1 \} s \qquad (29)$$

where the lower indices refer to the begin-, mid- and endpoint of the curve.

2. The reduction from arc to chord $\Delta s = d - \bar{s}$ is nearly always negligible.

The reduction of an angle consists of two reductions of directions viz. from arc to chord. This reduction is usually called the T - t reduction:

$$T - t = \mu = \tfrac{1}{2} \frac{\partial m}{\partial s} (\perp)s + \ldots \qquad (30)$$

where $\frac{\partial m}{\partial s}(\perp)$ stands for the rate of change of the scale factor in the direction perpendicular to the direction of the curve. The derivative should be taken at a distance 1/3 s of the point

ad 2. In some countries (e.g. Germany) the notation of the c.s. is ⟶ 

In some others (e.g. The Netherlands) the notation of the c.s. is ⟶ 

In order to avoid any ambiguity, one often prefers the names Northing and Easting for the c.n. ⟶ 

The origin of the planar {y,x} corresponds with the central point of projection and the y-axis with the central meridian. The point scale factor in the origin is fixed at unity and increases with growing distance from the origin. In most countries a point scale factor less than 1 is taken for the origin, in order to obtain a mean scale factor equal to 1 over the whole area. This latter scale factor m_o acts as a multiplicative parameter. Above that two additive parameters, called false Northing (fN) and false Easting (fE), are introduced to achieve positive c.n. over the whole region. This leads to the following formulae

$$\boxed{\begin{aligned} N &= m_o y + fN \\ E &= m_o x + fE \end{aligned}} \qquad (31)$$

2.2.3. Gauss conformal projection from ellipsoid to sphere.

On both surfaces isometric c.n. are introduced for which the following notation is used:

ellipsoid : {q, λ} ; geographic latitude: φ
principal radii of curvature: M, N

sphere : {ω, Λ} ; geographic latitude: Φ (the capital letter for φ)
radius: R

Be aware of the minor differences in the notation Φ and φ for both geographic latitudes.

It should be noted that {λ, Λ} refer to the meridian of Greenwich.

Gauss designed a conformal projection of the linear type:

in complex notation: $w = n\,z + a$

or
$$\omega + i\Lambda = (n_1 + in_2)(q + i\lambda) + (a_1 + ia_2) \qquad (32)$$
$$\omega = n_1 q - n_2 \lambda + a_1$$
$$\Lambda = n_1 \lambda + n_2 q + a_2$$

To fix the 5 parameters R, n_1, n_2, a_1, a_2, the following conditions are imposed:

condition 1:

The meridians and parallels on both surfaces are mapped onto each other:

$$\omega = f(q), \qquad \Lambda = g(\lambda) \qquad \text{and thus} \quad \boxed{n_2 = 0} \qquad (33)$$

The mapping formulae thus become:

$$\boxed{\begin{aligned} \omega &= n_1 q + a_1 \\ \Lambda &= n_1 \lambda + a_2 \end{aligned}} \qquad (34)$$

Condition 2:

The reference meridians have equal values $\lambda_0 = \Lambda_0$ and thus from (34)

$$\boxed{a_2 = \Lambda_0 - n_1 \lambda_0} \qquad (35)$$

If the reference meridans have zero value then $\boxed{a_2 = 0.}$

In order to formulate the remaining 3 conditions, the point scale factor is considered. Using formulae (28) one gets

$$m = \frac{n_1 R \cos \Phi}{N \cos \phi} \qquad (36)$$

Thus m is a function of latitude only.

The remaining conditions are such that we should have in the central point of projection (ϕ_0, λ_0):

Condition 3: $m = 1$

Condition 4: $\frac{dm}{d\Phi} = 0$

Condition 5: $\frac{d^2 m}{d\Phi^2} = 0$

These constraints mean that the overall variation of the point scale factor m is as small as possible over the whole area.

Condition 4 becomes: $$\frac{dm}{d\Phi} = \frac{\partial m}{\partial \Phi} + \frac{\partial m}{\partial \phi}\frac{d\phi}{d\Phi} \tag{37}$$

with $$\frac{d\phi}{d\Phi} = \frac{d\phi}{dq}\frac{dq}{d\omega}\frac{d\omega}{d\Phi}$$

Substituting (21), (22), (36) one gets

$$\frac{\sin \phi_o - n_1 \sin \Phi_o}{N_o \cos \phi_o} = 0$$

or $$\boxed{n_1 \sin \Phi_o = \sin \phi_o} \tag{38}$$

Differentiating (37) again and using (34) one gets for condition 5:

$$n_1 \cos \Phi_o = \cos \phi_o N_o^{\frac{1}{2}} / M_o^{\frac{1}{2}} \tag{39}$$

By squaring (38) and (39) one finds:

$$\boxed{n_1 = (1 + e'^2 \cos^4 \phi_o)^{\frac{1}{2}}} \tag{40}$$

From condition 3 and using (36) and (40):

$$\boxed{R = (M_o N_o)^{\frac{1}{2}}} \tag{41}$$

Finally from (34) one finds:

$$\boxed{a_1 = \omega_o - n_1 q_o} \tag{42}$$

The sequence of the computation is the following: first n_1 is computed with (40), then using (38), Φ_o can be found. The quantities ω_o and q_o are then obtained with (21) and (22) and finally a_1 and a_2 are found with (42) and (35).

The Gauss conformal projection is much used as the induced distortion of the geometric elements is very small. The length differences between corresponding geodesics on both surfaces are tabulated below

arcs in kilometres	absolute error in metres	relative error
50	0.000 014	1 : 370 0000 000
100	0.000 220	1 : 46 0000 000
200	0.003 500	1 : 5 7000 000
500	0.140 000	1 : 3700 000
800	0.900 000	1 : 890 000
1000	2.200 000	1 : 460 000
1500	11.000 000	1 : 140 000
2000	35.000 000	1 : 57 000
2500	85.000 000	1 : 29 000
3000	177.000 000	1 : 17 000

For nearly all practical surveying problems these reductions may safely by neglected, which means that most geometric problems may equally well be computed on the conformal sphere, as far as the geometric elements are concerned. Be cautious however that in the end the coordinates should be transferred from sphere to ellipsoid by applying the inverse mapping!

2.2.4.0. Computations on the conformal sphere.

The computations on the sphere can be divided in so-called direct and indirect computations, with the distinction that the former can be performed without linearisation

2.2.4.1. Direct computations.

Formulae from spherical trigonometry are applied to solve problems that are actually defined on an ellipsoid. Prior to the spherical computations and succeeding them, the transfer of the geographic c.n. in between both surfaces should be carried out. As these computations are the same for all surveying problems, they are outlined only in the first example. First of all one has to choose the central point of projection. Once its latitude is known, the 5

parameters of the Gauss mapping can be computed and the transfer of the coordinates can be performed.
One may choose either a fixed central point (e.g. for the whole working area of the radiopositioning system), or a variable central point (e.g. for each separate geodesic).

Example 1:
Bowring's solution for the direct and inverse problem on the ellipsoid (cf 2.3).
Bowring takes the centre of projection in one of the known endpoints of the geodesic, although the centre might also be chosen in between both endpoints.

Direct problem
Given ϕ_1, λ_1, α_1, s
Find ϕ_2, λ_2, α_2
Solution: With the aid of the mapping formulae $\{\phi_1, \lambda_1\}$ are transferred to $\{\Phi_1, \Lambda_1\}$. The geometric elements α_1 and s remain unchanged.
In radians: $\sigma = s/R$.
The direct problem is now solved on the sphere.
By applying in succession the cosinus- and sinusrule we obtain Φ_2, $\Lambda_2 (= \Lambda_1 + \Delta\Lambda)$ and α_2.
Finally $\{\Phi_2, \Lambda_2\}$ are transferred to $\{\phi_2, \lambda_2\}$ by applying the inverse mapping formulas.

Indirect problem
Given ϕ_1, λ_1, ϕ_2, λ_2
Find s, α_1, and α_2
Solution: an analoguous procedure is pursued.

Example 2:
Ballarin's method for the computation of hyperbolic patterns on the ellipsoid.
The elaboration of the spherical part leads to a trigonometric equation that can be solved in a direct way.

Let M be the master, R the red slave, L_R and Λ_R the observed red lane number and wavelength, G the green slave, L_G and Λ_G the observed green lane number and wavelength and P the observational point.

Referring to figure 12 it can be seen that the arc-length differences a_1 (= s_1 - s) and a_2 (= s_2 - s) can be computed from the observations by

$$a_1 = L_R \Lambda_R - b_1$$
$$a_2 = L_G \Lambda_G - b_2 \ .$$

The arc-lengths of the baselines b_1 and b_2 and their intersecting angle θ are computed by applying the algorithm for the direct and inverse problem on the sphere (see under example 1).
Introducing the bisector of θ as a reference for the directions, the problem is solved once the azimut α and arc-length s of the great circle from M to P are known.

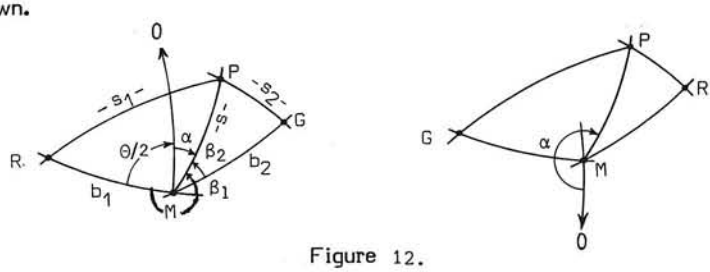


Figure 12.

If the positive direction of the reference is denoted by $\overline{MO}$, then:

$$\alpha(\overline{MO}) = \alpha(\overline{MG} + \tfrac{1}{2}\theta) \text{ with } \theta = \overline{MG} - \overline{MR}$$

In the spherical triangle (M,R,P) we have

$$\cos s_1 = \cos s \cos b_1 + \sin s \sin b_1 \cos \beta_1$$

as well as:

$$\cos s_1 = \cos s \cos a_1 - \sin s \sin a_1 \quad (\equiv \cos (s + a_1).$$

Equating the right-hand members yields, after division by sin s:

$$\cot s = \frac{\sin a_1 + \sin b_1 \cos \beta_1}{\cos a_1 - \cos b_1}$$

or: $$\cot s = \left(\frac{\sin a_1}{\sin b_1} + \cos \beta_1\right) \Big/ \left(\frac{\cos a_1 - \cos b_1}{\sin b_1}\right)$$

Elaborating the denominator:

$$\frac{\cos a_1 - \cos b_1}{\sin b_1} = \frac{2(\sin^2 \tfrac{1}{2}b_1 - \sin^2 \tfrac{1}{2}a_1)}{2 \sin \tfrac{1}{2}b_1 \cos \tfrac{1}{2}b_1} = \tan \tfrac{1}{2}b_1 - \frac{\sin a_1}{\sin b_1} \tan \tfrac{1}{2}a_1$$

Let $r_1 = \dfrac{\sin a_1}{\sin b_1}$ and $k_1 = \tan \tfrac{1}{2}b_1 - r_1 \tan \tfrac{1}{2}a_1$

then $$\cot s = \frac{r_1 + \cos \beta_1}{k_1} \qquad (43)$$

In like manner in spherical triangle (M,G,P):

$$\cot s = \frac{r_2 + \cos \beta_2}{k_2} \qquad (44)$$

From figure 12 it follows

$$\begin{aligned} \beta_1 &= 2\pi - (\theta/2 + \alpha) \Rightarrow \cos \beta_1 = \cos(\tfrac{\theta}{2} + \alpha) \\ \beta_2 &= \theta/2 - \alpha \qquad \Rightarrow \cos \beta_2 = \cos(\tfrac{\theta}{2} - \alpha) \end{aligned} \qquad (45)$$

After equating the right-hand members of (43) and (44) and taking into account (45), one gets a trigonometric equation in α :

$$\sin \alpha = n - m \cos \alpha \qquad (46)$$

with $$m = \frac{(k_1 - k_2)\cos \theta/2}{(k_1 + k_2)\sin \theta/2} \quad \text{en} \quad n = \frac{r_1 k_2 - r_2 k_1}{(k_1 + k_2)\sin \theta/2}$$

The solution of (46) can be found in a direct way:

$$\cos \alpha = \frac{m n \pm (1 + m^2 - n^2)^{\frac{1}{2}}}{1 + m^2} \qquad (47)$$

The - sign of the square root should be used if the observational point is situated in the 'outer' region of the chain.

figure 13.

Once α is known, s kan be computed with (43) or (44). With s and α the c.n. $\{\Phi,\Lambda\}$ of the observational point are computed by means of the algorithm for the direct problem on the sphere.

2.2.4.2. Indirect computations on the sphere.

As a rule geometric constructions on a sphere can not be solved in a direct way, since

- many survey constructions on the sphere (e.g. point resection) lead to trigonometric equations (being non-linear by nature) that can only be solved by a process of linearisation.
- most navigational systems provide an almost continuous stream of information, which necessitates a real-time elaboration and/or on-line plotting. It is then advantageous to linearise the conditions, taking the foregoing observational point as Taylor-point. The inevitable accumulation of discretisation errors is duly covered by regular updating.
- when redundant observations have been carried out, it is necessary to linearise the conditions in order to obtain the parameter equations of the adjustment. Remark. When the ellipsoid is used as a computing surface, indirect computations are a sheer necessity.

The basic geodetic observables are:

1. distance,
2. distance-difference,
3. (backward)azimut,
4. (backward)azimut difference ≡ angle,
5. (forward) azimut.

These observables are now related to geographical coordinates on the unit sphere by means of formulae of spherical trigonometry.
By the process of differentiation of these formulae one finds the corresponding linearised relations.

The cosine rule of spherical trigonometry gives: (see Appendix 1)

$$\cos b = \cos a \cos c + \sin a \sin c \cos B$$

Differentiating yields:

$$-\sin b\, db = (-\sin a \cos c + \cos a \sin c \cos B)da + (-\sin c \cos a + \cos c \sin a \cos B)dc + -\sin a \sin c \sin B\, dB$$

Substituting the 5-element rule gives:

$$-\sin b\, db = -\cos C \sin b\, da - \cos A \sin b\, dc - \sin a \sin c \sin B\, dB$$

or

$$\boxed{db = \cos C\, da + \cos A\, dc + \sin A \sin c\, dB} \qquad (48)$$

Using the sine-rule gives the equivalent formula:

$$db = \cos C\, da + \cos A\, dc + \sin C \sin a\, dB$$

The cotangent rule gives:

$$\cot c \sin a - \cot C \sin B = \cos a \cos B$$

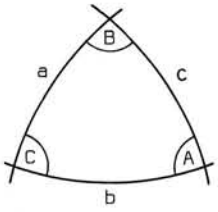


figure 14.

Differentiating yields:

$$-\sin^{-2}c \sin a\, dc + (\cot c \cos a + \sin a \cos B)da + \sin^{-2}C \sin B\, dC + +(\cos a \sin B - \cot C \cos B)dB = 0$$

Multiplying by sin c sin C:

$$-\sin^{-1}c \sin C \sin a\, dc + (\cos a \cos c \sin C + \sin a \sin c \sin C \cos B)da + +\sin^{-1}C \sin B \sin c\, dC + (\cos a \sin c \sin B \sin C - \sin c \cos C \cos B)dB = 0$$

Using the sine-rule and cosine-rule:

$$\boxed{\sin b\, dC = +\sin A\, dc - \sin C \cos b\, da - \cos A \sin c\, dB} \qquad (49)$$

These general formulae (48) and (49) are transferred to the polar triangle by making the following substitutions

$B \to \Lambda_j - \Lambda_i$ $\quad dB \to d\Lambda_j - d\Lambda_i$

$c \to \frac{1}{2}\pi - \Phi_j$ $\quad dc \to -d\Phi_j$

$A \to 2\pi - \alpha_{ji}$ $\quad dA \to -d\alpha_{ji}$

$b \to \sigma_{ij}$ $\quad db \to d\sigma_{ij}$

$C \to \alpha_{ij}$ $\quad dC \to d\alpha_{ij}$

$a \to \frac{1}{2}\pi - \Phi_i$ $\quad da \to -d\Phi_i$

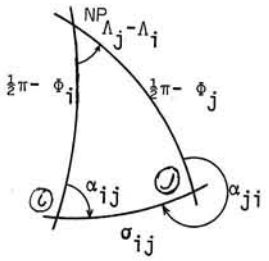


figure 15.

After replacing the notation d for the differential by the notation Δ for the difference (48) and (49) become:

$$\Delta\sigma_{ij} = -\cos\alpha_{ji}\,\Delta\Phi_j - \cos\alpha_{ij}\,\Delta\Phi_i - \sin\alpha_{ji}\cos\Phi_j\,\Delta\Lambda_j + \\ - \sin\alpha_{ij}\cos\Phi_i\,\Delta\Lambda_i \qquad (50)$$

$$\sin\sigma_{ij}\,\Delta\alpha_{ij} = \sin\alpha_{ji}\,\Delta\Phi_j + \sin\alpha_{ij}\cos\sigma_{ij}\,\Delta\Phi_i + \\ -\cos\alpha_{ji}\cos\Phi_j\,\Delta\Lambda_j + \cos\alpha_{ji}\cos\Phi_j\,\Delta\Lambda_i \qquad (51)$$

The sequence ij means: from point i to point j.

Now we introduce the following substitutions and approximations.
For a sphere with radius R we may substitute

$$s_{ij} = R\ \sigma_{ij}$$

For distances less than 100 km we may substitute $\cos\sigma_{ij} = 1$, $\sin\sigma_{ij} = \sigma_{ij}$.
Further we may replace, where necessary,

$$\alpha_{ji} = \pi + \alpha_{ij}\ , \qquad \Phi_i = \Phi_j = \Phi_k$$

Let point i be the point to be determined and point j and k known points, thus $\Delta\Phi_j = \Delta\Phi_k = \Delta\Lambda_j = \Delta\Lambda_k = 0$, then the following difference-equations are obtained.

ad 1) Distance s_{ij}

$$\Delta s_{ij} = -\cos \alpha_{ij} \underline{R\ \Delta\Phi_i} - \sin \alpha_{ij} \underline{R \cos \Phi_i\ \Delta\Lambda_i}$$

ad 2) Distance-difference $s_{ij} - s_{ik}$

$$\Delta(s_{ij} - s_{ik}) = (\cos \alpha_{ik} - \cos \alpha_{ij}) \underline{R\ \Delta\Phi_i} + (\sin \alpha_{ik} - \sin \alpha_{ij}) \underline{R \cos \Phi_i\ \Delta\Lambda_i}$$

$$= -2 \sin \tfrac{1}{2}(\alpha_{ik} - \alpha_{ij}) \{\sin \tfrac{1}{2}(\alpha_{ik} + \alpha_{ij}) \underline{R\ \Delta\Phi_i}$$

$$- \cos \tfrac{1}{2}(\alpha_{ik} + \alpha_{ij}) \underline{R \cos \Phi_i \Delta\Lambda_i}\}$$

Remark: The relation between lanenumber and distance-difference is, with i = obs. point, j = slave and k = master:

$$L = (s_{jk} + s_{ij} - s_{ik})/\Lambda$$

with $\Lambda\ \Delta L = \Delta(s_{ij} - s_{ik})$, with Λ is notation for wavelength. (Be careful as the notation for the spherical longitude is the same).

ad 3) Backward azimut α_{ij}

$$\Delta\alpha_{ij} = (\sin \alpha_{ij}/s_{ij}) \underline{R\ \Delta\Phi_i} - (\cos \alpha_{ij}/s_{ij}) \underline{R \cos \Phi_i\ \Delta\Lambda_i}$$

ad 4) Angle A_{jik} $(= \alpha_{ik} - \alpha_{ij})$

$$\Delta A_{jik} = (\sin \alpha_{ik}/s_{ik} - \sin \alpha_{ij}/s_{ij}) \underline{R\ \Delta\Phi_i} - (\cos \alpha_{ik}/s_{ik} -$$

$$-\cos \alpha_{ij}/s_{ij}) \underline{R \cos \Phi_i\ \Delta\Lambda_i}$$

ad 5) Forward azimut α_{ji}

$$\Delta\alpha_{ji} = -(\sin \alpha_{ji}/s_{ji}) \underline{R\ \Delta\Phi_i} + (\cos \alpha_{ji}/s_{ji}) \underline{R \cos \Phi_i\ \Delta\Lambda_i}$$

The equations ad 1) are applied in range-range-systems like Syledis, ad 1) and ad 5) in Artemis, ad 2) in hyperbolic systems like Decca, Hifix, Loran, ad 4) in circular systems like the sextant. For the so-called pseudo range-range system, ad 1) can be applied, provided that for each chain a time delay parameter Δt is introduced on the right-hand side of the equation.

The Δ -quantities on the left-hand side are equivalent with: 'computed' minus 'observed'. The computed values are obtained by using approximate values for the c.n. $\{\Phi_i, \Lambda_i\}$.

It should be noted that the coefficients of the increments $\underline{R\,\Delta\varphi_i}$ and $\underline{R\cos\varphi_i\Delta\lambda}_i$ need only to be evaluated in few significant digits.

Therefore the above difference equations might also be applied for ellipsoidal computation. However, be aware that in that case the 'computed' values on the left-hand side should be computed with the ellipsoidal algorithms for the direct and inverse problems.

The exact difference equations for the ellipsoid can be found by substituting in the relations 1) to 5): $R\ \Delta\Phi \rightarrow M\ \Delta\Phi$ and $R\cos\Phi\ \Delta\Lambda \rightarrow N\cos\Phi\ \Delta\Lambda$ and for the plane: $\underline{R\ \Delta\Phi} \rightarrow \Delta y$ and $\underline{R\cos\Phi\ \Delta\Lambda} \rightarrow \Delta x$.

2.2.5. Conformal mapping from ellipsoid onto plane.

In all mapping the longitude λ is referred to the central meridian of the area to be mapped and not to the Greenwich meridian. For the map projections of the 'North Sea countries" see appendix 2.

2.2.5.1 Mercator projection.

The mapping formula is, in complex form: $y + ix = a(q + i\lambda)$

splitting up in components:

$$\boxed{\begin{aligned} y &= a\,q \\ x &= a\,\lambda \end{aligned}}$$

point scale factor : $m = a/N\cos\phi$

meridian convergence : $\lambda = 0$

where a is equatorial radius.

Characteristics:

1. The p. takes its name from the latin surname of Gerhard Kremer, the inventor. Kremer was born in Flanders in 1512 and first used the p. for a map of the world in 1569. Gradually the mercator p. took over the role of the traditional, oldest, map projection, the equal-spaced 'plate carree': $\boxed{\begin{aligned} y &= a\phi \\ x &= a\lambda \end{aligned}}$.

2. The coordinate curves are mapped upon each other,
   thus meridians : $\lambda = c$ onto $x = c$
   and parallels : $q = c$ onto $y = c$
3. Distances along the equator are preserved in length, thus:
   $m = 1$ for $q = \phi = 0$.
4. The p. is the only projection on which the rhumbline is represented by a straight line. The rhumbline is defined on a surface of revolution as the curve having in each of its points the same angle with the local meridian.
   Proof: The solution of the s.d.e. for a rhumbline on an ellipsoid, having constant angle α, gives (cf. 2.3.2):

$$q_2 - q_1 = (\lambda_2 - \lambda_1) \operatorname{cotg} \alpha$$

   The application of the mapping formulae leads to:

$$y_2 - y_1 = (x_2 - x_1) \operatorname{cotg} \alpha$$

   which is the equation of a straight line.
5. The p. is much applied for navigational charts, but less used for geodetic purposes. Switzerland, Malaya and Serawak apply the mercator p in the oblique position of the cylinder, corresponding with the second step of a double-projection.

2.2.5.2. Lambert conformal conical projection.

The p. is applied in two variants, namely with one or two standard parallels.
Mapping formulae, in complex notation:

$$y + ix = -k\, e^{-l(q + i\lambda)} + n \quad (k,l,n \text{ are real parameters})$$

with $e^{i\lambda} = \cos \lambda + i \sin \lambda$

or

$$\boxed{\begin{aligned} y &= -ke^{-lq} \cos l\lambda + n \\ x &= ke^{-lq} \sin l\lambda \end{aligned}}$$

with $m = \dfrac{1}{N \cos \phi}\, k.l\, e^{-lq}$

$\gamma = -l\lambda$

The parameters assume the following values.
For one standard parallel ϕ_o

$$k = \frac{N_o \cos \phi_o}{e^{-l q_o}}$$

$$l = \sin \phi_o$$

For two standard parallels ϕ_1 and ϕ_2:

$$k = \frac{N_1 \cos \phi_1}{e^{-l q_1}} = \frac{N_2 \cos \phi_2}{e^{-l q_2}}$$

$$l = \frac{\ln(N_1 \cos \phi_1) - \ln(N_2 \cos \phi_2)}{q_2 - q_1}$$

For both variants: $n = k e^{-l q_r}$, where the subindex r stands for 'reference point', and q_r is the isometric latitude, corresponding to ϕ_r for that point. The parameter n acts as an additive constant.

There are a number of different conventions for the selection of ϕ_r.
For the one-standard parallel variant one often chooses $\phi_r = \phi_o$ but also $\phi_r = 0$.
For the two-standard parallel variant one often chooses $\phi_r = \phi_1$. For the unified world aeronautical charts, where a division in zones with 4^o width has been applied (cf. UTM), the reference latitude is taken 30' below the lower boundary of each zone.
Sometimes n is included in the false Northing.

Reduction formulae.
Reduction for lengths:

$$\Delta s = (m_o - 1)s + \frac{s}{6R^2}(y_1^2 + y_1 y_2 + y_2^2)$$

with R = mean radius of curvature.

Reduction for directions:

$$T-t = \mu = \frac{(2y_2 + y_1)(x_2 - x_1)}{6R^2} \quad , \mu \text{ in radians.}$$

Characteristics:

1. Johann Heinrich Lambert was a famous mathematician and astronomer, born in Alsace-Lorraine in 1728. The p. was first published in 1772.
2. The p. is very well suited to map areas, having their largest extension in east-west direction. As the main aeronautical and nautical routes have that direction, many navigational charts are made in this p.
3. The p. for 2 standard-parallels is tabulated for zones with zone-width of 4^{o}, from 0^{o} to 80^{o} latitude, on behalf of the world Aeronautical charts.
4. The p. is much used for geodetic purposes. The national coordinate systems in Belgium, France and Danmark are based on the p.
5. The polar stereographic p. and the mercator p. may be considered as special cases of the p. Geometrically the standard parallel is thought to be moved to the pole and equator respectively, transferring the cone into a plane and cylinder.

   Numerically the characteristic parameters k and l are found by substituting $\phi_o = \pi/2$ and $\phi_o = 0$ respectively. For the polar stereographic projection we find:

$$k = 2a(1 - e)^{\frac{1}{2}(e - 1)}(1 + e)^{-\frac{1}{2}(e + 1)}$$

$$l = 1$$

For a sphere: $l = 0$ and $k = 2a$.

2.2.5.3. The stereographic projection.

This p. can be seen as a special case of the Lambert conformal conical p., see 2.2.5.2 point 5. The p. is first used by Hipparchus, 160-126 B.C. and nowadays it is mainly applied for nautical and bathymetrical charts of polar regions and stellar charts.

For geodetic purposes it has especially been introduced in countries of circular shape. The national coordinate systems of Poland, Hungary and Holland are based on this p. To that end the projection is applied in the oblique position as second step of a double projection. That is why the spherical formulae of the p. are almost exclusively used, also for geodetic purposes.

The p. is the only conformal p. from sphere to plane for which the mapping formulas can be derived in a pure geometric way. The origin of the p. is the point on the sphere, opposite to the centre of projection, and the plane of projection coincides with the plane tangent to the sphere in the centre of projection.

The double projection actually consists of three modules, namely the two mappings and the coordinate transformation on the sphere in between them.

Module 1.
Conformal p. from ellipsoid to sphere, after Gauss (cf. 2.2.3) thus

$$\{\phi,\lambda\} \;\overset{\leftarrow}{\rightarrow}\; \{\Phi,\Lambda\}$$

Module 2.
Transformation from geographic to local geographic c.n., thus:

$$\{\Phi,\Lambda\} \;\overset{\leftarrow}{\rightarrow}\; \{\bar{\Phi},\bar{\Lambda}\}$$

The formulae are given below. The geographic longitude Λ is referred to the central meridian and the colatitude $\psi = \pi/2 - \phi$ is introduced. The centre of projection (Amersfoort) is taken as local pole and denoted by $\overline{NP}$ $\{\psi_o, 0\}$. An arbitrary point is denoted by P $\{\psi, \Lambda\}$. The cosine- and sine-rule of spherical trigonometry gives:

$$\cos\bar{\psi} = \cos\psi_o \cos\psi + \sin\psi_o \sin\psi \cos\Lambda$$

$$\sin\bar{\Lambda} = \sin\psi \sin\Lambda / \sin\bar{\psi}$$

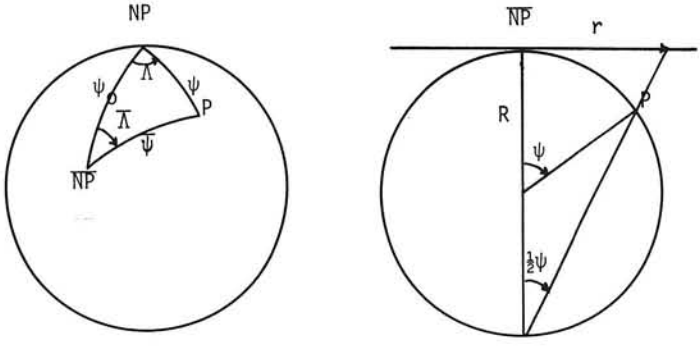


figure 16.

Module 3.
The mapping formulae are, with

$$r = 2R \tan\tfrac{1}{2}\bar{\psi} = 2R \tan(\pi/4 - \tfrac{1}{2}\bar{\Phi})$$

$$y = 2R \tan(\pi/4 - \tfrac{1}{2}\bar{\Phi}) \cos\bar{\Lambda}$$

$$x = 2R \tan(\pi/4 - \tfrac{1}{2}\bar{\Phi}) \sin\bar{\Lambda}$$

The map projection of the Netherlands.

By concatenating the above modules we get the following formulae:

$$x = 190066.91\,\lambda - 11831.00\,\phi\lambda - 114.20\,\phi^2\lambda - 32.39\,\lambda^3 - 2.33\,\phi^3\lambda + \\ - 0.61\,\phi\lambda^3$$

$$y = 309020.34\,\phi + 3638.36\,\lambda^2 + 72.92\,\phi^2 - 157.97\,\phi\lambda^2 + 59.77\,\phi^3 + \\ +0.09\,\lambda^4 - 6.45\,\phi^2\lambda^2 + 0.07\,\phi^4$$

The reverse mapping formulae are:

$$\phi = 3236.033\,y - 32.592\,x^2 - 0.247\,y^2 - 0.850\,x^2y - 0.065\,y^3 + \\ +0.005\,x^4 - 0.017\,x^2y^2$$

$$\lambda = 5261.305\,x + 105.979\,xy + 2.458\,xy^2 - 0.819\,x^3 + 0.056\,xy^3 + \\ -0.056\,x^3y$$

The dimension of (x,y) is meter, the dimension of (ϕ,λ) is 10^4'' (seconds of arc). In both sets of formulae (ϕ,λ) are the differences with respect to Amersfoort: $\phi = 52^\circ09'22''178$, $\lambda = 5^\circ23'15''500$. Recently the following false Easting and Northing have been introduced: fE = 155000 m. , fN = 463000 m.

Reductions. As a point scale factor $m_o = 0.9999079$ has been introduced for the central point of projection Amersfoort, the length-reduction in mm. per 100 m. , with (x,y) in km., amounts to:

$$\Delta s = -9.2 + \frac{x^2 + y^2}{1629}$$

For the reduction of a direction between points A and B we have:

$$\mu = \frac{y_A x_B - x_A y_B}{256}$$, with (x,y) in km. and μ in 0.0001 gon

2.2.5.4. The transverse mercator projection.

In the mercator projection the equator is mapped upon the x-axis in such a way that distances along the equator preserve their length.

In the transverse mercator projection the central meridian is mapped upon the y-axis in such a way that distances along the central meridian preserve their length. This constraint alone defines the mapping formulae.

The mapping formulae are far more complicated for the transverse case than for the normal case, since the central meridian, in contrast to the equator, is an ellipse instead of a circle.

The mapping formulae can not be expressed in closed form. They are given in a Taylor series and have the following form (λ is difference in longitude from central meridian in radians.

$$y = a_0 - a_2 \lambda^2 + a_4 \lambda^4 - a_6 \lambda^6 + \dots$$

$$x = a_1 \lambda - a_3 \lambda^3 + a_5 \lambda^5 - a_7 \lambda^7 + \dots$$

The first parameter a_0 is the meridian arc length of the point to be mapped.
For its computation see 2.1.2.

$$a_0 = \int_0^{\phi} M \, d\phi \qquad (M \text{ and } N \text{ are main radius of curvature})$$

For the remaining coefficients one finds

$$a_1 = N \cos \phi$$

$$a_2 = \tfrac{1}{2} N \cos^2\phi \, . \, t(-1)$$

$$a_3 = \frac{1}{6} N \cos^3\phi \, (-1 + t^2 - \eta^2)$$

$$a_4 = \frac{1}{24} N \cos^4\phi \, . \, t(5 - t^2 + 9\,\eta^2 + 4\,\eta^4)$$

$$a_5 = \frac{1}{120} N \cos^5\phi \, (5 - 18t^2 + t^4 + 14\,\eta^2 - 58t^2\,\eta^2)$$

$$a_6 = \frac{1}{720} N \cos^6\phi \, . \, t(-61 + 58t^2 - t^4 - 270\,\eta^2 + 330t^2\,\eta^2)$$

$$a_7 = \frac{1}{5040} N \cos^7\phi \, (-61 + 479t^2 - 179t^4 + t^6)$$

$t = \tan \phi$, $\eta = e' \cos \phi$.

For the reverse mapping one finds

$$q = b_0 - b_2x^2 + b_4x^4 - b_6x^6 + \ldots$$

$$\Delta\lambda = b_1x - b_3x^3 + b_5x^5 - b_6x^7 + \ldots$$

The first parameter b_0 is equal to the isometric latitude q_f of the footpoint Q_F (see figure 17).

The geographic latitude ϕ_f is obtained by applying the process of the so-called 'inversion of series' for the meridional arc length (cf. 2.1.2).

The remaining coefficients are

$$b_1 = (N \cos\phi)^{-1}$$

$$b_2 = \tfrac{1}{2}(N^2 \cos\phi)^{-1}t$$

$$b_3 = \frac{1}{6}(N^3 \cos\phi)^{-1}(1 + 2t^2 + \eta^2)$$

$$b_4 = \frac{1}{24}(N^4 \cos\phi)^{-1}t(5 + 6t^2 + \eta^2 - 4\eta^4)$$

$$b_5 = \frac{1}{120}(N^5 \cos\phi)^{-1}(5 + 28t^2 + 24t^4 + 6\eta^2 + 8t^2\eta^2)$$

$$b_6 = \frac{1}{720}(N^6 \cos\phi)^{-1}t(61 + 180t^2 + 120t^4 + 46\eta^2 + 48t^2\eta^2)$$

$$b_7 = \frac{1}{5040}(N^7 \cos\phi)^{-1}(61 + 662t^2 + 1320t^4 + 720t^6)$$

These parameters are computed with $\phi = \phi_f$.

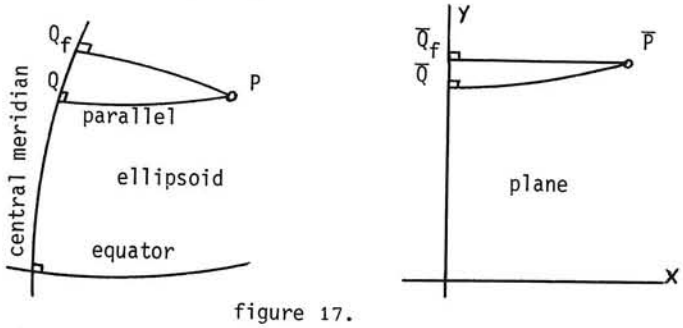


figure 17.

Point scale factor:

$$m = 1 + \tfrac{1}{2}\cos^2\phi\,(1 + \eta^2)\,\Delta\lambda^2 + \frac{1}{24}\cos^4\phi\,(5 - 4t^2 + 14\eta^2 - 28t^2\eta^2)\,\Delta\lambda^4$$

or expressed in y and x:

$$m = 1 - \frac{1}{2N^2}(1 + \eta^2)x^2 + \frac{1}{24N^4}(1 + 6\eta^2)x^4 + \ldots$$

Meridian convergence (in radians) :

$$\gamma = \lambda \sin \phi + \frac{\lambda^3}{3} \sin \phi \cos^2 \phi (1 + 3\eta^2) + \ldots$$

or $$\gamma = \frac{x}{N} \tan\phi \left(1 + \frac{1}{6}\left(\frac{x}{N}\right)^2 (1 - 2t^2 + 5\eta^2) \ldots\right)$$

Reductions.

Reduction of length:

$$\Delta s = \frac{s}{6R^2} (x_1^2 + x_1 x_2 + x_2^2)$$

For the UTM projection a scalefactor $m_o = 0.9996$ is introduced thus:

$$\Delta s = -0.0004\ s + \frac{s}{6R^2} (x_1^2 + x_1 x_2 + x_2^2)$$

one takes: R = 6378 km.

Reduction of direction:

$$\mu = \frac{(2x_2 + x_1)(y_2 - y_1)}{6R^2}$$

Characteristics.

In the German-speaking countries the t.m. projection is called after Gauss (1820) who designed the projection and Krüger (1912) who first published the theory in extension. In the English speaking countries one speaks of the transverse mercator- projection.

It is customary to divide the area to be mapped, into zones bounded by two meridians. In many countries the projection is basic to their coordinate systems. In Germany the Gauss-Krüger projection is applied for a division into 3 zones, each of 3^o width. In Great Britain, as well as in its former dominions, and in the USSR, the t.m. is also used for the national c.s. In Australia, Canada and New Zealand one uses a zone width of 5^o, 8^o and 4^o width respectively.

In USSR different zone divisions are in use: for the country as a whole: 6^o zones, for the separate republics: 3^o zones and for the larger townships: $< 3^o$ zones.

The p. got its fame when the UTM projection was introduced after world war II. Nowadays most countries dispose of two databases of coordinates, one refering to their traditional national map projection and based on a national datum, the other based on a more recent continental datum in combination with the U.T.M. The U.T.M. differs from t.m. only by a set of prescribed conventions of administrative nature.

Universal Transverse Mercator Projection (UTM)

Characteristics:

Transverse Mercator Projection
Zone width: 6^{o}, Central meridian: 3^{o}, 9^{o}, 15^{o}, ...
Latitude of origin: equator
False Easting: 500 000.00 m = central meridian
False Northing: 0.00 = equator
Scale factor k = 0.9996 = scale on central meridian
Easting = k.x, Northing = k.y, point scale = k.m
Latitude limits: 80^{o} South to 80^{o} North
Zone numbering: Starting with 1 for zone 180^{o} W to 174^{o} W
and increasing eastward to 60 for zone 174^{o}E to 180^{o}E
Zone 0^{o} to 6^{o}E (Northsea) has number 31.

Ellipsoid:
International, Hayford ellipsoid.
a = 6378388.00 m
b = 6356911.95 m
f = 1 : 297.00 = 0.00336700
e^{2} = 0.00672267

2.2.6. Transformation between overlapping planar coordinatesystems.

The problem of coordinate-transformation arises when

1. two datasets, based on national datums, have to be transferred into each other in the boundary region of two neighbouring countries,
2. a dataset, based on a national datum, has to be transferred into a continental datum that covers it, or into a world-datum (and vice versa).

The transformation may be looked upon as a datumtransformation (see 2.1), be it on a lower level, viz. the national 2d. planar (grid)coordinates.
The underlying principle is that a conformal relation may be assumed between both coordinate-sets, since either of them is based on a conformal mapping from ellipsoid onto the plane.
This conformal relation is represented by the following polynomial series:

$$w = \Sigma\alpha_k z^k \quad (= \alpha_0 + \alpha_1 z + \alpha_2 z^2 + \ldots.) \quad (k=1,2..n)$$

with $w = Y + iX$

$z = y + ix$

$\alpha_k = a_k + ib_k$

$\{Y,X\}$ and $\{y,x\}$ are the cartesian coordinates of both coordinatesystems and $i^2 = -1$.

For $n = 1$, the conformal transformation narrows down to a similarity-transformation, for $n > 1$ to a transformation that is 'locally similar'.
When the degree of the polynom equals n then (n+1) common points are needed to compute the unknown parameters $\{a_k, b_k\}$.

For the transformation in Holland from the national (RD) dataset to the UTM dataset (ED50), about 70 common stations were available of which 5 have been chosen for the evaluation of the parameters of the 4th degree polynomial.
The differences in the remaining common stations were used to prepare two maps, representing isocorrection-curves for both Δx and Δy. These maps can be used to evaluate the final small additional increments.
In order to computerize this last step, the interpolation can also be carried out by applying a socalled prediction-interpolation method,based on a prescribed variance-covariance matrix Q_{ik} for the increments $\{\Delta x,\Delta y\}$. (see Chapter 3).
The principle of this method is that the effect of an arbitrary known $\{\Delta x,\Delta y\}_i$, i=1,...n, on an unknown $\{\Delta x,\Delta y\}_p$, p=1,...m, is supposed to depend on the distance s_{ip} between point i and p. This effect is represented by Q_{ip}, the larger s_{ip}, the smaller Q_{ip} and the lesser the effect.
Various assumptions, based upon experiences, have been made for Q_{ik}. A reasonable choice would be:

$$Q_{ik} = 1/(1 + (s_{ik}/s_o)^2)$$

or also $$Q_{ik} = \exp(-\tfrac{1}{2}(s_{ik}/s_o)^2)$$

with s_o being a parameter to be selected.
e is the base of the natural system of logarithms (=2.71828..)
and $\exp(p) = e^p$.

Both relations can be represented by the following graphs:

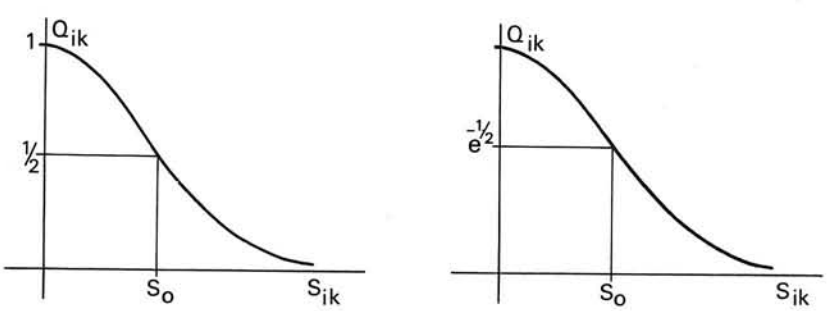


figure 18.

The interpolated values $(\Delta x)_p$, as well as $(\Delta y)_p$, in point p, are now obtained by:

$$(\Delta x)_p = Q_{pi} Q_{ik}^{-1} (\Delta x)_k$$

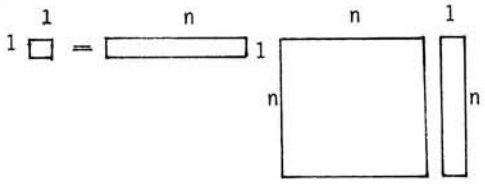
with summation over equal indices.

Substituting: $c_i = Q_{ik}^{-1} (\Delta x)_k$, we get

$$(\Delta x)_p = Q_{pi} c_i$$

The vector c_i assumes a constant value for the whole region.

In matrix structure:



In many cases there are only a few common points and sometimes there are no common points at all. In the case of regions of moderate extension it seems appropiate to take $n < 3$, even if the number of common points permits a higher degree of transformation. If no common points are available it is advised to create them by doing additional observations.

A few months ago (summer 1984), the embargo on the transformation-formulas in between RD and UTM has been raised in Holland.

The data and the isocorrection maps are given in an appendix 3.

2.3 Ellipsoidal computations.

2.3.1. Introduction.

The curve that is almost exclusively used on the ellipsoid is the geodesic, the shortest possible path between two points on the surface.

Any geometric network, having the geodesic as its structural element, can easily be transformed into a set of point coordinates once an algorithm has been developed for the direct and inverse problem, traditionally called 'the principal geodetic problem'.

The direct problem can be posed as follows: given the position in latitude and longitude, of a point on the ellipsoid, the 'standpoint', as well as the forward azimut and the length of the geodesic, find the position of the terminal point, the 'forepoint' and its back azimut.

The inverse problem is the converse of the direct problem: given the coordinates of both endpoints, find the length of the geodesic joining them as well as its forward and backward azimut at these endpoints.

These problems are thus comparable with the conversion from rectangular to polar coordinates in the plane.

Summarizing:

the direct problem : given $\phi_1, \lambda_1, \alpha_1, s$ find $\phi_2, \lambda_2, \alpha_2$

the inverse problem : given $\phi_1, \lambda_1, \phi_2, \lambda_2$ find α_1, α_2, s

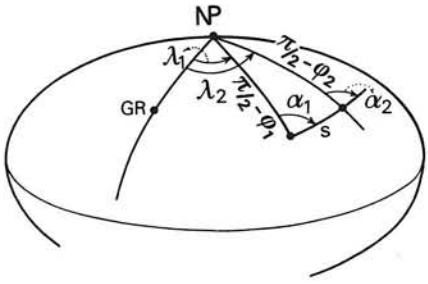


Figure 19.

The preoccupation of geodesists with this problem has a long and distinguished history. Gauss, Bessel, Helmert and Levallois are all prominent names associated with its solution. More recently Vincenty, Bowring, Sodano, Rainsford, Schödlbauer and many others gave solutions. A recent stocktaking comes to about 70 algorithms that have been developed up till now.

Happily however all these algorithms are based upon only three main methods of solution.
The basic principles of two of these methods are given in this chapter in order to enable the surveyor to criticize recent publications in which often only the straight-forward formulas are given without any reference to the basic principles. It appears that algorithms, based on the same method, show only minor differences caused by constraints such as the required accuracy and the computing hardware that is available.

The first group comprises the algorithms that are based on the evaluation of simultaneous differential equations. These equations are derived in section 2.3.2 and a numerical evaluation is given for small and large scale computers in section 2.3.3.

The second group is based on the elaboration of elliptic integrals that also govern the behaviour of the geodesic. The method dates back to Bessel and many geodesists have contributed to its numerical solution. This method, leading to fast and accurate algorithms, is dealt with in section 2.3.4.

In contrast to the above methods, that are ellipsoidal in the proper sense of the word, the third group is based on computations on a sphere that serves as a substitute for the ellipsoid. Although simple and attractive , the computations are no longer exact. For small regions however the approximation is good enough for most purposes. For this group the reader is referred to 2.2.4 in which also the linearized equations between the various observables in hydrographic surveying on the one hand and the station's position on the other hand, are derived. Together with the algorithms for the direct and inverse problem, these equations enable the surveyor to compute all kind of geometric configurations on the ellipsoid.
Anticipating on further explanation it seems useful to give already the main conclusions:

- for a general fast program, suitable for any length and demands of accuracy, Bessel's method is recommended;
- for a program, suitable for small-scale computers and limited accuracy and lengths, Bowring's method (belonging to the 3rd group) or Schödlbauer's method (belonging to the 1st group) seem most appropriate.

2.3.2. Differential equations of the geodesic.

Geographic coordinates $\{\phi , \lambda\}$ are, like all coordinates, numbers, associated with points on coordinate curves. Along these curves the relation between coordinate-differences and distances is obtained in the following way.

The simplest of both coordinate curves is the parallel. The parallel $\{\phi = \phi\}$ is a circle with constant radius (of curvature):

$$r = N \cos \phi$$

and thus

$$ds = N \cos \phi \, d\lambda$$

or

$$\frac{d\lambda}{ds} = \frac{1}{N \cos \phi}$$

This is the differential equation defining the parallel.

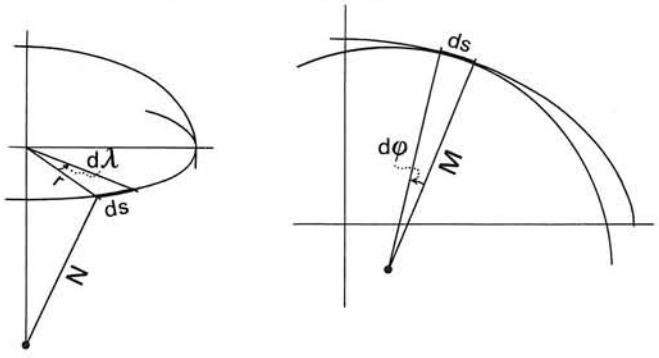


Figure 20.

The second coordinate curve is the meridian. The meridian $\{\lambda = \lambda\}$ is an ellipse with a varying radius of curvature: $M = a(1 - e^2)W^{-3}$ (see 2.1.2)

and thus $ds = M \, d\phi$

or $$\frac{d\phi}{ds} = \frac{1}{M}$$

This is a differential equation defining the meridian.

Now it is obvious that for any curve belonging to the family of curves that makes an angle α with the meridian, at least the following two differential equations hold:

$$\frac{d\phi}{ds} = \frac{\cos\alpha}{M} \qquad (52)$$

$$\frac{d\lambda}{ds} = \frac{\sin\alpha}{N\cos\phi} \qquad (53)$$

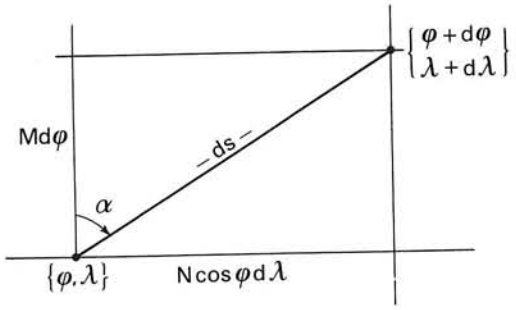


Figure 21.

The members of the family are distinguished from each other by specification of the variation of the azimut α with increasing s.
The most simple curve is the rhumbline, a curve on a surface of revolution that makes equal angles with the local meridian in all points of the curve.

Thus $\frac{d\alpha}{ds} = 0$ (54)

The formulas (52), (53) and (54) completely define the rhumbline. They form a set of simultaneous differential equations (s.d.e.) of the following general form:

$$\frac{dy^i}{dt} = f(y^i) \qquad \text{with } i = 1, 2, \ldots, n \qquad (55)$$

In our case n = 3 and $y^i = \{\phi, \lambda, \alpha\}$. The rhumbline plays (or rather played) a prominent part in marine navigation as it is the curve with constant steering angle.

For a geodesic the third d.e. is more complicated.
A kind of explanation can be given by considering the geodesic as "the most straight curve" on the surface.
In mathematical terminology: for any point holds that the projection of the geodesic, perpendicular to the tangent plane in that point has zero curvature: i.e. the tangents in two infinitesimal close points are coincident.

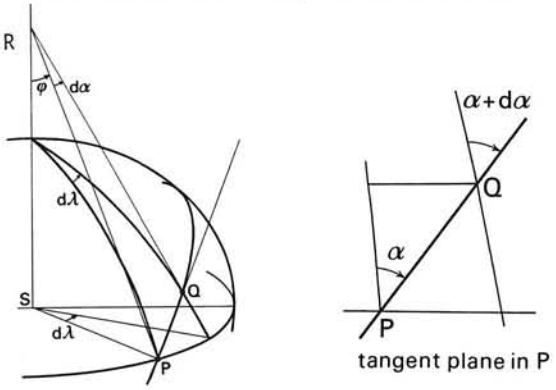


Figure 22.

From the above figure 2 2 it appears that the infinitesimal small angles dα and dλ are subtended by the same arc and since PS/PR = sin φ we have

$$d\alpha = \sin\phi \; d\lambda$$

or

$$\boxed{\frac{d\alpha}{ds} = \sin\phi \; \frac{d\lambda}{ds}} \qquad (56)$$

Equation (56) is called Bessel's d.e. for a geodesic. With (53) it gives the following (third) d.e.

$$\boxed{\frac{d\alpha}{ds} = \frac{\tan\phi}{N} \sin\alpha} \qquad (57)$$

The d.e. (52), (53) and (57) form a complete set of s.d.e. for the geodesic. For its solution see next section.

Finally the following important property for a geodesic is derived. From (52) and (53) follows

$$N \cos\phi \cos\alpha \frac{d\lambda}{ds} - M \sin\alpha \frac{d\phi}{ds} = 0 \qquad (58)$$

Substituting (56) in (58) gives

$$N \cos\phi \cos\alpha \frac{d\alpha}{ds} - M \sin\phi \sin\alpha \frac{d\phi}{ds} = 0 \qquad (59)$$

Integrating (59) and using $\frac{d(N \cos\phi)}{d\phi} = - M \sin\phi$ (which can easily be verified), one gets:

$$\frac{d(N \cos\phi \sin\alpha)}{ds} = 0$$

or $\boxed{N \cos\phi \sin\alpha = c \text{ (constant)}}$

or $r \sin\alpha = c \qquad (60)$

This is Claireaut's equation for the geodesic. It says that the product of the sine of the azimut of a geodesic and the appertaining radius of the parallel is a constant for all points of the geodesic.

It should be noted that both Bessel's and Claireaut's equation are valid on any surface of revolution.

2.3.3. The numerical solution of the s.d.e.

2.3.3.0. Introduction.

The s.d.e. for the geodesic are such that they can only be solved by methods of numerical analysis.

For an arbitrary point on a distance s from the starting point (1), the latitude, longitude and azimut can be expressed in a Taylor series:

$$\phi = \phi_1 + \Sigma \frac{1}{n!} \frac{d^n\phi}{ds^n} s^n$$

$$\lambda = \lambda_1 + \Sigma \frac{1}{n!} \frac{d^n\lambda}{ds^n} s^n \qquad n=1,2.... \qquad (61)$$

$$\alpha = \alpha_1 + \Sigma \frac{1}{n!} \frac{d^n\alpha}{ds^n} s^n$$

where all derivatives should be taken in the starting point. Expressions for the higher order derivatives can analytically be derived by successive differentiation of the first order derivatives. The larger the geodesic the higher the order of the derivatives that should be evaluated.

The basic principles of this approach are outlined in 2.3.3.2.

Unfortunately however, the Taylor series converge rather slowly and therefore a large number of terms should be taken into account, even for relatively short distances and low accuracy demands.

For the development of a general computing program that can handle geodesics of different lengths and demands of accuracy, one can best use one of the many general methods in numerical analysis that solve s.d.e. by introducing steps of varying size depending on the length of the geodesic. Such a general program, especially useful for large scale computers, will be dealt with in 2.3.3.1.

For small scale computers that are often applied in combination with a navigational survey system such as Loran, an algorithm will be given in 2.3.3.2 that is taylored for such a hardware configuration.

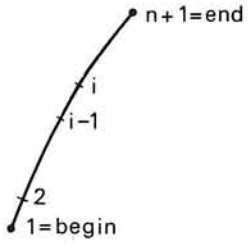
2.3.3.1. A general program for large scale computers.

The solution is given for the direct and inverse problem.

Direct problem: the total length s of the geodesic is split up into a number (n) of equal steps with length $\Delta s = s/n$.

For these steps it is assumed that the first order derivatives may be considered constant (and thus the higher order derivatives zero). One computes successively

$$\begin{aligned} \phi_i &= \phi_{i-1} + \frac{d\phi}{ds}\Delta s \\ \lambda_i &= \lambda_{i-1} + \frac{d\lambda}{ds}\Delta s \\ \alpha_i &= \alpha_{i-1} + \frac{d\alpha}{ds}\Delta s \end{aligned} \qquad (62)$$



where the derivatives should be taken in point (i-1).

In order to check whether the steplength has been small enough, the computation is repeated with half the steplength. After p repetitions when

$$|\phi_e^{(p)} - \phi_e^{(p-1)}| = |\lambda_e^{(p)} - \lambda_e^{(p-1)}| < \varepsilon \qquad (63)$$

the computation is ended. The tolerance ε reflects the accuracy that is required. A fair starting value for n is such that $s/n \tilde{=} 100$ km. The above mentioned method is called after Euler. It is the most simple one of the many methods that have been developed in numerical mathematics.

The most efficient one is probably Runga Kutta's method, that is based on the evaluation of first order derivatives in different points within each step. In any textbook on numerical analysis and in most software libraries Runga Kutta's method can be found.

Inverse problem: As a matter of fact, the above direct problem narrows down to the general problem of finding a particular solution for a set of s.d.e. that satisfies certain conditions given in the beginpoint. This problem is easier to solve than the inverse problem where the particular solution should satisfy conditions that are formulated in both begin- and endpoint.

In that latter case starting values for $\{\frac{d\phi}{ds}, \frac{d\lambda}{ds}, \frac{d\alpha}{ds}\}$ in the Taylor series are missing since α (and s) are unknown.
To meet this difficulty, the following iterative method, called 'turn in line method' or 'Einschwenkungsverfahren' has been developed. For the initial computation, approximate values $\alpha^{(1)}$ and $s^{(1)}$ are computed on a sphere with a radius equal to the mean radius of curvature of the ellipsoid in the working area. The value of R is not critical and any rough estimate e.g. R = 6370 km will do as well.

With these approximate values, the algorithm of the direct problem is applied. In general the coordinates $\{\phi^{(1)}, \lambda^{(1)}\}$ thus obtained, do not match with the given coordinates $\{\phi, \lambda\}$ of the endpoint.
From the differences $\Delta\phi^{(1)} (= \phi - \phi^{(1)})$ and $\Delta\lambda^{(1)} (= \lambda - \lambda^{(1)})$, the increments $\{\Delta\alpha^{(1)}, \Delta s^{(1)}\}$ can be computed. Refering to Chapter 3.4.2, one finds:

$$\begin{pmatrix} s\,\Delta\alpha^{(1)} \\ \Delta s^{(1)} \end{pmatrix} = \begin{pmatrix} -\sin\alpha_2^{(1)} & \cos\alpha_2^{(1)} \\ \cos\alpha_2^{(1)} & \sin\alpha_2^{(1)} \end{pmatrix} \begin{pmatrix} M\,\Delta\phi^{(1)} \\ N\cos\phi\,\Delta\lambda^{(1)} \end{pmatrix}$$

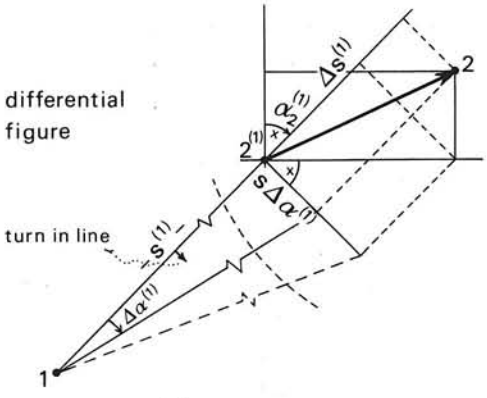


Figure 23.

Now improved approximations are found with

$$\alpha^{(2)} = \alpha^{(1)} + \Delta\alpha^{(1)}$$
$$s^{(2)} = s^{(1)} + \Delta s^{(1)}$$

and the computational process is repeated until the required accuracy is finally achieved.

2.3.3.2. Approximate solution for small scale computers.

The traditional solution of the s.d.e. is based on the evaluation of analytical expressions for the higher order derivatives

$$\{ \frac{d^n\phi}{ds^n} , \frac{d^n\lambda}{ds^n} , \frac{d^n\alpha}{ds^n} \} \qquad n = 2,3,..$$

The names of Legendre (direct problem) and Gauss (inverse problem) are associated with this approach.

The main characteristics will now briefly be outlined.

By differentiating the s.d.e. (1), (2) and (6) one finds the second order derivatives

$$\begin{aligned} \frac{d^2\phi}{ds^2} &= -\frac{V^4 t}{c^2}\sin^2\alpha - \frac{3V^4\eta^2 t}{c^2}\cos^2\alpha \\ \frac{d^2\lambda}{ds^2} &= \frac{2V^2 t}{c^2\cos\phi}\cos\alpha\sin\alpha \\ \frac{d^2\alpha}{ds^2} &= \frac{V^2}{c^2}(1 + 2t^2 + 2\eta^2)\cos\alpha\sin\alpha \end{aligned} \qquad (65)$$

with $\eta^2 = e'^2\cos^2\phi$; $t = \tan\phi$.

This process, although tedious and time consuming, can be continued for third and higher order derivatives.

In the Taylor series (61) the derivatives are now replaced by their corresponding expressions (65) and the series are then rearranged with respect to increasing powers of the terms (s cos α) and (s sin α).

Let $u = s\cos\alpha$

$v = s\sin\alpha$

with the azimut α taken in point (1) thus $\alpha = \alpha_1$,

then one gets:

$$\begin{aligned} \phi_2 - \phi_1 &= \Delta\phi = \Sigma f_{ik} u^i v^k \\ \lambda_2 - \lambda_1 &= \Delta\lambda = \Sigma g_{ik} u^i v^k \\ \alpha_2 - \alpha_1 &= \Delta\alpha = \Sigma h_{ik} u^i v^k \end{aligned} \qquad (66)$$

The series (66) are called Legendre's series and they give a solution for the direct problem. The coefficients f_{ik}, g_{ik}, h_{ik} are functions of the latitude and their values should be taken in the beginpoint, thus for $\phi = \phi_1$. The series expansion for $\Delta\alpha$ need not to be carried out, since α_2 can be computed with Claireaut's rule (60), once ϕ_2 is known.

By applying a mathematical process called 'inversion of series', Legendre's series (66) can be inverted. In addition to this, Gauss introduced the mid-latitude point to overcome the slow convergence of the inverted series and used this point as Taylor-point in the series expansion. Eventually he got his famous mid-latitude formulas for the solution of the inverse problem.

$$
\begin{aligned}
s\cos((\alpha_1 + \alpha_2)/2) &= u = \Sigma a_{ik}(\Delta\phi)^i(\Delta\lambda)^k \\
s\sin((\alpha_1 + \alpha_2)/2) &= v = \Sigma b_{ik}(\Delta\phi)^i(\Delta\lambda)^k \\
\alpha_2 - \alpha_1 &= \Delta\alpha = \Sigma c_{ik}(\Delta\phi)^i(\Delta\lambda)^k
\end{aligned}
\qquad (67)
$$

and thus: $s = (u^2 + v^2)^{\frac{1}{2}}$, $\alpha_1 + \alpha_2 = 2 \arctan v/u$.
The series (67) are called Gauss series. The coefficients a_{ik}, b_{ik}, c_{ik} are functions of the latitude and their values should be taken in the mid-latitude point

$$\phi = (\phi_1 + \phi_2)/2$$

The azimut α in $u = s\cos\alpha$ and $v = s\sin\alpha$ is taken equal to $\alpha = (\alpha_1 + \alpha_2)/2$. The coefficients a_{ik}, b_{ik}, c_{ik}, f_{ik}, g_{ik}, h_{ik} can be found in any handbook on mathematical geodesy.
In two recent articles Schödlbauer published a fast elaboration of the above formulas. The main characteristic of his method is the splitting up of the coefficients a_{ik} until h_{ik} into a dominating spherical part and a much smaller ellipsoidal part, only consisting of those terms, that contain powers of the eccentricity, notably $\eta^2 = e'^2\cos^2\phi$. In agreement with this division the series (66) and (67) are separated into 2 parts.
The contribution of the spherical part can now also be taken into account by applying the simple formulas of spherical trigonometry, which means a considerable reduction of computational effort.
This elegant proceeding will now be worked out for both problems.

The direct problem.

The spherical part of the series (66) is obtained by computing the direct problem on a sphere, using the given data. The radius of the sphere should be taken equal to the mean radius of curvature R, with $R = (MN)^{\frac{1}{2}} = c/V^2$, taken in point (1). Applying the cosine- and sine-rule in the triangle (NP, 1, 2) we get, with $s = \frac{S}{R}$:

cosine rule : $\sin \phi_2 = \sin \phi_1 \cos s + \cos \phi_1 \sin s \cos \phi_1$

sine rule : $\sin \Delta\lambda = \sin s \sin \alpha_1 / \cos \phi_2$

The ellipsoidal part of f_{ik} and g_{ik} appears to be (with the same notation as (66)):

$$f_{10} = \frac{4\eta^2 - \eta^4}{8R} \qquad g_{01} = \frac{4\eta^2 - 3\eta^4}{8R\cos\phi}$$

$$f_{20} = \frac{-3\eta^2 t}{2R^2} \qquad g_{11} = \frac{-(\eta^2 - \eta^4)t}{R^2\cos\phi}$$

$$f_{12} = \frac{\eta^2(-1 + 21t^2)}{12R^3} \qquad g_{03} = \frac{\eta^2 t^2}{2R^3\cos\phi} \qquad (68)$$

$$f_{03} = \frac{\eta^2(-1 + t^2)}{2R^3} \qquad g_{21} = \frac{-\eta^2(1 + 9t^2)}{6R^3\cos\phi}$$

$$f_{04} = \frac{-\eta^2 t^3}{2R^4}$$

$$f_{22} = \frac{\eta^2(17t + 15t^3)}{12R^4}$$

$R = (MN)^{\frac{1}{2}} = c/V^2$, taken in starting point (1). This point (1) serves as Taylorpoint in the series expansion, thus all coefficients should be computed for $\phi = \phi_1$.
By applying (66) and (68) one finds the ellipsoidal contribution to $\Delta\phi$ and $\Delta\lambda$.

The inverse problem.

The inverse problem is first worked out on a sphere using the given data. It can be done by applying the simple cosine- and sine-rule. By reasons of numerical stability it seems better to use Napier's and Delambre's formulas for the half-angles. The radius of the sphere is equal to N taken in the mid-latitude point.

With $p = (\lambda_2 - \lambda_1)/2$ $\quad z_1 = \cos p \cos q$

$q = (\phi_2 - \phi_1)/2$ $\quad z_2 = -\cos p \sin q$

$r = (\phi_2 + \phi_1)/2$ $\quad n_1 = -\sin p \sin r$

$n_2 = \sin p \cos r$

We get:

$$s = 2N \arctan\left(\frac{z_2^2 + n_2^2}{z_1^2 + n_1^2}\right)^{\frac{1}{2}}$$

$$\alpha_1 = \arctan\frac{z_2}{n_2} - \arctan\frac{z_1}{n_1}$$

$$\alpha_2 = \arctan\frac{z_2}{n_2} + \arctan\frac{z_1}{n_1} - \pi \qquad (69)$$

$$\alpha = \frac{\alpha_1 + \alpha_2}{2}$$

$$\Delta\alpha = \alpha_2 - \alpha_1$$

$$u = s \cos\alpha \; ; \quad v = s \sin\alpha$$

The ellipsoidal part of a_{ik}, b_{ik} and c_{ik} appears to be (with the same notation as (67)):

$$a_{10} = -N(\eta^2 - \eta^4) \qquad b_{21} = \frac{-N\cos\phi}{24}(\eta^2 + 9\eta^2 t^2)$$

$$a_{30} = \frac{-N}{8}(-\eta^2 + \eta^2 t^2) \qquad c_{21} = \frac{\cos\phi\, t}{12}\eta^2 \qquad (70)$$

$$a_{12} = \frac{N\cos^2\phi}{8}\eta^2 t^2 \qquad c_{03} = \frac{\cos^3\phi\, t}{12}\eta^2$$

These values should be taken in the mid-latitude point thus for $\phi = (\phi_1 + \phi_2)/2$. By applying (67) and (70) one finds the ellipsoidal part in u, v and $\Delta\alpha$.

SPHERICAL PART $\quad \eta^2 \equiv e'^2 \cos^2\bar{\varphi} \quad ; \; t = \tan\bar{\varphi}$

ϕ_2	$55^\circ0'0"$	λ_2	$10^\circ0'0"$	c	6398786.85	e'^2	0.006719219
ϕ_1	$45^\circ0'0"$	λ_1	$0^\circ0'0"$	$\bar{\phi}$	$50^\circ0'0"$	η^2	0.002776219
$\Delta\phi$	$10^\circ0'0"$	$\Delta\lambda$	$10^\circ0'0"$	t^2	1.42027663	N	6389923.08

$z_1 = \cos\frac{1}{2}\Delta\lambda \cos\frac{1}{2}\Delta\phi$	0.992403877	$z_2 = -\cos\frac{1}{2}\Delta\lambda \sin\frac{1}{2}\Delta\phi$	-0.086824089
$n_1 = -\sin\frac{1}{2}\Delta\lambda \sin\phi$	-0.066765172	$n_2 = \sin\frac{1}{2}\Delta\lambda \cos\phi$	0.056022632

$\alpha_1 = \arctan\frac{z_2}{n_2} - \arctan\frac{z_1}{n_1}$	$28^\circ9830025$	$\alpha = \frac{1}{2}(\alpha_1 + \alpha_2)$	$32^\circ8318465$
$\alpha_2 = \arctan\frac{z_2}{n_2} + \arctan\frac{z_1}{n_1} - \pi$	$36^\circ6806888$	$u = s\cos\alpha$	1111582.57
$s = 2N \arctan\left(\frac{z_2^2 + n_2^2}{z_1^2 + n_1^2}\right)^{\frac{1}{2}}$	1322894.63	$v = s\sin\alpha$	717240.82
		$\Delta\alpha = \alpha_2 - \alpha_1$	$7^\circ6976863$

ELLIPSOIDAL PART

a_{10}	-17690.576 m/rad	$a_{10}\Delta\phi$	-3087.59 m		
a_{30}	- 931.954 m/rad	$a_{30}\Delta\phi^3$	- 4.95 m		
a_{12}	1301.270 m/rad	$a_{12}\Delta\phi\Delta\lambda^2$	6.92 m		
		Δu	-3085.62 m	$u = u + \Delta u$	1108496.95 m
b_{21}	- 6548.371 m/rad	$b_{21}\Delta\phi^2\Delta\lambda$	- 34.81 m		
		Δv	- 34.81 m	$v = v + \Delta v$	717206.01 m
c_{21}	$0.00017723\ \text{rad}^{-2}$	$c_{21}\Delta\phi^2\Delta\lambda$	$0^\circ.0000538$		
c_{03}	$0.00007323\ \text{rad}^{-2}$	$c_{03}\Delta\lambda^3$	$0^\circ.0000223$		
		$\Delta\Delta\alpha$	$0^\circ.0000761$	$\Delta\alpha = \Delta\alpha + \Delta\Delta\alpha$	$7^\circ6977624$

EXACT: s = 1320284.31 m

$\alpha_1 = 29^\circ03'15"459$

$\alpha_2 = 36^\circ45'07"400$

$s = (u^2 + v^2)^{\frac{1}{2}}$	1320284.04 m
$\alpha = \arctan\frac{v}{u}$	$32^\circ9031762$
$\Delta\alpha =$	$7^\circ6977624$
$\alpha_1 = \alpha - \frac{1}{2}\Delta\alpha$	$29^\circ0542951$ $29^\circ03'15"462$
$\alpha_2 = \alpha + \frac{1}{2}\Delta\alpha$	$36^\circ7520574$ $36^\circ45'07"406$

SPHERICAL PART $t = \tan \bar{\phi}$, $\eta = e'.\cos \bar{\phi}$, $\bar{\phi} = (\phi_1 + \phi_2)/2$

ϕ_2 ϕ_1 $\Delta\phi$		λ_2 λ_1 $\Delta\lambda$		c $\bar{\phi}$ t^2		e'^2 η^2 N	

$z_1 = \cos \frac{1}{2}\Delta\lambda \cos \frac{1}{2}\Delta\phi$ $n_1 = -\sin \frac{1}{2}\Delta\lambda \sin \phi$		$z_2 = -\cos \frac{1}{2}\Delta\lambda \sin \frac{1}{2}\Delta\phi$ $n_2 = \sin \frac{1}{2}\Delta\lambda \cos \phi$	

$\alpha_1 = \arctan \frac{z_2}{n_2} - \arctan \frac{z_1}{n_1}$ $\alpha_2 = \arctan \frac{z_2}{n_2} + \arctan \frac{z_1}{n_1} - \pi$ $s = 2N \arctan \left(\frac{z_2^2 + n_2^2}{z_1^2 + n_1^2} \right)^{\frac{1}{2}}$		$\alpha = \frac{1}{2}(\alpha_1 + \alpha_2)$ $u = s \cos \alpha$ $v = s \sin \alpha$ $\Delta\alpha = \alpha_2 - \alpha_1$	

ELLIPSOIDAL PART

a_{10} a_{30} a_{12}		$a_{10}\Delta\phi$ $a_{30}\Delta\phi^3$ $a_{12}\Delta\phi\Delta\lambda^2$			
		Δu		$u = u + \Delta u$	
b_{21}		$b_{21}\Delta\phi^2\Delta\lambda$			
		Δv		$v = v + \Delta v$	
c_{21} c_{03}		$c_{21}\Delta\phi^2\Delta\lambda$ $c_{03}\Delta\lambda^3$			
		$\Delta\Delta\alpha$		$\Delta\alpha = \Delta\alpha + \Delta\Delta\alpha$	
				$s = (u^2 + v^2)^{\frac{1}{2}}$	
				$\alpha = \arctan \frac{v}{u}$ $\Delta\alpha =$	
				$\alpha_1 = \alpha - \frac{1}{2}\Delta\alpha$	
				$\alpha_2 = \alpha + \frac{1}{2}\Delta\alpha$	

2.3.4. Integral equations for the geodesic; Bessel's method.

A second fundamental algorithm dates back to the famous German geometer Bessel.

The characteristic feature of Bessel's method is the mapping of the geodesic from the ellipsoid onto a great circle on a sphere, in such a way that corresponding points have equal reduced latitudes on both surfaces.

In this method the radius of the sphere is immaterial and that is why unit radius is used.

For the ellipsoid holds: $\tan \beta = \frac{b}{a} \tan \phi$ (cf. Ch. 1.2) (71)

For the sphere the reduced and geographic latitude are identical since $a = b$.

An implicit property of the mapping after Bessel is the equality of corresponding azimuts.

Proof: Let P be an arbitrary point of a geodesic, P_o the point of intersection with the equator and P_m the point with highest latitude, then Claireaut's formula for the geodesic yields ,see fig. 25:

on the ellipsoid : $a \cos \beta \sin \alpha = a \cos \beta_m$ (since $\sin \alpha_m = 1$)

on the unit sphere: $\cos \beta \sin A = \cos \beta_m$

hence: $A = \alpha$

Since corresponding points have equal $\{\alpha,\beta\}$, it is not necessary to introduce a different notation for these quantities, as it is needed for the arclength and the longitude. (see fig. 24).

Let $\{l,\lambda\}$ be the longitudes on the ellipsoid and sphere resp., to be reckoned from the meridian through P_o and

$\{s,\sigma\}$ the arclengths to be reckoned from P_o as well.

Bessel attacked the problem by considering the differential relation between arclengths and longitudes in corresponding points on both surfaces.

For such points hold, (Consult table 2):

on the ellipsoid : $ds \cos\alpha = aV^{-1}d\beta$

on the unit sphere : $d\sigma \cos\alpha = d\beta$

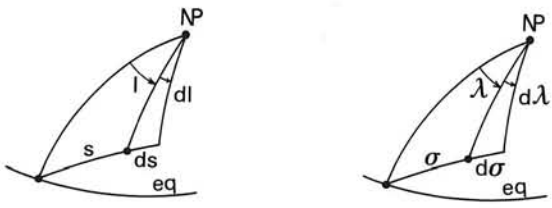
thus : $\boxed{ds = aV^{-1}d\sigma}$ (72)

on the ellipsoid : $ds \sin\alpha = a \cos\beta \, dl$

on the unit sphere : $d\sigma \sin\alpha = \cos\beta \, d\lambda$

thus : $\boxed{dl = V^{-1}d\lambda}$ (73)

(see fig. 24)



ellipsoid sphere

Figure 24.

How to solve these differential equations?
In order to be able to perform the integration of both d.e. with respect to the same independent variable σ, Bessel expressed V as a function of the constant β_m and the variable σ, and $d\lambda$ as a function of β_m, σ and $d\sigma$.

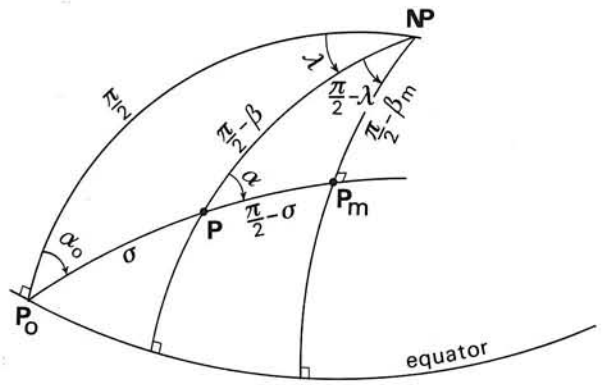


Figure 25.

Applying the formulae of the general spherical triangle (see Appendix 1) to the rectangular spherical triangle NP-P_m-P, we get for any point P:

$$\sin \beta = \sin \beta_m \sin \sigma \tag{74}$$

$$\cos \beta_m = \tan \lambda \cot \sigma \tag{75}$$

$$\cos \sigma = \cos \lambda \cos \beta \tag{76}$$

$$\cos \beta_m = \cos \beta \sin \alpha \tag{77}$$

Differentiation of (75) gives:

$$\cos^{-2}\lambda \, d\lambda = \cos \beta_m \cos^{-2}\sigma \, d\sigma$$

and in combination with (76) we get:

$$d\lambda = \cos \beta_m \cos^{-2}\beta \, d\sigma \tag{78}$$

Using appropiate expressions for V, a and e^2 from table 1 and 2, and introducing the auxiliary quantities

$$k = e' \sin \beta_m \tag{79}$$

$$c = \sin \beta_m \quad \text{and thus} \quad k/c = e' \tag{80}$$

one finds for (72) and (73) in integral form:

$$s = b\int_0^{\sigma}(1 + k^2\sin^2\sigma)^{\frac{1}{2}}\,d\sigma \qquad (81)$$

$$l = \frac{b}{a}\cos\beta_m\int_0^{\sigma}\frac{(1 + k^2\sin^2\sigma)^{\frac{1}{2}}}{1 - c^2\sin^2\sigma}\,d\sigma \qquad (82)$$

Both integrals are of the elliptic type and they can only be solved by expanding the integrands in a Taylor-series.
To that purpose the following socalled Wallis integrals are introduced, see 2.1:

$$W_{2p} = \int_0^{\sigma}\sin^{2p}\sigma\,d\sigma \qquad (p = 0,1,.....) \qquad (83)$$

$$I_{2p} = \int_0^{\sigma}\frac{\sin^{2p}\sigma\,d\sigma}{1 - c^2\sin^2\sigma} \qquad (p = 0,1,.....) \qquad (84)$$

For p = 0 one gets: $W_o = \sigma$ (85)

$$I_o = \lambda/\cos\beta_m \qquad (86)$$

Proof: $$I_o = \int_0^{\sigma}\frac{1}{1 - c^2\sin^2\sigma}\,d\sigma = \frac{1}{(1 - c^2)^{\frac{1}{2}}}\,\mathrm{atan}((1-c^2)^{\frac{1}{2}}\tan\sigma)$$

which can be checked by differentiating the result.

With $c = \sin\beta_m$ one gets $I_o = \frac{1}{\cos\beta_m}\,\mathrm{atan}(\cos\beta_m\tan\sigma)$

and using (75) one gets (86).

Applying the binomial expansion for the integrands of (81) and (82) and using the notation of (83) and (84) one gets:

for the arclength: $s = b\,\Sigma\binom{\frac{1}{2}}{p}\,k^{2p}W_{2p}$ (p=0,1,...) (87)

for the longitude: $l = \frac{b}{a}\cos\beta_m\,\Sigma\binom{\frac{1}{2}}{p}\,k^{2p}I_{2p}$ (p=0,1,...) (88)

with e.g.: $\binom{\frac{1}{2}}{3} = \frac{\frac{1}{2}(\frac{1}{2}-1)(\frac{1}{2}-2)}{3.2.1}$

By partial integration one can find recurrent relations for W_{2p} and I_{2p} (see 2.1).

$$\boxed{W_{2p} = \frac{2p-1}{2p} W_{2p-2} - \frac{1}{2p} \sin^{2p-1}\sigma \cos\sigma} \qquad (89)$$

$$I_{2p+2} = \frac{I_{2p} - W_{2p}}{c^2}$$

and by successive substitution :

$$I_{2p+2} = \frac{I_o}{c^{2p+2}} - \frac{W_o}{c^{2p+2}} - \frac{W_2}{c^{2p}} - \frac{W_4}{c^{2p-2}} - \ldots \frac{W_{2p}}{c^2} \qquad (90)$$

It was Levallois, 1952, who first introduced the above recurrent relations in Bessel's method.
Substituting (90) , for p=0,1,...., into (88) and collecting terms with I_o, W_o, W_2, ... one gets:

$$\begin{aligned} 1 = \frac{b}{a} \cos\beta\{ & I_o\left(1 + \left(\tfrac{1}{2}\right)\frac{k^2}{c^2} + \left(\tfrac{1}{2}\right)\frac{k^4}{c^4} + \left(\tfrac{1}{3}\right)\frac{k^6}{c^6} + ..\right) \\ & -W_o\left(+ \left(\tfrac{1}{2}\right)\frac{k^2}{c^2} + \left(\tfrac{1}{2}\right)\frac{k^4}{c^4} + \left(\tfrac{1}{3}\right)\frac{k^6}{c^6} + \right) \\ & -c^2 W_2\left(+ \left(\tfrac{1}{2}\right)\frac{k^4}{c^4} + \left(\tfrac{1}{3}\right)\frac{k^6}{c^6} + \right) \\ & -c^4 W_4\left(+ \left(\tfrac{1}{3}\right)\frac{k^6}{c^6} + ..\right)\} \end{aligned} \qquad (91)$$

From table I in section I it can be proved that

$$\frac{a}{b} = (1 + e'^2)^{\frac{1}{2}} = 1 + \left(\tfrac{1}{2}\right)e'^2 + \left(\tfrac{1}{2}\right)e'^4 + \left(\tfrac{1}{3}\right)e'^6 + \ldots \qquad (92)$$

By substituting (90), (86) and (92) into (91), we get:

$$1 = \lambda + \cos\beta_m(B_o W_o + B_2 W_2 + B_4 W_4 + \ldots)$$

with
$$B_o = \{\frac{b}{a} - 1\}$$
$$B_2 = \{\frac{b}{a}(1 + \left(\tfrac{1}{2}\right)e'^2) - 1\}c^2$$
$$B_4 = \{\frac{b}{a}(1 + \left(\tfrac{1}{2}\right)e'^2 + \left(\tfrac{1}{2}\right)e'^4) - 1\}c^4$$
$$B_6 = \{\frac{b}{a}(1 + \left(\tfrac{1}{2}\right)e'^2 + \left(\tfrac{1}{2}\right)e'^4 + \left(\tfrac{1}{3}\right)e'^6) - 1\}c^6$$
$$B_8 = \ldots\ldots\ldots\ldots$$

Hence the integration of (72) and (73) has finally led to

$$\boxed{\begin{aligned} s &= b\,(\sigma + A_2W_2 + A_4W_4 + A_6W_6 + \ldots\ldots\ ) \\ l &= \lambda + \cos\beta_m(B_0W_0 + B_2W_2 + B_4W_4 + \ldots.) \end{aligned}} \qquad (93)$$

with $A_{2p} = (\frac{1}{2}_p)\ k^{2p}$

and the recurrent relations:

$$\boxed{A_{2p} = \frac{3/2 - p}{p}\ k^2 A_{2p-2}} \quad (p=1,2,3,\ldots\ ) \qquad (94)$$

with $A_o = 1$

$$\boxed{B_{2p} = c^2 B_{2p-2} + \frac{b}{a} A_{2p}} \qquad (95)$$

with $B_o = \frac{b}{a} - 1$

Bessel's method has two main advantages.
First, the series (93) that give the solution, converge very rapidly, and second, the factors A_n, B_n and W_n in the separate terms of both series can be expressed in a recurrent form. That is why a versatile and efficient computerprogram can be developed, suitable for any length of the geodesic and any demand of accuracy. The running of the program is ruled by a tolerance factor ε that reflects the accuracy that is desired.

In geodesy and hydrography there are many variants based on Bessel's method. They are known by the names of Rainsford, Vincenty, Helmert, Jordan, Sodano a.o..
They make no use of recurrent relations and are distinguished by different prescribed accuracylevels and prevailing lengths of the geodesics.

THE IMPLEMENTATION OF BESSEL'S METHOD FOR THE DIRECT AND INVERSE PROBLEM

Characteristic for both problems is the computation of the boundary values σ_1 and σ_2 in the evaluation of the Wallis-integrals. The key formulas (93) are written in the following form:

$$\sigma = \frac{s}{b} - u \quad \text{with} \quad u = \Sigma A_{2p} W_{2p} \quad , p = 1,2,.. \tag{96}$$

$$\lambda = l - v \quad \text{with} \quad v = \cos \beta_m \, \Sigma B_{2p} W_{2p} \quad , p = 0,1.. \tag{97}$$

u and v are small in comparison to the first terms on the right hand side

Let $l(G)$, $\lambda(G)$ be the geographic longitude on the ellipsoid and sphere with respect to Greenwich meridian, then we have the following two cases:

THE DIRECT PROBLEM : given ϕ_1, $l_1(G)$, α_1, $\Delta s = s_2 - s_1$

find ϕ_2, $l_2(G)$, α_2

Apply the following formulas of spherical trigonometry, taking $P = P_1$:

- compute β_m with successively $\tan \beta_1 = \frac{b}{a} \tan \phi$ (71)
  $\cos \beta_m = \cos \beta_1 \sin \alpha_1$ (77)
- compute σ_1 with $\sin \sigma_1 = \sin \beta_1 / \sin \beta_m$ (74)
- compute s_1 with (93) and s_2 with $s_2 = s_1 + \Delta s$

Now σ_2 is computed from s_2 by applying an iteration process on (96), according to the following scheme:

$\sigma_2(0) = s_2/b$
$k = 1$
$u(0) = 0$
loop : $\sigma_2(k) = \sigma_2(k-1) - u(k-1)$
$u(k) = \Sigma A_{2p} W_{2p}$ (with argument $\sigma_2(k)$)
if $|u(k) - u(k-1)| < \varepsilon$, then goto end
$k = k+1$
go to loop
end : $\sigma_2 = \sigma_2(k)$

- compute β_2, ϕ_2 and α_2 using (74), (71) en (77):

$$\sin \beta_2 = \sin \beta_m \sin \sigma_2$$

$$\tan \phi_2 = \frac{a}{b} \tan \beta_2$$

$$\sin \alpha_2 = \cos \beta_m \,/\, \cos \beta_2$$

- compute ,for i=1,2, using (75) or (76), λ_i:

$$\tan \lambda_i = \cos \beta_m \tan \sigma_i$$

$$\cos \lambda_i = \cos \sigma_i \,/\, \cos \beta_i$$

- compute ,for i=1,2, using (93), l_i.
- compute $l_2(G) = l_1(G) + (l_2 - l_1)$

THE INVERSE PROBLEM: given ϕ_1, ϕ_2 and Δl $(=l_2 - l_1)$
find α_1, α_2 and Δs $(=s_2 - s_1)$

- initialize $\varepsilon = 10^{-8}$; $k = 0$; $\Delta v(k) = 0$
- loop : $\Delta\lambda(k) = \Delta l - \Delta v(k)$

evaluate on unit sphere, starting from known β_1, β_2, $\Delta\lambda(k)$, and using appropiate formulae of the spherical triangle : $\Delta\sigma$, $(= \sigma_2 - \sigma_1)$, α_1, α_2, β_m, σ_1 and σ_2 $(= \sigma_1 + \Delta\sigma)$.

$k = k + 1$

evaluate with series (97) and using arguments σ_1 and σ_2 : $\Delta v(k)$ $(= v_2 - v_1)$.

if $|v(k) - v(k-1)| < \varepsilon$ goto finish, else goto loop.

finish : evaluate with series (96), and using arguments σ_1 and σ_2 : s_2 and s_1 and from them Δs.
print α_1, α_2 and Δs

2.4 Literature.

(a rather arbitrary selection).

H. Wolf Die Grundgleichungen der dreidimensionale Geodäsie in elementarer Darstellung. ZfV 1963 S.225

Ordnance Survey Professional papers no 30 , on datumtransformation in the North Sea area

J.C.Blankenburg Doppler-European datum transformation parameters for the North Sea Phil. Trans. R. Soc. London A 294

T.O.Seppelin, World Geodetic System 1972 May 1974

Jordan Eggert Handbuch der Vermessungskunde Teil 4, 2er Halbband

P. Thomas Conformal projections in geodesy and cartography C. and Geod. Survey publ. 251, 1952

P. Richardus Map projections 1972

J.J.Levallois Note sur le calcul des grandes geodesiques. Publication of the I.G.N. 1952

E. Dorrer Direkte numerische Lösung der geodätische Hauptaufgaben. DGK Reihe C nr 90.

A.Schödlbauer Übertragung geografischer Koordinaten auf Bezugsellipsoiden AVN 4/1979 and 1980

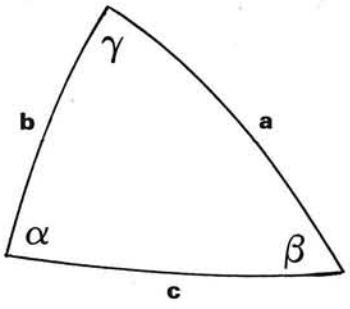
E.Sodano General non-iterative solution of the inverse and direct geodetic problems Bull Geod. no 75 1965

L.Pfeifer The use of Bowrings Algorithms for Hydrography and Navigation. The Hydrographic Journal no 31 1984.

B.R. Bowring The direct and inverse problem for short geodesics on the ellipsoid. Surveying and Mapping vol 41.

2.5 Appendices.

APPENDIX 1 SOME FORMULAE FOR THE SPHERICAL TRIANGLE.



Law of sines

$$\frac{\sin a}{\sin \alpha} = \frac{\sin b}{\sin \beta} = \frac{\sin c}{\sin \gamma}.$$

Law of cosines for sides

$$\cos a = \cos b \cos c + \sin a \sin b \cos \alpha$$
$$\cos b = \cos c \cos a + \sin c \sin a \cos \beta$$
$$\cos c = \cos a \cos b + \sin a \sin b \cos \gamma$$

Law of cosines for angles

$$\cos \alpha = -\cos \beta \cos \gamma + \sin \beta \sin \gamma \cos a$$
$$\cos \beta = -\cos \gamma \cos \alpha + \sin \gamma \sin \alpha \cos b$$
$$\cos \gamma = -\cos \alpha \cos \beta + \sin \alpha \sin \beta \cos c$$

For small values of a,b,c or α,β,γ both laws of cosines become inaccurate. They can then better be replaced by the following formulae.

$$\sin^2\tfrac{1}{2}a = \sin^2\tfrac{1}{2}(b - c) + \sin b \sin c \sin^2\tfrac{1}{2}\alpha$$
$$\sin^2\tfrac{1}{2}b = \sin^2\tfrac{1}{2}(c - a) + \sin c \sin a \sin^2\tfrac{1}{2}\beta$$
$$\sin^2\tfrac{1}{2}c = \sin^2\tfrac{1}{2}(a - b) + \sin a \sin b \sin^2\tfrac{1}{2}\gamma$$

$$\sin^2\tfrac{1}{2}\alpha = \cos^2\tfrac{1}{2}(\beta + \gamma) + \sin \beta \sin \gamma \sin^2\tfrac{1}{2}a$$
$$\sin^2\tfrac{1}{2}\beta = \cos^2\tfrac{1}{2}(\gamma + \alpha) + \sin \gamma \sin \alpha \sin^2\tfrac{1}{2}b$$
$$\sin^2\tfrac{1}{2}\gamma = \cos^2\tfrac{1}{2}(\alpha + \beta) + \sin \alpha \sin \beta \sin^2\tfrac{1}{2}c$$

The Five-Argument Formulae

$$\sin a \cos \beta = \cos b \sin c - \sin b \cos c \cos \alpha$$
$$\sin b \cos \gamma = \cos c \sin a - \sin c \cos a \cos \beta$$
$$\sin c \cos \alpha = \cos a \sin b - \sin a \cos b \cos \gamma$$

$$\sin a \cos \gamma = \cos c \sin b - \sin c \cos b \cos \alpha$$
$$\sin b \cos \alpha = \cos a \sin c - \sin a \cos c \cos \beta$$
$$\sin c \cos \beta = \cos b \sin a - \sin b \cos a \cos \gamma$$

$$\sin \alpha \cos b = \cos \beta \sin \gamma + \sin \beta \cos \gamma \cos a$$
$$\sin \beta \cos c = \cos \gamma \sin \alpha + \sin \gamma \cos \alpha \cos b$$
$$\sin \gamma \cos a = \cos \alpha \sin \beta + \sin \alpha \cos \beta \cos c$$

$$\sin \alpha \cos c = \cos \gamma \sin \beta + \sin \gamma \cos \beta \cos a$$
$$\sin \beta \cos a = \cos \alpha \sin \gamma + \sin \alpha \cos \gamma \cos b$$
$$\sin \gamma \cos b = \cos \beta \sin \alpha + \sin \beta \cos \alpha \cos c$$

The Four-Argument Formulae

$$\sin \alpha \operatorname{ctg} \beta = \operatorname{ctg} b \sin c - \cos c \cos \alpha$$
$$\sin \alpha \operatorname{ctg} \gamma = \operatorname{ctg} c \sin b - \cos b \cos \alpha$$
$$\sin \beta \operatorname{ctg} \gamma = \operatorname{ctg} c \sin a - \cos a \cos \beta$$
$$\sin \beta \operatorname{ctg} \alpha = \operatorname{ctg} a \sin c - \cos c \cos \beta$$
$$\sin \gamma \operatorname{ctg} \alpha = \operatorname{ctg} a \sin b - \cos b \cos \gamma$$
$$\sin \gamma \operatorname{ctg} \beta = \operatorname{ctg} b \sin a - \cos a \cos \gamma$$

Delambre's Formulae

$$\sin\frac{\gamma}{2}\sin\frac{a+b}{2} = \sin\frac{c}{2}\cos\frac{\alpha-\beta}{2}$$
$$\sin\frac{\gamma}{2}\cos\frac{a+b}{2} = \cos\frac{c}{2}\cos\frac{\alpha+\beta}{2}$$
$$\cos\frac{\gamma}{2}\sin\frac{a-b}{2} = \sin\frac{c}{2}\sin\frac{\alpha-\beta}{2}$$
$$\cos\frac{\gamma}{2}\cos\frac{a-b}{2} = \cos\frac{c}{2}\sin\frac{\alpha+\beta}{2}$$

$$\sin\frac{\alpha}{2}\sin\frac{b+c}{2} = \sin\frac{a}{2}\cos\frac{\beta-\gamma}{2}$$
$$\sin\frac{\alpha}{2}\cos\frac{b+c}{2} = \cos\frac{a}{2}\cos\frac{\beta+\gamma}{2}$$
$$\cos\frac{\alpha}{2}\sin\frac{b-c}{2} = \sin\frac{a}{2}\sin\frac{\beta-\gamma}{2}$$
$$\cos\frac{\alpha}{2}\cos\frac{b-c}{2} = \cos\frac{a}{2}\sin\frac{\beta+\gamma}{2}$$

$$\sin\frac{\beta}{2}\sin\frac{c+a}{2} = \sin\frac{b}{2}\cos\frac{\gamma-\alpha}{2}$$
$$\sin\frac{\beta}{2}\cos\frac{c+a}{2} = \cos\frac{b}{2}\cos\frac{\gamma+\alpha}{2}$$
$$\cos\frac{\beta}{2}\sin\frac{c-a}{2} = \sin\frac{b}{2}\sin\frac{\gamma-\alpha}{2}$$
$$\cos\frac{\beta}{2}\cos\frac{c-a}{2} = \cos\frac{b}{2}\sin\frac{\gamma+\alpha}{2}$$

Napier's Analogies

$$\operatorname{tg}\frac{a+b}{2} = \operatorname{tg}\frac{c}{2}\,\frac{\cos\frac{\alpha-\beta}{2}}{\cos\frac{\alpha+\beta}{2}}$$

$$\operatorname{tg}\frac{a-b}{2} = \operatorname{tg}\frac{c}{2}\,\frac{\sin\frac{\alpha-\beta}{2}}{\sin\frac{\alpha+\beta}{2}}$$

$$\operatorname{tg}\frac{b+c}{2} = \operatorname{tg}\frac{a}{2}\,\frac{\cos\frac{\beta-\gamma}{2}}{\cos\frac{\beta+\gamma}{2}}$$

$$\operatorname{tg}\frac{b-c}{2} = \operatorname{tg}\frac{a}{2}\,\frac{\sin\frac{\beta-\gamma}{2}}{\sin\frac{\beta+\gamma}{2}}$$

$$\operatorname{tg}\frac{c+a}{2} = \operatorname{tg}\frac{b}{2}\,\frac{\cos\frac{\gamma-\alpha}{2}}{\cos\frac{\gamma+\alpha}{2}}$$

$$\operatorname{tg}\frac{c-a}{2} = \operatorname{tg}\frac{b}{2}\,\frac{\sin\frac{\gamma-\alpha}{2}}{\sin\frac{\gamma+\alpha}{2}}$$

$$\operatorname{tg}\frac{\alpha+\beta}{2} = \operatorname{ctg}\frac{\gamma}{2}\,\frac{\cos\frac{a-b}{2}}{\cos\frac{a+b}{2}}$$

$$\operatorname{tg}\frac{\alpha-\beta}{2} = \operatorname{ctg}\frac{\gamma}{2}\,\frac{\sin\frac{a-b}{2}}{\sin\frac{a+b}{2}}$$

$$\operatorname{tg}\frac{\beta+\gamma}{2} = \operatorname{ctg}\frac{\alpha}{2}\,\frac{\cos\frac{b-c}{2}}{\cos\frac{b+c}{2}}$$

$$\operatorname{tg}\frac{\beta-\gamma}{2} = \operatorname{ctg}\frac{\alpha}{2}\,\frac{\sin\frac{b-c}{2}}{\sin\frac{b+c}{2}}$$

$$\operatorname{tg}\frac{\gamma+\alpha}{2} = \operatorname{ctg}\frac{\beta}{2}\,\frac{\cos\frac{c-a}{2}}{\cos\frac{c+a}{2}}$$

$$\operatorname{tg}\frac{\gamma-\alpha}{2} = \operatorname{ctg}\frac{\beta}{2}\,\frac{\sin\frac{c-a}{2}}{\sin\frac{c+a}{2}}$$

APPENDIX 2 MAP PROJECTIONS OF THE 'NORTH SEA COUNTRIES'

From : Rapport général no. 2, La neuvième assemblée générale de L'Union Geodésique et Geophysique Internationale, Bruxelles, 1951.

Country: Norway

Name of projection: Transverse Mercator

Use: Geodetic and Cartographic

Explanation: This projection is used for computation of rectangular coordinates of all triangulation in Norway, also for the modern topographical maps. The country is divided into 8 strips, each of which is referred to its central meridian, along which the scale is true.

Mapping equations:

$$X = X_0 + A_2L^2 + A_4L^4 + A_6L^6 + \ldots$$
$$Y = B_1L + B_3L^3 + B_5L^5 + \ldots\ldots$$

Where L is longitude from central meridian

For meaning of coefficients see: Jordan: "Handbuch der Vermessungskunde" Band III p 499, 1923 Edition

Spheroid: Modified Bessel $a = 6377492.018$ m, $f = \frac{1}{299.1528}$ Unit of length: Meter

Constants (including scale reduction and origin): The axes correspond to the following 8 central meridians:

4° 40' W Oslo 2° 20' " " 0° 00' " "	Origin at 58° latitude
2° 30' E "	" " 64° "
6° 10' " "	" " 66° "
10° 10' " "	" " 68° "
14° 10' " "	" " 68° "
18° 20' " "	" " 68° "

Oslo 10° 43' 22"5 E. Greenwich

Type of table (including accuracy and interval): There are two sets of tables, one giving rectangular coordinates for intervals 10' latitudes and 10' longitude referred to Oslo, the other one giving rectangular coordinates for intervals 2' latitude and 5' longitude referred to Greenwich.

The accuracy of the former is 0.01 m and of the latter 0.1 m.

Geographical extent of table:

Latitude 58°-64°	Range of longitude from axis 1°30'
" 64°-68°	" " " " " 2°0'
" 68°-71°	" " " " " 2°20'

Tables available upon request from: Not available

Country: Great Britain

Name of projection: Transverse Mercator

Use: Geodetic and Cartographic

Explanation: Used for all triangulation and mapping work in Great Britain. The basic triangulation is adjusted on the projection, giving plane rectangular coordinates direct.

Mapping equations: None used. All maps are grid sheets.

Spheroid: Airy

Unit of length: International metre

Constants (including scale reduction and origin):

Central Meridian	2° west of Greenwich
ϕ_o	49° north
False coordinates of origin	400,000 metres east, 100,000 metres south
Scale reduction	1/2500 approx.

(F_o = 0.9996012717. $\log_{10}$ = $\bar{1}$.99982680)

Type of table (including accuracy and interval): Tables designed for machine computations. Interval 1' of arc.

Computation	Tabulated to:
Rectangular co-ords from geographicals	0.001 metres
Geographicals from rectangular co-ords	0.0001
Convergence from rectangulars	0.001
Convergence from geographicals	0.001
Local scale factor from geographicals	0.0000001
Local scale factor from rectangulars	0.0000001
t - T (Projectional) correction	0.001

Geographical extent of tables: From 49° to 61° north and approximately 350 km E. and W of Central Meridian.

Tables available upon request from:

His Majesty's Stationery Office, London

Cost:

1. Constants, Formulae and Methods used in Transverse Mercator Projection — Price 1/-
2. Projection tables for the Transverse Mercator Projection of Great Britain — Price 4/-

Country: Germany (Deutsche Bundesrepublik)

Name of projection: Gauss-Krüger

Use: Geodäsie und Kartographie

Explanation: Diese Projektion dient der Darstellung des deutschen Reichsdreiecksnetzes (vorläufig und endgültig) und der angeschlossenen Ländernetze. Kartographisch wird sie als Grundlage für die deutsche Grundkarte 1:5,000 benutzt, und z.T. für 1:25,000.

Mapping equations: Siehe die einschlägigen Veröffentlichungen

Spheroid: Bessel Unit of length: Meter

Constants: (including scale reduction and origin)

Meridianstreifen Nr.	2	3	4
Mittelmeridian	6°	9°	12° ö.v. Gr.
X_o	0°	0°	0°
Y_o	2 500 000	3 500 000	4 500 000
Massstabsreduktion	0	0	0

Types of Tables (including accuracy and interval):

1) G. Thilo: Anweisungen und Tafeln zur Berechnung Gauss-Krügerscher Koordinaten. Berlin 1924.

2) Lips: Formeln und Tafeln zur Berechnung der ellipsoidischen, der konformen und der geographischen Koordinaten und der Rechenmaschinen. Stuttgart 1932.

3) M. K. Hristow: Die Gauss-Krügerschen Koordinaten auf dem Ellipsoid. Leipzig-Berlin 1943. Genauigkeit: 1 cm, Zur Umformung und für die beiden geodätischen Hauptaufgaben.

4) Blattneckenwerte der amtlichen deutschen Kartenwerke.

Die Tafeln sind vergriffen.

Country: Germany (Deutsche Bundesrepublik)

Name of projection: Preussische Polyeder-Projektion

Use: Kartographie

Explanation: Diese Projektion dient als Grundlage für die Kartenwerke 1:25,000; 1:50,000; 1:100,000 und teilweise 1:300,000.

Mapping equations: "Natürliche" Abbildung

Spheroid: Bessel Unit of length: Meter

Country: The Netherlands

Name of projection: Stereographic

Use: Geodetic and Cartographic

Explanation: This projection is used for local horizontal control surveys, as well as by all mapping agencies in the Netherlands. The primary triangulation is adjusted on the Bessel spheroid. The geographic positions of the primary triangulation points are first projected conformally on a sphere, with a radius equal to the mean radius of curvature $\sqrt{RH}$ of the spheroid at the central point. The geographic positions on the sphere are transformed to plane coordinates by stereographic projection.

Mapping equations:

ϕ Latitude) λ Longitude) on the spheroid

Ψ Latitude) l Longitude) on the sphere

Index 0 refers to the central point

$$l - l_0 = (\lambda - \lambda_0) + [1](\lambda - \lambda_0)$$

$$\Psi - \Psi_0 = (\phi - \phi_0) + [2](\phi - \phi_0) + [3](\phi - \phi_0)^2 + [4](\phi - \phi_0)^3$$

$$k_1 = [9] \frac{\cos\Psi}{\sin\frac{1}{2}(\Psi + \Psi_0)\cos\frac{1}{2}(\Psi - \Psi_0)}$$

$$\operatorname{tg}\tfrac{1}{2}p = \frac{\sin\frac{1}{2}(\Psi + \Psi_0)}{\cos\frac{1}{2}(\Psi - \Psi_0)} \operatorname{tg}\tfrac{1}{2}(l - l_0)$$

$$X = 2k_1 \operatorname{tg}\tfrac{1}{2}p \cos^2\tfrac{1}{2}p$$

$$Y = [11]\operatorname{tg}\tfrac{1}{2}(\Psi - \Psi_0) + X \operatorname{tg}\tfrac{1}{2}p$$

Spheroid: Bessel log a = 6.8046434636,5
log b = 6.8031892838,8

Unit of length: Using the international meter as unit of length a correction must be applied to the lengths of + log 4.2×10^{-7}

Constants (including scale reduction and origin): Central point is Amersfoort

ϕ_0 = 52° 9' 22".178 λ_0 = 5° 23' 15".500
Ψ_0 = 52° 7' 15.950 l_0 = idem

log r = 8.8050006.61 (r = radius of sphere)

log [1] = 6.667 -10
log [2] = 7.10111n-10
log [3] = 2.373 -10
log [4] = 6.275n -20
log [9] = 6.8049607
log [11] = 7.1059907

Scale reduction at central point: $10^7 \log m_0 = -400.0$

Scale reduction at an arbitrary point

$$\log m_1 = \log m_0 + 266.56 \times 10^{-10} (x_1^2 + y_1^2)$$

Tables available upon request from: Tables are not available

Literature: "De stereografische kaartprojectie in hare toepassing bij de Rijksdriehoeksmeting." Available upon request from: Rijkscommissie voor Geodesie, Delft, Holland.

Country: Belgium

Name of projection: Lambert

Use: Geodetic and Cartographic

Explanation: Used for mapping the sheets of the new map on scales 1/25,000, 1/50,000, and 1/100,000. The limits of the sheets are however the same as in our old map, for which the equivalent projection of Bonne was used. The basic triangulation is adjusted on the spheroid and geographic positions transformed to plane coordinates.

Mapping equations:

$$x = 150{,}000 + \rho \sin\theta \qquad y = 5{,}400{,}000 - \rho\cos\theta$$

With $\rho = C\left(\operatorname{tg}\frac{Z}{2}\right)^n$ $\qquad \theta = n\lambda$

$$\operatorname{tg}\frac{Z}{2} = \operatorname{tg}\left(\frac{\pi}{4} - \frac{\phi}{2}\right)\left(\frac{1 + e\sin\phi}{1 - e\sin\phi}\right)^{\frac{e}{2}}$$

ϕ, λ = latitude and longitude $\qquad e^2$ = (eccentricity)^2 = $\frac{a^2-b^2}{a^2}$

n and C are two constants computed by the formulas:

$$n = \frac{\log\cos\phi_1 - \log\cos\phi_2 - \log A_1 + \log A_2}{\log\operatorname{tg}\frac{Z_1}{2} - \log\operatorname{tg}\frac{Z_2}{2}} \qquad C = \frac{a\cos\phi_1}{A_1 n\left(\operatorname{tg}\frac{Z_1}{2}\right)^n} = \frac{a\cos\phi_2}{A_2 n\left(\operatorname{tg}\frac{Z_2}{2}\right)^n}$$

ϕ_1 ϕ_2 = latitude of the two fundamental parallels

Spheroid: International $\qquad$ Unit of length: Meter

Constants (including scale reduction and origin):

$\phi_1 = 49° 50'$

$\phi_2 = 51° 10'$

n = 0.771 642 1928

Log C = 7.063 180 0267

Type of table (including accuracy and interval): Table giving:

1) ρ as function of ϕ. Interval 1' , for $49° 23' < \phi < 51° 40'$

2) Sin θ and $2\sin^2\frac{\theta}{2}$ as function of λ. Interval 10" for $0° < \lambda < 2° 10'$

3) Values of the coefficients A,B,C,D in the formulas $\begin{cases} \Delta x = A\Delta\phi + B\Delta\lambda \\ \Delta y = C\Delta\phi + D\Delta\lambda \end{cases}$

4) Scale factor k and necessary elements to compute the convergence, θ and the "t-T correction" a) as function of ϕ and λ : intervals 1' for k and t-T, 1" for θ. b) as function of x and y: intervals 10,000 m. with the values of differences

Geographical extent of table:

$49° 23' < \phi < 51° 40'$
$-2° 10' < \lambda < 2° 10'$

The meridian of reference is that of Brussels, whose longitude = 4° 22' 04".71 East of Greenwich

Tables available upon request from:

Monsieur le Directeur Général de l'Institut
Géographique Militaire
2 Allée du Cloître
Bruxelles

Cost: Free

Country: France

Name of projection: Projections coniques conformes de Lambert (I;II;III;IV)

Use: Géodésique et cartographique

Mapping equations:

$$R = R_o \, e^{-(\mathcal{L} - Lo) \sin L_o}$$

$$\gamma = \sin L_o \times \Delta M$$

$$X = 600.000 + R \sin \gamma$$

$$Y = 200.000 + (R_o - R \cos \gamma)$$

$x^2 + y^2 = R^2$

$\mathcal{L}$ = latitude isometrique $\mathcal{L}(L)$

$$R_o = e_o \, N_o \, \text{cotg } Lo$$

Spheroid: Clarke 1880 Unit of length: mètre

a = 6.378.249^m,2
b = 6.356.515,0
α = 1/293.466

Constants (including scale reduction and origin):

	Lambert I Zone Nord	Lambert II Zone Centre	Lambert III Zone Sud	Lambert IV Corse
M_o	0^G Paris	0^G Paris	0^G Paris	0^G Paris
Parallèle Central L_o	55^G	52^G	49^G	46^G 85
sin L_o	0.76040.59656	0.72896.86274	0.69591.27966	0.67126.79323
R_o	5457.616.680	5999.695.770	6591.905.080	7053.300.180
Réduction d'échelle e_o	0.99987.73411	0.99987.74203	0.99987.74993	0.99994.47095

Type of table (including accuracy and interval):

1-Transformations en coordonnées géographiques (et vice versa) par les formules de base
L (R^2) à 0^{m}0001
R (L) au m/m|52^G50 [0,01] 57^G70| 50^G00 [0,01] 54^G50| 47^G00 [0,01] 51^G00 | 45^G80 [0,01] 47^G90
2-E (R^2)-E (L). tables de l'échelle locale à 8 décimales
3-Tables pour la réduction angulaire à la projection, entrée par L ou R^2

Geographical extent of table:
Extension des tables en latitude:

tables calculées | de 52^G50 à 57^G70 | de 50^G00 à 54^G50 | de 47^G00 à 51^G00 | de 45^G80 à 47^G90

Cette extension en Latitude est telle que chacun des ensembles géodésiques constituant la Géodesie Primordiale Française (en ce qui concerne le 1er ordre compl^{re}) puisse être calculé d'un seul tenant dans un système unique. Pour l'exploitation en 2e ordre et Géodésie Compl^{re}, l'extension officielle des zones est définie par le tableau d'assemblage joint. Il y a un recouvrement d'une feuille au 50.000e. [≅20 Km] entre zones adjacentes. La table IV concerns exclusivement la Corse.

Tables de Passage entre zones adjacentes. Pour le passage entre les coordonnées de systèmes adjacents dans une zone de recouvrement de 80 Km de hauteur-Formules de transformation conforme de degré deux. Intervalle de table: carré 20 Km. Précision: 0m,01; 4 Tables Lambert I ⇄ Lambert II; Lambert II ⇄ Lambert III.

Tables available:
Tables manuscrites: Peuvent être obtenues en tirage photocopié. Adresser les demandes au Bureau Technique de la 2e Direction de l'Institut Géographique National, 114, Av. Kléber, PARIS

Appendix 3.

Transformation from X, Y (RD) to Easting, Northing (UTM) for zone 31.

$$E = 663395.607 + (X - 155000) + \Delta X$$
$$N = 5781194.380 + (Y - 463000) + \Delta Y$$

$$\Delta X = A.P - B.Q + C.R - D.S + E.T - F.U + G.V - H.W$$
$$\Delta Y = B.P + A.Q + D.R + C.S + F.T + E.U + H.V + G.W$$

$x = (X - 155000).10^{-5}$
$P = x$
$R = P.x - Q.y$
$T = R.x - S.y$
$V = T.x - U.y$

$y = (Y - 463000).10^{-5}$
$Q = y$
$S = P.y + Q.x$
$U = R.y + S.x$
$W = T.y + U.x$

$A = -51.681$
$C = +20.172$
$E = + 2.075$
$G = + 0.075$

$B = +3290.525$
$D = + 1.133$
$F = + 0.251$
$H = - 0.012$

Transformation from Easting, Northing (UTM) to X, Y (RD) for zone 31.

$$X = E - 663395.607 + 155000 - \Delta X$$
$$Y = N - 5781194.380 + 463000 - \Delta Y$$

$$\Delta X = A.P - B.Q + C.R - D.S + E.T - F.U + G.V - H.W$$
$$\Delta Y = B.P + A.Q + D.R + C.S + F.T + E.U + H.V + G.W$$

$x = (E - 663395.607).10^{-5}$
$P = x$
$R = P.x - Q.y$
$T = R.x - S.y$
$V = T.x - U.y$

$y = (N - 5781194.380).10^{-5}$
$Q = y$
$S = P.y + Q.x$
$U = R.y + S.x$
$W = T.y + U.x$

$A = +56.619$
$C = +20.184$
$E = + 2.082$
$G = + 0.070$

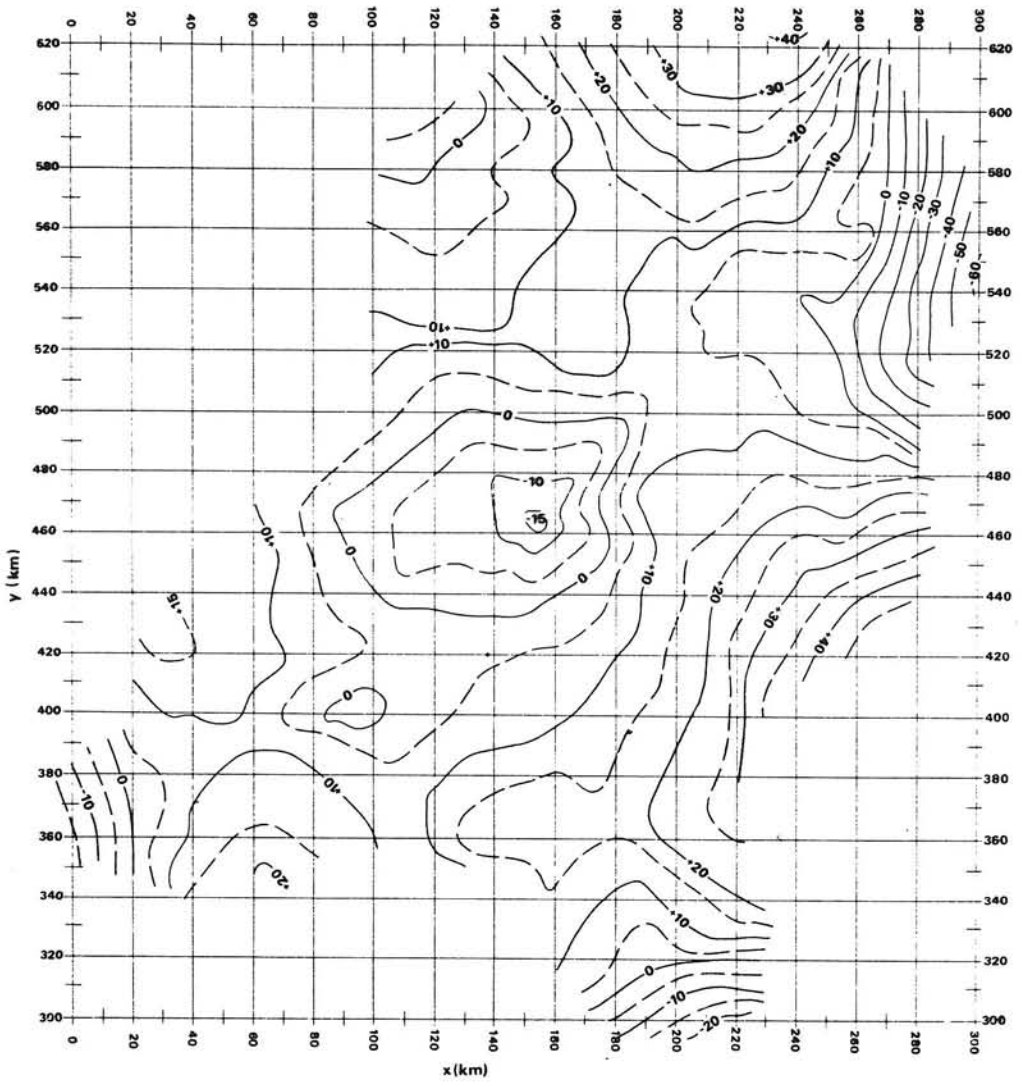
$B = +3290.362$
$D = - 0.861$
$F = - 0.023$
$H = - 0.025$

U.T.M.- CORRECTIETABEL VOOR ΔN (cm)

N = N' + ΔN

N is exacte U.T.M.- northing

N' is U.T.M.- northing volgens
4e graads transformatieformule



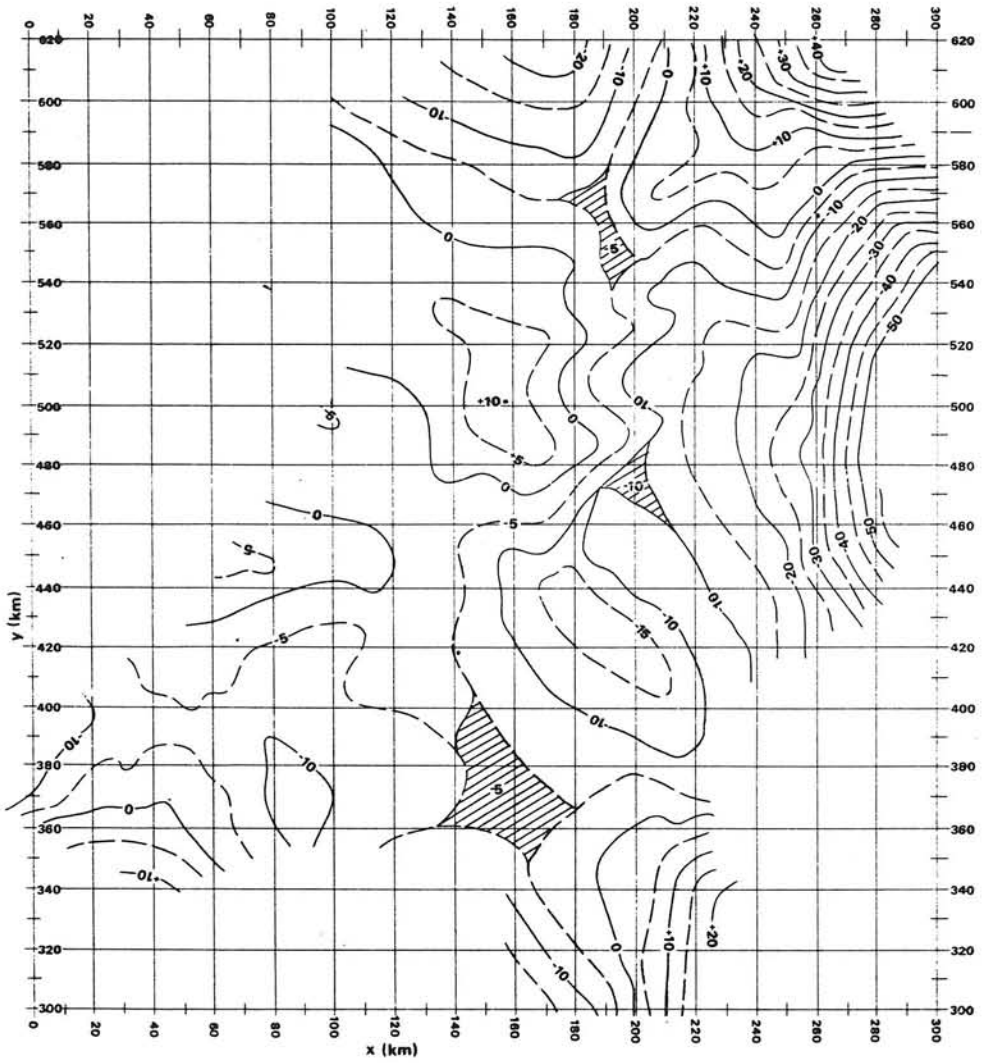
U.T.M.- CORRECTIETABEL VOOR ΔE (cm)

E=E'+ΔE

E is exacte U.T.M.-easting

E' is U.T.M.- easting volgens

4e graads transformatieformule



APPENDIX 4 TRANSFORMATION EQUATIONS FROM ED '50 TO WGS '84.

In order to obtain a better overall fit for Western Europe, regression equations are derived instead of a mean datum shift for the whole continent. When GPS becomes fully operational, WGS '84 will be its reference frame.

$$\Delta\phi = -2.65468 + 2.05138\ u + 0.42259\ v + 0.48984\ u^2 - 0.73355\ v^2 + \\ +0.90920\ u^3 + 2.53147\ u^2v - 0.72020\ uv^2 + 3.86471\ uv^3 - 0.59211\ u^5 + \\ -4.04947\ u^4v - 3.38467\ uv^5 + 4.77265\ u^4v^3 + 0.54156\ v^8 - 7.56917\ u^6v^9$$

$$\Delta\lambda = -4.45500 - 1.61659\ u + 1.95078\ v - 1.81975\ u^2 + 0.53202\ uv + \\ -0.65346\ u^3 - 3.67576\ uv^2 + 2.10356\ u^4 - 2.43915\ u^2v^2 + 1.38903\ v^4 + \\ +1.14509\ u^4v + 11.89961\ uv^4 - 8.48517\ uv^6 - 0.51702\ v^8 + \\ -3.31646\ u^9v^4 + 6.46701\ u^9v^9$$

$$\Delta h = +36.052 - 28.813\ u - 18.352\ v + 13.678\ u^2 + 3.769\ uv + 3.316\ v^3 + \\ +33.994\ u^7v + 5.416\ v^8 - 129.330\ u^6v^4 + 147.344\ u^6v^8$$

$\Delta\phi$ en $\Delta\lambda$ in seconds of arc ('') , h in meter (m).

ϕ = Geodetic latitude in degrees and decimal part of degree, positive North.

λ = Geodetic longitude in degrees and decimal part of degree, positive east from 0° to 180° and negative west from 0° to 180°.

h = Geodetic height.

ϕ(WGS '84) = ϕ(ED '50) + $\Delta\phi$

λ(WGS '84) = λ(ED '50) + $\Delta\lambda$

h(WGS '84) = h(ED '50) + Δh

$u = k(\phi - 52)$, $v = k(\lambda - 10)$, $k = 0.0523599$.

Testcase: $\phi = 46^\circ41'42''893$, $\lambda = 13^\circ54'54''098$.

$\Delta\phi = -3''115$, $\Delta\lambda = -3''717$, $\Delta h = 41.16$ m.

The ellipsoid is the Geodetic Reference System model of 1980 (see tabel 3)
A 2nd printing of this DMA Technical Report (30-9-87) gave slightly different results.

2.6 Exercises.

Exercise 1.

Given : The geodetic coordinates of station P in the national datum of the Netherlands.

$\varphi = 51^{o}59'13"310 = 51^{o}987031$

$\lambda = 4^{o}23'15"765 = 4^{o}387713$

$h = 30.10$ m

Find : The geodetic coordinates of P in the European datum E.D.'50.

Solution: There are two methods based on

1) the scheme on page
   and the example on page
2) the linearized equations on page and table on page

This last method is worked out below.

$$\begin{pmatrix} a\,\Delta\varphi \\ a\cos\varphi\,\Delta\lambda \\ \Delta h \end{pmatrix} = \begin{pmatrix} -0.78556 & -0.06027 & 0.61584 \\ -0.07651 & 0.99707 & 0 \\ 0.61403 & 0.04711 & 0.78787 \end{pmatrix} \begin{pmatrix} 676.52 \\ 121.57 \\ 595.27 \end{pmatrix}$$

$$+ \begin{pmatrix} 3.23838\cdot10^{-3} & 30943\cdot10^{2} \\ 0 & 0 \\ -0.99793 & 19794\cdot10^{2} \end{pmatrix} \begin{pmatrix} 990.85 \\ 48298\cdot10^{-9} \end{pmatrix}$$

$$= \begin{pmatrix} -172.18 \\ 69.46 \\ 890.13 \end{pmatrix} + \begin{pmatrix} 152.66 \\ 0 \\ -893.19 \end{pmatrix} = \begin{pmatrix} -19.52 \\ 69.46 \\ -3.06 \end{pmatrix}$$

$\Delta\varphi = -3.0617\ 10^{-6}$ rad. $= -0"631$

$\Delta\lambda = 1.7685\ 10^{-5}$ rad. $= +3"646$

$\Delta h = -3.06$ m

E.D.'50:

$\varphi = 51^{o}59'12"679$

$\lambda = 4^{o}23'19"411$

$h = 27.04$ m

Exercise 2.

Given : Points 1 and 2 on Bessel's ellipsoid.
1: $\varphi = 45^{o}0'0"$, $\lambda = 0^{o}0'0"$
2: $\varphi = 55^{o}0'0"$, $\lambda = 10^{o}0'0"$

Find : The length of the geodesic, joining both points and the azimuths in both endpoints, according to Bowring's method.
The solution according to Schödlbauer is given on page

Solution: 1) Find the parameters of the Gauss conformal projection, using $\varphi_o = 50^{o}0'0"$, $\lambda_o = 5^{o}0'0"$ as a centre of projection

2) Transform the ellipsoidal $\{\varphi,\lambda\}$ of both points into the spherical $\{\Phi,\Lambda\}$, using the isometric latitudes q and ω as intermediate quantities
3) Use the spherical formulas (represented in the upper part of the form for the computation after Schödlbauer).

Exercise 3.

Find : For the above points the length and the azimuth of the rhumbline joining both points.

Solution: The differential equations for the rhumbline are the formulae (51), (52) and (53) on page

With $dq = \frac{M}{N \cos \varphi} d\varphi$

one finds for (51): $\frac{dq}{ds} = \frac{\cos \alpha}{N \cos \varphi}$

(52) $\frac{d\lambda}{ds} = \frac{\sin \alpha}{N \cos \varphi}$

From (51) and (52) follows $\frac{dq}{d\lambda} = \cot \alpha$,
with α = constant.

Integration gives $\boxed{\alpha = \operatorname{arccot} \frac{q_2 - q_1}{\lambda_2 - \lambda_1}}$

Eq. (1) can be written as $M\, d\varphi = \cos\alpha\, ds$.

Integration gives $\int_{\varphi_1}^{\varphi_2} M\, d\varphi = \cos\alpha \int ds$

or $$\boxed{s = \frac{1}{\cos\alpha}\left\{\int_0^{\varphi_2} M\, d\varphi - \int_0^{\varphi_1} M\, d\varphi\right\}}$$

(see formulae on page

Exercise 4.

Given : The above points as well as the distances s_{13} = 1500 km and s_{23} = 800 km to a third point (both measured along the geodesic).

Find : The geographic coordinates $\{\varphi_3, \lambda_3\}$ with 0"001 accuracy.

Solution: Step 1: Find approximate coordinates $\{\varphi_3^{(1)}, \lambda_3^{(1)}\}$ by spherical computation, on a sphere with $R \approx 6380$ km.

Step 2: Find, for the approximate coordinates, the distances $s_{23}^{(1)}$ and $s_{13}^{(1)}$ by applying the exact algorithm after Bessel (2.3.4, page

Step 3: Formulate two linearized distance equations (according to
$\{\Delta\varphi^{(1)}, \Delta\lambda^{(1)}\}$.

Step 4: Return to step 2, until $|\Delta\varphi| = |\Delta\lambda|$ is less than 0"001.

3. ADJUSTMENT, TESTING AND FILTERING.

G.L. Strang van Hees

3.1 Matrices and least squares adjustment.

3.1.1. Matrices and determinants.

A matrix is an array of numbers with n rows and m columns, e.g.

$$\begin{matrix} \\ \uparrow \\ n \\ \downarrow \end{matrix} \begin{matrix} \leftarrow m \rightarrow \\ \begin{pmatrix} 2 & 4 & 3 & 2 \\ 1 & 5 & 6 & 3 \\ 2 & 4 & 8 & 1 \end{pmatrix} \end{matrix} \qquad n = 3,\ m = 4$$

An element of matrix A is a_{ij}
i = row number, j = column number.

$$\left(\begin{matrix} \downarrow i \\ \rightarrow \\ j \end{matrix} \quad a_{ij} \cdots \right)$$

A square matrix is a matrix with n = m.
A symmetric matrix is a square matrix which is symmetric with respect to the diagonal.
A diagonal matrix is a matrix with zero elements outside the diagonal.
The transpose of a matrix A is obtained by interchanging the rows and columns. It is denoted by A^* or A^T.

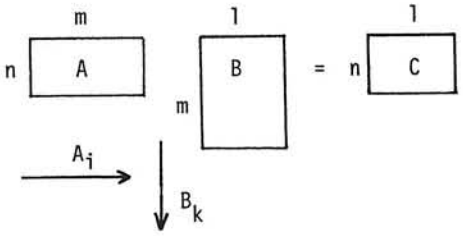
$$A = \begin{pmatrix} 2 & 4 & 6 \\ & & \\ 1 & 7 & 3 \end{pmatrix}, \qquad A^* = \begin{pmatrix} 2 & 1 \\ 4 & 7 \\ 6 & 3 \end{pmatrix}$$

A matrix with one column is called a vector and with one row a transposed vector, e.g. $x^* = (2\ 5\ 3\ 7)$.
Multiplication of matrices:

$$A_{nm} \cdot B_{ml} = C_{nl}$$

$$\sum_{j=1}^{m} a_{ij}\, b_{jk} = c_{ik}$$

(Row A_i).(column B_k) = c_{ik}



The same rule holds for vectors, e.g.:

$$x^* \cdot x = \sum_i (x_i^2)$$

Determinants.

The determinant of a square matrix is a number computed from the elements in the following way:

2 x 2 matrix A: Det. A = $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$ = ad - bc

3 x 3 matrix A:

$$\text{Det. } A = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = \underbrace{aei + bfg + cdh}_{\text{diagonal to right}} - \underbrace{ceg - bdi - afh}_{\text{diagonal to left}}$$

To compute the determinant of larger matrices we introduce subdeterminants or minors.

Each element a_{ij} is associated with a number A_{ij}, called its minor, which is the determinant of the matrix that is obtained after row i and column j have been from A.

E.g. in the 3 x 3 matrix above the minor of d is:

$$\text{minor d} = \text{Det.} \begin{vmatrix} b & c \\ h & i \end{vmatrix} = bi - ch .$$

The determinant of a n x n matrix A is the sum over one row or column of $\pm a_{ij}A_{ij}$, + if (i + j) = even, - if (i + j) is odd.

$$\text{Det. } A = \sum_{i \text{ or } j=1}^{n} (-1)^{(i+j)} a_{ij} A_{ij} \qquad (1)$$

Further $$\sum_{i=1}^{n} (-1)^{(i+k)} a_{ij} A_{ik} = 0 \quad \text{if } j \neq k \qquad (2)$$

Properties of determinants.

1. If two rows or columns are interchanged the value of the determinant is multiplied by -1 (change sign).
2. If two rows or columns are equal, the determinant is zero.
3. If a row or column is a lineair function of the other rows or columns, the determinant is zero.
4. If the elements of a row or column are multiplied by a constant, the determinant is also multiplied by this constant.
5. If the elements of a row or column are added to an other row or column, the determinant is not changed.
6. Det.(AB) = Det.(A) . Det(B).

Square matrices.

The dimension of a square matrix is its number of rows or columns.
A regular matrix is a matrix with determinant not equal zero (Det. $\neq 0$).
A singular matrix is a matrix with determinant equal zero (Det. = 0).
The rank of a matrix is the size of the largest minor with Det. $\neq 0$, or for a regular matrix the dimension of the matrix itself.
The trace of a matrix is the sum of the diagonal elements.
The unit matrix I is a diagonal matrix with all elements equal one on the diagonal.
Property: $I A = A$ and $A I = A$.

Inverse.

The inverse of a square regular matrix A, denoted by A^{-1}, is the matrix that satisfies the equations:

$$A A^{-1} = I \quad \text{and} \quad A^{-1} A = I \tag{3}$$

A regular matrix has only one unique inverse.

Computation of the inverse:
Rule of Cramer:
Suppose $A^{-1} = B^*$, then

$$b_{ij} = (-1)^{(i+j)} \frac{A_{ij}}{\text{Det. } A} \tag{4}$$

A_{ij} is the minor of a_{ij}.
A^{-1} is the transpose (!) of B (don't forget this).

This rule is practical for small matrices. For larger matrices fast computer-programs are available.

An orthogonal matrix is a square matrix for which

$$A \cdot A^* = \text{diagonal matrix.}$$

An orthonormal matrix is a square matrix for which

$$A \cdot A^* = I \quad \text{or:} \quad A^* = A^{-1}$$

The determinant of an orthonormal matrix is one (Det. A = 1).

Properties of matrices.

AB	$\neq$	BA	(unequal!)
$(AB)C$	=	$A(BC) = ABC$	
$(A+B)C$	=	$AC+BC$	
$C(A+B)$	=	$CA+CB$	
$(AB)^*$	=	$B^* A^*$	(change sequence)
$(AB)^{-1}$	=	$B^{-1}A^{-1}$	(change sequence)
$(A^{-1})^*$	=	$(A^*)^{-1}$	
$(ABC)^*$	=	$C^*B^*A^*$	
$k(A)$	=	(kA),	each element multiplied by a constant
ABA^*	=	C,	C is symmetric if B is symmetric, B and C are square.

Special case if B = I:

AA^* = C C is symmetric and square, A need not to be square. C is positive definite.

A symmetric matrix C is called positive definite if $x^* Cx > 0$ for each arbitrary vector x.

Example: $x = \begin{pmatrix} x \\ y \end{pmatrix}$, $C = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ then: $x^*Cx = ax^2 + 2bxy + cy^2$.

The diagonal elements of a positive definite matrix are all positive ($a_{ii} > 0$) and for each i and j is $a_{ii}a_{jj} > a_{ij}^2$.

C is positive definite if: $a > 0$, $c > 0$ and $ac > b^2$.

3.1.2. Linear equations.

Three linear equations with three unknowns x, y, z can be written as:

$$u = ax + by + cz$$
$$v = dx + ey + fz \quad \text{or:} \quad \begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$
$$w = gx + hy + iz$$

or: $u = Ax$ (5)

We can distinguish three cases:

1. Suppose A is a regular matrix. Solution:

$$x = ku + lv + mw$$
$$y = nu + ov + pw \quad \text{or:} \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} k & l & m \\ n & o & p \\ g & r & s \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix}$$
$$z = qu + rv + sw$$

or: $x = Bu$ (6)

substitution of (5) in (6) gives:

$$x = BAx \quad \text{or} \quad BA = I$$

so $B = A^{-1}$ (inverse) (7)

2. As second case we suppose that there are more unknowns than equations. Thus A is a $(n \times m)$ matrix with $n < m$. If there is a solution, this solution is not unique but one element of a set of solutions. The general solution is again:

$$x = Bu$$

but B satisfies the condition:

$$ABA = A \qquad (8)$$

Proof: $u = Ax = ABu = ABAx$.

B is called the general inverse of A and is a set of solutions. If A is a $n \times m$ matrix, $n < m$, and rank n, then:

$$B = PA^*(APA^*)^{-1} \quad \text{with} \quad \begin{matrix} & m \\ n & \boxed{A} \end{matrix} \qquad (9)$$

where P is an arbitrary regular m x m matrix. Check that (9) satisfies (8). (APA^*) is a regular n x n matrix and can be inverted. Each P gives an element of the set of solutions B.

3. The third case is that there are more equations than unknowns. Thus A is a n x m matrix with n > m. There is in general no solution that satisfies all equations. However, one can try to find a solution that fits as good as possible, that means it minimizes the sum of the squares of the deviations. It is called the least squares method (see adjustment).
A is a (n x m) matrix with n > m and rank m. The general inverse of A satisfies again (8), and can be computed with

$$B = (A^*PA)^{-1}A^*P \qquad \begin{matrix} & m \\ n & \boxed{A} \end{matrix} \qquad (10)$$

P is called the weight matrix that belongs to the vector u. u is often the result of a measurement. B satisfies again (8). (A^*PA) is a regular m x m matrix. The solution (6), x = Bu, does not satisfy (5) completely but gives a vector of deviation d.

$$u + d = Ax \quad \text{with} \quad x = Bu \qquad (11)$$

This solution minimizes the deviations d in the following way:

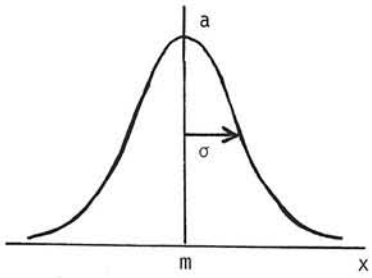
$$d^*Pd = \text{minimum.} \qquad (12)$$

If P is the unit matrix I, (12) reduces to

$$\sum_i d_i^2 = \text{minimum.}$$

3.1.3. Stochastic quantities.

In mathematics numbers are always exact, e.g. 5 means 5.000000.... However, the result of a measurement is never exact, it always is an estimation with a certain degree of uncertainty. This uncertainty can be described by a probability distribution. Usually the normal distribution or Gauss curve is assumed:



$$f(x) = a \cdot \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

m = mean value.

σ = standard deviation.

The results of a measurement should always be given as:

- estimation of the mean value;
- standard deviation.

When the standard deviation is not given, the results can hardly be interpreted, since we do not know its worth.

The square of σ is called the variance

$$\text{var}(x) = \sigma_x^2$$

Correlation.

Two different stochastic quantities can correlate, that means that a deviation in one quantity tends to change the deviation in the other quantity. The deviations are correlated. It is expressed mathematically by the covariance between two quantities x and y.

$$\text{cov}(x,y) = \sigma_{xy} = E\{(x-\bar{x})(y-\bar{y})\}$$

E stands for the Mathematical Expectation, that is the theoretical mean value over an infinite number of values x and y. $\bar{x}$, $\bar{y}$ are the mean values of x and y.

The correlation coefficient ρ is defined as:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Always is: $-1 \leq \rho \leq +1$.

$\sigma_{xy} = 0$ means no correlation.

In matrix notation the variances are also expressed as the product of two vectors.

$$\text{var}(x) = Q(x,x^*) \cdot \sigma^2$$

$$\text{cos}(x,y) = Q(x,y^*) \cdot \sigma^2$$

If x is a scalar, var(x) is a single number, but if x is a vector, var(x) is a matrix.

So the diffence between cov(x,y) and $Q(x,y^*)$ is only the proportional constant σ^2. Usually σ^2 is chosen such that the elements of Q become numbers around 1.
$Q(x,y^*)$ is called the matrix of weight coefficients, σ^2 is a proportional constant, called the variance factor.
$Q(x,y^*)$ can be written as a symbolic product of vector x times transposed vector y.

$$Q\left\{\begin{pmatrix}x_1\\x_2\end{pmatrix},\ (y_1,y_2)\right\} = \begin{pmatrix}Q(x_1,y_1) & Q(x_1,y_2)\\ Q(x_2,y_1) & Q(x_2,y_2)\end{pmatrix}$$

3.1.4. Law of variance and covariance propagation.

A function of stochastic quantities is again a stochastic quantity. The variance of this function can be expressed in the variances of the primary quantities.
The law of error propagation, or better covariance propagation consist of two steps:

1. Differentiation.
2. Multiplication.

Symbolic multiplication of two differentials gives the covariance: $dx.dy \equiv Q(x,y)$.
The best way to explain the covariance propagation is to demonstrate it by a number of examples.

a. $z = ax + by$, x,y,z stochastic; a,b constants, x,y,z are scalars
differentiate: $dz = adx + bdy$.
multiplicate : $Q(z,z) = a^2Q(x,x) + 2ab\ Q(x,y) + b^2\ Q(y,y)$.

b. $z = x \sin y$ x,y,z stochastic
differentiate: $dz = dx.\sin y + x \cos y\ dy$.
multiplicate :

$$Q(z,z) = \sin^2 y\ Q(x,x) + 2x \sin y \cos y\ Q(x,y) + x^2\cos^2 y\ Q(y,y)$$

c. $u = x \sin y$ x,y,u,v stochastic scalars
$v = x \cos y$
differentiate: $du = \sin y\ dx + x \cos y\ dy$
$dv = \cos y\ dx - x \sin y\ dy$
multiplicate:

$$Q(u,v) = \sin y \cos y\ Q(x,x) + x\ (\cos^2 y - \sin^2 y)\ Q\ (x,y) -$$

$$- x^2 \sin y \cos y\ Q(y,y)$$

d. Matrix equations.

The law of covariance propagation with matrices prescribes the multiplication of the differential equation with its transposed differential equation.

We have to realize that in matrix multiplications the sequence of the matrices may not be changed, $AB \neq BA$, and $(AB)^* = B^* A^*$. The dimensions of the matrices and vectors must of course satisfy the multiplication rule.

$y = Ax$, x, y column vectors (stochastic), A matrix non stochastic

differentiate: $dy = A\,dx$

multiplicate with transpose: $dy^* = dx^* A^*$

$$Q(y,y^*) = A \cdot Q(x,x^*)A^*$$

e. $u = Ax + By$
$v = Cx + Dy$
x, y, u, v are stochastic vectors, A,B,C,D matrices

differentiate:
$$du = A\,dx + B\,dy$$
$$du^* = dx^* A^* + dy^* B^*$$
$$dv = C\,dx + D\,dy$$
$$dv^* = dx^* C^* + dy^* D^*$$

multiplicate with transpose:

$$Q(u,u^*) = AQ(x,x^*)A^* + AQ(x,y^*)B^* + BQ(y,x^*)A^* + BQ(y,y^*)B^*$$

$$Q(u,v^*) = AQ(x,x^*)C^* + BQ(y,x^*)C^* + AQ(x,y^*)D^* + BQ(y,y^*)D^*$$

f. Example of a non linear matrix equation:

$z = x^* y$ x and y vectors, z a scalar

or $z = x_1y_1 + x_2y_2 + \cdots + x_ny_n$

differentiate: $dz = x^* .dy + dx^* .y$

As $dx^* .y$ is a 1 x 1 matrix it is equal to its transpose, so

$$dz = x^* .dy + y^* .dx$$
$$dz^* = dy^* .x + dx^* .y$$

multiplicate:

$$Q(z,z^*) = x^* Q(y,y^*)x + x^* Q(y,x^*)y + y^* Q(x,y^*)x + y^* Q(x,x^*)y.$$

3.1.5. Adjustment.

A surveyor does his measurements on land or at sea in usually difficult circumstances with many possibilities to make errors or less accurate measurements.

Therefore it is necessary to do more measurements to check the results.

It is usually impossible to go back later on and do some additional measurements, so it is very important to take enough redundant observations to check the system.

The redundant observations give the possibility to adjust the measurements with the following advantages:

1. To increase the precision of the computed unknowns.
2. To estimate the standard deviation of the observations and the unknowns.
3. To test the mathematical and stochastic model.
4. To find gross errors in the observations.
5. To compute the reliability of the system.

With the nowadays high precision instruments the increase of precision is often not the most important reason for carrying out redundant observations. The most important purpose is to detect gross-errors. In practice it turned out that gross-errors are not always found and these errors introduce large errors in the computed positions.

With statistical tests the gross-errors can be detected with a certain probability. On the other hand one can compute the effect of non detected errors on the resulting coordinates. This is called the reliability of the system. It expresses the internal check of the system.

Weights.

In the adjustment weights can be given to the observations, according to the precision of the measurements. Also correlation can be taken into account. The weight matrix P is the inverse of the variance matrix Q of the observations.

The use of a relatistic weight matrix is sometimes neglected, because the effect on the computed coordinates is rather small. However the effect on the computed precision of the coordinates is very important. A wrong weight matrix gives wrong standard deviations and the real precision may be much worse than the computed one. Also for testing on gross-errors it is important to use the correct weights. Especially it is dangerous to neglect correlation between the observations.

Types of adjustment.

It is possible to express the adjustment of the system in different ways. There are two main forms of adjustment, called:

1. adjustment with conditions;
2. adjustment with parameters.

Further all kinds of mixed forms are possible which can be reduced to the two main forms.

The adjustment with conditions is based on the conditions between the observations, e.g. if three angles of a triangle are observed, the condition is $\alpha + \beta + \gamma = 180^{o}$. However for a complex system of observations it is often difficult to find all the conditions.

The adjustment with parameters is based on the introduction of unknown parameters, e.g. the coordinates of points. All the observations are expressed as functions of the parameters.

For example, the three angles of a triangle. Introduce parameters x and y. The observations can be expressed in x and y as:

$$\alpha = x, \qquad \beta = y, \qquad \gamma = 180 - x - y$$

in matrix notation:

$$\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 180 \end{pmatrix}$$

This example shows that if we have 3 observations α, β, γ and one condition, we need $3 - 1 = 2$ parameters x, y. In general we have n observations and m parameters. The number of conditions is than $(n - m)$. This is equal to the number of redundant observations, $r = n - m$.

In the adjustment with conditions we have to invert a matrix of size (r x r), and in the adjustment with parameters a matrix of size (m x m).
As inversion is usually the most difficult part in the computation it is advantageous to use the conditions if $r < m$ and to use the parameters if $r > m$.

Notation.
Vectors are indicated by an undercast letter and matrices by an uppercast letter.

The observations are indicated by vector b ("Beobachtung") and the corrections to the observations by vector e.
The corrected observations after the adjustment are (b + e).
The unknown parameters are x.
The weight coefficient matrix of b is $Q(b,b^*) = Q$.
The weight matrix is $Q^{-1} = P$.
The variance factor is σ^2: $\text{cov}(b,b^*) = Q \cdot \sigma^2$.

In the following treatment of the adjustment it is assumed that the relations between the observations (conditions) and the expression of the parameters as function of the observations are linear equations. In this case the equations can be written as matrix equations. If the relations are non linear, they can be linearized, which will be explained later in this paper.

3.1.6. Adjustment with conditions.
The conditions between the observations b can be written in matrix form:

$$U(b + e) = u \qquad (13)$$

The corrections e are the unknowns which should be solved. U is a (r x n) matrix, with r conditions between n observations. (13) can also be written as:

$$Ue = u - Ub = t \qquad (14)$$

There are infinite solutions e that satisfy the equations (14). The least squares solution is the solution that minimizes the sum of the weighted squares of e. The weight matrix is P, so in matrix notation:

$$e^* P e = \text{minimum} \qquad (15)$$

This condition for e is sufficient to give an unique solution. Without derivation we give the result:

$$e = Q U^* (U Q U^*)^{-1} t, \text{(t follows from (14))} \qquad (16)$$

$(U Q U^*)$ is a regular (r x r) matrix, that has to be inverted. Check that (16) fulfils equation (14). If we define:

$$\boxed{N = U Q U^*} \qquad (17)$$

then: $$\boxed{e = Q U^* N^{-1} t} \qquad (18)$$

The minimum condition (15) can be written in a different form. Insert (18) in (15):

$$e^* P e = t^* N^{-1} U Q P Q U^* N^{-1} t \quad \text{(Note: } Q^* = Q \text{ and } PQ = I\text{)}$$

or: $$\boxed{e^* P e = t^* N^{-1} t} \qquad (19)$$

This is a good computational check.

Another check that should be carried out, is to insert the corrected observations (b + e) into the original conditions (13).

Covariances.

The covariances can be obtained by multiplying the weight coefficient matrices by the variance factor σ^2, e.g.:

$$\text{cov}(b, b^*) = \sigma^2 \cdot Q(b, b^*)$$

σ^2 was adopted before the adjustment, however an estimation of σ^2, called s^2, can be computed from the adjustment by:

$$\boxed{s^2 = \frac{e^* P e}{r}}$$

r is the number of redundant observations, s^2 can be compared with σ^2 to test if errors has been made (see testing procedure, chapter 3.2.3)

The weight coefficient matrices of t, e and (b + e) can be obtained by applying the law of error propagation:

$$Q(b, b^*) = Q \tag{20}$$

As u in (14) is non-stochastic, the covariance of t is equal to the covariance of (- Ub), or:

$$Q(t, t^*) = U Q U^* = N \tag{21}$$

and: $Q(b, t^*) = -Q U^*$

With (6):

$$Q(e, e^*) = Q U^* N^{-1} Q(t, t^*) N^{-1} U Q$$

or: $$\boxed{Q(e, e^*) = Q U^* N^{-1} U Q} \tag{22}$$

$$Q(b, e^*) = Q(b, t^*) N^{-1} U Q = -Q U^* N^{-1} U Q$$

So: $$Q(b, e^*) = -Q(e, e^*) \tag{22a}$$

And: $$Q(b + e, b^* + e^*) = Q(b, b^*) + Q(b, e^*) + Q(e, b^*) + Q(e, e^*)$$

or: $$\boxed{Q(b + e, b^* + e^*) = Q(b, b^*) - Q(e, e^*)} \tag{23}$$

This is a remarkable and important relation.
Insert (22) and we get the variance matrix of the corrected observations:

$$Q(b + e, b^* + e^*) = Q - Q U^* N^{-1} U Q \tag{24}$$

The second term is the improvement of the variance matrix due to the adjustment.

3.1.7. Adjustment with parameters.

The observational model is in the linear case:

$$b + e = A\,x + a \tag{25}$$

or

$$e = A\,x - (b - a) \tag{26}$$

A is a (n x m) matrix and is called the design matrix, a is a non-stochastic vector. The corrections e as well as the parameters x have to be determined. This is an underdetermined problem, with an infinite number of solutions. An unique solution is obtained if we introduce the following constraint, basic to the least squares method:

$$e^*\,P\,e = \text{minimum} \tag{27}$$

The solution is in that case:

$$x = (A^*\,P\,A)^{-1}\,A^*\,P(b - a) \tag{28}$$

let $\boxed{A^*\,P\,A = G}$

then $\boxed{x = G^{-1}\,A^*\,P(b - a).}$ (29)

Once the parameters x have been solved, the corrections e can be determined with (26).

A computational check is obtained by

$$A^*\,P\,e = 0 \tag{30}$$

which can be verified by substituting (26) and (29) for e.
The minimum condition (27) can also be written in another form:

$$\boxed{e^*\,P\,e = (b - a)^*\,P(b - a) - x^*\,G\,x} \tag{31}$$

This is also a good computational check.

The variance factor can be estimated from the adjustment by:

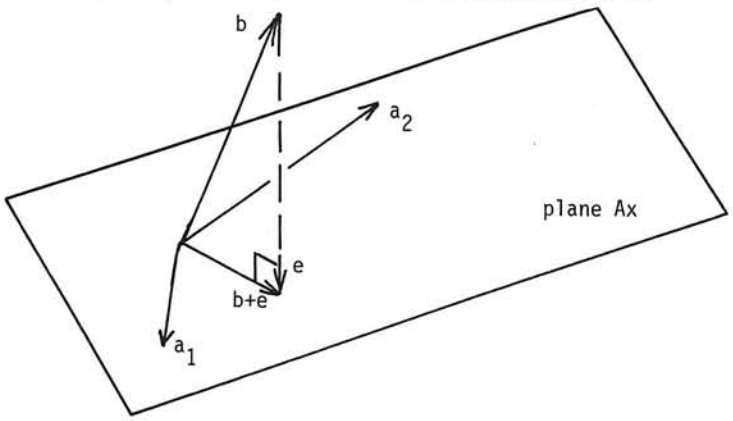
$$s^2 = \frac{e^* P e}{r}$$

where r = n - m, the number of redundant observations.

The adjustment with parameters can also be illustrated geometrically. The observation vector b is a vector with n elements and can be indicated as a vector in a n -dimensional space. The design matrix A is a (n x m) matrix. Each column is a vector with n elements in the n-dimensional space. The m column vectors span up a m-dimensional subspace. Each linear function Ax, with x an arbitrary vector, is a vector within this subspace. The corrected observation vector (b + e) is a vector in this subspace because

$$b + e = Ax \quad \text{(the constant vector a is included in b).}$$

Let us assume that the variance matrix of b, Q, and the weight matrix P are unit matrices. The geometrical concept is illustrated in the figure below.



b is a 3-element vector and A a (3 x 2) matrix, n = 3, m = 2, the two column vectors of A are a_1 and a_2. a_1 and a_2 span up a plane (2-dimensional subspace). b is a vector outside this plane, but (b + e) is a vector inside this plane. The correction vector e is the vector that connects b and (b + e). The least squares method solves the vector (b + e) such that the length of the vector e is minimum (condition (27) . Geometrically this solution can be illustrated as the perpendicular projection of b onto the plane Ax.

The vector e is perpendicular to the plane and therefore the internal (scalar) product of e with a_1 and a_2 is zero, or

$$A^* e = 0 .$$

The corrections are: $e = Ax - b$.

So: $A^* Ax = A^* b$.

$(A^* A)$ is a regular (m x m) matrix and can be inverted. The solution of x is:

$$x = (A^* A)^{-1} A^* b$$

This is the least squares solution of x. This equation is the same as equation (29) in case of P = I. This example shows how (29) can be derived. The corrected observations are

$$b + e = Ax = A(A^* A)^{-1} A^* b$$

Example:

Let: $b = \begin{pmatrix} 4 \\ 3 \\ 6 \end{pmatrix}$ and $A = \begin{pmatrix} 1 & 2 \\ 1 & 5 \\ 0 & 0 \end{pmatrix}$

As the third component of the A column vectors is zero, the plane spanned up by the column vectors of A is the x,y-plane.
The projection of b onto the x,y-plane is in this special case:

$b + e = \begin{pmatrix} 4 \\ 3 \\ 0 \end{pmatrix}$, (simply make the z component of b zero).

We will show that the least squares formulae give the same result:

$$A^* A = \begin{pmatrix} 1 & 1 & 0 \\ 2 & 5 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 1 & 5 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 7 \\ 7 & 29 \end{pmatrix}$$

$$(A^* A)^{-1} = \frac{1}{9}\begin{pmatrix} 29 & -7 \\ -7 & 2 \end{pmatrix}$$

$$x = \frac{1}{9}\begin{pmatrix} 29 & -7 \\ -7 & 2 \end{pmatrix}\begin{pmatrix} 1 & 1 & 0 \\ 2 & 5 & 0 \end{pmatrix}\begin{pmatrix} 4 \\ 3 \\ 6 \end{pmatrix} = \frac{1}{9}\begin{pmatrix} 42 \\ -3 \end{pmatrix}$$

$$b + e = Ax = \frac{1}{9}\begin{pmatrix} 1 & 2 \\ 1 & 5 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 42 \\ -3 \end{pmatrix} = \begin{pmatrix} 4 \\ 3 \\ 0 \end{pmatrix}$$

$$e = (b + e) - b = \begin{pmatrix} 4 \\ 3 \\ 0 \end{pmatrix} - \begin{pmatrix} 4 \\ 3 \\ 6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -6 \end{pmatrix}$$

Covariances.

The weight coefficient matrices of x, e and (b + e) can be obtained by the rules of variance propagation. From (29) follows:

$$Q(x, x^*) = G^{-1} A^* P\, Q(b,b^*) P\, A\, G^{-1}$$

$$= G^{-1} A^* P\, Q\, P\, A\, G^{-1} = G^{-1}$$

So: $\boxed{Q(x, x^*) = G^{-1}}$ with $G = A^*PA$. (32)

From (25) follows:

$$Q(b + e, b^* + e^*) = A\, Q(x, x^*)\, A^* = A\, G^{-1}\, A^* \quad (33)$$

Equation (23), derived for adjustment with conditions, is in general valid because the two types of adjustment are equivalent.

So: $Q(e, e^*) = Q(b, b^*) - Q(b + e, b^* + e^*)$

or: $\boxed{Q(e, e^*) = Q - A\, G^{-1}\, A^*}$ (34)

The correlation between b and x follows from (29)

$$Q(x, b^*) = G^{-1} A^* P\, Q(b, b^*) = G^{-1} A^* \quad (35)$$

and the correlation between x and e with (29)

$$Q(x, e^*) = G^{-1} A^* P\, Q(b, e^*)$$

with (22a):

$$Q(b, e^*) = -Q(e, e^*)$$

and with (34)

$$Q(x, e^*) = G^{-1} A^* P(A G^{-1} A^* - Q)$$

$$= G^{-1} A^* P A G^{-1} A^* - G^{-1} A^* P Q$$

$$= G^{-1} A^* - G^{-1} A^* = 0$$

So: $Q(x, e^*) = 0$. (36)

The corrections e are not correlated with the parameters x!

A review of the adjustment is given on the next page.

3.1.8. Standard ellipse.

The parameters in the adjustment are often coordinates. In the plane usually rectangular coordinates (x,y) or easting and northing are used. From the adjustment the covariance matrix of the parameters is obtained:

$$\text{cov}(x, x^*) = \sigma^2 . Q(x, x^*).$$

σ^2 is the variance factor.

For each point P(x,y) the covariance matrix can be written as a submatrix of $\text{cov}(x, x^*)$:

$$\text{cov}(x_p, y_p) = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

σ_x^2, σ_y^2 are the squares of the standard deviation.

The standard ellipse is an ellipse around point P that indicates the precision of the point. It can be computed as follows:

$$a^2 = \tfrac{1}{2}(\sigma_x^2 + \sigma_y^2) + \tfrac{1}{2}\sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4\,\sigma_{xy}^2}$$

$$b^2 = \tfrac{1}{2}(\sigma_x^2 + \sigma_y^2) - \tfrac{1}{2}\sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4\,\sigma_{xy}^2} \qquad (37)$$

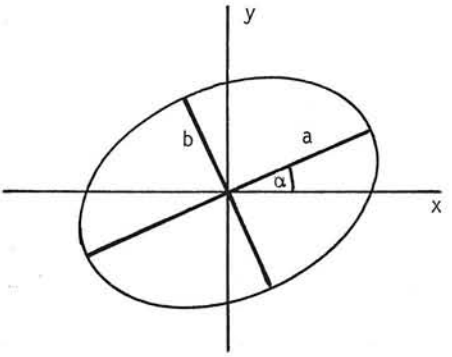
$$\tan 2\alpha = \frac{2\sigma_{xy}}{\sigma_x^2 - \sigma_y^2}$$

Summary of all formulas.

ADJUSTMENT

Conditions	Parameters
b: observations e: corrections P: weightmatrix = Q^{-1}, Q = Q(b,b) := means: equal by definition	b : observations, number: n x : unknown parameters, number: m $\sigma_b^2 = Q \cdot \sigma^2$ = variance
U(b+e) := u N := UQU^* M :=$=N^{-1}$ t := u-Ub k := Mt R := QU^*MUQ e = $QU^*k = QU^*Mt$ e = $QU^*M(u-Ub)$ e = -RPb+v (v=QU^*Mu) b+e = (I-RP)b+v Ue = t E = $e^*Pe=t^*k=k^*t=t^*Mt$ s^2 = E/(n-m) F = s^2/σ^2 RPR = R PRP = U*MU	b+e := Ax+a G := A^*PA H := G^{-1} d := $A^*P(b-a)$ x = Hd = $HA^*P(b-a)$ R := Q-AHA^* e = Ax-(b-a) e = $(AHA^*P-I)(b-a)$ e = -RP(b-a) b+e = $AHA^*P(b-a)+a$ A*Pe = 0 E = $e^*Pe=(b-a)^*P(b-a)-d^*x$ s^2 = E/(n-m) F = s^2/σ^2 RPR = R
covariances	covariances
Q(b,b*) = Q Q(t,t*) = N Q(k,k*) = M Q((b+e),t*) = Q((b+e),k*) = 0 Q((b+e),e*) = 0 Q(b,e*) = -Q(e,e*) Q(e,e*) = R Q((b+e),(b+e)*)= Q(b,b*)-Q(e,e*)= =Q-R Q(b,t*) = $-QU^*$ Q(b,k*) = $-QU^*M$ Q(e,t*) = $+QU^*$ Q(e,k*) = $+QU^*M$	Q(b,b*) = Q Q(d,d*) = G Q(x,x*) = H Q(e,d*) = Q(e,x*) = 0 Q((b+e),e*) = 0 Q(b,e*) = -Q(e,e*) Q(e,e*) = R Q((b+e),(b+e)*)= Q(b,b*)-Q(e,e*)=Q-R= =AHA* Q(b,d*) = A Q(b,x*) = AH Q((b+e),d*) = A Q((b+e),x*) = AH

a and b are the major and minor axis of the ellipse. α is the angle between the ellipse-axis and the x-axis.
If $\sigma_x > \sigma_y$, the major axis is close to the x-axis



3.1.9. Linearisation.

In the previous paragraph it is assumed that the conditions and the relations between the parameters and observations are linear. In practice this is often not the case. We have to make them linear, this process is called linearisation.

Adjustment with conditions.

In the linear case the conditions were (14):

$$U(b + e) = u \quad \text{or} \quad Ue = u - Ub = t \tag{14}$$

In the nonlinear case the conditions are:

$f(b + e) = u$. f indicates one or more nonlinear functions.

If the corrections are small this equation can be differentiated and written as:

$f(b) + \frac{\partial f}{\partial b} \cdot e = u$, $\frac{\partial f}{\partial b}$ is a matrix with elements of each function f differentiated to each observation b.

or $$\frac{\partial f}{\partial b} \cdot e = u - f(b) = t \tag{38}$$

This equation is linear in e and of the same form as (14), so:

$$U = \frac{\partial f}{\partial b} \quad \text{and} \quad t = u - f(b).$$

Adjustment with parameters.

The observation equation was in the linear case:

$$b + e = Ax + a \quad \text{or} \quad e = Ax - (b - a) \tag{26}$$

In the non-linear case each observation b can be expressed as a function of the parameters:

$$b + e = f(x)$$

To linearize this function, approximate estimations x_o of x should be inserted:

$$x = x_o + \Delta x .$$

f(x) can be linearized:

$$b + e = f(x_o) + \left(\frac{\partial f}{\partial x}\right)_o . \Delta x$$

or
$$e = \left(\frac{\partial f}{\partial x}\right)_o . \Delta x - (b - f(x_o)) \qquad (39)$$

This equation is of the same form as (26), so:

$$A = \left(\frac{\partial f}{\partial x}\right)_o \quad \text{and} \quad a = f(x_o)$$

The differences Δx are the new parameters.

Example:

l is a measured distance between two points A and B.

$$l + e = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2} = f(x) \qquad (40)$$

Insert approximate coordinates: $x_{A_o}, y_{A_o}, x_{B_o}, y_{B_o}$

$$l_o = \sqrt{(x_{B_o} - x_{A_o})^2 + (y_{B_o} - y_{A_o})^2} = f(x_o)$$

The approximate azimuth from A to B is α:

$$\sin \alpha = \frac{x_{B_o} - x_{A_o}}{l_o} , \quad \cos \alpha = \frac{y_{B_o} - y_{A_o}}{l_o}$$

Differentiation of (40) gives:

$$l + e = l_o - \sin \alpha . \Delta x_A + \sin \alpha . \Delta x_B - \cos \alpha . \Delta y_A + \cos \alpha . \Delta y_B$$

So: $A = (-\sin\alpha, +\sin\alpha, -\cos\alpha, +\cos\alpha)$

$x^* = (\Delta x_A, \Delta x_B, \Delta y_A, \Delta y_B)$

$b - a = l - l_o$

If there are more distances observed, A becomes a full matrix.
With these matrices the adjustment can be carried out in the same way as in the linear case.

3.1.10. Testing observational errors.

Errors in the observations can easily occur in large datasets. The cause of the errors can be:

- physical system errors, like instrumental errors,
- observational errors,
- processing errors, like writing or typing errors.

It is very important to have a computational test that automatically checks the observations. Therefore it is neccessary to do enough redundant observations, with the effect that each observation is checked by the other observations. If the system is well internally checked it has a high reliability, however, if the check is bad the reliability is low, even if the precision of the measurements is very good. Precision and reliability can both be expressed by a number.
Precision refers to the standard deviation of an observation and reliability to the magnitude of the possible error that can be found with a probability of 80%.
The precision of the observations is usually known before the observations are done. From previous similar observations and experiments the expected variance factor is adopted.
On the other hand, an estimation of the variance factor can be obtained after the observations have been carried out and the adjustment has been finished.
This estimation s^2 can be computed with the following formula:

$$s^2 = \frac{e^* P e}{r}$$

r is the number of redundant observations or the number of conditions. s^2 indicates the estimate of the variance factor after the adjustment (a posteriori) in contrary to the a priori adopted variance factor: σ^2.
The quotient s^2/σ^2 has a Fisher (F) distribution. Statistically, one can test, whether this quotient can be explained by the normal random noise in the

observations. One has to adopt a level of uncertainty, α, usually α = 5% or 1%. From α and the redundancy r the critical value $F_{r,\alpha}$ can be found in a table of the F-distribution.

Critical value $F_{r,\alpha}$ for α= 5% and α= 1%.

r	1	2	3	4	5	6	7	8	9	10
5%	3.84	3.00	2.60	2.37	2.21	2.09	2.00	1.94	1.88	1.83
1%	6.64	4.60	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

r	12	15	20	30	40	50	75	100	200	500
5%	1.75	1.66	1.57	1.46	1.40	1.35	1.28	1.24	1.17	1.11
1%	2.18	2.03	1.87	1.69	1.59	1.52	1.41	1.36	1.25	1.15

If: $\frac{s^2}{\sigma^2} > F_{r,\alpha}$

there is a chance of (1 - α) that some error in the observations has been made, or that the adopted σ^2 is not correct. However, one doesn't know which observation is not correct.

If r is small it is hard to find where the error has been made. Therefore it is advisable to take enough redundant observations, to check each observation sufficiently.

A simple method to find the error is to look at the corrections. The observation with the largest correction might be wrong. However, this method is not correct if the observations are correlated, if the redundancy r is small or if the adjustment system is complex. In that case one should use a systematic data-snooping procedure that tests each observation in the optimal way. This method is developped by Prof.Dr. W. Baarda and is described by Baarda in : A testing procedure for use in geodetic networks, Neth. Geod. Comm. Vol. 2, Nr. 5, 1968. A description that uses matrix notation is given by G. Strang van Hees in chapter 3.2.

3.1.11. Sequential adjustment.

In the adjustment with parameters, the observations b are expressed as functions of the unknown parameters x. The parameters are solved by the least squares method. In surveying the parameters are often the coordinates of the points. Let us suppose that, after the adjustment has been finished some new observations are done. Strictly speaking the whole adjustment has to be done again.

However, we will show that there is an other method which gives the same results. This method computes the improvement of the coordinates (parameters) due to the new observations.

The results of the first adjustment can be considered as estimates of the parameters and therefore be introduced as observations in the second adjustment, together with the new observations.

Suppose the parameters found in the first adjustment are indicated by x' and the variance matrix $Q_{x'x'}$. The new observations are b, with variance Q_{bb} and the parameters after the second adjustment are x with variance Q_{xx}.

The observation equations are, after eventually linearisation:

$$b + e = A\,x$$
$$x' + c = x$$

e = correction of b, c = correction of x'

or

$$\begin{pmatrix} b \\ x' \end{pmatrix} + \begin{pmatrix} e \\ c \end{pmatrix} = \begin{pmatrix} A \\ I \end{pmatrix} . x \qquad (41)$$

Least squares adjustment gives the solution: (compare (29) and (32)).

$$x = Q_{xx} . (A^*\ I) . \begin{pmatrix} Q_{bb} & 0 \\ 0 & Q_{x'x'} \end{pmatrix}^{-1} . \begin{pmatrix} b \\ x' \end{pmatrix}$$

with

$$Q_{xx} = \left((A^*\ I) \begin{pmatrix} Q_{bb} & 0 \\ 0 & Q_{x'x'} \end{pmatrix}^{-1} . \begin{pmatrix} A \\ I \end{pmatrix} \right)^{-1}$$

or

$$x = Q_{xx}\, A^*\, Q_{bb}^{-1}\, b + Q_{xx}\, Q_{x'x'}^{-1}\, x' \qquad (42)$$

$$Q_{xx} = (A^*\, Q_{bb}^{-1}\, A + Q_{x'x'}^{-1})^{-1} \qquad (43)$$

These formulas can be transformed by some manipulations:

from (43): $Q_{xx}^{-1} - Q_{x'x'}^{-1} = A^* Q_{bb}^{-1} A$

or: $$I - Q_{xx} Q_{x'x'}^{-1} = Q_{xx} A^* Q_{bb}^{-1} A = KA \qquad (44)$$

from (43) and (44) follows:

$$K = Q_{xx} A^* Q_{bb}^{-1} = (A^* Q_{bb}^{-1} A + Q_{x'x'}^{-1})^{-1} A^* Q_{bb}^{-1} \qquad (45)$$

This can be transformed with the remarkable relation:

$$K = (A^* Q_{bb}^{-1} A + Q_{x'x'}^{-1})^{-1} A^* Q_{bb}^{-1} = Q_{x'x'} A^* (Q_{bb} + A Q_{x'x'} A^*)^{-1} \qquad (46)$$

This formula can be checked by multiplying both sides, before and behind, with the matrices respectively: $(A^* Q_{bb}^{-1} A + Q_{x'x'}^{-1})$ and $(Q_{bb} + A Q_{x'x'} A^*)$.

Insert (46) in (45) and (44) and (45) in (42) gives:

with
$$\boxed{x = x' + K \cdot (b - A x')} \qquad (47)$$

$$\boxed{K = Q_{x'x'} A^* (Q_{bb} + A Q_{x'x'} A^*)^{-1}} \qquad (48)$$

K is called the gain matrix and gives the improvement of x' by the new observations b.

The variance matrix Q_{xx} expressed by (43) can also be written in an other form. Multiply (44) with $Q_{x'x'}$:

$$Q_{xx} = (I - Q_{xx} A^* Q_{bb}^{-1} A) Q_{x'x'}$$

from (45), (46) and (48) follows

$$Q_{xx} A Q_{bb}^{-1} = K. \qquad (49)$$

so $$\boxed{Q_{xx} = (I - K A) \cdot Q_{x'x'}} \qquad (50)$$

The corrections e are:

$$\boxed{e = Ax - b = (I - AK)(Ax' - b)} \qquad (51)$$

The improvement of the variance matrix is:

$$Q_{x'x'} - Q_{xx} = K\, A\, Q_{x'x'} \qquad (52)$$

Formulas (47), (48), (50), (52) form the computational framework of the sequential adjustment and as we will see also the base for the Kalman filter.

3.1.12. Kalman filter.

The Kalman filter is an application of sequential adjustment in case of a dynamic system or a time series. The second characteristic of Kalman filter is that it is a real time filter, that means it gives the adjusted results immediately after the new measurement and between the measurements. For example, a steaming ship is a dynamic system that moves, the position changes in time. After each position fix (observation) an adjustment is started that updates the coordinates.

The position, velocity, acceleration and possibly other properties of the dynamic system are described by the state vector, which is time dependent.

$$\text{State vector } x(t) = \begin{pmatrix} \text{position (t)} \\ \text{velocity (t)} \\ \text{acceleration (t)} \end{pmatrix}$$

The state vector is the vector with unknown parameters in our adjustment.

From the state vector on a previous time the state vector in the next time can be predicted. For example, if the position, velocity and acceleration of a ship are known on (t) the position and velocity and acceleration on $t+\Delta t$ can be predicted by

$$\begin{aligned} s(t+\Delta t) &= s(t) + \Delta t\, v(t) + \tfrac{1}{2}\, \Delta t^2\, a(t) \\ v(t+\Delta t) &= v(t) + \Delta t\, a(t) \\ a(t+\Delta t) &= a(t) \end{aligned}$$

We call this prediction of the state vector from a previous time the transition of the state vector. It is mathematical described by a matrix, the transition matrix T.

$$x'(t+\Delta t) = T \cdot x(t) \qquad (53)$$

$x'(t)$ is the predicted state vector.

The variance matrix of x' : $Q_{x'x'}$ can be computed by the law of error propagation. However, an extra component must be added due to the uncertainty of the transition matrix. This is the variance of the transition $Q_{trans}(t-1, t)$. So we get:

$$Q_{x'x'}(t+\Delta t) = T\, Q_{xx}(t)T^* + Q_{trans}(\Delta t) \qquad (54)$$

Next on time t new observations b are made (update) which are related to the state vector by the observation equation

$$b + e = Ax \qquad (e = \text{correction})$$

The predicted state vector x'(t) can be considered as an observation of x(t).

$$x' + c = x \qquad (c = \text{correction})$$

These observation equations have the same form as the observation equations in the sequential adjustment. (41).

Therefore the resulting formulas (47), (48), (50) and (52) are also valid for the Kalman filter, with the reminder that x' is the predicted state vector which follows from (53) and $Q_{x'x'}$ is computed with (54).

The problem of the Kalman filter is the computation of the matrices T and Q_{trans}. It is difficult to make a correct estimation.

The transition matrix T can also be computed from the differential equation of the system. This d.e. can be written as:

$$\frac{dx}{dt} = Fx, \qquad F \text{ is called the dynamic matrix.}$$

Differentiation gives: $\frac{d^2x}{dt^2} = F\,\frac{dx}{dt} = F \,.\, Fx.$

A Taylor expansion of x'(t+Δt) gives

$$x'(t+\Delta t) = x(t) + \frac{\partial x}{\partial t}\,\Delta t + \tfrac{1}{2}\,\frac{\partial^2 x}{\partial t^2}\,\Delta t^2 + \ldots$$

or $$x'(t+\Delta t) = (I + F\Delta t + \tfrac{1}{2}F.F\Delta t^2 + \ldots)x(t).$$

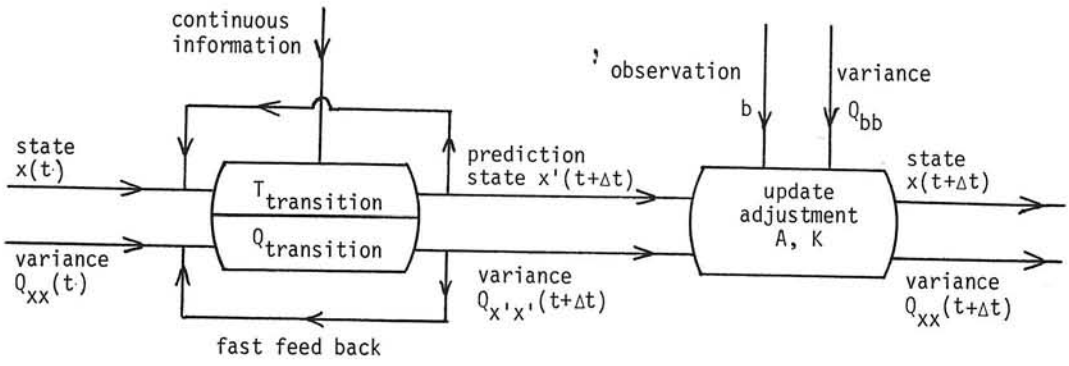
Comparison with (53) shows

$$\boxed{T = I + F\Delta t + \tfrac{1}{2}\,F.F\Delta t^2 + \ldots.} \qquad (55)$$

Sometimes it is easier to compute F than T, but then T can be computed with (55). The variance matrix Q_{trans} should also be estimated. It is dependent on the time interval Δt (see example).

An additional advantage of the Kalman filter is that the new observations b can be tested against the predicted state vector. This is important on sea where the position is updated with radiopositioning systems. A serious problem is to find lane-slips. The lanes can be controled by observing more than two lanes, but this check is not very reliable. An extra test of the position against the predicted position is very useful. This test can be incorporated in the computerprogram for Kalman filtering.

Kalman filter



Time (t-1) Time (t)

T	= transition matrix of state vector.
Q_{trans}	= transition matrix of variance matrix.
b(t)	= observation, update.
Q_{bb}	= variance of observation, (noise).
x(t)	= state vector.
$x'(t+\Delta t)$	= predicted state vector from "dead reckoning" (transition).
Q_{xx}	= variance matrix.

Summary of the formulas for the Kalman filter

Transition : $x'(t+\Delta t) = Tx(t)$

$$Q_{x'x'}(t+\Delta t) = T\, Q_{xx}(t) T^* + Q_{trans}$$

Update : $x(t+\Delta t) = x'(t+\Delta t) + K \cdot \left(b(t+\Delta t) - A\, x'(t+\Delta t)\right)$

$$Q_{xx}(t+\Delta t) = (I - K\, A) \cdot Q_{x'x'}(t+\Delta t)$$

with $K = Q_{x'x'}\, A^* (A\, Q_{x'x'}\, A^* + Q_{bb})^{-1}$ (gain-matrix)

Correction model: $b + e_b = A\, x$ adjustment model.
$x' + e_{x'} = x$ e = correction.

Testing with a Kalman filter.

An important advantage of the Kalman filter procedure is that one has the possibility to test the observations better than without the predicted parameters. The prediction x' gives a check on the observations b.
The corrections e can be computed with

$$e = Ax - b = \text{computed} - \text{observed}.$$

In case of a positioning system on sea is:

x the coordicates, computed from the adjustment,

Ax the computed lane number,

b the observed lane number.

The weight coefficient matrix of e is (34):

$$Q_{ee} = Q_{bb} - A\, Q_{xx}\, A^*$$

(49): $Q_{xx}\, A^*\, Q_{bb} = K$

So $A\, Q_{xx}\, A^* = A\, K\, Q_{bb}$

and $Q_{ee} = (I - A\, K) Q_{bb}$.

According to the testing theory (chapter 3.2), f is defined as:

$$f = -c^* Q_{bb}^{-1} e$$

where c is an unit vector: $c^* = (0,0,\ldots,0,1,0,\ldots)$, with 1 on the place of the observation to be tested. The weight coefficient matrix of f is:

$$Q_{ff} = c^* Q_{bb}^{-1} Q_{ee} Q_{bb}^{-1} c$$

or $$Q_{ff} = c^* Q_{bb}^{-1} (I - A K)c$$

The test-quantity is:

$$w = \frac{f}{\sigma\sqrt{Q_{ff}}}$$

which has a normal distribution with standard deviation one, N(0,1).
σ is the standard deviation of the unit weight, or σ^2 the variance factor.
The test is:

with $\alpha = 5\%$, if: $w > 1.96$; probably an error
with $\alpha = 1\%$, if: $w > 2.58$; probably an error.

In practice it is often assumed that the observations are free of correlation and have equal weights, Q_{bb} is diagonal. Then the formula for w simplifies to:

$$\boxed{w_i = \frac{-e_i}{\sigma\sqrt{1 - (A K)_{ii}}\sqrt{Q_{bb}}_{ii}}} \quad \text{with} \quad \boxed{-e_i = b-Ax = (I-AK)(b-Ax')}$$

i is the number of the observation to be tested.
$(A K)_{ii}$ means the i^{th} diagonal element of matrix (A K).

Example of a Kalman filter.

The position of a ship is determined with time intervals Δt. From the previous positions, the next position can be predicted by dead reckoning. If we assume that the ship has approximate constant speed, the acceleration is zero, then the covered distance s is:

$$s'(t) = s(t-1) + v(t-1)\Delta t.$$

The predicted speed v(t) is equal to v(t-1). The state vector $x = \begin{pmatrix} s \\ v \end{pmatrix}$.

$$\begin{pmatrix} s'(t+\Delta t) \\ v'(t+\Delta t) \end{pmatrix} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} s(t) \\ v(t) \end{pmatrix} \quad \text{or } x'(t+\Delta t) = T\, x(t)$$

The variance matrix $Q_{xx}(t) = \begin{pmatrix} Q_{ss} & Q_{sv} \\ Q_{sv} & Q_{vv} \end{pmatrix}$.

The mean acceleration is zero, however its variance Q_{aa} is not zero!
The transition variance matrix Q_{trans} is then:

$$Q_{trans} = \begin{pmatrix} \frac{1}{2}\Delta t^2 \\ \Delta t \end{pmatrix} Q_{aa} \begin{pmatrix} \frac{1}{2}\Delta t^2 \\ \Delta t \end{pmatrix}^* = Q_{aa} \begin{pmatrix} \frac{1}{4}\Delta t^4 & \frac{1}{2}\Delta t^3 \\ \frac{1}{2}\Delta t^3 & \Delta t^2 \end{pmatrix}$$

So

$$Q_{x'x'}(t+\Delta t) = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} Q_{ss} & Q_{sv} \\ Q_{sv} & Q_{vv} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \Delta t & 1 \end{pmatrix} + Q_{aa} \begin{pmatrix} \frac{1}{4}\Delta t^4 & \frac{1}{2}\Delta t^3 \\ \frac{1}{2}\Delta t^3 & \Delta t^2 \end{pmatrix}$$

or

$$Q_{x'x'}(t+\Delta t) = \begin{pmatrix} Q_{ss} + 2\,\Delta t\, Q_{sv} + \Delta t^2 Q_{vv} + \frac{1}{4}\Delta t^4 Q_{aa} & \vdots & Q_{sv} + \Delta t\, Q_{vv} + \frac{1}{2}\Delta t^3 Q_{aa} \\ \cdots\cdots & \vdots & \cdots\cdots \\ Q_{sv} + \Delta t\, Q_{vv} + \frac{1}{2}\Delta t^3 Q_{aa} & \vdots & Q_{vv} + \Delta t^2 Q_{aa} \end{pmatrix}$$

We can write this in short:

$$Q_{x'x'} = \begin{pmatrix} Q_{s's'} & Q_{s'v'} \\ Q_{s'v'} & Q_{v'v'} \end{pmatrix}$$

Next the predicted state vector x'(t) is updated by an observation b(t) of the position of the ship.

$$b + e = (1 \quad 0) \begin{pmatrix} s \\ v \end{pmatrix}$$

e = correction. So matrix A = (1 0).

The gain matrix K is:

$$K = Q_{x'x'} A^* (A\, Q_{x'x'} A^* + Q_{bb})^{-1}$$

Insert A and $Q_{x'x'}$:

$$A\, Q_{x'x'}\, A + Q_{bb} = Q_{s's'} + Q_{bb}$$

and $$K = \frac{1}{Q_{s's'} + Q_{bb}} \begin{pmatrix} Q_{s's'} \\ Q_{s'v'} \end{pmatrix}$$

The updated state vector is

$$x = x' + K \,.\, (b - Ax')$$

or $$s = s' + \frac{Q_{s's'}}{Q_{s's'} + Q_{bb}} (b - s')$$

$$v = v' + \frac{Q_{s'v'}}{Q_{s's'} + Q_{bb}} (b - s')$$

The variance matrix of the updated state is:

$$Q_{xx} = (I - KA)\, Q_{x'x'}$$

Insert I, K, A and $Q_{x'x'}$:

$$Q_{xx} = \frac{Q_{bb}}{Q_{s's'} + Q_{bb}} \begin{pmatrix} Q_{s's'} & \vdots & Q_{s'v'} \\ \cdots & \cdots & \cdots \\ Q_{s'v'} & \vdots & Q_{v'v'} + \dfrac{Q_{s's'}Q_{v'v'} - Q_{s'v'}^2}{Q_{bb}} \end{pmatrix}$$

This updated state vector and variance matrix can be used for the next prediction.

Test on big errors:

$$I - AK = \frac{Q_{bb}}{Q_{bb} + Q_{s's'}}$$

$$e = b - s = (I - AK)(b - s')$$

insert in the formula for w, the test quantity:

$$w = \frac{(b-s)}{\sigma} \frac{\sqrt{Q_{bb} + Q_{s's'}}}{Q_{bb}}$$

or expressed in s' instead of s:

$$w = \frac{(b-s')}{\sigma} \frac{1}{\sqrt{Q_{bb} + Q_{s's'}}}$$

If w is bigger than 2.0 or 2.6, dependending on the chosen $\alpha = 0.05$ or 0.01, the observation is probably bad.

We have considered an one-dimensional example. In practice the position and speed of a ship are two dimensional: northing and easting and the velocities in the two directions. However, if we assume that the observations of the position are not correlated in both directions, the northing and easting can be computed independently with the formulas given above.
In general there will exist a correlation and we have to solve for a state vector containing both northing and easting.

The transition is in that case:

$$\begin{pmatrix} N'(t+\Delta t) \\ E'(t+\Delta t) \\ V_N(t+\Delta t) \\ V_E(t+\Delta t) \end{pmatrix} = \begin{pmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} N(t) \\ E(t) \\ V_N(t) \\ V_E(t) \end{pmatrix}$$

The computation uses the same formulas as the example above.

3.1.13. Literature.

A nice and very instructive book on matrices, determinants and linear equations is:

Strang, Gilbert: Linear algebra and its applications, Academic Press, 1980.

On adjustment theory a very complete and readable book is:

Mikhail, Edward M.: Observatorions and least squares, Harper and Row, 1976.

A good overview of matrices and least squares adjustment is given in the appendices of:

Bomford, G.: Geodesy, Clarendon Press, Oxford, 1980.

The basic theory of testing geodetic networks is given in:

Baarda, W.: A testing procedure for use in geodetic networks, Neth. Geodetic Commission, Vol. 2, Nr. 5, 1968.

and also in:

Forty years of thoughts, festive book on the occasion of the 65 anniversary of Prof. W. Baarda, Geodetic Institute, Delft, 1982.

A nice book on surveying and adjustment theory with many examples worked out is:

Richardus, P.: Project surveying, second ed., A.A. Balkema, Rotterdam, 1984.

3.2 Testing geodetic networks.

G.L. Strang van Hees

3.2.1. Introduction.

Prof. W. Baarda developed a theory to test geodetic networks and to compute an objective measure to express the reliability of the network.

The general ideas can be formulated as follows:
A geodetic network is observed with more observations than the minimum necessary. The redundant observations give the possibility to adjust the network with the following advantages:

1. To increase the precision of the computed unknowns.
2. To estimate the standard-deviation of the observations and the unknowns.
3. To test the mathematical and stochastic model.
4. To find gross-errors in the observations.
5. To compute the reliability of the network.

With the nowadays high precision instruments the increase of precision is no longer the most important reason for measuring redundant observations. The most important purpose is to detect gross-errors. In practice it turned out that gross-errors are not always found and these errors deform the network considerably.
With statistical tests the gross-errors can be detected with a certain probability dependent on the probability parameter α (figure 1).
The marginal detectable gross-error in each observation can be expressed as a function of parameter β.
The effect of non-detected gross-errors on the unknown parameters is called the reliability.

In this paper the basic ideas of Baarda's theory are presented in matrix-notation.

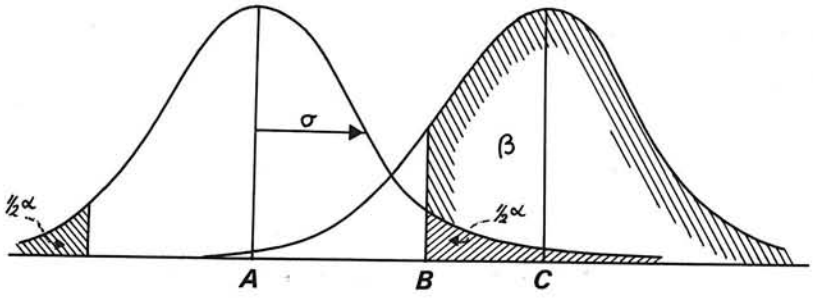


Figure 1.

Figure 1 gives the normal distribution of a stochastical quantity.
σ = standard deviation.
$\frac{AB}{\sigma}$ = critical value, dependent on α.
AC = marginal detectable error, dependent on α and β.
Usually α = 0.05, 0.01 or 0.001 and β = 0.80.
The chance to find a gross-error of size AC is β.

Notation.

The following notation is adopted.

A column vector is an undercast letter (a,b,c,...).
A row vector is a transpose of a column vector, indicated by (a^*,b^*,c^*,...).
A matrix is an uppercast letter (A,B,C,...).
A transpose matrix (A^*,B^*,C^*,...).
A scalar is also indicated by an undercast letter (a,b,c,...).
An element of a vector or matrix is indicated by an index

$$a^* = (a_1, a_2, a_3, \ldots, a_i, \ldots)$$

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots \\ A_{21} & A_{22} & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix}$$

The variance-covariance matrix of stochastic vectors a and b is indicated as:

$$\sigma_a^2 = Q(a, a^*).\sigma^2 \quad \text{and} \quad \sigma_{ab} = Q(a, b^*).\sigma^2$$

Q(a, a^*) is the weight coefficient matrix. It indicates a symbolic vector product:

$$Q(a,\ b^*) = Q\left(\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}, (b_1\ b_2\ \ldots\ b_n)\right) = \begin{pmatrix} Q(a_1,b_1) & Q(a_1,b_2) & \ldots \\ Q(a_2,b_1) & Q(a_2,b_2) & \ldots \\ \vdots & \vdots & \\ Q(a_n,b_1) & Q(a_n,b_2) & \ldots \end{pmatrix}$$

σ^2 is called the variance-factor and is a scale factor.
The law of error propagation can be expressed as a matrix multiplication. If the stochastic vector c is a linear function of the stochastic vectors a and b we get:

$$c = Aa + Bb \quad , \qquad c^* = a^*A^* + b^*B^*$$

$$Q(c,c^*) = Q\left((Aa + Bb),\ (a^*A^* + b^*B^*)\right)$$

$$Q(c,c^*) = AQ(a,a^*)A^* + AQ(a,b^*)B^* + BQ(b,a^*)A^* + BQ(b,B^*)B^*$$

or equivalent:

$$c = (A\ B)\begin{pmatrix} a \\ b \end{pmatrix}$$

$$Q(c,c^*) = (A\ B)Q\left(\begin{pmatrix} a \\ b \end{pmatrix}(a^*,b^*)\right)\begin{pmatrix} A^* \\ B^* \end{pmatrix}$$

or:

$$Q(c,c^*) = (A\ B)\begin{pmatrix} Q(a,a^*)Q(a,b^*) \\ Q(b,a^*)Q(b,b^*) \end{pmatrix}\begin{pmatrix} A^* \\ B^* \end{pmatrix}$$

3.2.2. Least squares adjustment.

We distinguish two types of adjustment: (see chapter 3.1.5)

1. With condition equations
2. With unknown parameters.

The formulas are collected in (3.1.8).

The following notation is used:

b = vector of observations, n elements.

e = vector of corrections, n elements. (not the error)

x = vector of estimates of the unknown parameters, m elements.

U = design matrix of conditions, u = constant term.

A = design matrix of parameter equations, a = constant term.

P = weight matrix of b = Q^{-1}.

Q = variance matrix of b.

$$Q = P^{-1} = Q(b,b^*)$$

n = number of observations.

m = number of parameters.

(n-m) = number of redundant observations.

: = means: by definition.

= means: can be computed.

The corresponding equations of the condition and parameter adjustment are written on the same line, so a comparision between the two types is possible.

In the formulas is assumed that the linearization of the equations has been done already beforehand.

So we start with linearized condition and correction equations.

3.2.3. Testing a network.

An intuitive way to test on gross-errors is to look at the corrections e. A faulty observation will probably get a big correction e_i. As test-quantity one can use e_i/σ_{e_i} which has a normal distribution with standard deviation = 1, N(0,1). However, this method is best only for uncorrelated observations.

In case of correlated observations Baarda derived that it is best to test (P.e), instead of e, the corrections multiplied with the weight matrix P. The vector (P.e) is most sensitive to an error in one of the observations (proof in Appendix 1).

Every observation must be tested separately. Baarda introduced the vector c, which is a unit vector with 1 on the place corresponding to the observation to be tested. $c^* := (0,0,\ldots 0,1,0\ldots 0)$, $c_i = 1$ if observation b_i is to be tested. It is also possible to test a combination of observations if one expects that the source of the gross-error effects more than one observation. In that case more elements of c are 1, e.g. $c^* = (0,\ldots 0,1,1,1,0\ldots 0)$. The choice of c determines the test-quantity, or with other words, c determines the alternative hypothesis that is to be tested.

If we define: $f := -c^*Pe$ then the testquantity is

$$w := \frac{f}{\sigma_f}$$

w has a normal distribution with standard deviation = 1, N(0,1). The variance of f can be computed with the law of error propagation:

$$Q(f,f^*) = c^*PQ(e,e^*)Pc = c^*PRPc$$

$$\sigma_f = \sigma\sqrt{c^*PRPc}$$

with σ^2 = variance factor

So:
$$\boxed{w = \frac{-c^*Pe}{\sigma\sqrt{c^*PRPc}}} \qquad (56)$$

In case of adjustment with conditions we have:

$$Pe = U^{*}k = U^{*}Mt$$

and

$$PRP = U^{*}MU$$

thus

$$w = \frac{-c^{*}U^{*}k}{\sigma\sqrt{c^{*}U^{*}MUc}}$$

This means that in case of adjustment with conditions the whole adjustment and testing can be carried out without computing matrix P. This is an advantage because usually matrix Q is given and the computation of P needs the inversion of Q which is a time consuming operation.
If c is successively the unit vector belonging to each observation, we get a w corresponding to that observation.
This process is called datasnooping.
w is tested for the normal distribution. The critical value w_0 is dependent on the choice of α:

α :	5%	1%	0.1%
w_0	1.96	2.58	3.29

$w > w_0$ means that there is a probability of $(1-\alpha)$ that the corresponding observation is wrong. As we don't want to reject to many observations, α = 1% turned out to be a good practical choice:

Test: if $w > w_0$ then: rejection.

F-test.

A second way to test the network is to test the estimation of the variance factor against the adopted variance factor.

$$s^2 = \frac{e^{*}Pe}{n-m} = \frac{E}{n-m}$$

(n-m) is the number of redundant observations.

$$F = \frac{s^2}{\sigma^2} \quad , \quad \text{test: if } F > F_0 \text{ then: rejection.}$$

F may be tested against a critical value F_0 dependent on the choice of α_F and on (n-m). σ^2 is supposed to be errorless

$$F_0 = F_0(\alpha_F, (n-m), \infty)$$

F_0 can be found in tables of the Fisher distribution.

In the next chapter is described that Baarda advocates to compute α_F in such a way that the chance to find an observation error is equal in the w-test as in the F-test.
As the F-test involves all the observations it is less sensitive to find one observation error. Therefore α_F becomes rather high under Baarda's assumption, if there are many redundant observations.
On the other hand it is possible to fix α_F independent of the w-test, to e.g. 5%. Then the F-test can be used to test on systematic errors, model errors and to test the adopted variance factor.

3.2.4. Reliability.

The influence of non-detected gross-errors on the result of the adjustment is called the reliability. The computed parameters (coordinates) may be wrong by a manyfold of the standard deviation if a gross error is not found. The better the observations are controlled the lesser the chance that a gross-error will slip through.

First we will consider the effect of an error Δ in one of the observations on the results of the adjustment.
If error Δ is made in observation b_i the change of observation vector b is:

$$\Delta b = c.\Delta \qquad (\Delta \text{ is a scalar, } \Delta b \text{ a vector})$$

c is defined as in the previous section: $c^* = (0,0,\ldots 0,1,0,\ldots 0)$ with $c_i = 1$ if Δ corresponds to b_i.
The change of the other quantities can easily be determined by the formulas of the adjustment.

$$\left.\begin{aligned} \Delta x &= HA^{*}P\ c.\Delta \\ \Delta e &= -RP\ c.\Delta \\ \Delta f &= c^{*}PRP\ c\ \Delta \\ \Delta w &= \sqrt{c^{*}PRP\ c}\ \frac{\Delta}{\sigma} \end{aligned}\right\} \qquad (57)$$

Δx is only valid for the adjustment with parameters. The other formulas are valid for both adjustment with parameters and with conditions. However, R is computed in a different way in both cases:

conditions: $R = QU^{*}MUQ$ and $PRP = U^{*}MU$

parameters: $R = Q - AHA^{*}$

The resulting R is the same in both cases.

Note that the changes Δx, Δc, Δf, Δw are non-stochastic, as they only depend on the initial change Δ, and not on the stochastic observations themselves.
In contrast ΔE and ΔF depend on the corrections e and are thus stochastic.

$$\Delta E = (e+\Delta e)^{*}P(e+\Delta e) - e^{*}Pe =$$

$$= \Delta e^{*}P\ \Delta e + \Delta e^{*}Pe + e^{*}P\Delta e =$$

$$= \Delta e^{*}P\ \Delta e + 2\Delta e^{*}Pe$$

However, the expectation of e (mean value) is zero, therefore the second term becomes zero if we consider the expectation of ΔE: $\overline{\Delta E}$.

$$\overline{\Delta E} = \Delta e^{*}P\ \Delta e = c^{*}PRPRPc\ \Delta^{2} =$$

$$= c^{*}PRP\ c.\Delta^{2} = \Delta w^{2}\sigma^{2}$$

$$\boxed{\overline{\Delta F} = \frac{\overline{\Delta E}}{\sigma^{2}(n-m)} = \frac{\Delta w^{2}}{n-m}} \qquad (58)$$

Up to now Δ was an arbitrary change in one observation. Now we will consider a special value of Δ, called ∇ (nabla).

∇ is chosen such that this change can be found in the w-test with a probability of β. Usually β is chosen to 80%.
The change in w that can be found with probability β is called $\nabla w(\alpha,\beta)$ (nabla w). In figure 1: AB depends on α and σ, BC on β and σ. If $\sigma = 1$ and $\beta = 80\%$ then BC = 0.84.

$\alpha = 5$ %, AB = 1.96, BC = 0.84, AC = 2.80 = ∇w
$\alpha = 1$ %, AB = 2.58, BC = 0.84, AC = 3.44 = ∇w
$\alpha = 0.1$ %, AB = 3.29, BC = 0.84, AC = 4.13 = ∇w

Next we compute backwards which changes in the other quantities corresponds to ∇w, by changing Δ into ∇ in formulas (57):

$$\left.\begin{aligned} \nabla &= \frac{1}{\sqrt{c^*PRPc}}\,\nabla w.\sigma \\ \nabla b &= c\nabla \\ \nabla x &= HA^*Pc\ \nabla \\ \nabla e &= -RPc\ \nabla \\ \nabla f &= c^*PRPc\ \nabla = \sqrt{c^*PRPc}\ \nabla w.\sigma \\ \overline{\nabla F} &= \frac{\nabla w^2}{n-m} \end{aligned}\right\} \qquad (59)$$

Note that $\overline{\nabla F}$ is independent of c, so for each alternative hypothesis we get the same $\overline{\nabla F}$.
On the other hand ∇x is a vector for each choice c_i an other one.
We combine these vectors x to a matrix with elements ∇x_{ki}, corresponding to observables b_i (i = 1 ... n) and parameter x_k (k = 1 ... m).
The marginal error ∇x_k in parameter x_k is the maximum value of ∇x_{ki} for all values i = 1 ... n.

The marginal detectable error in observation b_i is ∇b_i corresponding with c_i.
The formulas for ∇b_i and ∇x_k can be written as follows (see (59)):

$$\left.\begin{aligned} \nabla b_i &= \frac{\nabla w . \sigma}{\sqrt{(PRP)_{ii}}} \\ \nabla x_{ki} &= (HA^*P)_{ki}\ \nabla b_i \\ \nabla x_k &= \max(\nabla x_{ki}) \quad \text{over} \quad i = 1 \ldots n \ . \end{aligned}\right\} \qquad (59b)$$

∇x_k is also called the reliability of x_k.

Remark: If there are uncontrolled quantities introduced as observations, such like base points or scale parameters, ∇b_i and ∇x_k become infinitive. These parameters should be removed in the computation of x_k.
Baarda derived an other expression that gives an upperlimit of the reliability:

$$\frac{\nabla x_{ki}}{\sigma_{x_k}} \leqslant \frac{\sqrt{\nabla x_i^T G \nabla x_i}}{\sigma} \qquad \text{for each observation } b_i \ .$$

Here is ∇x_i the vector corresponding to b_i or c_i. The derivation of this formula is given in appendix 2. The right hand side is independent of k and can be evaluated as follows.

$$\nabla x G \nabla x = c^* PAHGHA^* Pc \; \nabla^2 =$$

$$= c^* PAHA^* Pc \; \nabla^2 =$$

$$= c^* P(Q-R)Pc \; \nabla^2 =$$

$$= (c^* PC - c^* PRPc)\nabla^2$$

and $$\nabla^2 = \frac{1}{c^* PRPc} \nabla w^2 . \sigma^2$$

So: $$\frac{\nabla x^* G \nabla x}{\sigma^2} = \left(\frac{c^* Pc}{c^* PRPc} - 1\right)\nabla w^2$$

Define $$\boxed{r_x := \sqrt{\frac{c^* Pc}{c^* PRPc} - 1} \; . \nabla w} \qquad (60)$$

Then $$\boxed{\nabla x_k \leq r_x \cdot \sigma_{x_k}} \qquad (61)$$

r_x is sometimes called the reliability of x and is dependent on the choice of c.
It is also possible to define r_x for a part of vector ∇x. $\nabla x'$ is a subvector of ∇x : $\nabla x' \subset \nabla x$.

Now $$r'_x := \frac{1}{\sigma} \sqrt{\nabla x'^* G' \nabla x'} \quad , \quad r'_x \leq r_x$$

and $$\nabla x'_k \leq r'_x \, \sigma_{x'_k}$$

This is important if we have different types of parameters, e.g. coordinates and orientation unknowns.
In a similar way one may define the reliability of the observations:

$$r_b = \frac{\nabla b}{\sigma_b} = \frac{\nabla b}{\sigma \sqrt{cQc}} \qquad (62)$$

Another quantity that expresses the reliability of the observations is defined as:

$$p = \frac{c^*PRPc}{c^*Pc}$$

or, if we consider the reliability of observation b_i:

$$\boxed{p_i = \frac{(PRP)_{ii}}{P_{ii}}} \quad . \tag{63}$$

Index ii means the i^e diagonal element of the matrix. It can be proved that:

$$0 \leq p_i \leq 1$$

For an observation that is well checked by the other observations, p comes close to 1. If the check is bad, p becomes small, and if an observation is not checked at all, p = 0:

good: $p > 0{,}7$, bad: $p < 0{,}4$, no check: $p = 0$.

Proof:

Variance of observation before adjustment: $Q_{bb} = Q$.
Variance of observation after adjustment : $Q_{b+e,b+e} = AHA$.
Variance of corrections : $Q_{e,e} = R$.

$$Q - R = AHA \quad \text{(see section adjustment theory).}$$

If the observation is well checked, $Q_{b+e,b+e} << Q$.
So R approaches Q.
As PQ = I (the identity matrix), PR becomes close to I or $\frac{(PRP)_{ii}}{P_{ii}}$ close to 1.
If the observation is hardly checked, $Q_{b+e,b+e} \approx 0$, thus R approaches zero, and $\frac{(PRP)_{ii}}{P_{ii}}$ comes close to zero.

Meaning of the reliability r_x and p.

To each observation belongs a quantity r_x and p. The magnitude of r_x and p is related to the capability of the whole network to check the corresponding observation on gross-errors.
If an observation is not checked, r_x will become infinitive and p zero.
The relation between r_x and p is, according to (60).

$$r_x = \sqrt{\frac{1}{p} - 1} \cdot \nabla w \quad . \tag{64}$$

On the other hand, the ability of a network to check an observation is also connected to the increase in precision of the adjusted observation.
The variance of an observation before the adjustment is the diagonal element of matrix Q, Q_{ii}.
The variance of the corrected observation after adjustment is
$Q(b+e,b+e) = Q_{ii} - R_{ii}$.

$$\frac{\sigma^2(b+e)}{\sigma^2(b)} = \frac{Q_{ii}-R_{ii}}{Q_{ii}} = 1 - \frac{R_{ii}}{Q_{ii}} \quad .$$

The improvement is:

$$q_i = \frac{R_{ii}}{Q_{ii}} \quad (0 < q_i < 1) \tag{65}$$

The relation between the reliability p and the improvement q is in general complicated, however for uncorrelated observations, with different weights, it is simple.

If P and Q are diagonal matrices, $P_{ii} = \frac{1}{Q_{ii}}$, and:

$$P_i = \frac{(PRP)_{ii}}{P_{ii}} = \frac{P_{ii}R_{ii}P_{ii}}{P_{ii}} = \frac{R_{ii}}{Q_{ii}} = q_i \quad .$$

So: $p_i = q_i$.

For correlated observations $p \neq q$, but in practice it turns out that the difference is small, so, the reliability p is about the same as the improvement of the observations. A positive correlation between the

observations can be interpretated as a stochastic relation between the observations which give an extra check on gross-errors. Therefore a positive correlation will increase the reliability and a negative correlation will decrease the reliability.

The relation between p and r_x is for α = 1%, thus ∇w = 3.44, according to (64):

p	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
r_x	∞	10.3	6.9	5.3	4.2	3.4	2.8	2.2	1.7	1.1	0

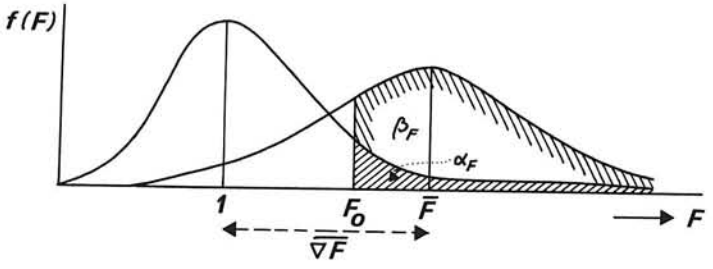
Next we consider $\overline{\nabla F}$.

The marginal change of F due to ∇ is:

$$\overline{\nabla F} = \frac{\nabla w^2}{n-m}$$

(n-m) is the number of redundant observations. Note that $\overline{\nabla F}$ is independent of c!

In the F-test a value of α_F was chosen. From $\overline{\nabla F}$ and α_F the value of β_F, corresponding to the shifted F-distribution can be computed.



With other words, $\overline{\nabla F}$, α_F and β_F are connected so that two of them determine the third. As $\overline{\nabla F}$ is determined by the choice of ∇w, one choice is left, α_F or β_F.

Baarda advocates to fix β_F = 80% and to apply the F-test with α_F which follows from this choice. This is called the B-method of testing. The disadvantage of this method is, that, if (n-m) becomes large, the F-distribution becomes narrower approximative proportional to $1/\sqrt{n-m}$ and $\overline{\nabla F}$ decreases proportional to $1/(n-m)$.
If β_F is fixed, then α_F increases if (n-m) increases. The consequence may be that α_F becomes so large that the test has no meaning anymore.

The second possibility is to fix α_F to e.g. 5%. Then β_F becomes smaller then 80%. This means that the F-test is less capable to detect observational errors than the w-test. This can also be understood intuitively. An observation error can best be tested by the w-test because this test is specifically designed for this kind of errors. The F-test is an overall test, which is less sensitive to an error in one observation, than to systematic errors or model errors, which disturb the whole network.

3.2.5. Literature:

Baarda, W.: A testing procedure for use in Geodetic networks, Neth. Geod. Comm., Vol. 2, No. 5, 1968.

Forty years of thoughts, festive book on the occasion of the 65 anniversary of Prof. W. Baarda, Geodetic Institute, Delft, 1982.

Mikhail, E.H.: Observations and least squares. Harper and Row, 1976.

3.2.6. Testing a network.

Summary: $c^* := (0,0,\ldots 0,1,0,\ldots 0)$

$R = Q(e,e^*) = Q\text{-}AHA^*$

$f := -c^*Pe$ (test quantity)

$Q(f,f^*) = c^*PRPc$

$$w = \frac{f}{\sigma_f} = \frac{-c^*Pe}{\sigma\sqrt{c^*PRPc}}$$

Test: $w \gtrless ?\ w_0(\alpha)$ = critical value.

A change Δ in one observation gives changes in the other quantities:

$\Delta b = c.\Delta$

$\Delta x = HA^*Pc.\Delta$

$\Delta e = -RPc.\Delta$

$\Delta f = c^*PRPc.\Delta$

$\Delta w = \sqrt{c^*PRPc}\ .\ \frac{\Delta}{\sigma}$

$\Delta E = \Delta e^*P\Delta e + 2\Delta e^*Pe$

$\overline{\Delta E} = c^*PRPc.\Delta^2 = \Delta w^2.\sigma^2$

$$\overline{\Delta F} = \frac{\overline{\Delta E}}{\sigma^2(n-m)} = \frac{\Delta w^2}{(n-m)}$$

Marginal detectable errors: choose $\nabla w(\alpha,\beta) = w_0 + 0.84$, if $\beta = 80\%$

$$\nabla b = \frac{c}{\sqrt{c^*PRPc}}\ \nabla w.\sigma, \qquad r_b = \frac{\nabla b}{\sigma_b}$$

$\nabla x = HA^*P.\nabla b$

$$r_x := \frac{1}{\sigma}\sqrt{\nabla x^* G \nabla x} = \sqrt{\frac{c^*Pc}{c^*PRPc} - 1}\ .\ \nabla w$$

$\nabla x_k \leqq r_x.\sigma_{x_k}$ for all elements x_k

$\overline{\nabla F} = \nabla w^2/(n-m) = f(\alpha_F,\beta_F)$ choose α_F or β_F, then follows the other.

Test: $F = \frac{s^2}{\sigma^2} \gtrless ?\ F_o(\alpha_F)$.

3.2.7. Appendix 1.

Proof of maximum sensitivity of w-test.

correction : e, covariance: $Q(e,e^*) = R$

observation: b, covariance: $Q(b,b^*) = Q = P^{-1}$

define : $f = -Pe$, covariance: $Q(f,f^*) = PRP = S$

define : $g = -P'e$, covariance: $Q(g,g^*) = P'RP' = T$

covariance: $Q(g,f^*) = P'RP = U$

P' is an arbitrary positive definite variance matrix. f_i, g_i, T_{ij}, U_{ij},... are elements of f, g, T, U,...

Define: $v = \frac{g_i}{\sigma_{g_i}}$ and $w = \frac{f_i}{\sigma_{f_i}}$, $v = w$ for $P' = P$.

Now we have to prove that $\Delta w \geq \Delta v$ due to a change Δ.

$$v = \frac{g_i}{\sigma_{g_i}} = \frac{c^*g}{\sigma\sqrt{cQ(g,g)c}} = \frac{-c^*P'e}{\sigma\sqrt{c^*P'RP'c}}$$

A change Δ gives (57): $\Delta e = -RPc\Delta$.

So:

$$\Delta v = \frac{c^*P'RPc}{\sqrt{c^*P'RP'c}} \cdot \frac{\Delta}{\sigma} = \frac{U_{ii}}{\sqrt{T_{ii}}} \cdot \frac{\Delta}{\sigma}$$

Further: $\frac{\sigma_{g_i f_i}}{\sigma_{g_i}\sigma_{f_i}} \leq 1 \Rightarrow Q(g,f^*_{ii}) < \sqrt{Q(g_i,g_i^*)Q(f_i,f_i^*)} \Rightarrow U_{ii} \leq \sqrt{T_{ii}\cdot S_{ii}}$

$$\Delta w = \sqrt{c^*PRPc} \cdot \frac{\Delta}{\sigma} = \sqrt{S_{ii}}\,\frac{\Delta}{\sigma} \quad \text{(follows from (57))}.$$

As: $\frac{U_{ii}}{\sqrt{T_{ii}}} \leq \sqrt{S_{ii}}$ we conclude:

$$\Delta v \leq \Delta w$$

An error Δ in an observation causes a maximal change in Δw. So:
w is the best test quantity (most sensitive).

Appendix 2.

Proof of formula: $\nabla x_k \leq r\sigma_{x_k}$.

We have to prove:

$$\nabla x^* G \nabla x \geq \frac{(\nabla x_k)^2}{Q(x_k, x_k^*)} \tag{66}$$

∇x_k is an element of the vector ∇x. If we introduce a unitvector $e_k(0,0,\ldots,1,0,\ldots 0)$: then is: $\nabla x_k = e_k^* \nabla x$ and $Q(x_k, x_k) = e_k^* G^{-1} e_k$. So (66) becomes:

$$(\nabla x^* G \nabla x)(e^* G^{-1} e) \geq (e^* \nabla x)^2 = (\nabla x^* e)^2$$

This is true if we prove the more general theorem:

$$\boxed{(a^* G a)(b^* G^{-1} b) \geq (a^* b)^2} \tag{67}$$

where: a and b are arbitrary vectors. G is an arbitrary positive definite matrix.

Proof: Develop G in the matrix of eigenvectors U and eigenvalues Λ.

$$G = U^* \Lambda U \qquad G^{-1} = U^* \Lambda^{-1} U.$$

U is an orthogonal matrix so: $U^* U = E$ (unit).
Λ is a diagonal matrix with only positive elements λ; on the diagonal, because G is positive definite.

Insert in (67):

$$(a^* U^* \Lambda U a)(b^* U^* \Lambda^{-1} U b) \geq (a^* U^* U b)^2$$

As a, b and U are arbitrary vectors and matrices, also Ua and Ub are arbitrary vectors.
Call $\bar{a} = Ua$ and $\bar{b} = Ub$. So we have to prove:

$$(\bar{a}^* \Lambda \bar{a})(\bar{b}^* \Lambda^{-1} \bar{b}) \geq (\bar{a}^* \bar{b})^2 \tag{68}$$

$$(\bar{a}^* \Lambda \bar{a}) = \sum_i a_i^2 \lambda_i \ , \quad (\bar{b}^* \Lambda^{-1} b) = \sum_k b_k^2 \cdot \frac{1}{\lambda_k}$$

The lefthand side of (68) becomes:

$$\sum_{i=1}^{n} (a_i^2 b_i^2) + \sum_{i=1}^{n} \sum_{k=i+1}^{n} (a_i^2 b_k^2 \frac{\lambda_i}{\lambda_k} + a_k^2 b_i^2 \frac{\lambda_k}{\lambda_i}) \tag{69}$$

The righthand side of (68) becomes:

$$\sum_{i=1}^{n} (a_i^2 b_i^2) + \sum_{i=1}^{n} \sum_{k=i+1}^{n} (2 a_i b_i a_k b_k) \tag{70}$$

Insert (69) and (70) in (68). The first term is left and right the same. Bring the second term from right to left:

$$\sum_{i=1}^{n} \sum_{k=i+1}^{n} (a_i^2 b_k^2 \frac{\lambda_i}{\lambda_k} - 2 a_i b_i a_k b_k + a_k^2 b_i^2 \frac{\lambda_k}{\lambda_i}) \geq 0$$

or

$$\sum_{i=1}^{n} \sum_{k=i+1}^{n} \left(a_i b_k \sqrt{\frac{\lambda_i}{\lambda_k}} - a_k b_i \sqrt{\frac{\lambda_k}{\lambda_i}} \right)^2 \geq 0 \tag{71}$$

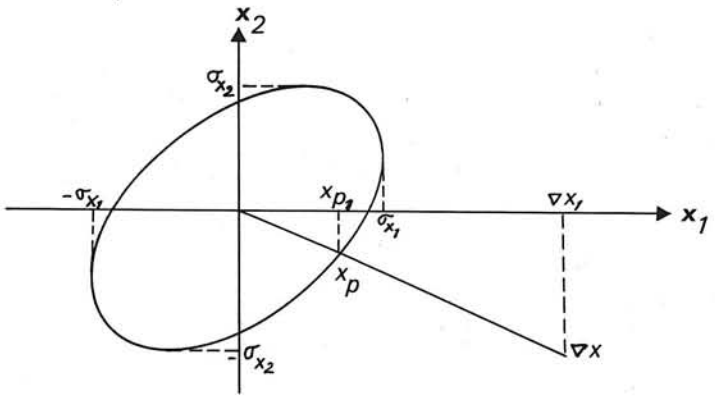
As a square is always positive, equation (71) is true, and therefore equation (67) too. So we proved (67).

Note: From (71) we see that λ_i and λ_k must be positive, so G positive definite.

Geometrical proof of $\nabla x_k \leq r\sigma_{x_k}$.

The equation of the n-dimensional error-ellipsoid is:

$$x^* G x = \sigma^2$$



In the figure is drawn the projection of the error-ellipsoid on the x_1, x_2-plane.
The projection of the ellipsoid on the x_i-axis is σ_{x_i}.

The vector ∇x represents a point in this space. ∇x_i is the projection of ∇x on the x_i-axis. The vector ∇x intersects the error-ellips in x_p.
Suppose $\frac{\nabla x}{x_p} = p$, a constant.
As x_p lays on the ellipsoid: $x_p^* G x_p = \sigma^2$.

So: $$\nabla x^* G \nabla x = p^2 \sigma^2 \qquad (72)$$

Further: $x_{p_i} \leq \sigma_{x_i}$, thus $\nabla x_i \leq p\sigma_{x_i}$ (73)

Combine (72) and (73):

$$\frac{\nabla x_i}{\sigma_{x_i}} \leq \frac{\sqrt{\nabla x^* G \nabla x}}{\sigma}$$

Appendix 3.

Correspondence between Baarda's notation and notation in this paper.

Baarda	This paper
x^i	b
ε^i	e
g^{ij}	Q
g_{ji}	P
u_i^ρ, u_j^τ	U
a_α^i, a_β^j	A
y^α, y^β	x
y_α, y_β	d
y^ρ, y^τ	-t
y_ρ, y_τ	-k
$g_{\alpha\beta}$	G
$g^{\alpha\beta}$	H
$g^{\rho\tau}$	N
$g_{\rho\tau}$	M
$\hat{\sigma}$	s
G^{ij}	$Q-R = AHA^*$
$\nabla_p^\sim x^i$	∇b
N_p	$c^*PRPc = \overline{f,f}$
$\sqrt{\lambda_0}, \sqrt{\lambda}$	∇w
$\sqrt{\bar{\lambda}}$	r_x
$\sqrt{F_{1-\alpha,1,\infty}}$	w_0
$F_{1-\alpha,b,\infty}$	F_0
b	n-m
c_p^i	c
$\nabla_{p,o}^\sim y^\alpha$	∇x

3.3 Adjustment of hyperbolic patterns.

G.L. Strang van Hees

3.3.1 Introduction.

To determine the position of a ship with hyperbolic position systems the minimum requirement is two position lines. However, because of the frequent occurence of lane slips it is necessary to observe at least three patterns to be able to check the observations. This redundant observation gives the possibility to compute an estimation of the standard deviation (the standard error-ellipse)of the position. The best estimation of the position is obtained by an adjustment. In this adjustment one should introduce the effect of the correlation between the patterns. The adjustment, testing and computation of the precision of the position is described. Some interesting geometrical features of the error figure of the observed lines of position (LOP) are derived.

For a hyperbolic positioning system at least three transmitters are necessary which give two position lines (LOP's). The third combination, slave-slave, gives a positioning-line going almost through the same point, because the third line is dependent on the other two. In a system of 4 transmitters usually three master-slave combinations are observed and the adjustment is executed with the assumption of correlation-free observations. However, as the master occurs in all the observed lane numbers there exists a correlation which may not be neglected. If we assume that the received signals of all the stations have the same standard deviation the correlation coefficient will be 0.5. This can be explained as follows:

Lane number 1 = range difference (Master - Slave 1) (L1 = M - S1)
Lane number 2 = range difference (Master - Slave 2) (L2 = M - S2)
Lane number 3 = range difference (Master - Slave 3) (L3 = M - S3)

The variances are obtained by the law of error propogation (see part I-4):

$$\sigma^2_{L1} = \sigma^2_M + \sigma^2_{S1} - 2\sigma_{M,S1} \quad \text{and} \quad \sigma_{L1,L2} = \sigma^2_M - \sigma_{M,S1} - \sigma_{M,S2} + \sigma_{S1,S2}$$

Suppose the slave signals have the same variance $\sigma^2_{S1} = \sigma^2_{S2} = \sigma^2_{S3} = \sigma^2_S$ and the crosscorrelation between the signals is zero: $\sigma_{M,S} = \sigma_{S1,S2} = \ldots = 0$ we get:

$$\sigma^2_L = {}^2_{L1} = {}^2_{L2} = {}^2_{L3} = {}^2_M + {}^2_S$$

$$\sigma_{L1,L2} = \sigma_{L1,L3} = \sigma_{L2,L3} = \sigma^2_M$$

The correlation coefficient is $\rho = \frac{\sigma_{L1,L2}}{\sigma_L^2} = \frac{\sigma_M^2}{\sigma_M^2 + \sigma_S^2}$

If $\sigma_M^2 = \sigma_S^2$ then $\rho = 0.5$.

However, as the slave signal is triggered by the Master signal the path through the atmosphere before a slave signal is received on the ship, goes from Master to Slave and Slave to ship. This is a longer path then the direct way Master-Ship and is therefore more disturbed by noise than the Master signal. Besides, the Slave signal is triggered by the Master signal, so the Slave signal contains more receiver noise than the Master signal.

So $\sigma_M^2 < \sigma_S^2$ and $0 < \rho < 0.5$. In practice a good estimate of ρ is:

$0.3 < \rho < 0.4$.

In this paper the effect of the correlation on the adjustment and on the testing of the observations is derived.

Further some peculiar geometrical relations of the errorfigure, that is the figure formed by the lanes, are derived.

3.3.2 The error figure of 4 transmitters.

Suppose we have 4 transmitters called 1, 2, 3 and 4. If we compute the position from the three transmitters 1, 2, 3 we get three lanes (1-2), (1-3), (2-3) which go through one point. As this position is computed without using transmitter 4 we will call this position "not 4": $\bar{4}$. In the same way we can compute the positions $\bar{1}$, $\bar{2}$ and $\bar{3}$. So we get a quadrangle $\bar{1}$, $\bar{2}$, $\bar{3}$, $\bar{4}$. (figure 1)

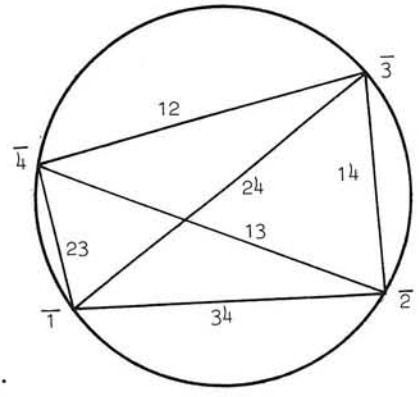


figure 1.

The line between points $\bar{1}$ and $\bar{2}$ is lane (3-4) because transmitters 1 and 2 are not used for points $\bar{1}$ and $\bar{2}$.
The error triangle which we get if we use transmitter 1 as a master consists of the LOP's 1-2, 1-3, 1-4 and is therefore triangle $\bar{2}$, $\bar{3}$, $\bar{4}$ thus not using 1. If we adopt transmitter 2 as the master the error triangle will be $\bar{1},\bar{3},\bar{4}$, a.s.o.
The form of figure 1 can be computed if we know approximate coordinates, thus without observations. If we observe 3 LOP's this means that the position (2 parameters) and the scale (1 parameter) of the figure are determined.

Now we will state two peculiar properties of the quadrangle $(\bar{1},\bar{2},\bar{3},\bar{4})$:

1. The four points $\bar{1}$, $\bar{2}$, $\bar{3}$, $\bar{4}$ are situated on a circle.
2. The lane widths are inversely proporsional to the opposite side length.

The proof is given in the appendix.

This means that if the form of the quadrangle is given the lane widths are also fixed and therefore the position of the adjusted point.

The adjustment is usually done for one triangle, combining the master with the three slaves. If we adjust without correlation between the lanes the position is always inside the triangle. We get a different position if we choose other master-slave and slave-slave combinations.
However, if we adjust with the proper correlation coefficient ρ the adjusted position is independent of the observed combinations of master-slave or slave-slave combinations. That means that the adjusted position corresponds to the position obtained from the best LOP combinations.

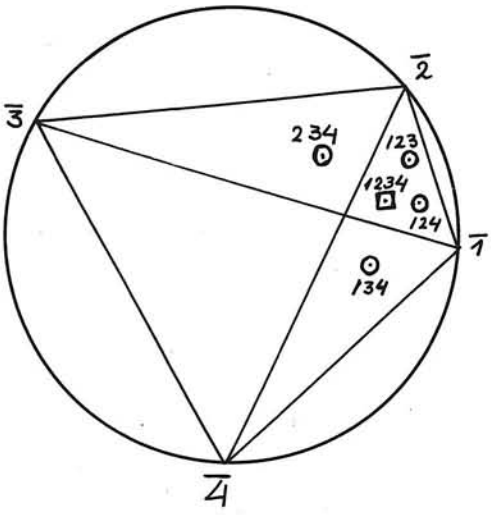
In the figures 2, 3, 4 examples are computed for different situations. It clearly shows that the position with correlation 0.5 fits better with the form of the quadrangle, than the positions computed without correlation.

At sea, on board of a pitching and rolling ship, one usually has no time for theoretical considerations. Therefore a simple and general rule should be followed. This is:
Use always Master-Slave combinations and compute the position with a fixed correlation ρ. In the next section the adjustment formulas are derived.

Adjustment of radiolocation systems

⊙ without correlation

⊡ with correlation 0.5



Senders: 1, 2, 3, 4.
$\bar{1}$ is position from the LOP's 2-3, 2-4, 3-4 thus not using 1. The same for $\bar{2}$, $\bar{3}$, $\bar{4}$.

Figure 2

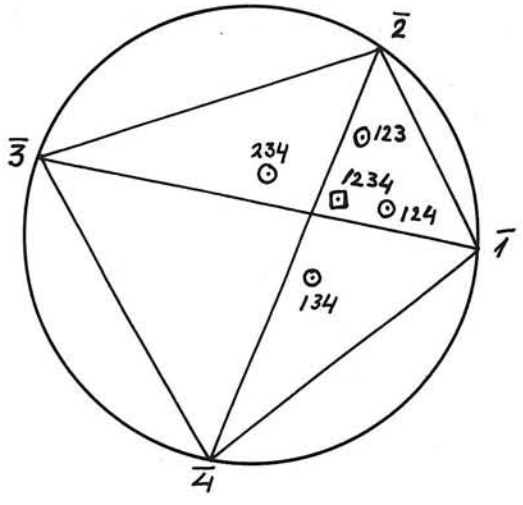


Figure 3

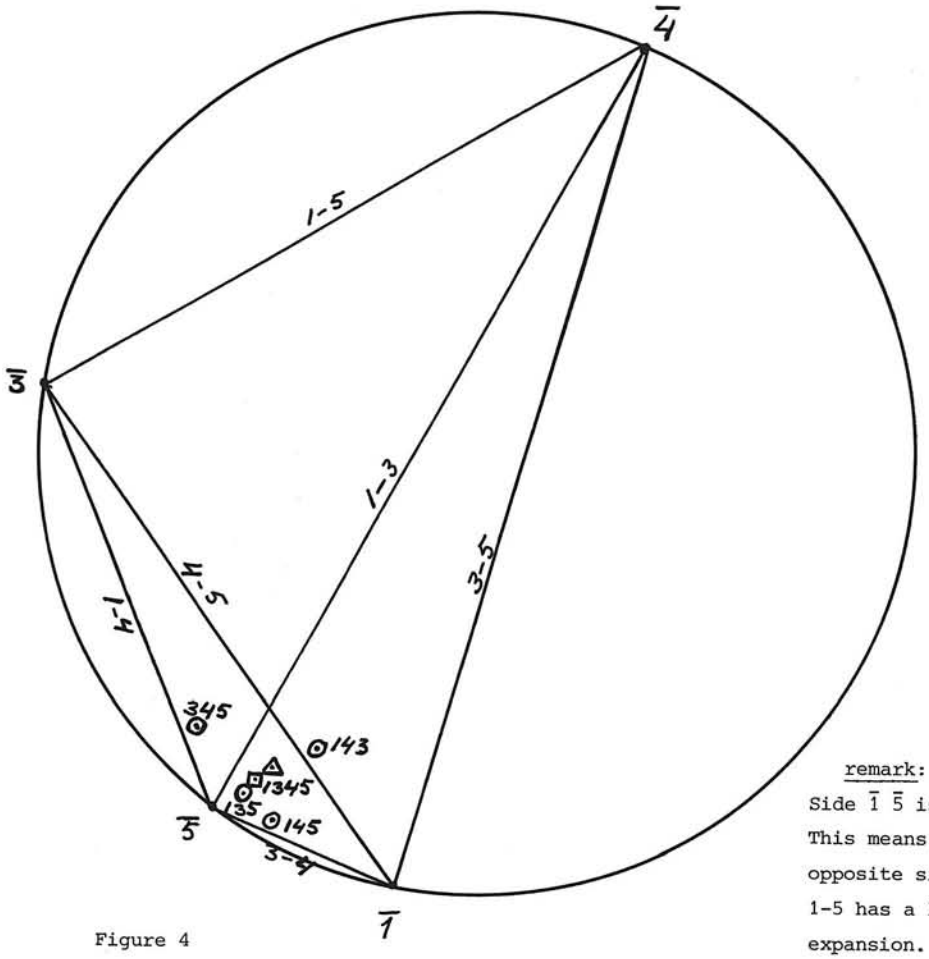


Figure 4

remark:
Side $\bar{1}\ \bar{5}$ is small. This means that the opposite side, LOP 1-5 has a large lane-expansion.

⊙ Position from one triangle computed without correlation between the lanes.

⊡ The same, but with correlation coefficient 0.5 (independent of the chosen triangle).

△ Computed from the lanes 3-4, 4-5 and 1-3 (Van Gein)

With correlation coefficient $0 < \rho < 0.5$ the position will move from ⊙ to ⊡ approximately linear with ρ.

3.3.3 Adjustment.

For adjustment computation on board the ship it is necessary to have a fast computer program for real time data processing. R. Nicolai developed a program which made use of optimal data processing for a small on board computer (Nicolai, 1982). Also A. Houtenbos developed a method of real time data processing which makes use of Kalman filtering. The new position measurements are combined with the previous measurements to determine the best estimate of the position (Houtenbos, 1982). The inversion of the correlation matrix takes much time and it is therefore better to invert the matrix analytically.

We call the lane observations minus the lane numbers computed from the approximate position: vector $b^* = (b_1, b_2, b_3)$

b^* = transpose vector of b.

The model is

$$b + \varepsilon = Ax$$

ε is the correction to the LOP numbers.

x is the difference vector between the approximate position and the adjusted position, x_1 = east, x_2 = north.

A is the matrix of coefficients:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \quad \text{with} \quad a_{i1} = \frac{\cos \alpha_i}{lw_i} \quad , \quad a_{i2} = \frac{-\sin \alpha_i}{lw_i}$$

α = azimuth of the LOP from the north

lw = lane width

If we observe only master-slave combinations, the covariance matrix of the observed lanes is:

$$Q = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

Least squares adjustment gives (see part I-5):

$$x = (A^* Q^{-1} A)^{-1} A^* Q^{-1} b.$$

We may simplify the computation if we compute a matrix R such that

$$Q^{-1} = cRR$$

c is a constant which allows for a convenient scale of R.

Suppose $\quad RA = \bar{A}$

$\quad\quad\quad Rb = \bar{b}$

then $x = (\bar{A}^{*}\bar{A})^{-1}\bar{A}\,\bar{b}$

This is the same formula as for correlation-free adjustment.

R can be computed analytically.

If we define r so that:

$$R = \begin{pmatrix} 1-r & -r & -r \\ -r & 1-r & -r \\ -r & -r & 1-r \end{pmatrix}$$

and solve r and c we get: $r = \frac{1}{3}\cdot\left(1-\sqrt{\frac{1-\rho}{1+2\rho}}\right)$ and $c = \frac{1}{1-\rho}$

ρ	0	0.1	0.2	0.3	0.4	0.5
r	0	0.04	0.08	0.11	0.14	0.17
c	1.0	1.1	1.2	1.4	1.7	2.0

thus $\bar{b} = Rb = b - r\cdot\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\cdot(b_1+b_2+b_3)$

$$\bar{A} = RA = A - r\cdot\begin{pmatrix} p & q \\ p & q \\ p & q \end{pmatrix} \quad \text{with} \quad \begin{cases} p = a_{11}+a_{12}+a_{13} \\ q = a_{21}+a_{22}+a_{23} \end{cases}$$

The modification of the computer programs is now rather simple. We have to subtract from each element of the observation vector b and the matrix A, r times the sum of the column vector. The further adjustment is the same as for correlation free observations.

ρ should be introduced in the program as a parameter that can be changed by hand. So we are able to compute the best position in each case by chosing ρ.

If we have observed n patterns of master-slave combinations instead of 3, the values of r and c become:

$$r = \frac{1}{n}\cdot\left(1-\sqrt{\frac{1-\rho}{1+(n-1)\rho}}\right) \quad \text{and} \quad c = \frac{1}{1-\rho}$$

The adjustment program may also be used for control on lane-slips. The correction ε can be computed by

$$\varepsilon = Ax - b \quad \text{or} \quad R\varepsilon = \bar{\varepsilon} = \bar{A}x - \bar{b}$$

The estimation of the standard derivation of the observations is

$$s^2 = \frac{\varepsilon^* Q^{-1} \varepsilon}{n-2} \quad \text{or} \quad s^2 = \frac{\bar{\varepsilon}^* \bar{\varepsilon}}{(1-\rho)(n-2)} \text{ because } Q^{-1} = \frac{R^* R}{1-\rho}$$

n is the number of lanes observed.

For the test on lane-slips we have to adopt a norm for the standard deviation (σ^2). The quotient (s^2/σ^2) has a F-distribution with (n-2) degrees of freedom.

If $$\frac{s^2}{\sigma^2} \geq F(n-2, \infty, 95\%)$$

we may conclude to a lane-slip with a probability of 95%. On the other hand, if s^2 is small, this will not be a guarantee that everything is alright.
Another test on lane slips is obtained by Kalman filtering (see part I-12 and IV-2).

For n = 3 is $F(n-2, \infty, 95\%) = 3.84$.
So we may conclude to a lane-slip if:

$$s > 1.96\ \sigma \simeq 2\sigma$$

The standard ellipse of the adjusted point may be computed by:

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} = \sigma^2 (A^* Q^{-1} A)^{-1} = \sigma^2 (1-\rho)(\bar{A}^* \bar{A})^{-1}$$

This leads to a standard ellipse with axis:

$$a^2 = \tfrac{1}{2}(\sigma_x^2 + \sigma_y^2) + \tfrac{1}{2}\sqrt{(\sigma_x^2 - \sigma_y^2) + 4\sigma_{xy}^2}$$

$$b^2 = \tfrac{1}{2}(\sigma_x^2 + \sigma_y^2) - \tfrac{1}{2}\sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4\sigma_{xy}^2}$$

$$\tan 2\theta = \frac{2\sigma_{xy}}{\sigma_x^2 - \sigma_y^2}$$

θ is the angle of the ellipse axis with the x-axis.

If $\sigma_x > \sigma_y$, then the major axis a is close to the x-axis.

3.3.4 Error figure for 5 transmitters.

If we have 5 transmitters we get:

$$\binom{5}{2} = \frac{5!}{2!.3!} = 10 \text{ position lines (LOP's)}$$

$$\binom{5}{3} = \frac{5!}{3!.2!} = 10 \text{ points where 3 LOP's cross.}$$

$$\binom{5}{4} = \frac{5!}{4!.1!} = 5 \text{ circles with 4 points on each.}$$

There are always 3 points on each LOP. So it is complicated to construct a figure which satisfies all these conditions. Figure 5 is an example.

There are only 5 degrees of freedom in the figure, 4 angles and 1 distance-ratio. e.g. (13-12):(12-23).

A remarkable condition is that all the circles go through one point. This can be proved by expressing the angles from this point to the other cross points as a function of the 4 independent angles.

The other cross points of the circles are just the 10 points of LOP crossings.

3.3.5 Conclusion

To adjust the position of a hyperbolic positioning system it is important to introduce correlation between the observed lane numbers. The effect of the correlation can easily be introduced in the computer program. A correlation between 0.3 and 0.4 for Master-Slave combinations seems to be best suited.

Error figure for 5 transmitters.

2-4 : Lane of Position (LOP)

24 : Point from transmitter combination (1,3,5) (not using 2 and 4).

Note: All circles go through one point.

Only 4 independent angles!

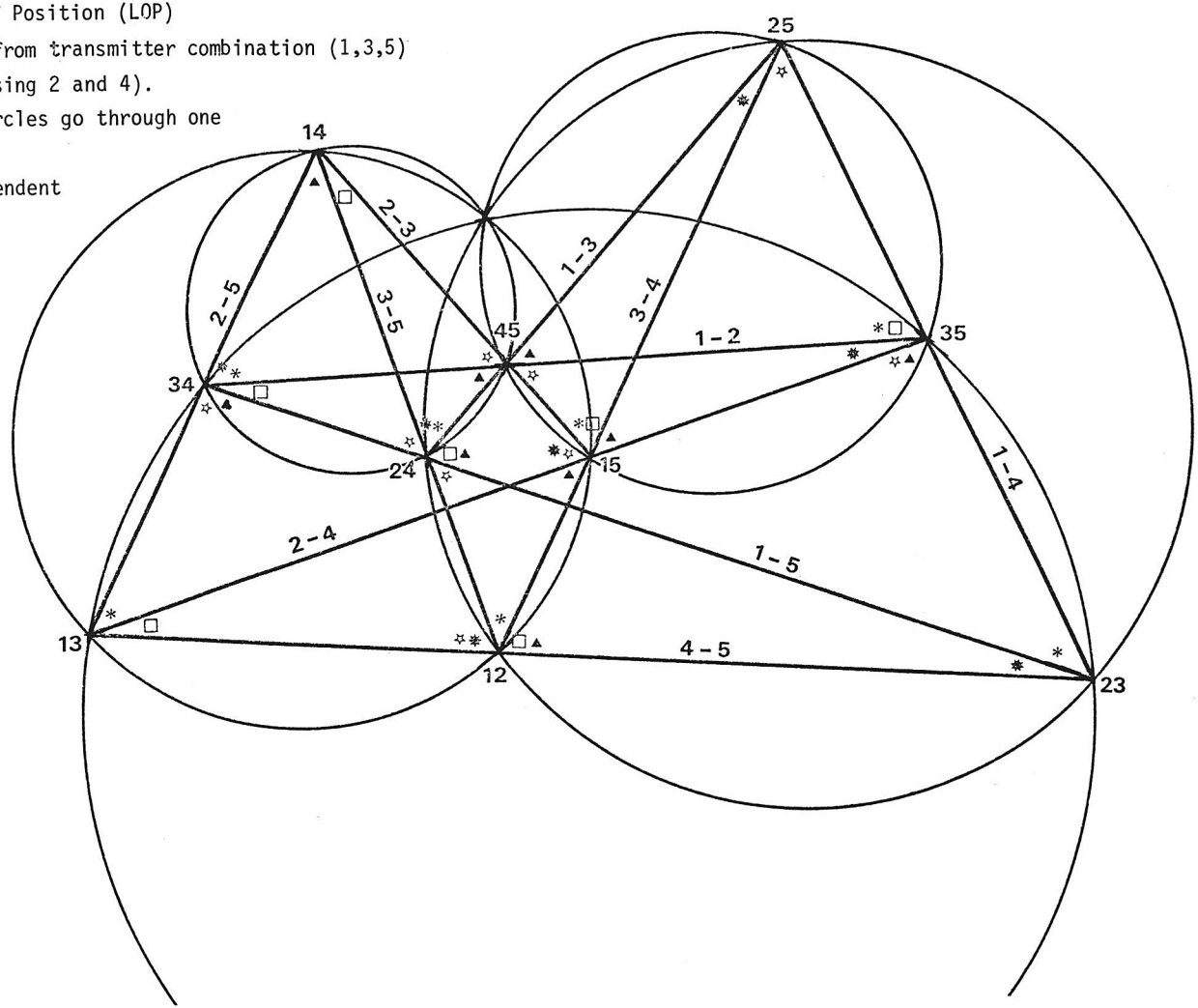


figure 5.

3.3.6 Literature

Van Gein, W.A.: Operational Properties and possibilities of Hifix/6. Hydrographic Service Royal Netherlands Navy, 1981.

Houtenbos, A.P.E.M.: Prediction, filtering and smoothing of offshore navigation data. In: "Forty years of thought" Geodetic Department, Delft, 1982.

Mikhail, E.M.: Observations and Least Squares, Harper and Row Publishers.

Nicolai, R.: Lane Control by least squares adjustment, Survey Department of Rijkswaterstaat Delft.

3.3.7. Appendix.

Proof of the geometrical properties of the error figure

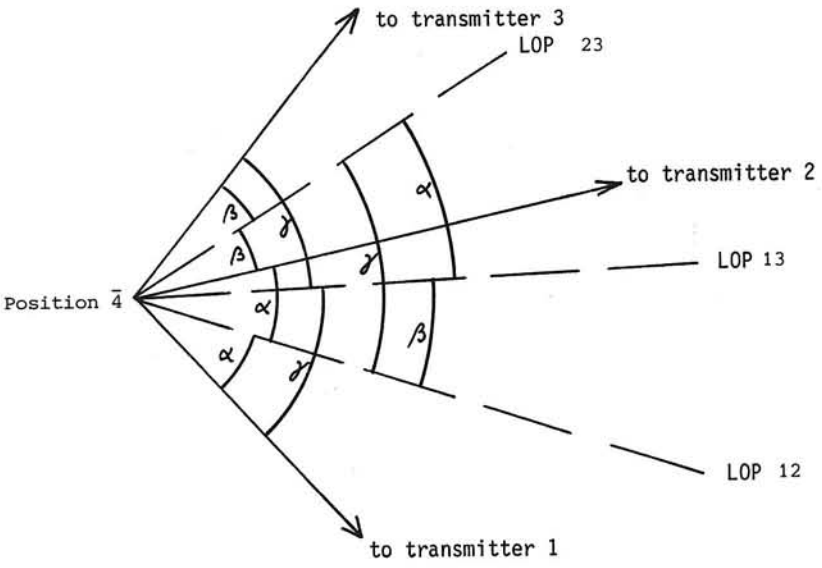


Figure 6

$$\text{lane width} = \frac{\text{wavelength } \lambda}{2 \cdot \sin \frac{1}{2}(\text{angle between the directions to the transmitters})}$$

So: $lw(12) = \lambda/2 \sin \alpha$

$lw(23) = \lambda/2 \sin \beta$

$lw(13) = \lambda/2 \sin \gamma$

The LOP's are the bissectrices between the directions to the transmitters. From the geometry of fig. 6 follows that the angles between the LOP's are half of the angles between the transmitters.

From figure 6 we see:

$$2\alpha + 2\beta = 2\gamma$$

or $$\alpha + \beta = \gamma$$

The angles between the LOP's are:

angle (12-23) = $\alpha + \beta = \gamma$ so lw(13) = $\lambda / 2 \sin(12\text{-}23)$
angle (12-13) = $\gamma - \alpha = \beta$ so lw(23) = $\lambda / 2 \sin(12\text{-}13)$
angle (13-23) = $\gamma - \beta = \alpha$ so lw(12) = $\lambda / 2 \sin(13\text{-}23)$

If we consider the quadrangle $\bar{1}$, $\bar{2}$, $\bar{3}$, $\bar{4}$, (figure 7) the lane width of e.g. (1-2) is inversely proportional to sin angle ($\bar{1}$, $\bar{3}$, $\bar{2}$) and, for the same reason, also to sin angle ($\bar{1}$, $\bar{4}$, $\bar{2}$).

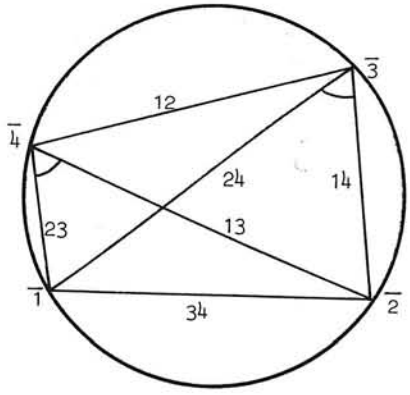


figure 7.

Thus angle ($\bar{1}$, $\bar{3}$, $\bar{2}$) = angle ($\bar{1}$, $\bar{4}$, $\bar{2}$)

<u>This means that point $\bar{4}$ lays on the circle through $\bar{1}$, $\bar{2}$, $\bar{3}$.</u>

However if the points $\bar{1}$, $\bar{2}$, $\bar{3}$, $\bar{4}$ are on a circle the sides are proportional to the sin of the opposite angles e.g. side ($\bar{1}$ $\bar{2}$) = 2Rsin($\bar{1}$ $\bar{3}$ $\bar{2}$).

<u>Thus, the lane widths are inversely proportional to the opposite sidelengths.</u>

Thus lane width (12) . side ($\bar{1}$ $\bar{2}$) = lane width (13) . side($\bar{1}$ $\bar{3}$) =
= lane width (14) . side ($\bar{1}$ $\bar{4}$) = lane width (23) . side ($\bar{2}\bar{3}$) =
= lane width (24) . side ($\bar{2}$ $\bar{4}$) = lane width (34) . side ($\bar{3}$ $\bar{4}$)

So if e.g. side ($\bar{1}$ $\bar{2}$) is large the lane width of LOP (12) is small.

3.4 Precision of radiopositioning systems.

G.L. Strang van Hees

3.4.1. Introduction.

Some concise formulas for the point precision of radiopositioning systems are derived.
The variance $(\sigma_x^2 + \sigma_y^2)$ of the point precision is expressed in nice symmetric formulas in three cases:
- range-range mode: formula (9);
- hyperbolic mode with 3 transmitters: formula (15);
- hyperbolic mode with 4 transmitters: formula (25).

The variables in these formulas are the angles between the transmitters, seen from the receiver (ship).

Two types of radiopositioning systems are distinguised:
a. range systems;
b. hyperbolic systems.
The observation equations in matrix notation are:

$$b + e = AX \tag{1}$$

b is the vector with observations. In range-mode: b are the distances to the transmitter. In hyperbolic mode: b are the distance differences between master and slaves or two slave stations.
e is the vector with corrections.
A is the design matrix.
X is the vector with the coordinates of the receiver station.
In range mode :

$$A = \begin{pmatrix} -\sin\alpha_1 & -\cos\alpha_1 \\ -\sin\alpha_2 & -\cos\alpha_2 \\ \vdots & \vdots \\ -\sin\alpha_n & -\cos\alpha_n \end{pmatrix} \tag{2}$$

In hyperbolic mode:

$$A = \begin{pmatrix} \sin\alpha_1 - \sin\alpha_m & \cos\alpha_1 - \cos\alpha_m \\ \sin\alpha_2 - \sin\alpha_m & \cos\alpha_2 - \cos\alpha_m \\ \vdots & \vdots \\ \sin\alpha_n - \sin\alpha_m & \cos\alpha_n - \cos\alpha_m \end{pmatrix} \qquad (3)$$

$\alpha_1 \ldots \alpha_n$ are the azimuths from the receiver (ship) to the transmitters. α_m is the azimuth to the Master station. For this purpose it is sufficient to compute the azimuths on the sphere with the following formula

$$\tan\alpha = \frac{\sin(\lambda_t - \lambda_r)}{\tan\varphi_t \cos\varphi_r - \sin\varphi_r \cos(\lambda_t - \lambda_r)} \qquad (4)$$

α = azimuth from receiver (r) to transmitter (t). (φ_t,λ_t), (φ_r,λ_r) are latitude and longitude of transmitter and receiver.
Q is the variance matrix between the observations.

In the range-mode Q can be taken equal to the unit matrix. In the hyperbolic mode a correlation between the observations should be taken into account, because the master-station is involved in all the range differences.
As derived in 3.3.3 the Q matrix is, in case of Master-Slave combinations:

$$Q = \begin{pmatrix} 1 & \rho & \rho & \ldots \\ \rho & 1 & \rho & \ldots \\ \rho & \rho & 1 & \ldots \\ \vdots & \vdots & \vdots & \end{pmatrix} \qquad (5)$$

where

$$\rho = \frac{\sigma^2_{Master}}{\sigma^2_{Master} + \sigma^2_{Slave}}$$

In case of Slave-Slave-combinations Q can also be expressed in ρ. This expression can be found by using the law of covariance propagation. However, the final precision of the point is independent of the combinations used, if the correct correlations are taken into account.
The standard ellipse of the point precision can be computed from the adjustment:

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} = \sigma^2 (A^* \, Q^{-1} \, A)^{-1} \quad .$$

σ is the standard deviation of the observation. In range mode: σ is the standard deviation of the measured ranges in meter. In hyperbolic mode: σ is the standard deviation of the ranges differences between master and slave, this is the wavelength λ times the precision of the observed lane number, e.g.: $\sigma = 0.01\ \lambda$, if the precision is 0.01 lane.

An indication of the point precision is the trace of the variance matrix. This trace is equal to a^2+b^2 with a and b the axis of the standard ellipse.

$$\sigma_p^2 = \sigma_x^2 + \sigma_y^2 = \sigma^2 \cdot \text{trace}(A^* \, Q^{-1} \, A)^{-1} \qquad (6)$$

The objective of this paper is to derive an expression of σ_p, for range and for hyperbolic mode.

3.4.2. Range mode.

In the range mode is Q the unit matrix. A is defined by (2)

$$A^*A = \begin{pmatrix} \Sigma \sin^2 \alpha_i & \Sigma \sin \alpha_i \cos \alpha_i \\ \Sigma \sin \alpha_i \cos \alpha_i & \Sigma \cos^2 \alpha_i \end{pmatrix}$$

Determinant is:

$$D = \sum_{i=1}^{n} \sin^2 \alpha_i \cdot \sum_{i=1}^{n} \cos^2 \alpha_i - \left(\sum_{i=1}^{n} \sin \alpha_i \cdot \cos \alpha_i \right)^2 \quad .$$

After some manipulations with trigonometric functions this expression can be transformed to:

$$D = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sin^2(\alpha_i - \alpha_j) \quad . \qquad (7)$$

The inverse of (A^*A) is

$$(A^*A)^{-1} = \frac{1}{D} \begin{pmatrix} \Sigma \cos^2 \alpha_i & -\Sigma \sin \alpha_i \cdot \cos \alpha_i \\ -\Sigma \sin \alpha_i \cdot \cos \alpha_i & \Sigma \sin^2 \alpha_i \end{pmatrix}$$

The trace is:

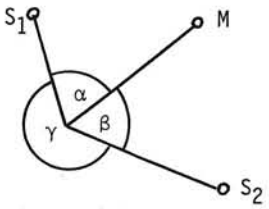
$$\text{Trace} = \frac{1}{D}\left(\sum_{i=1}^{n}\cos^2\alpha_i + \sum_{i=1}^{n}\sin^2\alpha_i\right) = \frac{n}{D} \qquad (8)$$

Insert (7) and (8) in (6):

$$\sigma_P^2 = \sigma_x^2 + \sigma_y^2 = \frac{n \cdot \sigma^2}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sin^2(\alpha_i-\alpha_j)} \qquad (9)$$

This rather simple final formula gives the point precision in a range-mode system.
It is dependent on the angles $(\alpha_i-\alpha_j)$ between the transmitters as seen from the receiver station P (the ships position). E.g. for a three range system is:

$$\sigma_P^2 = \frac{3 \cdot \sigma^2}{(\sin^2\alpha + \sin^2\beta + \sin^2\gamma)}$$



3.4.3. Hyperbolic mode.

The precision of a computed position in a hyperbolic system depends on the variances of the master and slave signals, σ_M^2 and σ_S^2. This is expressed by the parameter (3.3.1)

$$\rho = \frac{\sigma_M^2}{\sigma_M^2 + \sigma_S^2}$$

In general is $\sigma_S^2 > \sigma_M^2$ and so: $0 < \rho < 0.5$. For the calculation of the coordinates it is important to choose ρ as good as possible. However, one gets a good impression of the precision if one assumes $\sigma_S^2 = \sigma_M^2$ or $\rho = 0.5$. This assumption means that the Master and Slave stations may be interchanged and therefore nice symmetric formulas are obtained.

3.4.3.1. Hyperbolic system with 3 transmitters.

The design matrix is given by (3). For one Master and two Slaves A becomes a 2 x 2 matrix which can be written in short:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{10}$$

The inverse of Q (5), is:

$$Q^{-1} = \frac{1}{1 - \rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \tag{11}$$

$$A^* Q^{-1} A = \frac{1}{1 - \rho^2}\begin{pmatrix} a & c \\ b & d \end{pmatrix}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} =$$

$$= \frac{1}{1 - \rho^2}\begin{pmatrix} a^2 + c^2 - 2\rho ac & ab + cd - \rho(ad+bc) \\ ab + cd - \rho(ad+bc) & b^2 + d^2 - 2\rho bd \end{pmatrix}$$

This matrix should be inverted according to (5). The determinant of $(A^* Q^{-1} A)$ is obtained as the product of the determinants of the components:

$$\mathrm{Det}(A^* Q^{-1} A) = \frac{1}{(1-\rho^2)^2} \cdot \det(A) \cdot \det(Q^{-1}) \cdot \det(A)$$

$$\mathrm{Det}(A^* Q^{-1} A) = \frac{(ad-bc)^2}{1 - \rho^2}$$

So the inverse is:

$$(A^* Q^{-1} A)^{-1} = \frac{1}{(ad-bc)^2}\begin{pmatrix} b^2 + d^2 - 2\rho bd & \cdots \\ \cdots & a^2 + c^2 - 2\rho ac \end{pmatrix}$$

The trace is

$$\mathrm{trace} = \frac{a^2 + b^2 + c^2 + d^2 - 2\rho(ac+bd)}{(ad-bc)^2} \tag{12}$$

We now insert (3) into (12) and choose $\rho = 0.5$. After some manipulations with trigonometric functions we obtain:

$$a^2 + b^2 = 2 - 2\cos(\alpha_3 - \alpha_1)$$
$$c^2 + d^2 = 2 - 2\cos(\alpha_2 - \alpha_3)$$

$$ac + bd = \cos(\alpha_1 - \alpha_2) - \cos(\alpha_2 - \alpha_3) - \cos(\alpha_3 - \alpha_1) + 1$$

or

$$a^2 + b^2 + c^2 + d^2 - 2\rho(ac+bd) =$$

$$= 3 - \cos(\alpha_1-\alpha_2) - \cos(\alpha_2-\alpha_3) - \cos(\alpha_3-\alpha_1) \qquad (13)$$

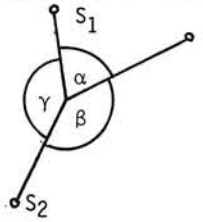
$$ad - bc = \sin(\alpha_1-\alpha_2) + \sin(\alpha_2-\alpha_3) + \sin(\alpha_3-\alpha_1) \qquad (14)$$

Note the cycle in the indices!
Inserting (12), (13), (14) into (6) gives finally:

$$\sigma_x^2 + \sigma_y^2 = \frac{\sigma^2(3-\cos(\alpha_1-\alpha_2) - \cos(\alpha_2-\alpha_3) - \cos(\alpha_3-\alpha_1))}{(\sin(\alpha_1-\alpha_2) + \sin(\alpha_2-\alpha_3) + \sin(\alpha_3-\alpha_1))^2} \qquad (15)$$

This symmetric formula expresses the point precision as a function of the angles between the transmitters as seen from the receiver (the ship). Expressed in the angles α, β, γ (see figure) it becomes:

$$\sigma_p^2 = \frac{\sigma^2(3 - \cos\alpha - \cos\beta - \cos\gamma)}{(\sin\alpha + \sin\beta + \sin\gamma)^2} \qquad (16)$$



Note that the direction of counting α, β, γ is important, because $\sin\alpha = -\sin(-\alpha)$ a.s.o.
σ is the precision of the observed lane number, multiplied by the wavelength λ:

$$\sigma = \sigma_{lane} \cdot \lambda \quad .$$

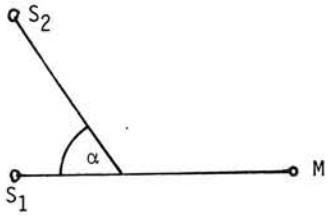
Example 1: $\sigma_{lane} = 0.01$

$$\alpha = 60^0, \ \beta = 60^0, \ \gamma = 240^0$$

$$\sigma_p = 0.01\ \lambda \cdot \sqrt{\frac{10}{3}} = 0.018\ \lambda \quad .$$

Example 2: On a baseline is:

$$\sigma_P = \frac{\sigma\sqrt{3 - \cos\alpha + \cos\alpha + 1}}{\sin\alpha + \sin\alpha + 0} = \frac{\sigma}{\sin\alpha}$$



3.4.3.2. Hyperbolic system with 4 transmitters.

The variance matrix of the coordinates is in case of one Master and 3 Slaves (3.3.3):

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} = \sigma^2 \cdot (1-\rho)(\bar{A}^*\bar{A})^{-1} \qquad (17)$$

with $\bar{A} = RA$. Again we choose: $\rho = \frac{1}{2}$. This gives:

$$R = \frac{1}{6}\begin{pmatrix} 5 & -1 & -1 \\ -1 & 5 & -1 \\ -1 & -1 & 5 \end{pmatrix} \qquad (18)$$

$$A = \begin{pmatrix} \sin\alpha_1 - \sin\alpha_4 & \cos\alpha_1 - \cos\alpha_4 \\ \sin\alpha_2 - \sin\alpha_4 & \cos\alpha_2 - \cos\alpha_4 \\ \sin\alpha_3 - \sin\alpha_4 & \cos\alpha_3 - \cos\alpha_4 \end{pmatrix} \qquad (19)$$

α_i = azimuth from receiver to the transmitter i = 1, 2, 3, 4. In short notation:

$$\bar{A} = \begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix} \qquad (20)$$

$$(\bar{A}^*\bar{A}) = \begin{pmatrix} a^2 + c^2 + e^2 & ab + cd + ef \\ ab + cd + ef & b^2 + d^2 + f^2 \end{pmatrix}$$

The determinant becomes after some manipulations:

$$\text{Det} = (ad-bc)^2 + (af-be)^2 + (cf-de)^2 \tag{21}$$

According to (17) is:

$$\sigma_x^2 + \sigma_y^2 = \tfrac{1}{2}\sigma^2 \cdot \text{trace}(\bar{A}^*\bar{A})^{-1} \tag{22}$$

$$\text{trace}(\bar{A}^*\bar{A})^{-1} = \frac{a^2 + b^2 + c^2 + d^2 + e^2 + f^2}{(ad-bc)^2 + (af-be)^2 + (cf-de)^2} \tag{23}$$

The a, b, c, d, e, f are now replaced by the expressions that follow from (18) and (19).

This gives a rather complicated expression, but after a lot of trigonometric reductions this expression can be simplified.
The following notation is used:

$$S_{ik} = \sin(\alpha_i - \alpha_k) \quad , \quad C_{ik} = \cos(\alpha_i - \alpha_k) \tag{24}$$

Note that $S_{ik} = -S_{ki}$ and $C_{ik} = C_{ki}$

therefore the sequence of the indices is important for S_{ik}!
The resulting final formula for the point precision is

$$\sigma_x^2 + \sigma_y^2 = \sigma^2 \frac{(6 - C_{12} - C_{13} - C_{14} - C_{23} - C_{24} - C_{34})}{K_1 + K_2 + K_3 + K_4} \tag{25}$$

with
$$K_1 = (S_{23} + S_{34} + S_{42})^2$$
$$K_2 = (S_{13} + S_{34} + S_{41})^2$$
$$K_3 = (S_{12} + S_{24} + S_{41})^2$$
$$K_4 = (S_{12} + S_{23} + S_{31})^2$$

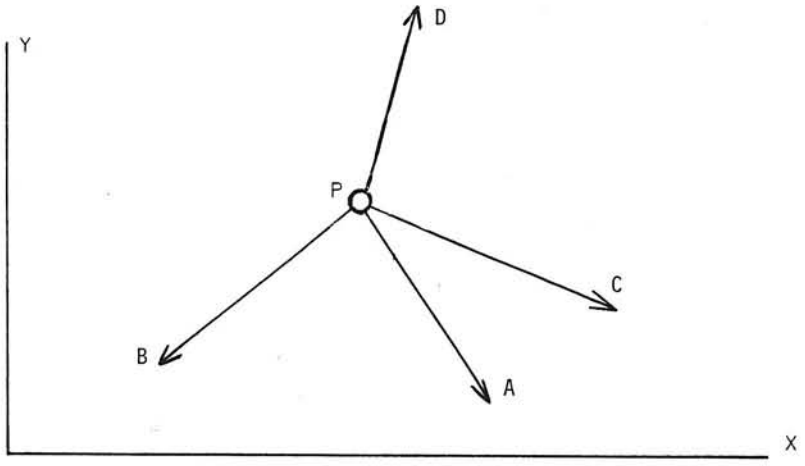
with $\sigma = \lambda \cdot \sigma$ (lane number), λ = wavelength of the transmitted signal.
Again this is a symmetric expression in 1, 2, 3 and 4, as it should be.
If the positions of the transmitters and receiver (ship) are given, it is possible to calculate the azimuths with (4) and from this the point precision with (25).

3.4.4.

Excercise Least squares adjustment.

Excercise 1.

Suppose we have 4 transmitters A, B, C, and D. From a ship in point P the distances to the 4 transmitters are measured (range-range method). The distances are reduced to the plane of the map projection. The coordinates of the transmitters are given in this map projection. Hence, the computation and adjustment can be carried out in the x,y-plane. P' is the approximate position of the ship.



Given coordinates:

	Latitude	Longitude	
A:	52.00	4.00	
B:	52.50	2.00	
C:	52.80	3.80	
D:	55.00	4.00	
P':	53.00	3.00	approximate coordinates

Measured distances:

PA:	130165 m
PB:	87305 m
PC:	58085 m
PD:	231770 m

The measured distances are assumed to be free of correlation and with the same precision.

Compute the least squares estimate of the coordinates of P and the corrections to the distances.
Compute the estimated precision of the distances and the standard ellipse of point P.

Excercise 2.

Suppose A is the master transmitter and B, C and D are slaves.
Now the distance differences are measured instead of the distances itself.

PA - PB = +42860 m
PA - PC = +72080 m
PA - PD =-101605 m

In practice this is the case wit a hyperbolic positioning system. Compute again the position and the standard deviations by least squares adjustment.

As the distance PA is involved in all three distance differences, these observations are correlated. Usually this correlation is neglected. You can carry out the adjustment with and without the correlation and see the differences. Especially if the intersection of the lanes is bad, it is important to take the correlation into account.
With a correlation coefficient $\rho = 0.5$ the solution is independent of what transmitter is chosen as a Master station.

```
EXERCISE LEAST SQUARES ADJUSTMENT
*********************************

PROGRAM DESIGN:  IR.G.L. STRANG VAN HEES


APPROXIMATE COORD.   LAT = 53.00000       LON =  3.00000


TRANSMITTER          LAT = 52.00000       LON =  4.00000
MEASURED DISTANCE        = 130165.00 METER

DISTANCE FROM APPR.POINT = 130175.37     AZIMUTH FROM APPR.POINT = 148.2709

TRANSMITTER          LAT = 52.50000       LON =  2.00000
MEASURED DISTANCE        =  87305.00 METER

DISTANCE FROM APPR.POINT =  87297.62     AZIMUTH FROM APPR.POINT =-129.1594

TRANSMITTER          LAT = 52.80000       LON =  3.80000
MEASURED DISTANCE        =  58085.00 METER

DISTANCE FROM APPR.POINT =  58084.47     AZIMUTH FROM APPR.POINT = 112.1922

TRANSMITTER          LAT = 55.00000       LON =  4.00000
MEASURED DISTANCE        = 231770.00 METER

DISTANCE FROM APPR.POINT = 231788.79     AZIMUTH FROM APPR.POINT =  15.9743


ADJUSTMENT RANGE-RANGE MODE
***************************


   4 OBSERVATIONS,         2 UNKNOWNS.

ST.DEV.=10.00      ALPHA=0.050      CRITICAL VALUE= 2.00

OBSERVATIONS

-10.365       7.384       0.530     -18.786


WEIGHT-COEFFICIENTS: Q(I,J), (AS VECTOR)

0.100E+01  0.000E+00  0.100E+01  0.000E+00  0.000E+00  0.100E+01  0.000E+00  0

DESIGN MATRIX A(I,J)

1     -0.526       0.775      -0.926      -0.275
2      0.851       0.631       0.378      -0.961
```

NR.	OBSERV.	ST.DV.	CORREC.	OBS+CO.	ST.DV.	W-TEST	CODE	MA
1	-10.365	11.842	11.308	0.942	8.151	1.316	0	4
2	7.384	11.842	3.706	11.090	8.580	0.454	0	4
3	0.530	11.842	-6.282	-5.752	8.632	-0.775	0	4
4	-18.786	11.842	9.970	-8.816	8.118	1.156	0	4

MARGINAL ERROR = ERROR IN OBSERVATION WHICH CAN BE FOUND WITH 80% PROBABILIT

REL.= RELIABILITY = MARGINAL ERROR DIVIDED BY THE STANDARD DEVIATION, FOR OB

R = VARIANCE OF THE CORRECTIONS

R/Q = IMPROVEMENT, (0 .LE. R/Q .LE. 1), IF (R/Q .GT. 0.7) THEN GOOD.

PRP/P=RELIABILITY, (0 .LE.PRP/P.LE. 1), IF (PRP/P.GT. 0.7) THEN GOOD.

ESTIMATION STANDARD DEVIATION WEIGHT-UNIT: 11.842

APRIORI STANDARD DEVIATION WEIGHT-UNIT: 10.000

ESTIMATION VARIANCE FACTOR: 140.233098

APRIORI VARIANCE FACTOR: 100.000000

F-TEST: 1.402 CRITICAL VALUE:

SOLUTION:

NR.	X(I)	STAND.DEV.	MARG.ERR.
1	8.913	8.802	24.929
2	6.619	8.006	20.434

VARIANCE MATRIX OF UNKNOWNS, Q(X,X).

0.552 0.011 0.457

ADJUSTED COORDINATES RANGE-RANGE METHOD.

LON P = 3.000133 SDX = 7.43 M COV(X,Y) = 1.08

LAT P = 53.000060 SDY = 6.76 M COV(X,Y) = 1.08

A = 7.44 M B = 6.75 M ANGLE WITH X-AXIS = 6.38

ADJUSTMENT HYPERBOLIC MODE, CORRELATION 0.0
****

ST.DEV.=15.00 ALPHA=0.050 CRITICAL VALUE= 2.00

OBSERVATIONS

17.749 10.895 -8.421

WEIGHT-COEFFICIENTS: Q(I,J), (AS VECTOR)

1.000 0.000 1.000 0.000 0.000 1.000

DESIGN MATRIX A(I,J)

1	1.301	-0.400	0.251
2	-0.219	-0.473	-1.812

NR.	OBSERV.	ST.DV.	CORREC.	OBS+CO.	ST.DV.	W-TEST	CODE	MA
1	17.749	15.000	-5.957	11.792	14.157	-1.201	0	12
2	10.895	15.000	-16.268	-5.373	6.455	-1.201	0	4
3	-8.421	15.000	4.965	-3.456	14.419	1.201	0	15

MARGINAL ERROR = ERROR IN OBSERVATION WHICH CAN BE FOUND WITH 80% PROBABILIT
REL.= RELIABILITY = MARGINAL ERROR DIVIDED BY THE STANDARD DEVIATION, FOR OB
R = VARIANCE OF THE CORRECTIONS
R/Q = IMPROVEMENT, (0 .LE. R/Q .LE. 1), IF (R/Q .GT. 0.7) THEN GOOD.
PRP/P=RELIABILITY, (0 .LE.PRP/P.LE. 1), IF (PRP/P.GT. 0.7) THEN GOOD.

ESTIMATION STANDARD DEVIATION WEIGHT-UNIT: 18.022

APRIORI STANDARD DEVIATION WEIGHT-UNIT: 15.000

ESTIMATION VARIANCE FACTOR: 324.790110

APRIORI VARIANCE FACTOR: 225.000000

F-TEST: 1.444 VALUE:

SOLUTION:

NR.	X(I)	STAND.DEV.	MARG.ERR.
1	9.607	11.085	89.199
2	3.236	8.139	79.202

VARIANCE MATRIX OF UNKNOWNS, Q(X,X).

0.546 0.085 0.294

LON P = 3.000144 SDX = 11.09 M COV(X,Y) = 19.02
LAT P = 53.000029 SDY = 8.14 M COV(X,Y) = 19.02

A = 11.34 M B = 7.77 M ANGLE WITH X-AXIS = 16.94

ADJUSTMENT HYPERBOLIC MODE, CORRELATION 0.5
****

ST.DEV.=15.00 ALPHA=0.050 CRITICAL VALUE= 2.00

OBSERVATIONS

17.749 10.895 -8.421

WEIGHT-COEFFICIENTS: Q(I,J), (AS VECTOR)

1.000 0.500 1.000 0.500 0.500 1.000

DESIGN MATRIX A(I,J)

1	1.301	-0.400	0.251
2	-0.219	-0.473	-1.812

NR.	OBSERV.	ST.DV.	CORREC.	OBS+CO.	ST.DV.	W-TEST	CODE	MA
1	17.749	15.000	-12.110	5.640	11.297	-1.227	0	12
2	10.895	15.000	-17.488	-6.593	4.681	-1.227	0	4
3	-8.421	15.000	-6.413	-14.834	14.060	1.227	0	15

MARGINAL ERROR = ERROR IN OBSERVATION WHICH CAN BE FOUND WITH 80% PROBABILIT
REL.= RELIABILITY = MARGINAL ERROR DIVIDED BY THE STANDARD DEVIATION, FOR OB
R = VARIANCE OF THE CORRECTIONS
R/Q = IMPROVEMENT, (0 .LE. R/Q .LE. 1), IF (R/Q .GT. 0.7) THEN GOOD.
PRP/P=RELIABILITY, (0 .LE.PRP/P.LE. 1), IF (PRP/P.GT. 0.7) THEN GOOD.

ESTIMATION STANDARD DEVIATION WEIGHT-UNIT: 18.407

APRIORI STANDARD DEVIATION WEIGHT-UNIT: 15.000

ESTIMATION VARIANCE FACTOR: 338.814558

APRIORI VARIANCE FACTOR: 225.000000

F-TEST: 1.506 VALUE:

SOLUTION:

NR.	X(I)	STAND.DEV.	MARG.ERR.
1	5.848	8.466	89.199
2	8.996	7.559	90.875

VARIANCE MATRIX OF UNKNOWNS, Q(X,X).

0.319 -0.027 0.254

LON P = 3.000087 SDX = 8.47 M COV(X,Y) = -6.16
LAT P = 53.000081 SDY = 7.56 M COV(X,Y) = -6.16

A = 8.60 M B = 7.41 M ANGLE WITH X-AXIS = -20.16

3.5 Kalman filtering.

J.C. de Munck

3.5.1. Introduction.

The accuracy of the estimate of an unknown quantity can often be improved by repeating measurements many times and averaging the results. If, however, the unknown quantities and the measuring quantities are part of a time dependent process this procedure is not feasible in its simple form. Nevertheless, for such a process, like position fixing with a moving vehicle, measurements at successive moments may be combined to obtain better knowledge of positions.

A Kalman filter is a procedure to find an "optimal" estimate for some unknowns at any moment based on all measurements up to that moment and on some a priori knowledge at the start of the process. For the Kalman procedure the process has to be known to some extent. For a moving ship this process may therefore consist of the dynamic equations of the motion based on the known forces on the ship from the screw, the rudder, the currents and the wind. The steering of the ship may also be incorporated in this model. Also some stochastic terms are needed because of lack of complete knowledge of the forces.
In an other set of equations, the observation model, the relation between the measurements and the unknowns are expressed, also with some noise terms because of the uncertainties in the measurements.

3.5.2. The models.

The dynamic model can be constructed in different ways. It will often contain some differential equations based on the theory of mechanics, on emperical rules or on intuition. Of all possible differential equations only the linear ones can be solved more or less straightforwardly. So one will nearly always try to define the dynamical model in the form of such equations if not in a still more elementary form.

Now each set of simultaneous differential equations with one independent variable (most often the time t) can be written in the form:

$$\dot{\underline{X}}(t) = F(t)\underline{X}(t) + G(t)\underline{Y}(t), \qquad (1)$$

where $\underline{X}(t)$, the so-called state vector, is a column vector of time dependent unknowns, such as components of position and velocity, course, etc.;
$\underline{Y}(t)$ is a column vector of time dependent steering and noise variables;
$F(t)$ is a square matrix expressing the (linear) relations between the unknowns and their time derivates in absence of steering and noise variables;

and the dot above the $\underline{X}$ vector indicates the time derivative:

$$\dot{\underline{X}} \triangleq \frac{d}{dt}\,\underline{X}$$

For the purpose of the Kalman filter the last term is often split up:

$$G(t)\underline{Y}(t) = B(t)\underline{U}(t) + C(t)\underline{W}(t)\,, \qquad (2)$$

where $\underline{U}(t)$ includes the steering components (for a ship for instance rudder angle and screw power);

$\underline{W}(t)$ the noise of the dynamic model, representing the lack of knowledge of the system;

and $B(t)$ and $C(t)$ matrices of time dependent coefficients.

In order to keep the problem relatively simple we will suppose the steering to be negligible. Then we can eliminate $G(t)$, $\underline{Y}(t)$, $B(t)$ and $\underline{V}(t)$ and we can write:

$$\dot{\underline{X}}(t) = F(t)\underline{X}(t) + C(t)\underline{W}(t) \qquad (3)$$

This equation could be solved straightforwardly if the matrices $F(t)$ and $C(t)$ were at least locally independent of time and if the noise vector $\underline{W}$ should be known. Although $\underline{W}$ is essentially not fully known, the differential equation may be solved. With sufficient knowledge of the initial conditions an estimate of the state values $\underline{X}$ can be found as function of the time t.

If the matrices F and C are independent of time equation (3) can be written as:

$$\dot{\underline{X}}(t) = F\underline{X}(t) + C\underline{W}(t) \qquad (4)$$

Example 1.

A ship sails approximately a straight line with constant velocity in a small region (flat earth). The movement is disturbed by small random forces, giving random variations in the two components of the accelaration. We can write:

$$\begin{aligned}
\dot{E}(t) &= 0.E(t) + \dot{E}(t) + 0.N(t) + 0.\dot{N}(t) \\
\ddot{E}(t) &= 0.E(t) + 0.\dot{E}(t) + 0.N(t) + 0.\dot{N}(t) + w_E(t) \\
\dot{N}(t) &= 0.E(t) + 0.\dot{E}(t) + 0.N(t) + \dot{N}(t) \\
\ddot{N}(t) &= 0.E(t) + 0.\dot{E}(t) + 0.N(t) + 0.\dot{N}(t) + w_N(t)
\end{aligned} \qquad (5)$$

where E(t) is easting, N(t) northing and $w_E(t)$ and $w_N(t)$ are stochastic accelerations in East and in North direction and where the state vector has the form:

$$\underline{X}(t) = \{E(t)\ \dot{E}(t)\ N(t)\ \dot{N}(t)\}^T.$$

Equation (5) has exactly the form of (4) if:

$$F = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad C = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \qquad \text{and } \underline{W}(t) = \begin{pmatrix} w_E(t) \\ w_N(t) \end{pmatrix}$$

If we know the state vector at a certain time t_o, the solution of (5) is very simple if the noise acceleration $\underline{W}$ can be assumed to be constant during a time interval (t_o,t):

$$\left.\begin{aligned} E(t) &= E(t_o) + (t-t_o)\dot{E}(t_o) + \tfrac{1}{2}(t-t_o)^2 w_E \\ \dot{E}(t) &= \dot{E}(t_o) + (t-t_o)w_E \\ N(t) &= N(t_o) + (t-t_o)\dot{N}(t_o) + \tfrac{1}{2}(t-t_o)^2 w_N \\ \dot{N}(t) &= \dot{N}(t_o) + (t-t_o)w_N \end{aligned}\right\} \qquad (6)$$

which can easily be checked by substitution of E(t), $\dot{E}(t)$, N(t) and $\dot{N}(t)$ into (5). (6) can also be written as

$$\underline{X}(t) = \begin{pmatrix} 1 & t-t_o & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & t-t_o \\ 0 & 0 & 0 & 1 \end{pmatrix} \underline{X}(t_o) + \begin{pmatrix} \tfrac{1}{2}(t-t_o)^2 & 0 \\ t-t_o & 0 \\ 0 & \tfrac{1}{2}(t-t_o)^2 \\ 0 & t-t_o \end{pmatrix} \underline{W} \qquad (7)$$

So for this system the position and the velocity at any moment t can be computed from the position and velocity at the moment t_o if the disturbances are known.
The coefficient matrix of $\underline{X}(t_o)$ is called the (state) transition matrix $\Phi(t,t_o)$.

For the more general linear system (3) the solution may be written as:

$$\underline{X}(t) = \Phi(t,t_o)\underline{X}(t_o) + \int_{t_o}^{t} \Phi(t,\tau)C(\tau)\underline{W}(\tau)d\tau \qquad (8)$$

where $\Phi(t,t_0)$ is the transition matrix which is the coefficient matrix to find the state at the time t from the state at time t_0 if no disturbances ($\underline{U}$ or $\underline{W}$) were present.

This solution can be checked by substitution of (8) into (3), using the properties of the transition matrix given in table 1.
So if the transition matrix $\Phi(t,t_0)$ is known the state vector at any moment t can be found from the state vector at the moment t_0 as long as (3) holds and if the disturbances $\underline{W}(t)$ are known for any moment from t_0 to t.
The transition matrix for a linear first order system like (3) obeys some interesting properties, shown in table 1.

1	$\Phi(t,t_0) = \frac{\partial \underline{X}(t)}{\partial \underline{X}(t_0)}$ i.e.: $\varphi_{i,j}(t,t_0) = \frac{\partial x_i(t)}{\partial x_j(t_0)}$ $\varphi_{i,j}(t,t_0) \approx \frac{\Delta x_i(t)}{\Delta x_j(t_0)}$
2	$\Phi^{-1}(t,t_0)$ exists
3	$\Phi(t_2,t_1) = \Phi^{-1}(t_1,t_2)$
4	$\Phi(t_3,t_1) = \Phi(t_3,t_2)\Phi(t_2,t_1)$
5	$\dot{\Phi}(t,t_0) \triangleq \frac{d\Phi(t,t_0)}{dt} = F(t)\Phi(t,t_0)$
Properties of the state transition matrix Φ of the homogeneous linear first order matrix differential equation: $\dot{\underline{X}}(t) = F(t)\ \underline{X}(t)$	

Table 1.

If the matrix F is independent of the time t, the transition matrix $\Phi(t,t_0)$ may be written as:

$$\Phi(t,t_o) = e^{(t-t_o)F} \qquad (9)$$

where for any square matrix A the exponential form is defined as

$$e^A \triangleq I + A + \quad AA + \quad AAA + \ldots \qquad (10)$$

if I is the unit matrix of the same dimension as A.

Therefore $\Phi(t,t_o)$ may be approximated by a limited number of terms from:

$$\Phi(t,t_o) = I + (t-t_o)F + \frac{1}{2!}(t-t_o)^2F^2 + \frac{1}{3!}(t-t_o)^3F^3 + \ldots \qquad (11)$$

Example 2.

Substituting the matrix F of example 1 into (11) and substituting the resulting Φ and the matrix C of the same example into (8) one finds indeed equation (7).

After a certain time the dynamics will no longer be known with sufficient accuracy: the noise terms will become too large, or even the model will not hold well enough. The state can then be updated by measurements which can be made continuously, but will often be made at discrete moments t_k.

These measurements have to provide at least some information about the state vector. So relations have to exist between the measurements and state components. If these relations are linear the observational model can be expresses as:

$$\underline{Z}_k = H_k \underline{X}_k + \underline{V}_k \qquad (12)$$

where $\underline{Z}_k \triangleq \underline{Z}(t_k)$ is the column vector of measurements,

$\underline{X}_k \triangleq \underline{X}(t_k)$

H_k is the matrix of the coefficients expressing the linear relations between the state components and the measurements

$\underline{V}_k \triangleq \underline{V}(t_k)$ is the vector of measurement errors.

Example 3.

If, with the same dynamic model of example 1, where $\underline{X} = (E\ \dot{E}\ N\ \dot{N})^T$, one measures the easting and the northing, (respectively Z_E and Z_N) one has a very simple though not quite realistic case. The measurement model may then be written as:

$$\underline{Z}_k = \begin{pmatrix} Z_{Ek} \\ Z_{Nk} \end{pmatrix} = \begin{pmatrix} E_k \\ N_k \end{pmatrix} + \begin{pmatrix} (V_E)_k \\ (V_N)_k \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \underline{X}_k + \underline{V}_k = H\underline{X}_k + \underline{V}_k$$

■

Example 4.

Often non linear relations exist between the state and the measurements. This is for instance the case if the distances are measured to two transmitters T' and T" of which T' is drawn in figure 1.

The measurements can then be written as:

$$Z'_k = \sqrt{(E_k-E')^2 + (N_k-N')^2} + v'_k$$

and

$$Z''_k = \sqrt{(E_k-E'')^2 + (N_k-N'')^2} + v''_k$$

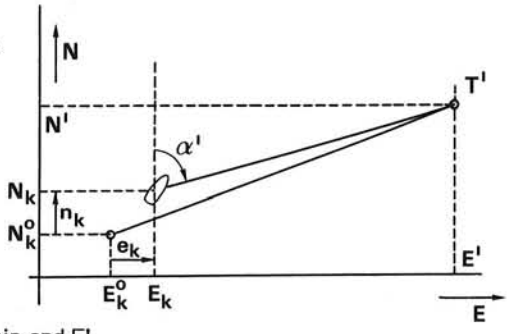


Figure 1: Linearization of a measurement model.

if E_k and N_k are the coordinates of the ship and E', N', E" and N" the coordinates of both transmitters; v'_k and v''_k are the measuring errors.

In vector form:

$$\underline{Z}_k = \begin{pmatrix} Z'_k \\ Z''_k \end{pmatrix} = \begin{pmatrix} \sqrt{(E_k-E')^2 + (N_k-N')^2} \\ \sqrt{(E_k-E'')^2 + (E_k-E'')^2} \end{pmatrix} + \begin{pmatrix} v'_k \\ v''_k \end{pmatrix} \qquad (13)$$

To find a linear form a point E_k^o, N_k^o in the neighbourhood of the ship is taken, and one will try to make the calculations with the small difference values e_k, n_k, z_k and z_k defined by (see also figure 1):

$$e_k \triangleq E_k - E_k^o \qquad z'_k \triangleq Z'_k - Z'^o_k \triangleq Z'_k - \sqrt{(E_k^o-E')^2 + (N_k^o-N'_k)^2} + v'_k$$

$$n_k \triangleq N_k - N_k^o \qquad z'_k \triangleq Z''_k - Z''^o_k \triangleq Z''_k - \sqrt{(E_k^o-E'')^2 + (N_k^o-N'')} + v''_k$$

Substitution in (13) and development in a Taylor series with omission of all terms of second and higher order in e_k and n_k yields:

$$\underline{Z}_k \approx \begin{pmatrix} Z_k^{o\,'} + \dfrac{E_k^o - E'}{Z_k^{o\,'}} e_k + \dfrac{N_k^o - N'}{Z_k^{o\,'}} n_k \\ Z_k^{o\,''} + \dfrac{E_k^o - E''}{Z^{o\,''}} e_k + \dfrac{N_k^o - N''}{Z^{o\,''}} n_k \end{pmatrix} + \underline{V}_k$$

or by introducing the heading α_k and α_k' (see figure 1):

$$\underline{z}_k = \begin{pmatrix} e_k \sin \alpha_k' - n_k \cos \alpha_k' \\ e_k \sin \alpha_k'' - n_k \cos \alpha_k'' \end{pmatrix} + \underline{V}_k$$

If the point E_k^o, N_k^o, or better $E^o(t)$, $N^o(t)$ is considered as a moving point the whole state vector can be taken as differences: $\underline{x}_k = (e_k\ \dot{e}_k\ n_k\ \dot{n}_k)^T$. In this case one finds:

$$\underline{z}_k = \begin{pmatrix} \sin \alpha_k' & 0 & -\cos \alpha_k' & 0 \\ \sin \alpha_k'' & 0 & -\cos \alpha_k'' & 0 \end{pmatrix} \underline{x}_k + \underline{V}_k \qquad (14)$$

$$\underline{z}_k = H_k \underline{x}_k + \underline{V}_k \qquad (15)$$

N.B. 1. If the movement of the point E_k^o, N_k^o is known a priori (the planned track), all the matrices H_k may be calculated beforehand.

N.B. 2. The measurements $\underline{Z}$ are expressed in metres or another measure of length and not in lanes.

3.5.3. The Kalman filter.

Now we come to the Kalman filter, an algorithm to find some optimal estimation $\hat{\underline{X}}_k$ of the state at the time t_k and of its covariance matrix P_k, from the values X_{k-1} and P_{k-1} at the preceeding time t_{k-1} and from the last measurements $\underline{Z}_k$.

So if the algorithm can be started somehow at t_o the state may be calculated at each new moment $t_k (k > 0)$ without retaining all measurements.

For optimal estimation in the least squares sense it is possible to rove the undermentioned formulae for appropriate processes.

Let us assume the simple linear process with constant coefficients:

$$\dot{\underline{X}}(t) = F\,\underline{X}(t) + C\,\underline{W}(t) \tag{4}$$

As one can learn from the theory of stochastic processes it is quite reasonable to work with the - physically not very realistic - assumption, that the disturbances $\underline{W}(t)$ are constant during the time interval between the measurements $t_{k-1} < t < t_k$. In this case for the process (4) the solution (8) for this interval may written as:

$$\underline{X}(t) = \Phi(t, t_{k-1})\,\underline{X}(t_{k-1}) + \Gamma(t,t_{k-1})\,\underline{W}_{k-1} \tag{16}$$

with $$\Gamma(t,t_{k-1}) \triangleq \int_{t_{k-1}}^{t} \Phi(t_k, \tau) C_{k-1}\, d\tau \tag{17}$$

So $\Gamma(t,t_{k-1})$ is a coefficient matrix giving the influences of the disturbances $\underline{W}_{k-1}$ in the interval $t_{k-1} < t < t_k$, on the state $\underline{X}(t_k)$.
If the mathematical expectation of the disturbances $\underline{W}(t)$ is zero, i.e. if $E\{\underline{W}(t)\} = 0$, one finds for the estimate $\underline{X}_{k,k-1}$ just before the measurements at t_k:

$$\hat{\underline{X}}_{k,k-1} = \Phi_{k-1}\,\hat{\underline{X}}_{k-1} \tag{18}$$

where $\hat{\underline{X}}_{k-1}$ is the "best" estimate of the state at the moment t_{k-1}, including the measurements at t_{k-1}
and $\Phi_{k-1} \triangleq \Phi(t_k, t_{k-1})$ is the transition matrix.

If no correlation exists between disturbances at different moments, i.e. if $E\{\underline{W}_k\,\underline{W}_l^T\} = 0$ for $k \neq l$, (so-called white noise) the covariance matrix P_k of the state $\hat{\underline{X}}_k$ can be estimated from the covariance matrix P_{k-1} of the preceding state $\hat{\underline{X}}_{k-1}$. One finds for the process (16):

$$P_{k,k-1} = \Phi_{k-1}\,P_{k-1}\,\Phi_{k-1}^T + \Gamma_{k-1}\,Q_{k-1}\,\Gamma_{k-1}^T \tag{19}$$

where Q_{k-1} is the covariance matrix of the disturbances in the interval $t_{k-1} < t < t_k$:

$$Q_{k-1} = E\{W_{k-1}\,W_{k-1}^T\}$$

and $$\Gamma_{k-1} \triangleq \Gamma(t_k,t_{k-1}) \quad .$$

Example 5.

Let in example 1 the East- and the North disturbing accelerations be independent, with standard deviation σ_a (for example in m s^{-2}). Then

$$Q_k = Q = \sigma_a^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

As already found in example 2 the transition matrix may be written as

$$\Phi(t,\tau) = \begin{pmatrix} 1 & (t-\tau) & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & (t-\tau) \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{for } t_{k-1} \le \tau \le t \le t_k$$

Substitution of $\Phi(t_k, \tau)$ and the C of example 1 into (17) yields:

$$\Gamma_{k-1} = \int_{\tau=t_{k-1}}^{t_k} \Phi(t_k,\tau) C \, d\tau =$$

$$= \int_{t_{k-1}}^{t_k} \begin{pmatrix} 1 & t_k-\tau & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & t_k-\tau \\ 0 & 0 & 0 & 1 \end{pmatrix} d\tau \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} T & +\frac{1}{2}T^2 & 0 & 0 \\ 0 & T & 0 & 0 \\ 0 & 0 & T & +\frac{1}{2}T^2 \\ 0 & 0 & 0 & T \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} +\frac{1}{2}T^2 & 0 \\ T & 0 \\ 0 & +\frac{1}{2}T^2 \\ 0 & T \end{pmatrix} \text{ with } T = t_k - t_{k-1}$$

The covariance matrix $P_{k,k-1}$ of $\hat{\underline{X}}_{k,k-1}$ is found by substituting $\Phi_{k-1} = \Phi(t_k,t_{k-1})$, Γ_{k-1} and $Q_{k-1} = Q$ into (19):

$$\begin{pmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{pmatrix} P_{k-1} \begin{pmatrix} 1 & 0 & 0 & 0 \\ T & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & T & 1 \end{pmatrix} + \begin{pmatrix} \frac{1}{4}T^4 & \frac{1}{2}T^3 & 0 & 0 \\ \frac{1}{2}T^3 & T & 0 & 0 \\ 0 & 0 & \frac{1}{4}T^4 & \frac{1}{2}T^3 \\ 0 & 0 & \frac{1}{2}T^3 & T^2 \end{pmatrix} \sigma_a^2$$

In literature one sometimes finds instead of $\frac{1}{4}T^4$ in the last matrix $\frac{1}{3}T^4$. The difference is caused by a slightly different assumption about the stochastics of the noise $\underline{W}$.

Let the noise of the initial state vector $\underline{X}_0$ be given by:

$$P_0 = \begin{pmatrix} \sigma_P^2 & 0 & 0 & 0 \\ 0 & \sigma_S^2 & 0 & 0 \\ 0 & 0 & \sigma_P^2 & 0 \\ 0 & 0 & 0 & \sigma_S^2 \end{pmatrix}$$

, where σ_P^2 is the variance of the components of the position, and σ_S^2 is the variance of the components of the velocity.

then substitution into the last equation gives:

$$P_{1,0} = \begin{pmatrix} p_{11} & p_{12} & 0 & 0 \\ p_{12} & p_{22} & 0 & 0 \\ 0 & 0 & p_{11} & p_{12} \\ 0 & 0 & p_{12} & p_{22} \end{pmatrix}$$

with $p_{11} = \sigma_P^2 + \sigma_S^2 T^2 + \frac{1}{4}\sigma_a^2 T^4$

$p_{12} = \sigma_S^2 T + \frac{1}{2}\sigma_a^2 T^3$

$p_{22} = \sigma_S^2 + \sigma_a^2 T^2$

Some quantitative estimates:

1st Doppler satellite fixes each 5000^s (= T).
A low estimate for the effective disturbing acceleration over such an interval:
$\sigma_a = 0{,}01$ m s^{-2} (0,001 x gravity).
Square roots of the contributions to the diagonal elements of $P_{k,k-1}$:
$\frac{1}{2}\ \sigma_a T^2 = 12{,}5 \cdot 10^4$ m = 125 km in position;
$\sigma_a T = 50$ m s^{-1} = 100 knot in velocity.
So clearly for this case our model is not adequate. One can hope for a better result if the dynamic model is aided by course and velocity information.

2nd Let the time interval T be 1^s, and let the standard deviation σ_a be 1 m s^{-2} (= 10% of gravity).
Here $\frac{1}{2}\ \sigma_a T^2 = 0{,}5$ m and $\sigma_a T = 1$ m s^{-1} = 2 knot.

3rd For an interval, $T = 100^s$ σ_a will be much smaller than for the shorter interval because of the averaging effect. Say $\sigma_a = 0{,}1$ m s^{-2}. If this is real one finds $\frac{1}{2}\ \sigma_a T^2 = 500$ m and $\sigma_a T = 10$ m s^{-1} = 20 knot. A very weak dynamic model.

At the moments t_k new measurements $\underline{Z}_k$ are performed for which (12) holds:

$$\underline{Z}_k = H_k \underline{X}_k + \underline{V}_k \qquad (12)$$

The disturbances $\underline{V}_k$ are now supposed to form a zero mean, white sequence with a covariance matrix R_k, i.e. a sequence with mathematical expectation zero and without correlation between different moments:

$$E\{\underline{V}_k\} = 0$$

$$E\{\underline{V}_k \underline{V}_l\} = \begin{cases} (0) \text{ for } l \neq k \\ R_k \text{ for } l = k \end{cases} \qquad (20)$$

With these assumptions, using the theory of least square adjustment, one can find for our process the relations to calculate the "optimal" estimate $\hat{\underline{X}}_k$ of the state just after the last measurements $\underline{Z}_k$ and their covariance matrix P_k:

The derivation is given in Part I-11,12.

$$\hat{\underline{X}}_k = \hat{\underline{X}}_{k,k-1} + K(\underline{Z}_k - H_k \hat{\underline{X}}_{k,k-1}) \qquad (21)$$

$$P_k = (I - K_k H_k) P_{k,k-1} \qquad (22)$$

where K_k is the so-called gain matrix:

$$K_k \triangleq P_{k,k-1} H_k^T (H_k P_{k,k-1} H_k^T + R_k)^{-1} \qquad (23)$$

and I = unit matrix of the same dimensions as P_k.

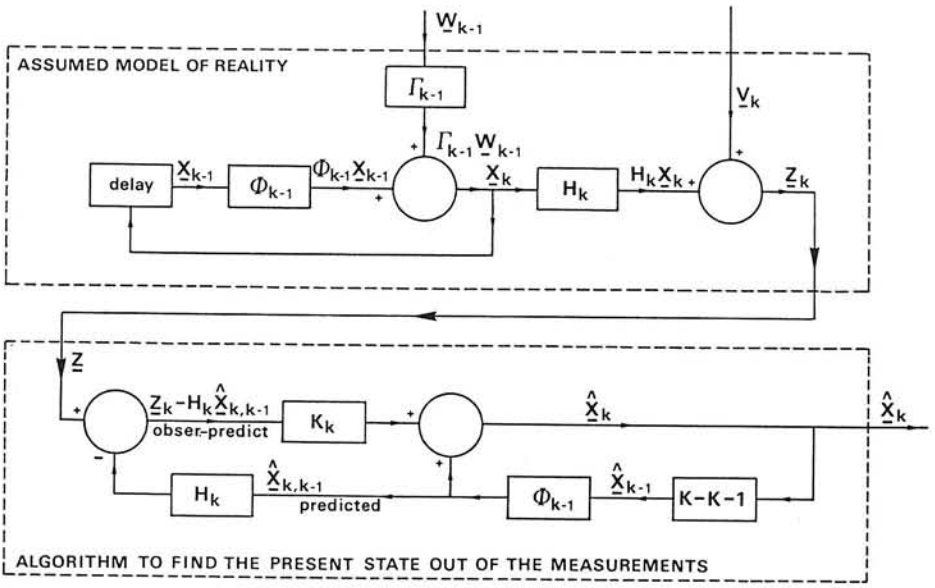


Fig. 2. The assumed model of reality (ship) and the Kalman filter as block-diagrams.

The matrix $(\underline{Z}_k - H_k \underline{\hat{X}}_{k,k-1})$ in (21) can be seen as the difference between the observations $\underline{Z}_k$ and the predicted values of this observations, $H_k \underline{\hat{X}}_{k,k-1}$. This difference can be used to check the appropriateness of the filter.

In figure 2 the ideas about the Kalman filter are presented in the form of a block diagram. The upper part reflects the assumed models of reality. The dynamics of the process are in the transitionmatrix Φ, and in the added noise $\underline{W}$ transformed by into fluctuations of the state. The observations $\underline{Z}$ are assumed to be derived by transformation of the state by the H matrix with added noise $\underline{V}$.
The lower part is a diagram of the Kalman filter algorithm. The new state $\underline{X}_{k,k-1}$ is predicted from the best estimation of the earlier state $\underline{X}_{k-1}$. From this predicted state an observation vector is calculated and compared with the real observations $\underline{Z}_k$. The best estimate $\underline{\hat{X}}_k$ of the new state is found as the sum of the predicted state and the by K "weighted" difference between real and predicted observations.

Table 2 presents a recapitulation of the Kalman filter equations and figure 3 gives a simplified flow diagram of the calculations.

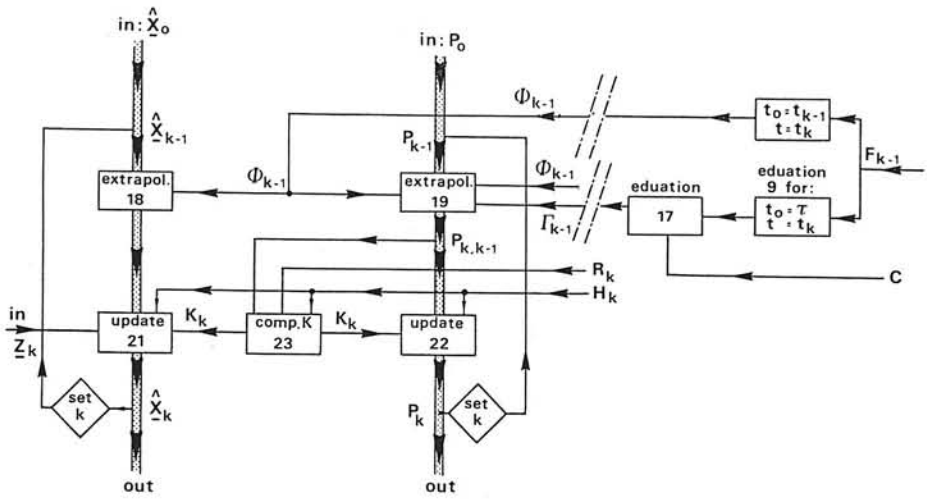


Fig. 3. Information flow diagram for Kalman filter.
(Note that no information flows from the state to the covariance calculations, i.e. if Q, R, H and Γ and Φ or F and C are known a priori, all matrices P can also be computed a priori).

Dynamic model	$\dot{\underline{X}}(t) = F\,\underline{X}(t) + C\,\underline{W}$	(4)
Transition matrix	$\Phi_{k-1} = \Phi(t_k, t_{k-1}) = e^{(t_k - t_{k-1})F}$	(9)
Measurement model	$\underline{Z}_k = H_k \underline{X}_k + \underline{V}_k$	(12)

Dynamic noise and measurement noise independent, zero mean, white with covariance matrices Q and R

Initial conditions (k-1 = 0)
State $\hat{\underline{X}}_o$ covariance matrix P_o

Extrapolation

state estimate	$\hat{\underline{X}}_{k,k-1} = \Phi_{k-1}\,\hat{\underline{X}}_{k-1}$	(18)
covariances	$P_{k,k-1} = \Phi_{k-1} P_{k-1} \Phi_{k-1}^T + \Gamma_{k-1} Q_{k-1} \Gamma_{k-1}^T$	(19)
with	$\Gamma_{k-1} = \int_{t_{k-1}}^{t_k} \Phi(t_k, \tau) C \, d\tau$	(17)
Filter gain	$K_k = P_{k,k-1} H_k^T \{H_k P_{k,k-1} H_k^T + R_k\}^{-1}$	(23)

Update

state estimate	$\hat{\underline{X}}_k = \hat{\underline{X}}_{k,k-1} + K_k \{\underline{Z}_k - H_k \hat{\underline{X}}_{k,k-1}\}$	(21)
covariances	$P_k = \{I - K_k H_k\} P_{k,k-1}$	(22)

Table 2. Recapitulation of Kalman filter equations.

Example 6.

Now, assume in the process of example 5 the measurements of example 3 to be done as independent ones with standard deviations σ_m. Then one finds from example 1:

$$\underline{X}_k = \begin{pmatrix} E_k \\ \dot{E}_k \\ N_k \\ \dot{N}_k \end{pmatrix} \quad F = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad C = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\underline{W}_k = \begin{pmatrix} (w_E)_k \\ (w_N)_k \end{pmatrix} \text{ and } \Phi_{k-1} = \Phi(t_k, t_{k-1}) = \begin{pmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

for $T = t_k - t_{k-1}$, and from example 3:

$$\underline{V}_k = \begin{pmatrix} (v_E)_k \\ (v_N)_k \end{pmatrix} \quad \underline{Z}_k = \begin{pmatrix} Z_{Ek} \\ Z_{Nk} \end{pmatrix} \text{ and } H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

from example 5:

$$P_o = \begin{pmatrix} \sigma_P^2 & 0 & 0 & 0 \\ 0 & \sigma_s^2 & 0 & 0 \\ 0 & 0 & \sigma_P^2 & 0 \\ 0 & 0 & 0 & \sigma_s^2 \end{pmatrix} \qquad \text{and } Q = \sigma_a^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and from equation (20):

$$R_k = E(\underline{V}_k \underline{V}_k) = \sigma_m^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Let the initial state at the moment t_o be estimated as:

$$\hat{\underline{X}}_o = \begin{pmatrix} 0 \\ S_o \\ 0 \\ S_o \end{pmatrix}$$

i.e. we have chosen the origin in the estimated initial point and the estimate of the initial velocity in North-East direction.

Now with table 2 we can obtain the covariances at the moment t_1.
With formula (19) we have found already in example (5):

$$P_{1,0} = \begin{pmatrix} P_{11} & P_{12} & 0 & 0 \\ P_{12} & P_{22} & 0 & 0 \\ 0 & 0 & P_{11} & P_{12} \\ 0 & 0 & P_{12} & P_{22} \end{pmatrix} \quad \text{with} \quad \begin{aligned} P_{11} &= \sigma_p^2 + \sigma_s^2 T^2 + \tfrac{1}{4}\sigma_a^2 T^4 \\ P_{12} &= \sigma_s^2 T + \tfrac{1}{2}\sigma_a^2 T^3 \\ P_{22} &= \sigma_s^2 + \sigma_a^2 T^2 \end{aligned}$$

Using (23) the gain matrix is then derived as:

$$K_1 = P_{1,0} \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \left\{ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} P_{1,0} \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} + \sigma_m^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{-1} \right\} =$$

$$K_1 = \frac{1}{p_{11} + \sigma_m^2} \begin{pmatrix} p_{11} & 0 \\ p_{12} & 0 \\ 0 & p_{11} \\ 0 & p_{12} \end{pmatrix} \qquad (24)$$

With formula (22) one can now find the covariances of the new state:

$$P_1 = \frac{1}{p_{11} + \sigma_m^2} \begin{pmatrix} M & 0 \\ 0 & M \end{pmatrix} \quad \text{with}$$

$$M = \begin{pmatrix} \sigma_m^2 p_{11} & \vdots & \sigma_m^2 p_{12} \\ \cdots & \vdots & \cdots \\ \sigma_m^2 p_{12} & \vdots & \sigma_p^2\sigma_s^2 + \sigma_p^2\sigma_a^2 T^2 + \tfrac{1}{4}\sigma_a^2\sigma_s^2 T^4 + p_{22}\sigma_m^2 \end{pmatrix}$$

Although this formula for P_1 looks quite complicated, it can easily illustrate some characteristics of our filter:

1. If the measurements are assumed to be very inaccurate, i.e. if σ_m is very large, P_1 tends to become idential to its estimate without new measurements: $P_1 \to P_{1,0}$.

2. If the dynamics are assumed to be very weak, i.e. if σ_a is very large, the variance of the new position becomes σ_m^2 (= the variance of the measurement) and the variance of the new velocity components becomes $\sigma_s^2 + 4(\sigma_p^2 + \sigma_m^2)T^{-2}$.
3. If the measurements are very good, i.e. if σ_m tends to zero, the variance of the new position components becomes σ_m^2.

Note that for these calculations no measurement results have been used. In this case, where the matrices Φ, H, Q, R and Γ are independent of the observations, the covariance matrices P can be calculated beforehand.
The new state vector can be estimated with equation (18):

$$\underline{\hat{X}}_{1,0} = \Phi_0 \underline{\hat{X}}_0 = \begin{pmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} 0 \\ s_o \\ 0 \\ s_o \end{pmatrix} = \begin{pmatrix} s_o T \\ s_o \\ s_o T \\ s_o \end{pmatrix}$$

By introducing the observations, the state vector can be updated with equation (21):

$$\underline{\hat{X}}_1 = \underline{\hat{X}}_{1,0} + \frac{1}{p_{11} + \sigma_m^2} \begin{pmatrix} p_{11} & 0 \\ p_{12} & 0 \\ 0 & p_{11} \\ 0 & p_{12} \end{pmatrix} \left\{ \underline{Z}_1 - \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \underline{\hat{X}}_{1,0} \right\}$$

$$\underline{\hat{X}}_1 = \begin{pmatrix} \frac{p_{11}}{p_{11} + \sigma_m^2} Z_{E1} + \frac{\sigma_m^2}{p_{11} + \sigma_m^2} s_o T \\ \frac{p_{12}^T}{p_{11} + \sigma_m^2} \frac{Z_{E1}}{T} + \frac{p_{11} + \sigma_m^2 - p_{12}^T}{p_{11} + \sigma_m^2} s_o \\ \frac{p_{11}}{p_{11} + \sigma_m^2} Z_{N1} + \frac{\sigma_m^2}{p_{11} + \sigma_m^2} s_o T \\ \frac{p_{12}^T}{p_{11} + \sigma_m^2} \frac{Z_{N1}}{T} + \frac{p_{11} + \sigma_m^2 - p_{12}^T}{p_{11} + \sigma_m^2} s_o \end{pmatrix}$$

Note that the components of this estimated state vector consist of a weighted mean of the directly or indrectly observed value and the extrapolated value.
It might be interesting for the reader to look for the consequences of the magnitudes of the different variances on the updated state vector like it has been done for the updated covariance matrix P_1.

Example 7.

Here we consider a process analogue to example 6, but in one dimension:

$$\underline{X}_k = \begin{bmatrix} E_k \\ \dot{E}_k \end{bmatrix} \quad F = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad C = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad t_k - t_{k-1} = T = 40 \text{ s}, \quad t_0 = 0$$

$$\underline{W}_k = (w_E)_k \quad \Phi(t,\tau) = \Phi(t-\tau) = \begin{bmatrix} 1 & t-\tau \\ 0 & 1 \end{bmatrix}$$

$$Z_k = (Z_{Ek}) \quad H = (1\ 0) \quad \underline{V}_{Ek} = (V_{Ek})$$

$$\hat{\underline{X}}_0 = \begin{bmatrix} \hat{E}_0 \\ \hat{\dot{E}} \end{bmatrix} = \begin{bmatrix} 0 \\ 2{,}5 \text{ m/s} \end{bmatrix} \quad P_0 = \begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \sigma_s^2 \end{bmatrix} = \begin{bmatrix} 5^2 \text{ m}^2 & 0 \\ 0 & (0{,}2)^2 \text{ m}^2/\text{s}^2 \end{bmatrix}$$

$$Q = \sigma_a^2 = (0{,}01)^2 \text{ m}^2/\text{s}^2 \qquad R = \sigma_m^2 = 5^2 \text{ m}^2 \text{ or } 200 \text{ m}^2 .$$

The variances σ_p^2, σ_s^2, σ_a^2 and σ_m^2 are chosen so that for an extrapolation of 10 s their influences are of the same order of magnitude.

Now we shall calculate the state vector $\hat{X}(t)$ and its covariance matrix $P(t)$ not only around the moments t_k (= $k \cdot T$), but also at intermediate moments, say t = 10 s, 20 s, 30 s, 50 s, ...

Before the first measurement (at t_1 = 40 s) we find with the epolation formula:

$$\hat{\underline{X}}(t) = \Phi(t-t_0)\hat{\underline{X}}_0 \qquad t_0 = 0 < t < t_1$$

$$\left.\begin{aligned} \hat{E}(t) &= \hat{E}_0 + t\,\hat{\dot{E}}_0 \\ \hat{\dot{E}}(t) &= \hat{\dot{E}}_0 \end{aligned}\right\} \qquad (1)$$

and

$$P(t) = \Phi(t-t_0)P_0\Phi^T(t-t_0) + \Gamma(t)Q\Gamma^T(t)$$

with

$$\Gamma(t) = \int_{t_0}^{t} \Phi(t-\tau) C \, d\tau$$

$$= \int_{0}^{t} \begin{bmatrix} 1 & t-\tau \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} d\tau = \begin{bmatrix} \frac{1}{2} t^2 \\ t \end{bmatrix} .$$

So that:

$$P(t) = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \sigma_s^2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ t & 1 \end{bmatrix} + \begin{bmatrix} \frac{1}{2} t^2 \\ t \end{bmatrix} \sigma_0^2 \begin{bmatrix} \frac{1}{2} t^2 & t \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_p^2 + t^2 \sigma_s^2 & t\sigma_s^2 \\ t\sigma_s^2 & \sigma_s^2 \end{bmatrix} + \begin{bmatrix} \frac{1}{4} t^4 & \frac{1}{2} t^3 \\ \frac{1}{2} t^3 & t^2 \end{bmatrix} \sigma_a^2$$

$$P(t) = \begin{bmatrix} \sigma_p^2 + \sigma_s^2 t^2 + \frac{1}{4} \sigma_a^2 t^4 & \sigma_s^2 t + \frac{1}{2} \sigma_a^2 t^3 \\ \sigma_s^2 t + \frac{1}{2} \sigma_a^2 t^3 & \sigma_s^2 + \sigma_a^2 t^2 \end{bmatrix} \quad \text{for } t < t_1 \qquad (2)$$

Analogue to example 6 we find for $t = t_1 = T$:

$$K_1 = \frac{1}{\sigma_p^2 + \sigma_s^2 T^2 + \frac{1}{4} \sigma_a^2 T^4 + \sigma_m^2} \begin{bmatrix} \sigma_p^2 + \sigma_s^2 T^2 + \frac{1}{4} \sigma_a^2 T^4 \\ \sigma_s^2 T + \frac{1}{2} \sigma_a^2 T^3 \end{bmatrix} = \begin{bmatrix} k_{1,1} \\ k_{1,2} \end{bmatrix}$$

and for the covariance immediately after the first measurement:

$$P_1 = \frac{M_1}{\sigma_p^2 + \sigma_s^2 T^2 + \frac{1}{4} \sigma_a^2 T^4 + \sigma_m^2} \qquad (3)$$

with

$$M_1 = \begin{bmatrix} \sigma_m^2\sigma_p^2 + \sigma_m^2\sigma_s^2T^2 + \frac{1}{4}\sigma_m^2\sigma_a^2T^4 & \sigma_m^2\sigma_s^2T + \frac{1}{1}\sigma_m^2\sigma_a^2T^3 \\ \sigma_m^2\sigma_s^2T + \frac{1}{2}\sigma_m^2\sigma_a^2T^3 & \sigma_s^2\sigma_m^2 + \sigma_s^2\sigma_p^2 + \sigma_a^2\sigma_m^2T^2 + \sigma_a^2\sigma_p^2T^2 + \frac{1}{4}\sigma_a^2\sigma_s^2T^4 \end{bmatrix}$$

$$= \begin{bmatrix} m_{1,11} & m_{1,12} \\ m_{1,12} & m_{1,22} \end{bmatrix}$$

or

$$P_1 = \begin{bmatrix} P_{1,11} & P_{1,12} \\ P_{1,12} & P_{1,22} \end{bmatrix} .$$

For the estimation of the state vector just after the first measurement we find:

$$\hat{\underline{X}}_1 = \hat{\underline{X}}_{1,0} + K_1\{z_1 - (1\ 0)\hat{\underline{X}}_{1,0}\}$$

with

$$\hat{\underline{X}}_{1,0} = \begin{bmatrix} \hat{E}_0 + T\hat{\dot{E}}_0 \\ \hat{\dot{E}}_0 \end{bmatrix}$$

so that

$$\hat{\underline{X}}_1 = \begin{bmatrix} \hat{E}_0 + T\hat{\dot{E}}_0 + k_{1,1}\left[z_1 - \hat{E}_0 - T\hat{\dot{E}}_0\right] \\ \hat{\dot{E}}_0 + k_{1,2}\left[z_1 - \hat{E}_0 - T\hat{\dot{E}}\right] \end{bmatrix} = \begin{bmatrix} \hat{E}_1 \\ \hat{\dot{E}}_1 \end{bmatrix} . \qquad (4)$$

Between the first and the second measurement we find ($t_1 < t < t_2$):

$$\hat{\underline{X}}(t) = \Phi(t - t_1)\hat{\underline{X}}_1$$

$$\hat{\underline{X}}(t) = \begin{bmatrix} 1 & t-t_1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{E}_1 \\ \hat{\dot{E}}_1 \end{bmatrix} = \begin{bmatrix} 1 & t-T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{E}_1 \\ \hat{\dot{E}}_1 \end{bmatrix}$$

$$\hat{\underline{X}}(t) = \begin{bmatrix} \hat{E}_1 - T\hat{\dot{E}}_1 + t\hat{\dot{E}}_1 \\ \hat{\dot{E}}_1 \end{bmatrix} \qquad (5)$$

So again in this interval the position changes linearly with time and the velocity is constant in the interval (as to be expected with our model). For the covariances we find in this interval ($t_1 < t < t_2$):

$$P(t) = \Phi(t-t_1)P_1\Phi^T(t-t_1) + \Gamma(t)\sigma_a^2\Gamma^T(t)$$

with $\Gamma(t) = \int_{t_1}^{t} \Phi(t-\tau)C\,d\tau$

$$\Gamma(t) = \int_{T}^{t} \begin{bmatrix} 1 & t-\tau \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} d\tau = \begin{bmatrix} \frac{1}{2}(t-T)^2 \\ t-T \end{bmatrix}$$

So that in this interval the extrapolation gives:

$$P(t) = \begin{bmatrix} 1 & t-T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} m_{1,11} & m_{1,12} \\ m_{1,12} & m_{1,22} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ t-T & 1 \end{bmatrix} \cdot \frac{1}{\sigma_p^2 + \sigma_s^2 T^2 + \frac{1}{4}\sigma_a^2 T^2 + \sigma_m^2}$$

$$+ \begin{bmatrix} \frac{1}{4}(t-T)^4 & \frac{1}{2}(t-T)^3 \\ \frac{1}{2}(t-T)^3 & (t-T)^2 \end{bmatrix} \sigma_a^2$$

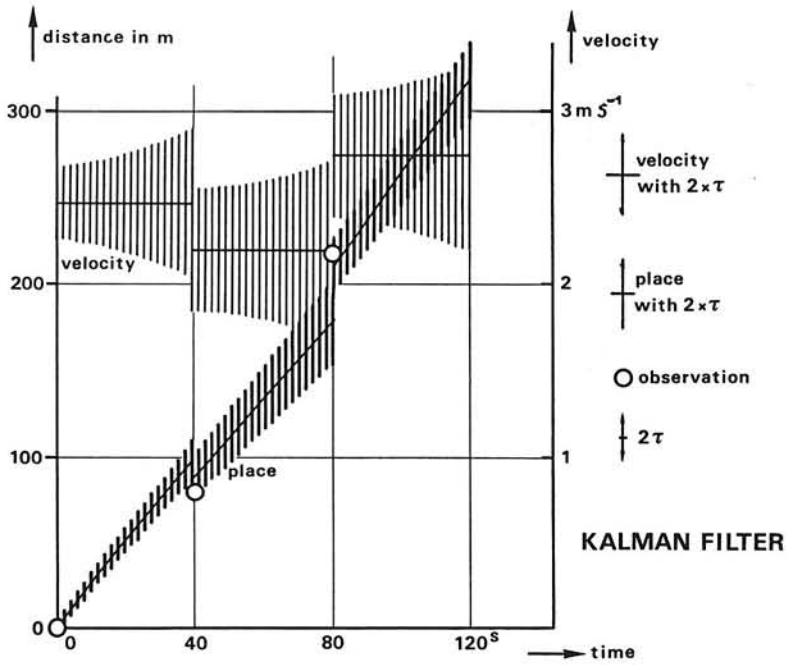
$$= \begin{bmatrix} m_{1,11} + 2m_{1,12}(t-T) + m_{1,22}(t-T)^2 & m_{1,12} + m_{1,22}(t-T) \\ m_{1,12} + m_{1,22}(t-T) & m_{1,22} \end{bmatrix} x$$

$$\times \frac{1}{\sigma_p^2 + \sigma_s^2 T^2 + \frac{1}{4}\sigma_a^2 T^4 + \sigma_m^2} + \begin{bmatrix} \frac{1}{4}(t-T)^4 & \frac{1}{2}(t-T)^3 \\ \frac{1}{2}(t-T)^3 & (t-T)^2 \end{bmatrix} \sigma_a^2 \qquad (6a)$$

$$= \begin{bmatrix} p_{1,11} + 2p_{1,12}(t-T) + p_{1,22}(t-T)^2 & p_{1,12} + p_{1,22}(t-T) \\ p_{1,12} + p_{1,22}(t-T) & p_{1,22} \end{bmatrix} +$$

$$+ \begin{bmatrix} \frac{1}{4}(t-T)^4 & \frac{1}{2}(t-T)^3 \\ \frac{1}{2}(t-T)^3 & (t-T)^2 \end{bmatrix} \sigma_a^2 \qquad (6b)$$

In the figure and in the table a numerical example is given.



$\sigma_p^2 = 25\ m^2$ $\sigma_s^2 = 0{,}04\ m^2s^{-2}$ $\sigma_a^2 = 0{,}0001\ m^2s^{-4}$ $\sigma_m^2 = 200\ m^2$

$\hat{E}_0 = 0$ $\dot{\hat{E}}_0 = 2{,}5\ ms^{-1}$

$Z_1 = 80\ m$ $Z_2 = 220\ ms^{-1}$

s t	m^2 p_{11}	m $\sqrt{p_{11}}=\sigma_p$	m^2s^{-4} p_{12}	m^2s^{-2} p_{22}	ms^{-1} $\sqrt{p_{22}}=\sigma_{\dot{E}}$	m $\hat{E}$	ms^{-1} $\dot{\hat{E}}$
0	25	5	0	0	0	0	2
10	29	5,4	0,45	0,05	0,22	25	2,5
20	45	6,7	1,2	0,08	0,28	50	2,5
30	81	9,0	2,6	0,13	0,36	75	2,5
40^-	153	12,4	4,8	0,20	0,45	100	2,5
40^+	87	9,3	2,7	0,13	0,37	91,3	2,23
50	154	12,4	4,1	0,14	0,37	113,6	2,23
60	251	15.8	5,7	0,17	0,41	135,9	2,23
70	386	19,6	9,8	0,22	0,47	158,2	2,23
80^-	575	24,0	11,1	0,29	0,54	180,5	2,23
80^+	148	12,2	2,9	0,13	0,36	209,8	2,79
120^-	652	25,5	11,3	0,29	0,54	321,6	2,79

3.5.4. Literature.

Recommended for further study for Kalman filtering.

(1) C. de Witt: "Sea Navigation and Stochastics", Manual for lecture course T.H. Delft, Dept. Mathematics and Informatics, 1982.

(2) A. Gelb: "Applied Optimal Estimation", M.I.T. press 1974, (Kalman filtering, smoothing and predicting with applications and with many examples).

(3) A.H. Jazwinsky: "Stochastic Processing and Filtering Theory", Acad. Press 1970, (with thorough mathematical background).

(4) A. Papoulis: "Probability, Random Variables and Stochastic Processes", McGraw-Hill, 1965, (gives very clear the difficult mathematical background of the stochastical processes).

(5) A.P.E.M. Houtenbos: "Prediction filtering and smoothing of offshore navigation", Hydrogr. J. no. 25 (July 1982), p. 5-16, Errate: Hydr. J. 28 (april 1983, p. 33) (gives a simple Kalman filter described in geodetic terms).

3.5.5. Exercises Kalman.

1. Transition matrix.

Show, using the properties of table 1 that

$$\underline{X}(t) = \Phi(t,t_o)\ \underline{X}(t_o) \qquad (a)$$

satisfies the homogeneous linear differential equation

$$\dot{\underline{X}}(t) = F(t)\ \underline{X}(t) \qquad (b)$$

Solution.

Substitute (a) in (b):

$$\dot{\underline{X}}(t) = \dot{\Phi}(t,t_o)\ \underline{X}(t_o) + \Phi(t,t_o)\ \dot{\underline{X}}(t_o) \qquad \left[\underline{X}(t_o) \text{ is independent of } t, \dot{\underline{X}}(t_o) \quad \frac{d\underline{X}(t_o)}{dt} = 0.\right]$$

$$= \dot{\Phi}(t,t_o)\ \underline{X}(t_o) \qquad \text{[property 5 of table 1]}$$

$$= F(t)\ \Phi(t,t_o)\ \underline{X}(t_o) \qquad \text{[with equation (a)]}$$

$$= F(t)\ \underline{X}(t) \qquad \text{q.e.d.}$$

2. Exponential solution.

Show that for the F and the C of example 1, equation (11) gives indeed the solution (7), given the general solution (8).

$$F = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \qquad C = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}$$

$\underline{W}$ is independent of t between t_o and t.

$$\Phi(t,t_o) = I + (t-t_o)F + \frac{1}{2!}(t-t_o)^2 F^2 + \frac{1}{3!}(t-t_o)^3 F^3 + \ldots \qquad (11)$$

$$\underline{X}(t) = \Phi(t,t_o)\ \underline{X}(t_o) + \int_{t_0}^{t} \Phi(t,\tau)\ C(\tau)\ \underline{W}(\tau)\ d\tau \qquad (8)$$

$$\underline{X}(t) = \begin{pmatrix} 1 & t-t_o & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & t-t_o \\ 0 & 0 & 0 & 1 \end{pmatrix} \underline{X}(t_o) + \begin{pmatrix} \frac{1}{2}(t-t_o)^2 & 0 \\ t-t_o & 0 \\ 0 & \frac{1}{2}(t-t_o)^2 \\ 0 & t-t_o \end{pmatrix} \underline{W} \qquad (7)$$

Solution

$$F^2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad F^3 = FF^2 = (0)$$

$$F^4 = FF^3 = (0)$$

etc.

Substituting into (11):

$$\phi(t,t_o) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + (t-t_o) \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & t-t_o & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & t-t_o \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Substituting into (8):

$$\underline{X}(t) = \begin{pmatrix} 1 & t-t_o & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & t-t_o \\ 0 & 0 & 0 & 1 \end{pmatrix} \underline{X}(t_o) + \begin{pmatrix} \int d\tau & \int (\tau-t_o) d\tau & 0 & 0 \\ 0 & \int d\tau & 0 & 0 \\ 0 & 0 & \int d\tau & \int (\tau-t_o) d\tau \\ 0 & 0 & 0 & \int d\tau \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \underline{W}$$

(the limits of the integrals are t_o and t)

or

$$\underline{X}(t) = \begin{pmatrix} 1 & t-t_o & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & t-t_o \\ 0 & 0 & 0 & 1 \end{pmatrix} \underline{X}(t_o) + \begin{pmatrix} t-t_o & \frac{1}{2}(t-t_o)^2 & 0 & 0 \\ 0 & t-t_o & 0 & 0 \\ 0 & 0 & t-t_o & \frac{1}{2}(t-t_o)^2 \\ 0 & 0 & 0 & t-t_o \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \underline{W}$$

or

$$\underline{X}(t) = \begin{pmatrix} 1 & t-t_o & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & t-t_o \\ 0 & 0 & 0 & 1 \end{pmatrix} \underline{X}(t_o) + \begin{pmatrix} \frac{1}{2}(t-t_o)^2 & 0 \\ t-t_o & 0 \\ 0 & \frac{1}{2}(t-t_o)^2 \\ 0 & t-t_o \end{pmatrix} \underline{W} \qquad \text{q.e.d.}$$

3. $\frac{\text{Good}}{\text{Bad}}$ dynamics, $\frac{\text{bad}}{\text{good}}$ measurements.

Show, or make at least plausible, that for bad dynamics and good measurements indeed $\hat{\underline{X}}_k \rightarrow K_k \underline{Z}_k$ and
for good dynamics and bad measurements $\hat{\underline{X}}_k \rightarrow \hat{\underline{X}}_{k,k-1}$.

Solution.

a. For good measurements the elements of R are small.
So (23) becomes:

$$K_k \approx P_{k,k-1} H_k^T \{H_k P_{k,k-1} H_k^T\}^{-1}$$

$$= P_{k,k-1} H_k^T (H_k^T)^{-1} P_{k,k-1}^{-1} H_k^{-1}$$

$$= H_k^{-1}$$

Substitution into (21) gives:

$$\hat{\underline{X}}_k \approx \hat{\underline{X}}_{k,k-1} + H_k^{-1} \underline{Z}_k - \hat{\underline{X}}_{k,k-1} = H_k^{-1} \underline{Z}_k = K_k \underline{Z}_k \qquad \text{q.e.d.}$$

b. For bad measurements the elements of R are large.
Then the elements of the filtergain K are small (see equation 23).
Then from equation (21) follows:

$$\hat{\underline{X}}_k \approx \hat{\underline{X}}_{k,k-1} \qquad \text{q.e.d.}$$

4. Kalman filter with measurement of one position line at a time.

A Kalman filter can be used even if no complete fix is carried out at any moment. In such a case the dynamics are essential to find a solution. Also the assumption of an initial state vector with its precision P_o is essential for the Kalman procedure.
Show this for a process with the same $\hat{\underline{X}}_o$, P_o, F, C, $\underline{W}_k$, Q, ϕ_{k-1} as in the examples 5 and 6, with a constant time-interval $t_k - t_{k-1} = T$, but with $\underline{Z}_1 = Z_e, \underline{Z}_2 = Z_n$, $\underline{V}_1 = v_1$ and $\underline{V}_2 = v_2$, and

with variances $\sigma_m^2 = E\{v_1^2\} = E\{v_2^2\}$. Find expressions for the update of the state vector after the second measurement, $\hat{\underline{X}}_2$, and its covariance matrix P_2, and discuss the results.

$$\hat{\underline{X}}_0 = \begin{pmatrix} 0 \\ S_0 \\ 0 \\ S_0 \end{pmatrix}, \quad P = \begin{pmatrix} \sigma_p^2 & 0 & 0 & 0 \\ 0 & \sigma_s^2 & 0 & 0 \\ 0 & 0 & \sigma_p^2 & 0 \\ 0 & 0 & 0 & \sigma_s^2 \end{pmatrix}, \quad Q = \sigma_a^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \underline{X}_k = \begin{pmatrix} E_k \\ \dot{E}_k \\ N_k \\ \dot{N}_k \end{pmatrix}$$

$$F = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad \underline{W}_k = \begin{pmatrix} (W_e)_k \\ (W_n)_k \end{pmatrix}, \quad \phi_{k-1} = \begin{pmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Solution.

For this process: $H_1 = (1 \quad 0 \quad 0 \quad 0)$ $\qquad H_2 = (0 \quad 0 \quad 1 \quad 0)$

$R_1 = R_2 = \sigma_m^2$ (one element) $\qquad \Gamma_1 = \Gamma_2 = \begin{pmatrix} \frac{1}{2}T^2 & 0 \\ T & 0 \\ 0 & \frac{1}{2}T^2 \\ 0 & T \end{pmatrix}$ like in example 5

For $P_{1,0}$ the same form is found as in example 5:

$$P_{1,0} = \begin{pmatrix} P_{11} & P_{12} & 0 & 0 \\ P_{12} & P_{22} & 0 & 0 \\ 0 & 0 & P_{11} & P_{12} \\ 0 & 0 & P_{12} & P_{22} \end{pmatrix} \quad \text{with} \quad \begin{aligned} P_{11} &= \sigma_p^2 + \sigma_s^2 T + \tfrac{1}{4}\sigma_a^2 T^4 \\ P_{12} &= \sigma_s^2 T + \tfrac{1}{2}\sigma_a^2 T^3 \\ P_{22} &= \sigma_s^2 + \sigma_a^2 T^2 \end{aligned}$$

with equation 23 one finds:

$$K_1 = P_{1,0} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \left\{ (1 \quad 0 \quad 0 \quad 0) \; P_{1,0} \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \sigma_m^2 \right\}^{-1}$$

(the form between $\{\}$ is a one-element-matrix or just a number).

or

$$K_1 = \frac{1}{P_{11} + \sigma_m^2} \begin{pmatrix} P_{11} \\ P_{12} \\ 0 \\ 0 \end{pmatrix}$$

With equation (22) the updated covariance matrix is found:

$$P_1 = (I - K_1 H_1) P_{1,0}$$

or

$$P_1 = \begin{pmatrix} \frac{\sigma_m^2}{P_{11}+\sigma_m^2} & 0 & 0 & 0 \\ \frac{-P_{12}}{P_{11}+\sigma_m^2} & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} P_{11} & P_{12} & 0 & 0 \\ P_{12} & P_{22} & 0 & 0 \\ 0 & 0 & P_{11} & P_{12} \\ 0 & 0 & P_{12} & P_{22} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{12} & a_{22} & 0 & 0 \\ 0 & 0 & P_{11} & 0 \\ 0 & 0 & 0 & P_{22} \end{pmatrix}$$

with:

$$a_{11} = \frac{\sigma_m^2 P_{11}}{P_{11}+\sigma_m^2}, \quad a_{12} = \frac{\sigma_m^2 P_{12}}{P_{11}+\sigma_m^2} \text{ and } a_{22} = \frac{\sigma_p^2\sigma_s^2+\sigma_p^2\sigma_a^2 T^2+\frac{1}{4}\sigma_a^2\sigma_s^2 T^4+P_{22}\sigma_m^2}{P_{11}+\sigma_m^2}$$

If one of the standard deviations σ_p, σ_s or σ_a is infinite, at least one of the diagonal elements of P_1 will become infinite as well, preventing the estimation of the whole state vector $\underline{X}_1$.

The extrapolated $P_{2,1}$ can now be found with equation (19):

$$P_{2,1} = \phi_1 P_1 \phi_1^T + \Gamma_1 Q_1 \Gamma_1^T$$

$$= \begin{pmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{12} & a_{22} & 0 & 0 \\ 0 & 0 & P_{11} & 0 \\ 0 & 0 & 0 & P_{22} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ T & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & T & 1 \end{pmatrix} + \sigma_a^2 \begin{pmatrix} \frac{1}{4}T^4 & \frac{1}{2}T^3 & 0 & 0 \\ \frac{1}{2}T^3 & T^2 & 0 & 0 \\ 0 & 0 & \frac{1}{4}T^4 & \frac{1}{2}T^3 \\ 0 & 0 & \frac{1}{2}T^3 & T^2 \end{pmatrix} =$$

$$\begin{pmatrix} b_{11} & b_{12} & 0 & 0 \\ b_{12} & b_{22} & 0 & 0 \\ 0 & 0 & b_{33} & b_{34} \\ 0 & 0 & b_{34} & b_{44} \end{pmatrix}$$

with:

$$b_{11} = a_{11} + 2a_{12}T + a_{22}T^2 + \frac{1}{4}\sigma_a^2 T^4 \qquad b_{33} = p_{11} + \frac{1}{4}\sigma_a^2 T^4$$

$$b_{12} = a_{12} + a_{22}T + \frac{1}{2}\sigma_a^2 T^3 \qquad b_{34} = \frac{1}{2}\sigma_a^2 T^3$$

$$b_{22} = a_{22} + \sigma_a^2 T^2 \qquad b_{44} = p_{22} + \sigma_a^2 T^2$$

The new gain is found with equation (23):

$$K_2 = P_{2,1}H_2^T \{H_2P_{2,0}H_2^T + R_2\}^{-1} \quad \text{with } R_2 = \sigma_m^2 \text{ and } H_2 = (0 \;\; 0 \;\; 1 \;\; 0)$$

$$K_2 = \frac{1}{b_{33}+\sigma_m^2} \begin{pmatrix} 0 \\ 0 \\ b_{33} \\ b_{34} \end{pmatrix}$$

With equation (22):

$$P_2 = \{I - K_2H_2\} P_{2,1} = \left\{ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} - \frac{1}{b_{33}+\sigma_m^2} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & b_{33} & 0 \\ 0 & 0 & b_{34} & 0 \end{pmatrix} \right\} P_{2,1}$$

or

$$P_2 = \begin{pmatrix} b_{11} & b_{12} & & 0 \\ b_{12} & b_{22} & & 0 \\ 0 & 0 & \frac{b_{33}\sigma_m^2}{b_{33}+\sigma_m^2} & \frac{b_{34}\sigma_m^2}{b_{33}+\sigma_m^2} \\ 0 & 0 & \frac{b_{34}\sigma_m^2}{b_{33}+\sigma_m^2} & \frac{b_{33}b_{44}-b_{34}^2+\sigma_m^2 b_{44}}{b_{33}+\sigma_m^2} \end{pmatrix}$$

The precision of $\hat{\underline{X}}_2$ is determined by the diagonal elements of P_2. It appears that for very badly known dynamics ($\sigma_a \rightarrow$) or for a very weak estimation of the initial velocity ($\sigma_s \rightarrow$) at least one of the diagonal elements of P_2 becomes infinite, so that the state vector cannot be found in this cases.

Although the question is now answered, we will als devellop a form for the state vector estimate $\hat{\underline{X}}_2$.

The extrapolation $\hat{\underline{X}}_{1,0}$ of the state vector is found with equation (18):

$$\hat{\underline{X}}_{1,2} = \Phi_0\hat{\underline{X}}_0 = \begin{pmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ S_0 \\ 0 \\ S_0 \end{pmatrix} = \begin{pmatrix} S_0T \\ S_0 \\ S_0T \\ S_0 \end{pmatrix}$$

First update:

$$\hat{\underline{X}}_1 = \hat{\underline{X}}_{1,0} + \frac{1}{p_{11}+\sigma_m^2}\begin{pmatrix} p_{11} \\ p_{12} \\ 0 \\ 0 \end{pmatrix}\{Z_e - (1 \quad 0 \quad 0 \quad 0)\ \hat{\underline{X}}_{1,0}\}$$

or:

$$\hat{\underline{X}}_1 = \begin{pmatrix} \dfrac{\sigma_m^2}{p_{11}+\sigma_m^2}\ S_0T + \dfrac{p_{11}}{p_{11}+\sigma_m^2}\ Z_e \\ \dfrac{p_{11}+\sigma_m^2-p_{12}T}{p_{11}+\sigma_m^2}\ S_0 + \dfrac{p_{12}T}{p_{11}+\sigma_m^2}\ \dfrac{Z_e}{T} \\ S_0T \\ S_0 \end{pmatrix}$$

Second extrapolation:

$$\hat{\underline{X}}_{2,1} = \Phi\hat{\underline{X}}_1 = \begin{pmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{pmatrix}\hat{\underline{X}}_1 = \begin{pmatrix} \dfrac{p_{11}+2\sigma_m^2-p_{12}T}{p_{11}+\sigma_m^2}\ S_0T + \dfrac{p_{11}+p_{12}T}{p_{11}+\sigma_m^2}\ Z_e \\ \dfrac{p_{11}+\sigma_m^2-p_{12}T}{p_{11}+\sigma_m^2}\ S_0 + \dfrac{p_{12}T}{p_{11}+\sigma_m^2}\ \dfrac{Z_e}{T} \\ 2S_0T \\ S_0 \end{pmatrix}$$

Second update:

$$\hat{\underline{X}}_2 = \hat{\underline{X}}_{2,1} + \{\underline{Z}_2 - H_2\hat{\underline{X}}_{2,1}\} = (I - K_2H_2)\hat{\underline{X}}_{2,1}\underline{Z}_2$$

$$= \begin{pmatrix} \alpha_1.2S_0T + (1-\alpha_1)2Z_e \\ \alpha_2.S_0 + (1-\alpha_2)Z_e/T \\ \alpha_3.2S_0T + (1-\alpha_3)Z_n \\ \alpha_4.S_0 + (1-\alpha_4)Z_n/T \end{pmatrix}$$

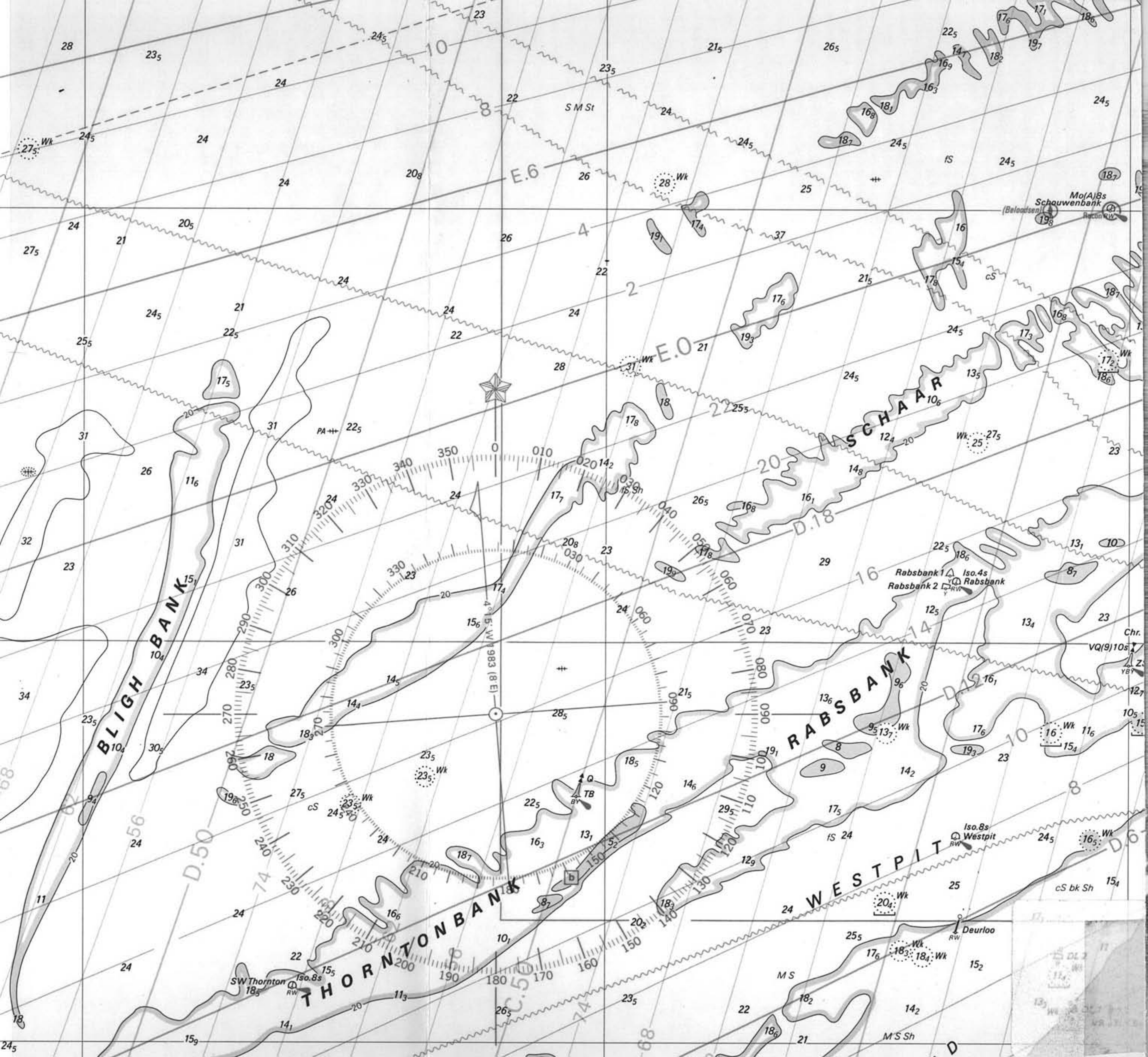
with $\alpha_1 = \dfrac{\frac{1}{2}p_{11}+\sigma_m^2-\frac{1}{2}p_{12}T}{p_{11}+\sigma_m^2}$

$$\alpha_2 = \frac{p_{11}+\sigma_m^2-p_{12}T}{p_{11}+\sigma_m^2}$$

$$\alpha_3 = \frac{\sigma_m^2}{b_{33}+\sigma_m^2}$$

$$\alpha_4 = \frac{b_{33}+\sigma_m^2-2b_{34}T}{b_{33}+\sigma_m^2}$$

So these estimated states appear to be weighted means of estimates from the initial states and estimates from the observations.
If the measurements are estimated to be very bad (large σ_m) the first terms appear to be accounted for to 100% ($\alpha \rightarrow 1$).



BLIGH BANK
THORNTONBANK
RABSBANK
SCHAAR
WESTPIT
Schouwenbank
Mo(A)8s
Racon
(Beloodsen)
Rabsbank 1
Rabsbank 2
Iso.4s
Rabsbank
Iso.8s
Westpit
Deurloo
SW Thornton
Iso.8s
Chr.
VQ(9)10s
S M St
PA
cS
fS
fS Sh
cS bk Sh
M S
M S Sh
Wk
TB
4°15' W 1983 (8'E)
E.6
E.0
D.18
D.50
D.12
D.6
C.50