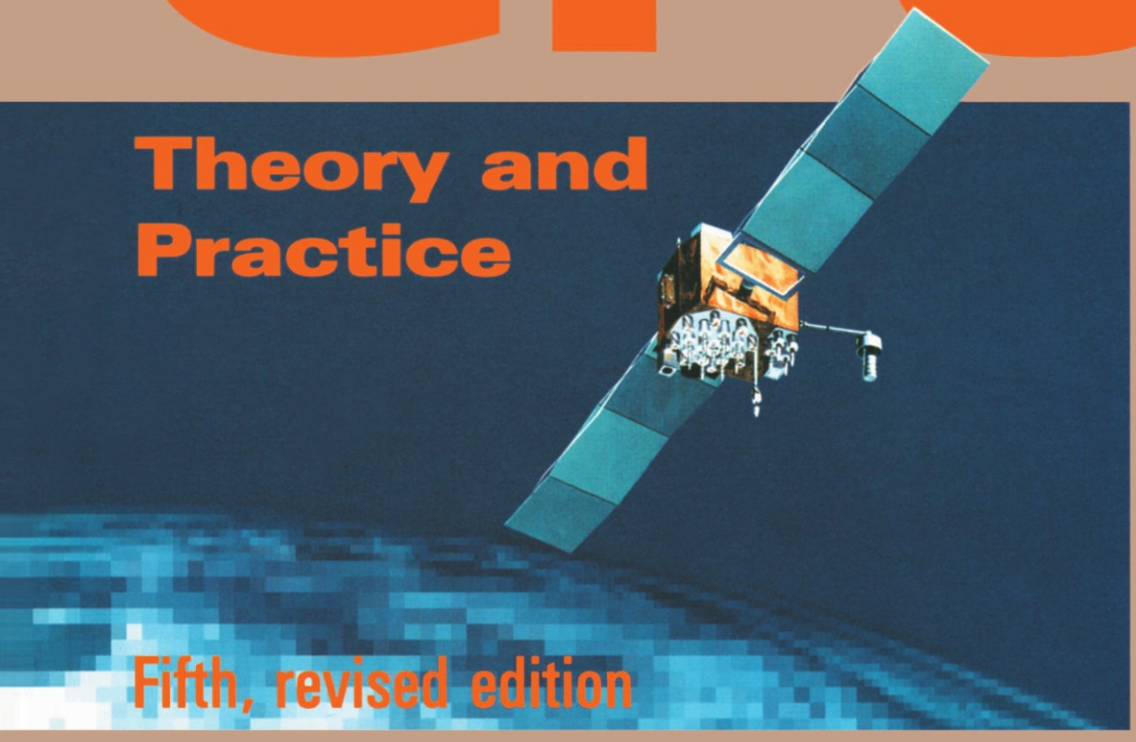
B. Hofmann-Wellenhof,
H. Lichtenegger, and J. Collins

GPS

Theory and Practice



Fifth, revised edition

Springer-Verlag Wien GmbH

B. Hofmann-Wellenhof,
H. Lichtenegger, J. Collins

Global Positioning System

Theory and Practice

Fifth, revised edition

Springer-Verlag Wien GmbH

Dr. Bernhard Hofmann-Wellenhof
Dr. Herbert Lichtenegger
Abteilung für Positionierung und Navigation, Technische Universität Graz
Graz, Austria

Dr. James Collins
GPS Services, Inc.
Rockville, Maryland, U.S.A.

Ursprünglich erschienen bei Springer-Verlag Wien New York 2001

Camera-ready copies provided by the authors
Cover illustration: photograph courtesy of Rockwell International
Printed by Novographic Druck G.m.b.H., A-1230 Wien
Printed on acid-free and chlorine-free bleached paper
SPIN 10772968

With 45 Figures

ISBN 978-3-211-83534-0 ISBN 978-3-7091-6199-9 (eBook)
DOI 10.1007/978-3-7091-6199-9

We dedicate this book to

Benjamin William Remondi

Foreword

This book is dedicated to Dr. Benjamin William Remondi for many reasons. The project of writing a Global Positioning System (GPS) book was conceived in April 1988 at a GPS meeting in Darmstadt, Germany. Dr. Remondi discussed with me the need for an additional GPS textbook and suggested a possible joint effort. In 1989, I was willing to commit myself to such a project. Unfortunately, the timing was less than ideal for Dr. Remondi. Therefore, I decided to start the project with other coauthors. Dr. Remondi agreed and indicated his willingness to be a reviewer.

I selected Dr. Herbert Lichtenegger, my colleague from the Technical University Graz, Austria, and Dr. James Collins from Rockville, Maryland, U.S.A.

In my opinion, the knowledge of the three authors should cover the wide spectrum of GPS. Dr. Lichtenegger is a geodesist with broad experience in both theory and practice. He has specialized his research to geodetic astronomy including orbital theory and geodynamical phenomena. Since 1986, Dr. Lichtenegger's main interest is dedicated to GPS. Dr. Collins retired from the U.S. National Geodetic Survey in 1980, where he was the Deputy Director. For the past ten years, he has been deeply involved in using GPS technology with an emphasis on surveying. Dr. Collins was the founder and president of Geo/Hydro Inc. My own background is theoretically oriented. My first chief, Prof. Dr. Peter Meissl, was an excellent theoretician; and my former chief, Prof. Dr.mult. Helmut Moritz, fortunately, still is.

It is appropriate here to say a word of thanks to Prof. Dr.mult. Helmut Moritz, whom I consider my mentor in science. He is – as is probably widely known – one of the world's leading geodesists and is currently president of the International Union for Geodesy and Geophysics (IUGG). In the fall of 1984, he told me I should go to the U.S.A. to learn about GPS. I certainly agreed, although I did not even know what GPS meant. On the same day, Helmut Moritz called Admiral Dr. John Bossler, at that time the Director of the National Geodetic Survey, and my first stay in the U.S.A. was arranged. Thank you, Helmut! I still remember the flight where I started to read the first articles on GPS. I found it interesting but I did not understand very much. Benjamin W. Remondi deserves the credit for providing my GPS instruction. He was a very patient and excellent teacher. I benefited enormously, and I accepted his offer to return to the U.S.A. several times. Aside from the scientific aspect, our families have also become friends.

The selection of topics is certainly different from the original book con-

ceived by Dr. Remondi. The primary selection criteria of the topics were: relevancy, tutorial content, and the interest and expertise of the authors. The book is intended to be a text on GPS, recognizing the tremendous need for textual materials for professionals, teachers, and for students. The authors believe that it was not necessary to dwell on the latest technical advances. Instead, concepts and techniques are emphasized.

The book can be employed as a classroom text at the senior or graduate levels, depending on the level of specialization desired. It can be read, selectively, by professional surveyors, navigators, and many others who need to position with GPS.

May 1992 B. Hofmann-Wellenhof

Preface

The contents of the book are partitioned into 13 chapters, a section of references, and a detailed index which should immediately help in finding certain topics of interest.

The first chapter is a historical review. It shows the origins of surveying and how global surveying techniques have been developed. In addition, a short history on the Global Positioning System (GPS) is given.

The second chapter is an overview of GPS. The system is explained by means of its three segments: the space segment, the control segment, and the user segment.

The third chapter deals with the reference systems, such as coordinate and time systems. The celestial and the terrestrial reference frames are explained in the section on coordinate systems, and the transformation between them is shown. The definition of different times is given in the section on time systems, together with appropriate conversion formulas.

The fourth chapter is dedicated to satellite orbits. This chapter specifically describes GPS orbits and covers the determination of the Keplerian and the perturbed orbit, as well as the dissemination of the orbital data.

The fifth chapter covers the satellite signal. It shows the fundamentals of the signal structure with its various components and the principles of the signal processing.

The sixth chapter deals with the observables. The data acquisition comprises code and phase pseudoranges and Doppler data. The chapter also contains the data combinations, both the phase combinations and the phase/code range combinations. Influences affecting the observables are described: the atmospheric and relativistic effects, the impact of the antenna phase center, and multipath.

The seventh chapter is dedicated to surveying with GPS. This chapter defines the terminology used and describes the planning of a GPS survey, surveying procedures, and in situ data processing.

The eighth chapter covers mathematical models for positioning. Models for observed data are investigated. Therefore, models for point positioning and relative positioning, based on various data sets, are derived.

The ninth chapter comprises the data processing and deals with cycle slip detection and repair. This chapter also discusses phase ambiguity resolution. The method of least squares adjustment is assumed to be known to the reader and, therefore, only a brief review (including the principle of Kalman filtering) is presented. Consequently, no details are given apart from

the linearization of the mathematical models, which are the input for the adjustment procedure.

The tenth chapter links the GPS results to a local datum. The necessary transformations are given. The combination of GPS and terrestrial data is also considered.

The eleventh chapter treats software modules. The intent of this chapter is not to give a detailed description of existing software and how it works. This chapter should help the reader decide which software would best suit his purposes. The very short sections of this chapter try to cover the variety of features which could be relevant to the software.

The twelfth chapter describes some applications of GPS. Global, regional, and local uses are mentioned, as well as the installation of control networks. The compatibility of GPS with other systems, such as Inertial Navigation Systems (INS) and the Global Navigation Satellite System (GLONASS), the Russian equivalent to GPS, is shown.

The thirteenth chapter deals with the future of GPS. Both critical aspects, such as selective availability and anti-spoofing, are discussed, along with positive aspects such as the combination of GPS with GLONASS and the International Maritime Satellite Communication Organization (INMARSAT). Also, some possible improvements in the hardware and software technology are suggested.

The hyphenation is based on Webster's Dictionary. Therefore, some deviations may appear for the reader accustomed to another hyphenation system. For example, the word "measurement", following Webster's Dictionary, is hyphenated mea-sure-ment; whereas, following The American Heritage Dictionary, the hyphenation is meas-ure-ment. The Webster's hyphenation system also contains hyphenations which are sometimes unusual for words with a foreign language origin. An example is the word "parameter". Following Webster's Dictionary, the hyphenation is pa-ram-e-ter. The word has a Greek origin, and one would expect the hyphenation pa-ra-me-ter.

Symbols representing a vector or a matrix are underlined. The inner product of two vectors is indicated by a dot "$\cdot$". The outer product, cross product, or vector product is indicated by the symbol "$\times$". The norm of a vector, i.e., its length, is indicated by two double-bars "$\|\|$".

Many persons deserve credit and thanks. Dr. Benjamin W. Remondi of the National Geodetic Survey at Rockville, Maryland, was a reviewer of the book. He has critically read and corrected the full volume. His many suggestions and improvements, critical remarks and proposals are gratefully acknowledged.

A second technical proofreading was performed by Dipl.-Ing. Gerhard Kienast from the section of Surveying and Landinformation of the Technical

University Graz. He has helped us with constructive critique and valuable suggestions.

Nadine Collins kindly read and edited the book in its final form, improving the flow and grammar of the text.

The index of the book was produced using a computer program written by Dr. Walter Klostius from the section of Surveying and Landinformation of the Technical University Graz. Also, his program helped in the detection of spelling errors.

The book is compiled based on the text system LaTeX. Some of the figures included were also developed with LaTeX. The remaining figures are drawn by using Autocad 11.0. The section of Physical Geodesy of the Technical University Graz deserves the thanks for these figures. Dr. Norbert Kühtreiber has drawn two of these figures, and the others were carefully developed by Dr. Konrad Rautz. This shows that theoreticians are also well-suited for practical tasks.

We are also grateful to the Springer Publishing Company for their advice and cooperation.

Finally, the inclusion by name of a commercial company or product does not constitute an endorsement by the authors. In principle, such inclusions were avoided whenever possible. Only those names which played a fundamental role in receiver and processing development are included for historical purposes.

May 1992 B. Hofmann-Wellenhof H. Lichtenegger J. Collins

Preface to the fifth edition

Today, writing a book on GPS is a completely different matter compared to the task eight years ago when this book was released for the first time. In these eight years, a tremendous improvement in GPS knowledge has occurred which would impact a conceptually new book. Topics have evolved which were not discussed a few years ago, like the third civil frequency, and by contrast, other topics have lost much of the importance of the earlier GPS years. The World Wide Web (WWW) is another feature with strong impact. Some GPS information may be more reliably downloaded in updated form from the internet than read in a book. Examples are the steadily changing satellite constellations and the RINEX format.

In the fifth edition, the afore-mentioned aspects have been considered to some extent. However, the overall structure of the former editions is retained since this was the request of many reviewers. In addition, the most recent advances in GPS technology like the decision to switch off Selective Availability (SA), the GPS modernization covering the new signal structure, improvements in the space and the control segment, and the augmentation of GPS by satellite-based and ground-based systems leading to future Global Navigation Satellite Systems (GNSS) are considered.

An exterior restriction was the limited number of pages of the book to keep a reasonable price. Figuratively, the more recent advances had to be implemented in exchange against subjects which might be considered well known. For this reason, we also had to reduce the list of references. With some exceptions like books and monographs, we treated this subtle matter by omitting all references "older" than 1990. This was relatively easy in cases where more recent publications of the authors were available but more critical if this was not the case. Thus, to avoid eliminating names of distinguished GPS pioneers, the person's name in connection with his fundamental idea is mentioned accordingly.

The authors are very grateful to Steve Nerem from the University of Texas at Austin for reporting us several errors we were not aware of. Steve wrote us in detail how to implement the corrections.

Similarly, the authors would like to thank the geodetic group of the Delft University of Technology, The Netherlands, headed by Peter Teunissen, for supporting us with many ideas and publications concerning the ambiguity resolution. Furthermore, Paul de Jonge and Christian Tiberius wrote us on the rank deficiency problem in some of our models. This letter arrived briefly before the release of the fourth edition and could not be implied adequately. The fifth edition fully encompasses this aspect. Thanks also to Kees de Jong for providing the figure on the SA discontinuity.

The authors would also like to thank Jan Kouba from the Geodetic Survey of Canada at Ottawa, Shin-Chan Han from the Ohio State University, Maria Tsakiri from the Curtin University of Technology at Perth, Australia, and Werner Oberegger, Austria, for some hints and advices. Thanks also to Walter Klostius, Norbert Kühtreiber, and Klaus Legat, all Graz University of Technology, for some help with the subject index and the figures.

The authors consider this edition the final one in the present structure; if another update should be required, the book will be written from the scratch. Your ideas for a future book and your advice is appreciated and encouraged.

August 2000 B. Hofmann-Wellenhof H. Lichtenegger J. Collins

Contents

Abbreviations

AFB	Air Force Base
AOC	Auxiliary Output Chip
AROF	Ambiguity Resolution On-the-Fly
A-S	Anti-Spoofing
AVL	Automatic Vehicle Location
C/A	Coarse/Acquisition
CBIS	Central Bureau Information System
CDMA	Code Division Multiple Access
CEP	Celestial Ephemeris Pole
CEP	Circular Error Probable
CGSIC	Civil GPS Service Interface Committee
CHAMP	Challenging Mini-satellite Payload
CIGNET	Cooperative International GPS Network
CIO	Conventional International Origin
CODE	Center for Orbit Determination in Europe
CORS	Continuously Operating Reference Station
CRF	Celestial Reference Frame
CSOC	Consolidated Space Operations Center
DARC	Data Radio Channel
DGPS	Differential GPS
DLL	Delay Lock Loop
DMA	Defense Mapping Agency
DME	Distance Measuring Equipment
DoD	Department of Defense
DOP	Dilution of Precision
DOSE	Dynamics of Solid Earth
DoT	Department of Transportation
DRMS	Distance Root Mean Square (error)
ECEF	Earth-Centered-Earth-Fixed
EGM	Earth Gravitational Model
EGNOS	European Geostationary Navigation Overlay Service
EOP	Earth Orientation Parameters
ERS	Earth Remote Sensing (satellite)
FAA	Federal Aviation Administration
FDMA	Frequency Division Multiple Access
FGCC	Federal Geodetic Control Committee
FOC	Full Operational Capability

FRP	Federal Radionavigation Plan
FTP	File Transfer Protocol
GBAS	Ground-Based Augmentation System
GDOP	Geometric Dilution of Precision
GEO	Geostationary Orbit (satellite)
GIC	GPS Integrity Channel
GIM	Global Ionosphere Map
GIS	Geographic Information System
GLONASS	Global Navigation Satellite System
GNSS	Global Navigation Satellite System
GOTEX	Global Orbit Tracking Experiment
GPS	Global Positioning System
GRS	Geodetic Reference System
HDOP	Horizontal Dilution of Precision
HIRAN	High Range Navigation (system)
HOW	Hand-Over Word
IAG	International Association of Geodesy
IAT	International Atomic Time
IAU	International Astronomical Union
ICRF	IERS (or International) Celestial Reference Frame
IERS	International Earth Rotation Service
IF	Intermediate Frequency
IGEB	Interagency GPS Executive Board
IGEX	International GLONASS Experiment
IGS	International GPS Service (for Geodynamics)
ILS	Instrument Landing System
INMARSAT	International Maritime Satellite (organization)
INS	Inertial Navigation System
IOC	Initial Operational Capability
ION	Institute of Navigation
IRM	IERS (or International) Reference Meridian
IRP	IERS (or International) Reference Pole
ISU	International System of Units
ITRF	IERS (or International) Terrestrial Reference Frame
ITS	Intelligent Transportation System
ITU	International Telecommunication Union
IUGG	International Union for Geodesy and Geophysics
IVHS	Intelligent Vehicle/Highway System
IWV	Integrated Water Vapor
JD	Julian Date
JPL	Jet Propulsion Laboratory

JPO	Joint Program Office
LAAS	Local Area Augmentation System
LEO	Low Earth Orbit (satellite)
LEP	Linear Error Probable
LORAN	Long-Range Navigation (system)
MEDLL	Multipath Estimating Delay Lock Loop
MEO	Mean Earth Orbit (satellite)
MIT	Massachusetts Institute of Technology
MITES	Miniature Interferometer Terminals for Earth Surveying
MJD	Modified Julian Date
MLS	Microwave Landing System
MRSE	Mean Radial Spherical Error
NAD	North American Datum
NAGU	Notice Advisories to GLONASS Users
NANU	Notice Advisories to Navstar Users
NASA	National Aeronautics and Space Administration
NAVSTAR	Navigation System with Timing and Ranging
NGS	National Geodetic Survey
NIMA	National Imagery and Mapping Agency
NIS	Navigation Information Service
NMEA	National Marine Electronics Association
NNSS	Navy Navigation Satellite System (or TRANSIT)
NSWC	Naval Surface Warfare Center
OCS	Operational Control System
OEM	Original Equipment Manufacturer
OTF	On-the-Fly
OTR	On-the-Run
PCMCIA	PC Memory Card International Association
PDD	Presidential Decision Directive
PDOP	Position Dilution of Precision
PLL	Phase Lock Loop
PPS	Precise Positioning Service
PRC	Pseudorange Correction
PRN	Pseudorandom Noise
RAIM	Receiver Autonomous Integrity Monitoring
RDS	Radio Data System
RF	Radio Frequency
RINEX	Receiver Independent Exchange (format)
RRC	Range Rate Correction
RTCM	Radio Technical Commission for Maritime (services)
RTK	Real-Time Kinematic

SA	Selective Availability
SBAS	Satellite-Based Augmentation System
SD	Selective Denial
SEP	Spherical Error Probable
SERIES	Satellite Emission Range Inferred Earth Surveying
SINEX	Software Independent Exchange (format)
SLR	Satellite Laser Ranging
SNR	Signal-to-Noise Ratio
SPOT	Satellite Probatoire d'Observation de la Terre
SPS	Standard Positioning Service
SV	Space Vehicle
TACAN	Tactical Air Navigation
TCAR	Three-Carrier Ambiguity Resolution
TDOP	Time Dilution of Precision
TEC	Total Electron Content
TLM	Telemetry (word)
TOPEX	(Ocean) Topography Experiment
TOW	Time-of-Week (count)
TRF	Terrestrial Reference Frame
TT	Terrestrial Time
TVEC	Total Vertical Electron Content
UERE	User Equivalent Range Error
UHF	Ultra High Frequency
URL	Uniform Resource Locator
USCG	U.S. Coast Guard
USGS	U.S. Geological Survey
USNO	U.S. Naval Observatory
UT	Universal Time
UTC	Universal Time Coordinated
UTM	Universal Transverse Mercator (projection)
VDOP	Vertical Dilution of Precision
VHF	Very High Frequency
VLBI	Very Long Baseline Interferometry
VOR	VHF Omnidirectional Range (equipment)
VRS	Virtual Reference Station
WAAS	Wide Area Augmentation System
WADGPS	Wide Area Differential GPS
WGS	World Geodetic System
WRC	World Radio Conference
WWW	World Wide Web

Numerical constants

GPS signal frequencies		
$f_0 =$	10.23 MHz	fundamental frequency
L1 =	1 575.42 MHz	primary carrier frequency
L2 =	1 227.60 MHz	secondary carrier frequency

WGS-84			
a	=	6 378 137.0 m	semimajor axis of ellipsoid
f	=	1/298.257 223 563	flattening of ellipsoid
ω_E	=	$7\,292\,115 \cdot 10^{-11}$ rad s^{-1}	angular velocity of the earth
GM	=	$3\,986\,004.418 \cdot 10^{8}$ m^3 s^{-2}	earth's gravitational constant
b	=	6 356 752.314 25 m	semiminor axis of ellipsoid
e^2	=	$6.694\,379\,990\,13 \cdot 10^{-3}$	first numerical eccentricity
e'^2	=	$6.739\,496\,742\,26 \cdot 10^{-3}$	second numerical eccentricity

Various constants			
c	=	299 792 458 m s^{-1}	velocity of light
ω_E	=	$7\,292\,115.1467 \cdot 10^{-11}$ rad s^{-1}	untruncated angular velocity
CD	=	January 6.0, 1980	Civil date of GPS standard epoch
JD	=	2 444 244.5	Julian date of GPS standard epoch

1 Introduction

1.1 The origins of surveying

Since the dawn of civilization, man has looked to the heavens with awe searching for portentous signs. Some of these men became experts in deciphering the mystery of the stars and developed rules for governing life based upon their placement. The exact time to plant the crops was one of the events that was foretold by the early priest astronomers who in essence were the world's first surveyors. Today, it is known that the alignment of such structures as the pyramids and Stonehenge was accomplished by celestial observations and that the structures themselves were used to measure the time of celestial events such as the vernal equinox.

Some of the first known surveyors were Egyptian surveyors who used distant control points to replace property corners destroyed by the flooding Nile River. Later, the Greeks and Romans surveyed their settlements. The Dutch surveyor Snell van Royen was the first who measured the interior angles of a series of interconnecting triangles in combination with baselines to determine the coordinates of points long distances apart. Triangulations on a larger scale were conducted by the French surveyors Picard and Cassini to determine a baseline extending from Dunkirk to Collioure. The triangulation technique was subsequently used by surveyors as the main means of determining accurate coordinates over continental distances.

The chain of technical developments from the early astronomical surveyors to the present satellite geodesists reflects man's desire to be able to master time and space and to use science to further his society. The surveyor's role in society has remained unchanged from the earliest days; that is to determine land boundaries, provide maps of his environment, and control the construction of public works.

1.2 Development of global surveying techniques

The use of triangulation (later combined with trilateration and traversing) was limited by the line of sight. Surveyors climbed to mountain tops and developed special survey towers to extend this line of sight usually by small amounts. The series of triangles was generally oriented or fixed by astronomic points where special surveyors had observed selected stars to determine the position of that point on the surface of the earth. Since these

B. Hofmann-Wellenhof et al., *Global Positioning System*

astronomic positions could be in error by hundreds of meters, each continent was virtually (positionally) isolated and their interrelationship was imprecisely known.

1.2.1 Optical global triangulation

Some of the first attempts to determine the interrelationship between the continents were made using the occultation of certain stars by the moon. This method was cumbersome at best and was not particularly successful. The launch of the Russian Sputnik satellite in 1957, however, had tremendously advanced the connection of the various geodetic world datums. In the beginning of the era of artificial satellites, an optical method, based (in principle) on the stellar triangulation method developed in Finland as early as 1946, was applied very successfully. The worldwide satellite triangulation program, often called the BC-4 program (after the camera that was used), for the first time determined the interrelationships of the major world datums. This method involved photographing special reflective satellites against a star background with a metric camera that was fitted with a specially manufactured chopping shutter. The image that appeared on the photograph consisted of a series of dots depicting each star's path and a series of dots depicting the satellite's path. The coordinates of selected dots were precisely measured using a photogrammetric comparator, and the associated spatial directions from the observing site to the satellite were then processed using an analytical photogrammetric model. Photographing the same satellite from a neighboring site simultaneously and processing the data in an analogous way yields another set of spatial directions. Each pair of corresponding directions forms a plane containing the observing points and the satellite, and the intersection of at least two planes results in the spatial direction between the observing sites. In the next step, these oriented directions were used to construct a global network with the scale being derived from several terrestrial traverses. An example is the European baseline running from Tromsø in Norway to Catania on Sicily. The main problem in using this optical technique was that clear sky was required simultaneously at a minimum of two observing sites separated by some 4 000 km, and the equipment was massive and expensive. Thus, optical direction measurement was soon supplanted by the electromagnetic ranging technique because of all-weather capability, greater accuracy, and lower cost of the newer technique.

1.2.2 Electromagnetic global trilateration

First attempts to (positionally) connect the continents by electromagnetic techniques was by the use of an electronic ranging system (HIRAN) devel-

oped during World War II to position aircraft. Beginning in the late 1940s, HIRAN arcs of trilateration were measured between North America and Europe in an attempt to determine the difference in their respective datums. A significant technological breakthrough occurred in 1957 after the launch of Sputnik, the first artificial satellite, when scientists around the world experienced that the Doppler shift in the signal broadcast by a satellite could be used as an observable to determine the exact time of closest approach of the satellite. This knowledge, together with the ability to compute satellite ephemerides according to Kepler's laws, led to the present capability of instantaneously determining precise position anywhere in the world.

The immediate predecessor of today's modern positioning system is the Navy Navigation Satellite System (NNSS), also called TRANSIT system. This system was composed of six satellites orbiting at altitudes of about 1 100 km with nearly circular polar orbits. The TRANSIT system was developed by the U.S. military, primarily, to determine the coordinates of vessels and aircraft. Civilian use of this satellite system was eventually authorized, and the system became widely used worldwide both for navigation and surveying.

Some of the early TRANSIT experiments by the former U.S. Defense Mapping Agency (DMA) and the U.S. Coast & Geodetic Survey showed that accuracies of about one meter could be obtained by occupying a point for several days and reducing the observations using the postprocessed precise ephemerides. Groups of Doppler receivers in translocation mode could also be used to determine the relative coordinates of points to submeter accuracy using the broadcast ephemerides. This system employed essentially the same Doppler observable used to track the Sputnik satellite; however, the orbits of the TRANSIT satellites were precisely determined by tracking them at widely spaced fixed sites.

1.3 History of the Global Positioning System

The Global Positioning System (GPS) was developed to replace the TRANSIT system because of two major shortcomings in the earlier system. The large time gaps in coverage were the main problem with TRANSIT. Since nominally a satellite passed overhead every 90 minutes, users had to interpolate their position between "fixes" or passes. The second problem with the TRANSIT system was its relatively low navigation accuracy.

In contrast, GPS answers the questions "What time, what position, and what velocity is it?" quickly, accurately, and inexpensively anywhere on the globe at any time.

1.3.1 Navigating with GPS

One aim of navigation is instantaneous positioning and velocity determination. As stated, one of the main problems with the TRANSIT system was the fact that it was not able to provide continuous positioning.

Satellite constellation
To provide a continuous global positioning capability, a scheme to orbit a sufficient number of satellites to ensure that (at least) four were always electronically visible was developed for GPS. Several schemes were proposed and it was found that 24 evenly spaced satellites placed in circular 12-hour orbits inclined 55° to the equatorial plane would provide the desired coverage for the least expense. In any event, this constellation provides the minimum of four satellites in good geometric position 24 hours per day anywhere on the earth. There are often more than the minimum number of satellites available for use and it is during these periods that surveyors perform special surveys. In fact, assuming a 10° elevation angle, there are brief periods where up to 10 GPS satellites are visible on the earth.

Point positioning
The GPS satellites are configured, primarily, to provide the user with the capability of determining his position, expressed for example by latitude, longitude, and elevation. This is accomplished by the simple resection process using the distances measured to satellites.

Consider the satellites frozen in space at a given instant. The space coordinates $\underline{\varrho}^S$ relative to the center of the earth of each satellite (see also Fig. 1.1) can be computed from the ephemerides broadcast by the satellite by an algorithm presented in Chap. 4. If the ground receiver defined by its geocentric position vector $\underline{\varrho}_R$ employed a clock that was set precisely to GPS system time (Sect. 3.3) the true distance or range ϱ to each satellite could be accurately measured by recording the time required for the (coded) satellite signal to reach the receiver. Each range defines a sphere (more precisely: surface of a sphere) with its center at the satellite. Hence, using this technique, ranges to only three satellites would be needed since the intersection of three spheres yields the three unknowns (e.g., latitude, longitude, and height) which could be determined from the three range equations

$$\varrho = \|\underline{\varrho}^S - \underline{\varrho}_R\| \,. \tag{1.1}$$

GPS receivers apply a slightly different technique. They typically use an inexpensive crystal clock which is set approximately to GPS time. Thus, the clock of the ground receiver is offset from true GPS time, and because

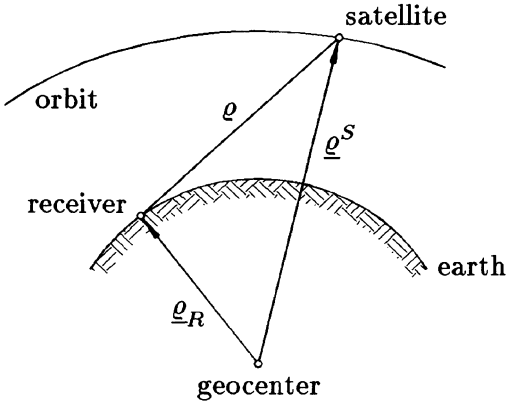


Fig. 1.1. Principle of satellite positioning

of this offset, the distance measured to the satellite differs from the "true" range. These distances are called pseudoranges R since they are the true range plus a range correction $\Delta\varrho$ resulting from the receiver clock error or bias δ. A simple model for the pseudorange is

$$R = \varrho + \Delta\varrho = \varrho + c\,\delta \tag{1.2}$$

with c being the velocity of light.

Four simultaneously measured pseudoranges are needed to solve for the four unknowns; these are three components of position plus the clock bias. Geometrically, the solution is accomplished by a sphere being tangent to the four spheres defined by the pseudoranges. The center of this sphere corresponds to the unknown position and its radius equals the range correction.

It is worth noting that the range error $\Delta\varrho$ could be eliminated in advance by differencing the pseudoranges measured from one site to two satellites or two different positions of one satellite. In the second case, the resulting range difference or delta range corresponds to the observable in the TRANSIT system. In both cases, the delta range defines a hyperboloid with its foci placed at the two satellites or the two different satellite positions for the geometric location of the receiver.

Considering the fundamental equation (1.1), one can conclude that the accuracy of the position determined using a single receiver essentially is affected by the following factors:

- accuracy of each satellite position,
- accuracy of pseudorange measurement,
- geometry.

Systematic errors in the satellite position and eventual satellite clock biases in the pseudoranges can be reduced or eliminated by differencing the pseudoranges measured from two sites to the satellite. This interferometric approach has become fundamental for GPS surveying as demonstrated below. However, no mode of differencing can overcome poor geometry.

A measure of satellite geometry with respect to the observing site is a factor known as Geometric Dilution of Precision (GDOP). In a geometric approach, this factor is inversely proportional to the volume of a body. This body is formed by points obtained from the intersection of a unit sphere with the vectors pointing from the observing site to the satellites. More details and an analytical approach on this subject are provided in Sect. 9.6.

Velocity determination
The determination of the instantaneous velocity of a moving vehicle is another goal of navigation. This can be achieved by using the Doppler principle of radio signals. Because of the relative motion of the GPS satellites with respect to a moving vehicle, the frequency of a signal broadcast by the satellites is shifted when received at the vehicle. This measurable Doppler shift is proportional to the relative radial velocity. Since the radial velocity of the satellites is known, the radial velocity of the moving vehicle can be deduced from the Doppler observable. A minimum of four Doppler observables is required to solve for the three components of the vehicle's velocity vector and, possibly, one frequency bias.

In summary, GPS was designed to solve many of the problems inherent to the TRANSIT system. Above all, GPS provides 24 hours a day instantaneous global navigation. The system as originally designed, however, did not include provision for the accurate surveying that is performed today. This surveying use of GPS resulted from a number of fortuitous developments described below.

1.3.2 Surveying with GPS

From navigation to surveying
As previously described, the use of near-earth satellites for navigation was demonstrated by the TRANSIT system. In 1964, I. Smith filed a patent describing a satellite system that would emit time codes and radio waves that would be received on earth as time delayed transmissions creating hyperbolic lines of position. This concept would become important in the treatment of GPS observables to compute precise vectors. A few years later, another patent was filed by R. Easton further refining the concept of comparing the

phase from two or more satellites.

In 1972, C. Counselman along with colleagues from the Massachusetts Institute of Technology (MIT) reported on the first use of interferometry to track the Apollo 16 Lunar Rover module. The principle they described is in essence the same technique they used later in developing the first geodetic GPS receiver and corresponds to differencing pseudoranges measured from two receivers to one satellite. The present use of the GPS carrier phase to make millimeter vector measurements dates from work by the MIT group using Very Long Baseline Interferometry (VLBI) performed between 1976 and 1978 where they proved that (for short lines) millimeter accuracy was obtainable using the interferometric technique.

In 1978, the Miniature Interferometer Terminals for Earth Surveying (MITES) were proposed. They detail how a satellite system can be used for precise surveying. This concept was further refined to include the Navigation System with Timing and Ranging (NAVSTAR). Also, the codeless technique that later became important in developing high-accuracy dual frequency receivers was described for the first time.

Observation techniques

When referring to high accuracy, GPS surveying implies the precise measurement of the vector between two (or more) GPS receivers.

The observation technique where both receivers involved remain fixed in position is called static surveying. The static method formerly required hours of observation and was the technique that was primarily used for early GPS surveys. A second technique where one receiver remains fixed, while the second receiver moves is called kinematic surveying. In 1986, B. Remondi first demonstrated that subcentimeter vector accuracies could be obtained between a pair of GPS survey instruments with as little as a few seconds of data collection using this method. B. Remondi also first developed another survey technique which is denoted as pseudokinematic technique. In this technique, a pair of receivers occupies a pair of points for two brief periods that are separated in time. This method, also denoted intermittent static, snapshot static, or reoccupation has demonstrated accuracies comparable to the static method.

Data processing is performed in two conceptually different ways. (1) Relative positioning is the technique where simultaneously observed data are processed; strictly speaking, the result is not obtainable instantaneously due to the required simultaneity. (2) The differential positioning technique involves placing one receiver at a fixed site of known position. Comparing computed ranges with measured pseudoranges, the reference site can transmit corrections to a roving receiver to improve its measured pseudoranges.

This technique provides instantaneous positioning and is, thus, mainly applied in kinematic surveys.

Hardware developments
The following sections contain reference to various terms that are more fully described in subsequent chapters. These are the C/A-code (Coarse/Acquisition) and P-code (Precision) which are basically code bits that are modulated on the two carrier signals broadcast by the GPS satellites. Code correlation as well as codeless techniques strip these codes from the carrier so that the phase of the (reconstructed) carrier can be measured. Brand names mentioned in this section are included for historical purposes since they represent the first of a certain class or type of receiver.

An interferometric technology for codeless pseudoranging was developed by P. MacDoran at the Jet Propulsion Laboratory (JPL). This Satellite Emission Range Inferred Earth Surveying (SERIES) technique was later improved for commercial geodetic applications. The culmination of the VLBI interferometric research applied to earth orbiting satellites was the production of a "portable" codeless GPS receiver that could measure short baselines to millimeter accuracy and long baselines to one part per million (ppm). This performance of the receiver, trade-named the Macrometer Interferometric Surveyor^{TM}, was demonstrated by the former U.S. Federal Geodetic Control Committee (FGCC).

A parallel development was being carried out by the DMA in cooperation with the U.S. National Geodetic Survey (NGS) and the U.S. Geological Survey (USGS). In 1981, these agencies developed specifications for a portable dual frequency code correlating receiver that could be used for precise surveying and point positioning. Texas Instruments Company was awarded the contract to produce a receiver later trade-named the TI-4100. The NGS geodesists C. Goad and B. Remondi developed software to process its carrier phase data in a manner similar to the method used by the MIT group (i.e., interferometrically).

The physical characteristics of the TI-4100 were significantly different from the Macrometer. The TI-4100 was a dual frequency receiver that used the P-code to track a maximum of four satellites, while the original Macrometer was a rack mounted codeless single frequency receiver that simultaneously tracked up to six satellites. There were also significant logistic differences in performing surveys using these two pioneer instruments. The TI-4100 received the broadcast ephemerides and timing signals from the GPS satellites so units could be operated independently, while the Macrometer required that all units be brought together prior to the survey and after the survey so that the time of the units could be synchronized. Also, the Macrometer

required that the ephemerides for each day's tracking be generated at the home office prior to each day's observing session.

The next major development in GPS surveying occurred in 1985 when manufacturers started to produce C/A-code receivers that measured and output the carrier phase. The first of this class of receivers was trade-named the Trimble 4000S. This receiver required the data to be collected on an external (i.e., laptop) computer. The 4000S was the first of the generic C/A-code receivers that eventually were produced by a host of manufacturers. The first Trimble receivers were sold without processing software; however, the company soon retained the services of C. Goad who produced appropriate vector computation software which set the standard for future software developers.

Today's GPS receivers include all features of the early models and additionally have expanded capabilities. By far the major portion of receivers produced today are the C/A-code single frequency type. For precise geodetic work, however, the more expensive dual frequency receivers are the standard.

Software developments

The development of GPS surveying software has largely paralleled the development of hardware.

The NGS has been one of the primary organizations in the world in developing independent GPS processing software. As previously mentioned, C. Goad and B. Remondi pioneered this development.

The NGS first produced processing software that used the Macrometer phase measurements and the precise ephemerides produced by the U.S. Naval Surface Warfare Center (NSWC). Other Macrometer users had to apply the processing software developed by the Macrometer manufacturer which required the use of specially formatted ephemerides produced (and sold) by them. The NGS software was also adapted for the TI-4100 format data and finally for other receivers that were subsequently used.

The original software developed by both the NGS and manufacturers computed individual vectors one at a time. These vectors were then combined in a network or geometric figure, and the coordinates of all points were determined using least squares adjustment programs.

The NGS and the Macrometer manufacturer eventually developed processing software that simultaneously determined all vectors observed during a given period of time (often called session). The second generation multibaseline software included the ability to determine corrections to the satellite orbits and was often called orbital relaxation software. This technique was pioneered by G. Beutler's group at the Bernese Astronomical Institute. To-

day, relaxation software is no longer applied since precise orbital information is available to the public.

Today, a huge number of scientific as well as commercial software is available. A review of software features is given in Chap. 11.

Ephemerides service

The first GPS surveys performed in late 1982 using Macrometers depended on orbital data derived from a private tracking network. Later, the broadcast ephemerides were used to supplement this private tracking data. The TI-4100 receiver obtained the ephemerides broadcast by the satellites so that processing programs could use this ephemerides to process vectors. The NSWC originally processed the military ephemerides obtaining "precise" postprocessed ephemerides which were turned over to the NGS for limited distribution to the public.

Today, various organizations around the world provide satellite tracking data from points that are referenced to a global datum. The tracking stations collect code range and phase data for both frequencies for all satellites. The data are processed and archived on a daily basis. Precise ephemerides are now available within a few days to the public upon request.

2 Overview of GPS

2.1 Basic concept

The Global Positioning System is the responsibility of the Joint Program Office (JPO), a component of the Space and Missile Center at El Segundo, California. In 1973, the JPO was directed by the U.S. Department of Defense (DoD) to establish, develop, test, acquire, and deploy a spaceborne positioning system. The present Navigation System with Timing and Ranging (NAVSTAR) Global Positioning System (GPS) is the result of this initial directive.

The Global Positioning System was conceived as a ranging system from known positions of satellites in space to unknown positions on land, at sea, in air and space. Effectively, the satellite signal is continually marked with its (own) transmission time so that when received the signal transit period can be measured with a synchronized receiver. The original objectives of GPS were the instantaneous determination of position and velocity (i.e., navigation), and the precise coordination of time (i.e., time transfer). A detailed definition given by W. Wooden in 1985 reads:

"The Navstar Global Positioning System (GPS) is an all-weather, space-based navigation system under development by the Department of Defense (DoD) to satisfy the requirements for the military forces to accurately determine their position, velocity, and time in a common reference system, anywhere on or near the Earth on a continuous basis."

Since the DoD is the initiator of GPS, the primary goals were military ones. But the U.S. Congress, with guidance from the President, directed DoD to promote its civil use. This was greatly accelerated by employing the Macrometer for geodetic surveying. This instrument was in commercial use at the time the military was still testing navigation receivers so that the first productive application of GPS was to establish high-accuracy geodetic networks.

As previously stated, GPS uses pseudoranges derived from the broadcast satellite signal. The pseudorange is derived either from measuring the travel time of the (coded) signal and multiplying it by its velocity or by measuring the phase of the signal. In both cases, the clocks of the receiver and the satellite are employed. Since these clocks are never perfectly synchronized, instead of true ranges "pseudoranges" are obtained where the synchronization error (denoted as clock error) is taken into account, cf. Eq. (1.2). Consequently, each equation of this type comprises four unknowns: the three

B. Hofmann-Wellenhof et al., *Global Positioning System*

point coordinates contained in the true range, and the clock error. Thus, four satellites are necessary to solve for the four unknowns. Indeed, the GPS concept assumes that four or more satellites are in view at any location on earth 24 hours a day. The solution becomes more complicated when using the measured phase. This observable is ambiguous by an integer number of signal wavelengths so that the model for phase pseudoranges is augmented by an initial bias, also called integer ambiguity.

The all-weather global system managed by the JPO consists of three segments:

- the space segment consisting of satellites which broadcast signals,
- the control segment steering the whole system,
- the user segment including the many types of receivers.

2.2 Space segment

2.2.1 Constellation

The GPS satellites have nearly circular orbits with an altitude of about 20 200 km above the earth and a period of approximately 12 sidereal hours. The constellation and the number of satellites used have evolved from earlier plans for a 24-satellite and 3-orbital plane constellation, inclined 63° to the equator. Later, for budgetary reasons, the space segment was reduced to 18 satellites, with three satellites in each of six orbital planes. This scheme was eventually rejected, since it did not provide the desired 24-hour worldwide coverage. In about 1986, the number of satellites planned was increased to 21, again three each in six orbital planes, and three additional active spares. The spare satellites were designated to replace malfunctioning "active" satellites. The present nominal constellation consists of 24 operational satellites deployed in six evenly spaced planes (A to F) with an inclination of 55° and with four satellites per plane. Furthermore, up to four active spare satellites for replenishment will be operational (Graviss 1992).

With the full constellation, the space segment provides global coverage with four to eight simultaneously observable satellites above 15° elevation at any time of day. If the elevation mask is reduced to 10°, occasionally up to 10 satellites will be visible; and if the elevation mask is further reduced to 5°, occasionally 12 satellites will be visible.

2.2.2 Satellites

General remarks

The GPS satellites, essentially, provide a platform for radio transceivers, atomic clocks, computers, and various ancillary equipment used to operate

the system. The electronic equipment of each satellite allows the user to measure a pseudorange R to the satellite, and each satellite broadcasts a message which allows the user to determine the spatial position $\underline{\varrho}^S$ of the satellite for arbitrary instants. Given these capabilities, users are able to determine their position $\underline{\varrho}_R$ on or above the earth by resection (Fig. 1.1). The auxiliary equipment of each satellite, among others, consists of solar panels for power supply and a propulsion system for orbit adjustments and stability control.

The satellites have various systems of identification: launch sequence number, assigned pseudorandom noise (PRN) code, orbital position number, NASA catalogue number, and international designation. In agreement with the satellite navigation message and to avoid any confusion, the PRN number will be used throughout this textbook.

Satellite categories

There are six classes or types of GPS satellites. These are the Block I, Block II, Block IIA, Block IIR, Block IIF, and Block III satellites. Detailed information on launch dates, orbital position (designation letter for orbital plane plus number) and operational periods can be found on the Web site of the U.S. Coast Guard Navigation Center under http://www.navcen.uscg.mil/gps/geninfo/constell.htm.

Eleven Block I satellites (weighing 845 kg) were launched in the period between 1978 to 1985 from Vandenberg AFB, California, with Atlas F launch vehicles. With the exception of one booster failure in 1981, all launches were successful. Today, none of the original Block I satellites is in operation. Considering the 4.5-year design life of these satellites, however, it is remarkable that some of the Block I satellites were operational for more than 10 years.

The Block II constellation is slightly different from the Block I constellation since the inclination of their orbital planes is 55° compared to the former 63° inclination. Apart from orbital inclination, there is an essential difference between Block I and Block II satellites related to U.S. national security. Block I satellite signals were fully available to civilian users, while some Block II satellite signals are restricted.

The first Block II satellite, costing approximately $ 50 million and weighing more than 1 500 kg, was launched on February 14, 1989 from the Kennedy Space Center, Cape Canaveral AFB in Florida, using a Delta II Rocket. The design life of the Block II satellites is 7.5 years. Individual satellites, however, remained operational more than 10 years.

The Block IIA satellites ("A" denotes advanced) are equipped with mutual communication capability. Some of them carry retroreflectors and can be tracked by Laser ranging. The first Block IIA satellite was launched on

November 26, 1990. Today, no distinction is made between Block II and Block IIA satellites.

The Block IIR satellites ("R" denotes replenishment or replacement) weigh more than 2 000 kg and the $ 42 million cost are about the same as for the Block II's. The first Block IIR satellite was successfully launched on July 23, 1997 and 19 more launches are expected to follow. These satellites have a design life of 10 years. They are equipped with improved facilities for communication and intersatellite tracking. Satellites launched after 2005 will also transmit additional signal components.

The Block IIF satellites ("F" denotes follow on) will weigh more than 2 000 kg and will be launched from 2007 onwards. These satellites will have a design life of 15 years. They will be equipped with improved on-board capabilities (such as inertial navigation systems) and an augmented signal structure. An artist's rendering of a Block IIF satellite is shown on the front cover of this textbook.

Presently, the DoD undertakes studies for the next generation of GPS satellites, called Block III satellites. These satellites are expected to carry GPS into 2030 and beyond.

Satellite signal

The actual carrier broadcast by the satellite is a spread spectrum signal that makes it less subject to intentional (or unintentional) jamming. The spread spectrum technique is commonly used today by such diverse equipment as hydrographic positioning ranging systems and wireless local area network systems.

The key to the system's accuracy is the fact that all signal components are precisely controlled by atomic clocks. The Block II satellites have four on-board time standards, two rubidium and two cesium clocks. The long-term frequency stability of these clocks reaches a few parts in 10^{-13} and 10^{-14} over one day. The future hydrogen masers will have an even better stability of 10^{-14} to 10^{-15} over one day. These highly accurate frequency standards being the heart of GPS satellites produce the fundamental L-band frequency of 10.23 MHz. Coherently derived from this fundamental frequency are (presently) two signals, the L1 and the L2 carrier waves generated by multiplying the fundamental frequency by 154 and 120 respectively yielding

$$\text{L1} = 1575.42 \text{ MHz},$$
$$\text{L2} = 1227.60 \text{ MHz}.$$

These dual frequencies are essential for eliminating the major source of error, i.e., the ionospheric refraction (Sect. 6.3.2).

The pseudoranges that are derived from measured travel times of the signal from each satellite to the receiver use two pseudorandom noise (PRN) codes that are modulated (superimposed) onto the two base carriers.

The first code is the C/A-code (Coarse/Acquisition-code) which is available for civilian use. The C/A-code, designated as the Standard Positioning Service (SPS), has an effective wavelength of approximately 300 m. The C/A-code is presently modulated upon L1 only and is purposely omitted from L2. This omission allows the JPO to control the information broadcast by the satellite and, thus, denies full system accuracy to nonmilitary users.

The second code is the P-code (Precision-code) which has been reserved for U.S. military and other authorized users. The P-code, designated as the Precise Positioning Service (PPS), has an effective wavelength of approximately 30 m. The P-code is modulated on both carriers L1 and L2. Unlimited access to the P-code was permitted until the system was declared fully operational.

In addition to the PRN codes, a data message is modulated onto the carriers consisting of status information, satellite clock bias, and satellite ephemerides. A detailed signal description is given in Sect. 5.1.

It is worth noting that the present signal structure will be improved in the near future (Sect. 13.2).

2.2.3 Operational capabilities

There are two operational capabilities: (1) Initial Operational Capability (IOC) and (2) Full Operational Capability (FOC).

IOC was attained in July 1993 when 24 (Block I/II/IIA) GPS satellites were operating and were available for navigation. Officially, IOC was declared by the DoD on December 8, 1993.

FOC was achieved when 24 Block II/IIA satellites were operational in their assigned orbits and the constellation was tested for operational military performance. Even though 24 Block II and Block IIA satellites were available since March 1994, FOC was not declared before July 17, 1995.

2.2.4 Denial of accuracy and access

Two techniques are known for denying civilian users full use of the system. The first is Selective Availability (SA) and the second is Anti-spoofing (A-S).

Selective availability

Originally, the accuracy expected from C/A-code pseudorange positioning was in the range of some 400 m. Field tests achieved the surprising level of navigation accuracy of 15–40 m for positioning and a fraction of a meter

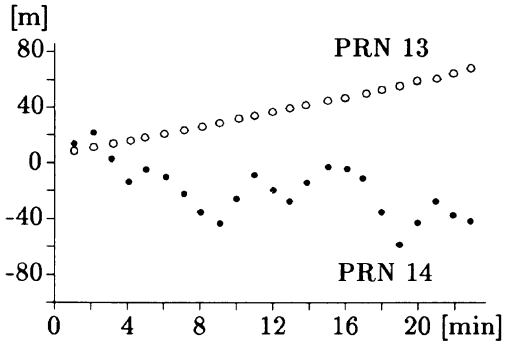


Fig. 2.1. Satellite clock behavior of PRN 13 (without SA) and of PRN 14 (with SA) on day 177 of 1992 after Breuer et al. (1993)

per second for velocity. The goal of SA was to deny this navigation accuracy to potential adversaries by dithering the satellite clock (δ-process) and manipulating the ephemerides (ε-process).

The δ-process is achieved by dithering the fundamental frequency of the satellite clock. The satellite clock bias has a direct impact on the pseudorange which is derived from a comparison of the satellite clock and the receiver clock. Since the fundamental frequency is dithered, code and carrier pseudoranges are affected in the same way. In Fig. 2.1, the different behavior of satellite clocks with and without SA is shown. With SA activated, there are variations of the pseudoranges with amplitudes of some 50 m and with periods of some minutes. When pseudoranges are differenced between two receivers, the dithering effect is eliminated.

The ε-process is the truncation of the orbital information in the transmitted navigation message so that the coordinates of the satellites cannot accurately be computed. The error in satellite position roughly translates to a like position error of stand-alone receivers. For baselines, the relative satellite position errors are (approximately) equal to the relative baseline errors. In Fig. 2.2, the behavior of the radial orbit error with and without SA is shown. In the case of SA, there are variations with amplitudes between 50 m and 150 m and with periods of some hours. The orbital errors cause pseudorange errors with similar characteristics. Thus, these errors are highly reduced when pseudoranges are differenced between two receivers,

SA has been in force since March 25, 1990. According to the specifications of the DoD, the accuracy for stand-alone receivers was degraded to 100 m for horizontal position and to 156 m for height. These specifications also implied a velocity error of $0.3\,\mathrm{m\,s^{-1}}$ and an error in time of 340 ns. All numbers are given at the 95% probability level. At the 99.99% probability

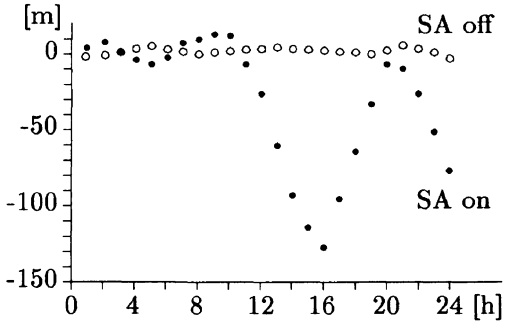


Fig. 2.2. Radial orbit error of PRN 21 on day 177 of 1992 with SA on and on day 184 of 1992 with SA off after Breuer et al. (1993)

level, the predictable accuracy decreased to 300 m for horizontal position and to 500 m for height (Department of Defense 1995).

Due to the undermined military effectiveness of SA by applying differential techniques, a joint recommendation of the U.S. National Academy of Public Administration and a committee of the National Research Council has proposed that SA should immediately be turned to zero and deactivated after some years (CGSIC 1995). The official answer to this proposal was released on March 29, 1996 in form of the Presidential Decision Directive (PDD) on GPS. The PDD expressed the intention to discontinue the use of SA within a decade in a manner that allows adequate time and resources for the military forces to prepare fully for operations without SA. In addition, the permanent Interagency GPS Executive Board (IGEB) was established. This board is commonly chaired by the DoD and the Department of Transportation (DoT) to balance military and civil interests. The full text of the public release statement on the PDD is published, e.g., in *GPS World* 1996, 7(5), page 50.

Somehow surprisingly, SA was turned off on May 2, 2000 at about 4:00 Universal Time (UT) after an announcement of the White House one day before. The benefits for civilian users are discussed in a fact sheet released by the U.S. Department of Commerce (2000). A prediction of the world after SA is given in Conley and Lavrakas (1999) and first experiences with SA off are discussed in Conley (2000), Jong (2000). One impressive result is presented in Fig. 2.3. Although the accuracy for stand-alone receivers is improved by a factor of ten, it must be kept in mind that despite turning off SA military advantages are ensured by new developments. One of these developments is Selective Denial (SD) which will deny access to the GPS signal for unauthorized users in regions of interest by ground-based jammers.

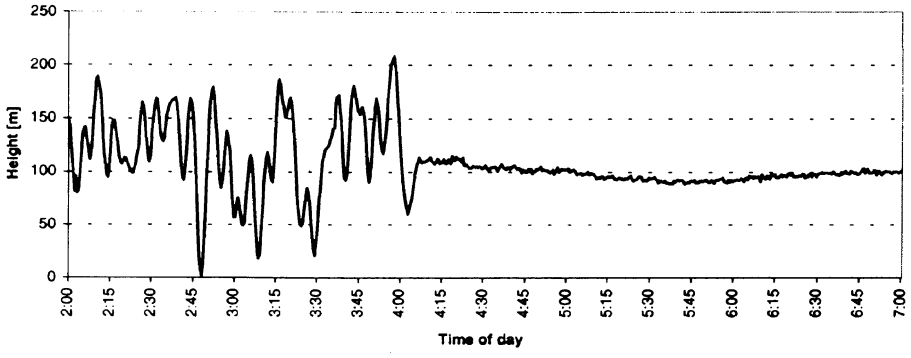


Fig. 2.3. Height variation in the station Kootwijk (The Netherlands) during the SA transition on May 2, 2000 (courtesy K. de Jong, Delft)

Anti-spoofing
The design of GPS includes the ability to essentially "turn off" the P-code or invoke an encrypted code as a means of denying access to the P-code to all but authorized users. The rationale for doing this is to keep adversaries from sending out false signals with the GPS signature to create confusion and cause users to misposition themselves.

A-S is accomplished by the modulo 2 sum of the P-code and an encrypting W-code. The resulting code is denoted as the Y-code. Thus, when A-S is active, the P-code on the L1 and the L2 carrier is replaced by the unknown Y-code. Note that A-S is either on or off. A variable influence of A-S (as was the case with SA) cannot occur.

For testing purposes, A-S was first turned on over the weekend of August 1, 1992 and later for several periods. It was expected that A-S would be switched on permanently when FOC had been attained; however, A-S was permanently implemented on January 31, 1994. In accordance with the DoD policy, no advance announcement of the implementation date was made.

The future signal structure will provide the C/A-code on both the L1 and the L2 carrier. Instead of the Y-code, a new military split-spectrum signal, denoted as M-code, will be introduced. This feature will make A-S superfluous.

2.3 Control segment

The Operational Control System (OCS) consists of a master control station, monitor stations, and ground control stations. The main operational tasks of the OCS are: tracking of the satellites for the orbit and clock determination and prediction, time synchronization of the satellites, and upload of the data

message to the satellites. The OCS was also responsible for imposing SA on the broadcast signals. The OCS performs many nonoperational activities, such as procurement and launch activities, that will not be addressed here.

Note that the control segment will be improved within the next ten years during the GPS modernization process.

2.3.1 Master control station

The location of the master control station was first at Vandenberg AFB, California, but has been moved to the Consolidated Space Operations Center (CSOC) at Shriver AFB (formerly known as Falcon AFB), Colorado Springs, Colorado. CSOC collects the tracking data from the monitor stations and calculates the satellite orbit and clock parameters using a Kalman estimator. These results are then passed to one of the three ground control stations for eventual upload to the satellites. The satellite control and system operation is also the responsibility of the master control station.

2.3.2 Monitor stations

There are five monitor stations located at: Hawaii, Colorado Springs, Ascension Island in the South Atlantic Ocean, Diego Garcia in the Indian Ocean, and Kwajalein in the North Pacific Ocean. Each of these stations is equipped with a precise atomic time standard and receivers which continuously measure pseudoranges to all satellites in view. Pseudoranges are measured every 1.5 seconds and, using the ionospheric and meteorological data, they are smoothed to produce 15-minute interval data which are transmitted to the master control station.

The tracking network described above is the official network for determining the broadcast ephemerides as well as modeling the satellite clocks. The data of up to 14 additional sites operated by the National Imagery and Mapping Agency (NIMA) are used to compute the precise ephemerides. Other tracking networks exist. These networks generally have no part in managing the system. A private tracking network was operated by the manufacturer of the Macrometer during the early 1980s. Today, more globally oriented tracking networks are operated. More details on this subject are provided in Sect. 4.4.1.

2.3.3 Ground control stations

These stations collocated with the monitor stations at Ascension, Diego Garcia, and Kwajalein are the communication links to the satellites and mainly consist of the ground antennas. The satellite ephemerides and clock

information, calculated at the master control station and received via communication links, are uploaded to each GPS satellite via S-band radio links. Formerly, uploading to each satellite was performed every eight hours; then the rate has been reduced to once (or twice) per day (Remondi 1991b). If a ground station becomes disabled, prestored navigation messages are available in each satellite to support a prediction span so that the positioning accuracy degrades quite gradually. The durations of positioning service of the satellites without contact from the OCS are given in Table 2.1.

Table 2.1. Positioning service without contact from the control segment

Block	Duration
I	3–4 days
II	14 days
IIA	180 days
IIR	>180 days

2.4 User segment

2.4.1 User categories

Military user
Strictly speaking, the term "user segment" is related to the DoD concept of GPS as an adjunct to the national defense program. Even during the early days of the system, it was planned to incorporate a GPS receiver into virtually every major defense system. It was envisioned that every aircraft, ship, land vehicle, and even groups of infantry would have an appropriate GPS receiver to coordinate their military activities. In fact, many GPS receivers were used as planned during, e.g., the 1991 Gulf War under combat conditions. In this war, SA which had been previously invoked was turned off so that troops could use more readily available civilian receivers. Handheld C/A-code receivers were particularly useful in navigating the desert.

There are various other military uses that have been proposed. One example is a receiver that can be connected to four antennas. When the antennas are placed in a fixed array (e.g., corners of a square), the attitude of the array can be determined in addition to its position. For example, placing antennas on the bow, stern, and port and starboard points of a ship would result in the determination of pitch, roll, yaw, and position of the vessel.

Civilian user
The civilian use of GPS occurred several years ahead of schedule in a manner not envisioned by the system's planners. The primary focus in the first few years of the system's development was on navigation receivers. As previously described in Chap. 1, the SERIES technique at JPL and the development of the Macrometer by C. Counselman started the GPS surveying revolution. The primary concept of using an interferometric rather than Doppler solution model meant that GPS could be used for not only long line geodetic measurements but also for the most exacting short line land survey measurements.

Today, GPS receivers are routinely being used to conduct all types of land and geodetic control surveys, and to precisely position photo-aircraft to reduce the amount of ground control needed for mapping.

The nonsurveyor civilian uses of GPS outnumber the survey uses of the system. One of the major uses of GPS is for fleet management and control. Several cities have equipped emergency vehicles with receivers and computers with screens that display the cities' road system. The location of each emergency vehicle can be sent to a dispatcher by radio link so that disposition of the resources are known, and vehicles can be rerouted when necessary. Similar systems are used to track trains and freight hauling vehicles. Probably, all aircraft and vessels will be equipped with GPS in the near future.

GPS is also being used by hikers and boaters to determine their locations. Some manufacturers are presently offering a combined system of GPS and computer graphics for use in automobiles at the cost of a good high-fidelity music system.

2.4.2 Receiver types

The uses of GPS described in the previous section are just a sample of the applications of this system. The diversity of the uses is matched by the type of receivers available today. This section will give an overview of the equipment marketed today; however, more details will be provided in Sect. 5.2. Based on the type of observables (i.e., code pseudoranges or carrier phases) and on the availability of codes (i.e., C/A-code, P-code, or Y-code), one can classify GPS receivers into four groups: (1) C/A-code pseudorange, (2) C/A-code carrier phase, (3) P-code carrier phase, and (4) Y-code carrier phase measuring instruments.

C/A-code pseudorange receivers
With this type of receiver, only code pseudoranges using the C/A-code are

measured. The receiver is usually a hand-held device powered by flashlight batteries. Typical devices output the three-dimensional position either in longitude, latitude, and height or in some map projection system (e.g., UTM coordinates and height). Receivers with four or more channels are preferred for applications where the receiver is in motion since simultaneous satellite ranges can be measured to produce more accurate positions. On the other hand, a single channel receiver would still be adequate for applications where the receiver is at a fixed location and the range measurements can be sequentially determined. The basic multichannel C/A-code pseudorange receiver is the type of receiver that is mostly used by hikers, boaters, and in automobiles.

C/A-code carrier receivers
With this type of receiver, code ranges and carrier phases from the L1 carrier only are obtained because the C/A-code is not modulated on L2. This means that no dual frequency data are available.

Most of the receivers for surveying in the early stage of GPS used the C/A-code to acquire and lock on to the L1 carrier. Most instruments have a minimum of four independent receiver channels and some of the more recent designs have 12 channels. These receivers perform all the functions of the previously described models and, in addition, store the time-tagged code range and carrier phase in some type of memory. Early models used laptop computers and magnetic tapes to store the measured data. Later models store measurement data in memory chips and PCMCIA cards.

This type of receiver has been augmented to measure the phases of the L2 carrier by the use of some codeless technique. The drawback is that the signal-to-noise ratio (SNR) of the L2 measurements is considerably lower than the C/A-code measurements on L1. Normally, the L2 phase is used in combination with the L1 measurement to reduce the ionospheric effect on the signal and, thus, provide a more accurate vector determination (especially for long lines).

These receivers can be used for all types of precise surveys including static, kinematic, and pseudokinematic methods.

P-code receivers
This type of receiver uses the P-code and is able to lock on to the L1 and L2 carrier. In the absence of A-S, the observables are derived by first correlating the signals with a replica of the P-code. After removing the P-code from the received satellite signal, phase measurements can be performed. One of the first receivers for surveying, point positioning, and navigation was the P-code TI-4100 receiver, completed in 1984. This receiver was developed more

from a military perspective than a civilian one and only military-related development would have attempted this. Manufacturers of civilian receivers were able to justify P-code work around 1989–1990. In the fall of 1991, the two main advantages of the P-code receiver were demonstrated by another P-code receiver during FGCC tests. The first is its capability to measure long (100 km) lines to an accuracy of a few centimeters. The second advantage is that P-code instruments can measure moderate length lines (20 km) to an accuracy of a few millimeters with as little as some minutes of data using techniques based on a linear combination of the measured phases of L1 and L2.

With A-S activated, in the emitted signal the P-code is replaced by the unknown Y-code. Thus, traditional P-code correlation technique can no longer be applied. However, this type of receiver can operate in a codeless or quasi-codeless mode providing carrier phase data and code pseudoranges for the L2 frequency without knowledge of the Y-code. The L2 tracking is accomplished using four techniques which differ in their performance characteristics: signal squaring, cross correlation, code correlation followed by squaring, and the Z-tracking technique. More details on these techniques can be found in Sect. 5.2.

Y-code receivers
This type of receiver provides access to the P-code with A-S invoked. Thus, the code ranges and phases can be derived from L1 and L2 signals by the P-code correlation technique. The access to the P-code is achieved by installing Auxiliary Output Chips (AOC) in each receiver channel which allow the decryption of the Y-code. However, only users authorized by the DoD have access to the AOC.

2.4.3 Information services

Several governmental and private information services have been established to provide GPS status information and data to the civilian users. Generally, the information contains constellation status reports, scheduled outages, and the DoD Notice Advisories to Navstar Users (NANU). Orbital data are provided in the form of an almanac suitable for making GPS coverage and satellite visibility predictions, and as precise ephemerides suitable for making the most precise vector computations. General information is also provided by listing the various GPS papers and meetings.

The official source for civilian information is the Navigation Information Service (NIS), formerly the GPS Information Center. This service is run by the U.S. Coast Guard (USCG) and includes 24-hour operation of a

telephone information service. In the U.S., call (703) 313- to enter the service and dial the extension 5900 for live information, the extension 5907 for GPS status voice recording and the extension 5920 for FAX. Information by the USCG Navigation Center is also disseminated via Internet. The e-mail address reads webmaster@smtp.navcen.uscg.mil and the Uniform Resource Locator (URL) in the World Wide Web (WWW) is included in Table 2.2. More details can be found in Department of Defense and Department of Transportation (2000).

Comprehensive information including precise ephemerides, satellite clock parameters, and other data is provided by the Central Bureau Information System (CBIS) of the International GPS Service for Geodynamics (IGS) located at the U.S. Jet Propulsion Laboratory (JPL). The CBIS is accessible through Internet and offers file transfer capability by anonymous file transfer protocol (FTP). More details on the CBIS can be found in, e.g., Gurtner (1995).

Outside the U.S., GPS information sources are also available. Among them are: the Australian Surveying and Land Information Group (AUSLIG), the Canadian Space Geodesy Forum (CANSPACE), the German GPS Information and Observation System (GIBS), and the Russian Interstate Navigation Information Center (INIC) to name a few. The actual coordinates of the information services are published regularly, for example, in the monthly magazine *GPS World.* Some Internet addresses are given in Table 2.2 providing a variety of links to other GPS related sites in the Internet. These include links to manufacturers, associations, governments, and universities. Some of them offer also basic information or tutorials for novice GPS users. A comprehensive overview of such tutorials can be found on the Web site http://www.gpsy.com/gpsinfo.

Table 2.2. GPS information services

Agency	Location	WWW address
AUSLIG	Australia	http://www.auslig.gov.au
CBIS	U.S.	http://igscb.jpl.nasa.gov
GIBS	Germany	http://gibs.leipzig.ifag.de
NIS	U.S.	http://www.navcen.uscg.mil

3 Reference systems

3.1 Introduction

The basic equation which relates the range ϱ with the instantaneous position vector $\underline{\varrho}^S$ of a satellite and the position vector $\underline{\varrho}_R$ of the observing site reads

$$\varrho = \|\underline{\varrho}^S - \underline{\varrho}_R\| \,. \tag{3.1}$$

In Eq. (3.1), both vectors must be expressed in a uniform coordinate system. The definition of a three-dimensional Cartesian system requires a convention for the orientation of the axes and for the location of the origin.

For global applications such as satellite geodesy, equatorial coordinate systems are appropriate. According to Fig. 3.1, a space-fixed or celestial system $\underline{X}_i^0$ and an earth-fixed or terrestrial system $\underline{X}_i$ must be distinguished where $i = 1, 2, 3$. The earth's rotational vector $\underline{\omega}_E$ serves as $\underline{X}_3$-axis in both cases. The $\underline{X}_1^0$-axis for the space-fixed system points towards the vernal equinox and is, thus, the intersection line between the equatorial and the ecliptic plane. The $\underline{X}_1$-axis of the earth-fixed system is defined by the intersection line of the equatorial plane with the plane represented by the Greenwich meridian. The angle Θ_0 between the two systems is called Greenwich sidereal time. The $\underline{X}_2$-axis being orthogonal to both the $\underline{X}_1$-axis and the $\underline{X}_3$-axis completes a right-handed coordinate frame.

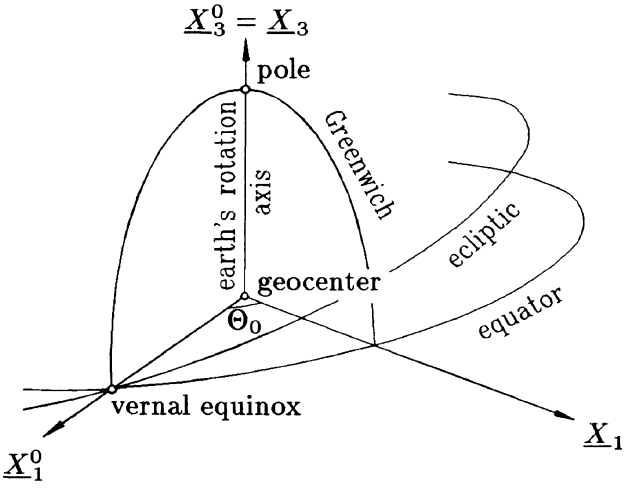


Fig. 3.1. Equatorial coordinate systems

B. Hofmann-Wellenhof et al., *Global Positioning System*

A coordinate system whose origin is located at the barycenter is at rest with respect to the solar system. It is, therefore, an inertial system which conforms to Newtonian mechanics. In a geocentric system, however, accelerations are present because the earth is orbiting the sun. Thus, in such a system the laws of general relativity must be taken into account. But, since the main relativistic effect is caused by the gravity field of the earth itself, the geocentric system is better suited for the description of the motion of a satellite close to the earth. Note that the axes of a geocentric coordinate system remain parallel because the motion of the earth around the sun is described by revolution without rotation.

The earth's rotational vector denoted by $\underline{\omega}_E$ oscillates due to several reasons. The basic differential equations describing the oscillations follow from classical mechanics and are given by

$$\underline{M} = \frac{d\underline{N}}{dt} \tag{3.2}$$

$$\underline{M} = \frac{\partial \underline{N}}{\partial t} + \underline{\omega}_E \times \underline{N} \tag{3.3}$$

where $\underline{M}$ denotes a torque vector, $\underline{N}$ is the angular momentum vector of the earth, and t indicates time (Moritz and Mueller 1988: Eqs. (2-54) and (2-59)). The symbol "$\times$" indicates a vector product. The torque $\underline{M}$ originates mainly from the gravitational forces of the sun and moon; therefore, it is closely related to the tidal potential. Equation (3.2) is valid in a (quasi-) inertial system such as $\underline{X}_i^0$ and Eq. (3.3) holds for the rotating system $\underline{X}_i$. The partial derivative expresses the temporal change of $\underline{N}$ with respect to the earth-fixed system, and the vector product considers the rotation of this system with respect to the inertial system. The earth's rotational vector $\underline{\omega}_E$ is related to the angular momentum vector $\underline{N}$ by the inertia tensor $\underline{C}$ as

$$\underline{N} = \underline{C}\,\underline{\omega}_E\,. \tag{3.4}$$

Introducing for the earth's rotational vector $\underline{\omega}_E$ its unit vector $\underline{\omega}$ and its norm $\omega_E = \|\underline{\omega}_E\|$, the relation

$$\underline{\omega}_E = \omega_E\,\underline{\omega} \tag{3.5}$$

can be formed.

The differential equations (3.2) and (3.3) can be separated into two parts. The oscillations of $\underline{\omega}$ are responsible for the variations of the $\underline{X}_3$-axis and are considered in the subsequent section. The oscillations of the norm ω_E cause variations in the speed of rotation which are treated in the section on time systems.

Considering only the homogeneous part (i.e., $\underline{M} = \underline{0}$) of Eqs. (3.2) and (3.3) leads to free oscillations. The inhomogeneous solution gives the forced oscillation. In both cases, the oscillations can be related to the inertial or to the terrestrial system. A further criterion for the solution concerns the inertia tensor. For a rigid earth and neglecting internal mass shifts, this tensor is constant; this is not the case for a deformable earth.

3.2 Coordinate systems

3.2.1 Definitions

Oscillations of axes

The oscillation of $\underline{\omega}$ with respect to the inertial space is called nutation. For the sake of convenience, the effect is partitioned into the secular precession and the periodic nutation. The oscillation with respect to the terrestrial system is named polar motion. A simplified representation of polar motion is given in Fig. 3.2. The image of a mean position of $\underline{\omega}$ is denoted by P in this polar plot of the earth. The free oscillation results in a motion of the rotational axis along a circular cone, with its mean position as axis, and an aperture angle of about $0.''4$. On the earth, this motion is represented by a 6 m radius circle around P. The image of an instantaneous position of the free oscillating earth's rotational axis is denoted by R_0. The period of the free motion amounts to about 430 days and is known as the Chandler period. The forced motion can also be described by a cone. In Fig. 3.2, this cone is mapped by the circle around the free position R_0. The radius of this circle is related to the tidal deformation and amounts to approximately 0.5 m. The nearly diurnal period of the forced motion corresponds to the tesseral part of the tidal potential of second degree.

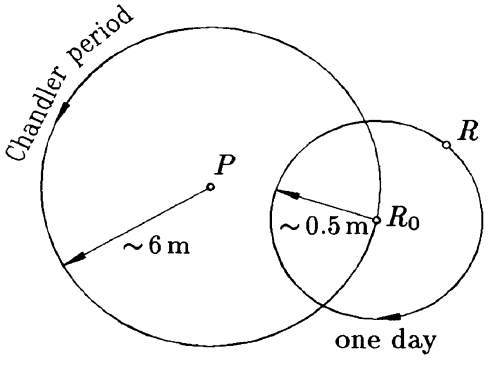


Fig. 3.2. Polar motion of the earth's rotational axis

The respective motions of the angular momentum axis, which is within 0.″001 of the rotational axis, are very similar. The free motion of the angular momentum axis deserves special attention because the forced motion can be removed by modeling the tidal attractions. The free polar motion is long-periodic and the free position in space is fixed since for $\underline{M} = \underline{0}$ the integration of Eq. (3.2) yields $\underline{N}$ = constant. By the way, this result implies the law of conservation of angular momentum as long as no external forces are applied. Because of the above mentioned properties, the angular momentum axis is appropriate to serve as a reference axis and the scientific community has named its free position in space Celestial Ephemeris Pole (CEP). A candidate for serving as reference axis in the terrestrial system is the mean position of the rotational axis denoted by P (Fig. 3.2). This position is called Conventional International Origin (CIO).

Conventional Celestial Reference System
By convention, the $\underline{X}_3^0$-axis is identical to the position of the angular momentum axis at a standard epoch denoted by J2000.0 (Sect. 3.3.3). The $\underline{X}_1^0$-axis points to the associated vernal equinox. This equinox is defined, for example, by a set of fundamental stars (Wielen et al. 1999). Since this system is defined conventionally and the practical realization does not necessarily coincide with the theoretical system, it is called (conventional) Celestial Reference Frame (CRF). Sometimes the term "quasi-inertial" is added to point out that a geocentric system is not rigorously inertial because of the accelerated motion of the earth around the sun. One example of such a celestial reference frame is that established by the International Earth Rotation Service (IERS) (McCarthy 1996). This frame is called ICRF where the first letter indicates the IERS origin. The ICRF is kinematically defined by a set of over 500 extragalactic objects, mostly quasars and galactic nuclei.

Conventional Terrestrial Reference System
Again by convention, the $\underline{X}_3$-axis is identical to the mean position of the earth's rotational axis as defined by the CIO. The $\underline{X}_1$-axis is associated with the mean Greenwich meridian. This system is named the (conventional) Terrestrial Reference Frame (TRF) and is defined by a set of terrestrial control stations serving as reference points. Most of the reference stations are equipped with Satellite Laser Ranging (SLR) or Very Long Baseline Interferometry (VLBI) capabilities.

An example for a terrestrial reference frame is the one produced by the IERS and called International Terrestrial Reference Frame (ITRF) (McCarthy 1996). The $\underline{X}_3$-axis is defined by the IERS Reference Pole (IRP) and the $\underline{X}_1$-axis lies in the IERS Reference Meridian (IRM). The ITRF is

realized by a number of terrestrial sites where temporal effects (plate tectonics, tidal effects) are also taken into account. Thus, ITRF is regularly updated and the acronym is supplemented by the last two digits of the last year whose data were used in the formation of the frame. Since mid-1999, the ITFR97 has been the operative version (Boucher et al. 1999).

Another terrestrial reference frame is the World Geodetic System 1984 (WGS-84). This geocentric system was originally realized by the coordinates of about 1 500 terrestrial sites which have been derived from TRANSIT observations. Associated to this frame is a geocentric ellipsoid of revolution, originally defined by the four parameters: semimajor axis a, normalized second degree zonal gravitational coefficient $\bar{C}_{2,0}$, truncated angular velocity of the earth ω_E, and earth's gravitational constant μ. This frame has been used for GPS since 1987.

The comparison of the original WGS-84 and ITRF revealed remarkable differences (Malys and Slater 1994):

1. The WGS-84 was established through Doppler observations from the TRANSIT satellite system, while ITRF is based on SLR and VLBI observations. The accuracy of the TRANSIT reference stations was estimated to be in the range of 1 to 2 meters, while the accuracy of the ITRF reference stations is at the centimeter level.

2. The numerical values for the original defining parameters differ from those in the ITRF. The only significant difference, however, was in the earth's gravitational constant $d\mu = \mu_{\mathrm{WGS}} - \mu_{\mathrm{ITRF}} = 0.582 \cdot 10^8\,\mathrm{m^3\,s^{-2}}$, which resulted in measurable differences in the satellite orbits.

Based on this information, the former DMA has proposed to replace the μ-value in the WGS-84 by the standard IERS value and to refine the coordinates of the GPS tracking stations. The revised WGS-84, valid since January 2, 1994, has been given the designation WGS-84 (G730) where the number 730 denotes the GPS week number when DMA has implemented the refined system (Bock 1996).

In 1996, NIMA (the successor of DMA) has implemented a revised version of the frame denoted as WGS-84 (G873) and being valid since September 29, 1996. The frame is realized by monitor stations with refined coordinates. The associated ellipsoid is now defined by the four parameters listed in Table 3.1 which are slightly different compared to the respective ITRF values. The current WGS-84 (G873) frame and the ITRF97 show unsignificant systematic differences of less than 2 cm. Hence, they are virtually identical. For more details on the frame, the reader is referred to National Imagery and Mapping Agency (1997).

Table 3.1. Parameters of the WGS-84 ellipsoid

Parameter and Value	Description
$a = 6\,378\,137.0$ m	Semimajor axis of the ellipsoid
$f = 1/298.257\,223\,563$	Flattening of the ellipsoid
$\omega_E = 7\,292\,115 \cdot 10^{-11}$ rad s^{-1}	Angular velocity of the earth
$\mu = 3\,986\,004.418 \cdot 10^{8}$ m^3 s^{-2}	Earth's gravitational constant

The transformation between the various terrestrial reference frames is generally performed by three-dimensional similarity transformations. This conformal transformation contains seven parameters and is given by

$$\underline{X}_{\mathrm{TRF}_1} = \underline{c} + \mu\, \underline{R}\, \underline{X}_{\mathrm{TRF}_2} \tag{3.6}$$

where $\underline{X}_{\mathrm{TRF}_1}$ denotes the three-dimensional position vector of a site in one and $\underline{X}_{\mathrm{TRF}_2}$ denotes the corresponding vector represented in another terrestrial coordinate reference frame. The vector $\underline{c} = [c_1, c_2, c_3]^T$ is the translation vector between the two coordinate frames, μ is a scale factor, and $\underline{R}$ is an orthonormal matrix. The latter is composed of three successive rotations α_i about the coordinate frame axes.

A vector $\underline{X}$ in the terrestrial reference frame can be represented by Cartesian coordinates X, Y, Z as well as by ellipsoidal coordinates φ, λ, h. The rectangular coordinates are often called Earth-Centered-Earth-Fixed (ECEF) coordinates. Details on the transformation of Cartesian and ellipsoidal (i.e., geodetic) coordinates are provided in Sect. 10.2.1.

3.2.2 Transformations

General remarks

The transformation between the Celestial Reference Frame (CRF) and the Terrestrial Reference Frame (TRF) is performed by means of rotations. For an arbitrary vector $\underline{x}$, the transformation is given by

$$\underline{x}_{\mathrm{TRF}} = \underline{R}^M \underline{R}^S \underline{R}^N \underline{R}^P \underline{x}_{\mathrm{CRF}} \tag{3.7}$$

where

$\underline{R}^M$ ... rotation matrix for polar motion
$\underline{R}^S$ ... rotation matrix for sidereal time
$\underline{R}^N$ ... rotation matrix for nutation
$\underline{R}^P$ ... rotation matrix for precession.

The CRF, defined at the standard epoch J2000.0, is transformed into the instantaneous or true system at observation epoch by applying the corrections due to precession and nutation. The $\underline{X}_3^0$-axis of the true CRF represents the free position of the angular momentum axis and, thus, points to the CEP. Rotating this system about the $\underline{X}_3^0$-axis and through the sidereal time by the matrix $\underline{R}^S$ does not change the position of the CEP. Finally, the CEP is rotated into the CIO by $\underline{R}^M$ which completes the transformation.

The rotation matrices in Eq. (3.7) are composed of the elementary matrices $\underline{R}_i\{\alpha\}$ describing a positive rotation of the coordinate system about the $\underline{X}_i$-axis and through the angle α. As it may be verified from any textbook on vector analysis, the rotation matrices are given by

$$\begin{aligned}
\underline{R}_1\{\alpha\} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{bmatrix} \\
\underline{R}_2\{\alpha\} &= \begin{bmatrix} \cos\alpha & 0 & -\sin\alpha \\ 0 & 1 & 0 \\ \sin\alpha & 0 & \cos\alpha \end{bmatrix} \\
\underline{R}_3\{\alpha\} &= \begin{bmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} .
\end{aligned} \tag{3.8}$$

Note that these matrices are consistent with right-handed coordinate systems. The rotation angle α has a positive sign for clockwise rotation as viewed from the origin to the positive $\underline{X}_i$-axis.

Precession

A graphic representation of precession is given in Fig. 3.3. The position of the mean vernal equinox at the standard epoch t_0 is denoted by E_0 and the position at the observation epoch t is denoted by E. The precession matrix

$$\underline{R}^P = \underline{R}_3\{-z\}\,\underline{R}_2\{\vartheta\}\,\underline{R}_3\{-\zeta\} \tag{3.9}$$

is composed of three successive rotation matrices where z, ϑ, ζ are the precession parameters. Explicitly, performing the multiplication,

$$\underline{R}^P = \begin{bmatrix} \cos z\cos\vartheta\cos\zeta - \sin z\sin\zeta & -\cos z\cos\vartheta\sin\zeta - \sin z\cos\zeta & -\cos z\sin\vartheta \\ \sin z\cos\vartheta\cos\zeta + \cos z\sin\zeta & -\sin z\cos\vartheta\sin\zeta + \cos z\cos\zeta & -\sin z\sin\vartheta \\ \sin\vartheta\cos\zeta & -\sin\vartheta\sin\zeta & \cos\vartheta \end{bmatrix} \tag{3.10}$$

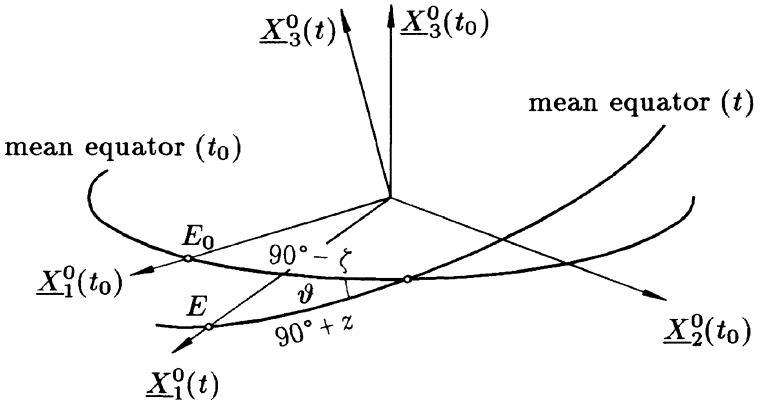


Fig. 3.3. Precession

is obtained. The precession parameters are computed from the time series

$$\begin{aligned} \zeta &= 2306.''2181\,T + 0.''30188\,T^2 + 0.''017998\,T^3 \\ z &= 2306.''2181\,T + 1.''09468\,T^2 + 0.''018203\,T^3 \\ \vartheta &= 2004.''3109\,T - 0.''42665\,T^2 - 0.''041833\,T^3\,, \end{aligned} \tag{3.11}$$

as given in Seidelmann (1992), Table 3.211.1. The parameter T represents the timespan expressed in Julian centuries of 36 525 mean solar days between the standard epoch J2000.0 and the epoch of observation. For a numerical example, consider an observation epoch J1990.5 which corresponds to $T = -0.095$. Substituting T into Eq. (3.11), the numerical values $\zeta = -219.''0880$, $z = -219.''0809$, and $\vartheta = -190.''4134$ are obtained. The substitution of these values into Eq. (3.10) gives the following numerical precession matrix:

$$\underline{R}^P = \begin{bmatrix} 0.999997318 & 0.002124301 & 0.000923150 \\ -0.002124301 & 0.999997744 & -0.000000981 \\ -0.000923150 & -0.000000981 & 0.999999574 \end{bmatrix}.$$

Nutation

A graphic representation of nutation is given in Fig. 3.4. The mean vernal equinox at the observation epoch is denoted by E and the true equinox by E_t. The nutation matrix $\underline{R}^N$ is composed of three successive rotation matrices where both the nutation in longitude $\Delta\psi$ and the nutation in obliquity $\Delta\varepsilon$ can be treated as differential quantities:

$$\underline{R}^N = \underline{R}_1\{-(\varepsilon + \Delta\varepsilon)\}\,\underline{R}_3\{-\Delta\psi\}\,\underline{R}_1\{\varepsilon\}\,. \tag{3.12}$$

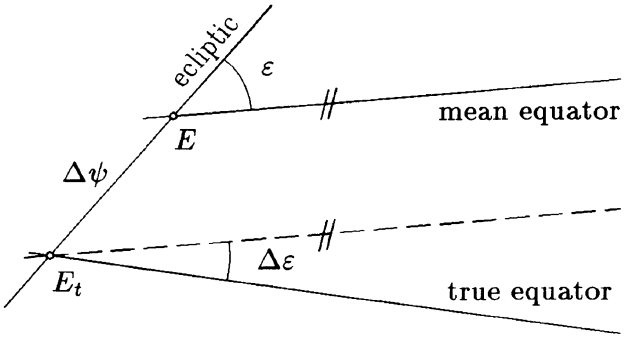


Fig. 3.4. Nutation

Explicitly,

$$\underline{R}^N = \begin{bmatrix} 1 & -\Delta\psi\cos\varepsilon & -\Delta\psi\sin\varepsilon \\ \Delta\psi\cos\varepsilon & 1 & -\Delta\varepsilon \\ \Delta\psi\sin\varepsilon & \Delta\varepsilon & 1 \end{bmatrix} \tag{3.13}$$

is obtained. The mean obliquity of the ecliptic ε has been determined (Seidelmann 1992: p. 114), as

$$\varepsilon = 23°26'21.''448 - 46.''8150\,T - 0.''00059\,T^2 + 0.''001813\,T^3 \tag{3.14}$$

where T is the same time factor as in Eq. (3.11). The nutation parameters $\Delta\psi$ and $\Delta\varepsilon$ are computed from the harmonic series:

$$\begin{aligned} \Delta\psi &= \sum_{i=1}^{106} a_i \sin\Big(\sum_{j=1}^{5} e_j\, E_j\Big) = -17.''2\,\sin\Omega_m + \ldots \\ \Delta\varepsilon &= \sum_{i=1}^{64} b_i \cos\Big(\sum_{j=1}^{5} e_j\, E_j\Big) = \quad 9.''2\,\cos\Omega_m + \ldots \end{aligned} \tag{3.15}$$

The amplitudes a_i, b_i as well as the integer coefficients e_j are tabulated, for example, in Seidelmann (1992), Table 3.222.1. The five fundamental arguments E_j describe mean motions in the sun-earth-moon system. The mean longitude Ω_m of moon's ascending node is one of the arguments. The moon's node retrogrades with a period of about 18.6 years and this period appears in the principal terms of the nutation series.

Sidereal time

The rotation matrix for sidereal time $\underline{R}^S$ is

$$\underline{R}^S = \underline{R}_3\{\Theta_0\}\,. \tag{3.16}$$

The computation of the apparent Greenwich sidereal time Θ_0 is shown in the section on time systems (Sect. 3.3.2).

The WGS-84 system is defined by a uniform angular velocity ω_E. Consequently, the mean sidereal time must be used in the case of GPS for the rotation angle in Eq. (3.16).

Polar motion

The previous computations yield the instantaneous CEP. The CEP must still be rotated into the CIO. This is achieved by means of the pole coordinates x_P, y_P which define the position of the CEP with respect to the CIO (Fig. 3.5). The pole coordinates are determined by the IERS and are available upon request (Gambis and Ray 2000). The rotation matrix for polar motion $\underline{R}^M$ is given by

$$\underline{R}^M = \underline{R}_2\{-x_P\}\,\underline{R}_1\{-y_P\} = \begin{bmatrix} 1 & 0 & x_P \\ 0 & 1 & -y_P \\ -x_P & y_P & 1 \end{bmatrix}. \tag{3.17}$$

The rotation matrices $\underline{R}^S$ and $\underline{R}^M$ are often combined to a single matrix $\underline{R}^R$ for earth rotation:

$$\underline{R}^R = \underline{R}^M \underline{R}^S. \tag{3.18}$$

In the case of GPS, the space-fixed coordinate system is already related to the CEP; hence, $\underline{R}^R$ is the only rotation matrix which must be applied for the transformation into the terrestrial system. For most practical purposes, the effect of polar motion is negligible.

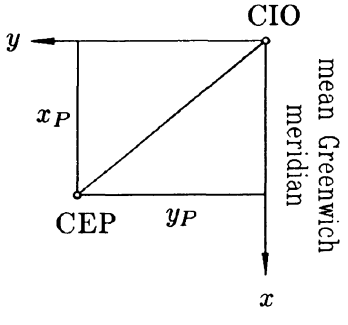


Fig. 3.5. Pole coordinates

3.3 Time systems

3.3.1 Definitions

Several time systems are in current use. They are based on various periodic processes such as earth rotation and are listed in Table 3.2.

Solar and sidereal times
A measure of earth rotation is the hour angle which is the angle between the meridian of a celestial body and a reference meridian (preferably the Greenwich meridian). Universal Time (UT) is defined by the Greenwich hour angle augmented by 12 hours of a fictitious sun uniformly orbiting in the equatorial plane. Sidereal time is defined by the hour angle of the vernal equinox. Taking the mean equinox as the reference leads to mean sidereal time and using the true equinox as a reference yields true or apparent sidereal time. Both, solar and sidereal time are not uniform since the angular velocity ω_E is not constant. The fluctuations are partly due to changes in the polar moment of inertia exerted by tidal deformation as well as other mass transports. Another factor is due to the oscillations of the earth's rotational axis itself. In this case, the universal time corrected for polar motion is denoted by UT1.

Dynamic times
The time systems derived from planetary motions in the solar system are called dynamic times. The Barycentric Dynamic Time (BDT) is an inertial time system in the Newtonian sense and provides the time variable in the equations of motion. The quasi-inertial Terrestrial Dynamic Time (TDT) was formerly called ephemeris time and serves for the integration of the differential equations for the orbital motion of satellites around the earth. In 1991, the International Astronomical Union (IAU) introduced the term

Table 3.2. Time systems

Periodic process	Time system
Earth rotation	Universal Time (UT)
	Greenwich Sidereal Time (Θ_0)
Earth revolution	Terrestrial Dynamic Time (TDT)
	Barycentric Dynamic Time (BDT)
Atomic oscillations	International Atomic Time (IAT)
	UT Coordinated (UTC)
	GPS Time

Terrestrial Time (TT) to replace TDT. Furthermore, the terminology of coordinate times according to the theory of general relativity was introduced. More details on this subject are given in Seidelmann and Fukushima (1992).

Atomic times
In practice, the dynamic time system is achieved by the use of atomic time scales. The UTC system is a compromise. The unit of the system is the atomic second, but to keep the system close to UT1 and approximate civil time, integer leap seconds are inserted at distinct epochs. The GPS time is also related to the atomic time system. GPS time is referenced to UTC as maintained by the U.S. Naval Observatory (USNO). The GPS time system nominally has a constant offset of 19 seconds with IAT and was coincident with UTC at the GPS standard epoch 1980, January $6.^{d}0$.

3.3.2 Conversions

The conversion between the times derived from earth rotation (i.e., the mean solar time corrected for polar motion UT1 and the apparent sidereal time Θ_0) is achieved by the formula

$$\Theta_0 = 1.002\,737\,9093\ \mathrm{UT1} + \vartheta_0 + \Delta\psi \cos\varepsilon\,. \tag{3.19}$$

The first term on the right side of Eq. (3.19) accounts for the different scales of solar and sidereal time and the quantity ϑ_0 represents the actual sidereal time at Greenwich midnight (i.e., 0^{h} UT). The third term describes the projection of $\Delta\psi$ onto the equator and considers the effect of nutation. The mean sidereal time follows from Eq. (3.19) by neglecting the nutation term and is a part of the navigation message broadcast by the GPS satellites (Sect. 4.4.2).

A time series has been determined for ϑ_0 as

$$\begin{aligned}\vartheta_0 = {} & 24\,110.^{s}54841 + 8\,640\,184.^{s}812866\,T_0 \\ & + 0.^{s}093104\,T_0^2 - 6.^{s}2 \cdot 10^{-6}\,T_0^3\end{aligned} \tag{3.20}$$

where T_0 represents the timespan expressed in Julian centuries of 36 525 mean solar days between the standard epoch J2000.0 and the day of observation at 0^{h} UT (Seidelmann 1992: p. 50).

The time UT1 is related to UTC by the quantity dUT1 which is time dependent and is reported by the IERS:

$$\mathrm{UT1} = \mathrm{UTC} + \mathrm{dUT1}\,. \tag{3.21}$$

When the absolute value of dUT1 becomes larger than $0.^{s}9$, a leap second is inserted into the UTC system.

Instead of the dynamic time system itself, the atomic time system serves as reference in GPS. The following relations are defined:

$$\begin{aligned} \mathrm{IAT} &= \mathrm{GPS} + 19.^{\mathrm{s}}000 && \text{constant offset} \\ \mathrm{IAT} &= \mathrm{TDT} - 32.^{\mathrm{s}}184 && \text{constant offset} \\ \mathrm{IAT} &= \mathrm{UTC} + 1.^{\mathrm{s}}000\, n && \text{variable offset as leap seconds are substituted.} \end{aligned} \tag{3.22}$$

The actual integer n is reported by the IERS. In June 2000, for example, the integer value was $n = 32$ and, thus, GPS time differed by exactly 13 seconds from UTC at this time.

3.3.3 Calendar

Definitions
The Julian Date (JD) defines the number of mean solar days elapsed since the epoch 4713 B.C., January $1.^{\mathrm{d}}5$.

The Modified Julian Date (MJD) is obtained by subtracting 2 400 000.5 days from JD. This convention saves digits and MJD commences at civil midnight instead of noon. For the sake of completeness, Table 3.3 with the Julian date for two standard epochs is given. This table enables, for example, the calculation of the parameter T for the GPS standard epoch. Subtracting the respective Julian dates and dividing by 36 525 (i.e., the number of days in a Julian century) yields $T = -0.199\,876\,7967$.

Date conversions
The relations for date conversions are taken from Montenbruck (1984) and are slightly modified so that they are only valid for an epoch between March 1900 and February 2100.

Let the civil date be expressed by integer values for the year Y, month

Table 3.3. Standard epochs

Civil date	Julian date	Explanation
1980 January $6.^{\mathrm{d}}0$	2 444 244.5	GPS standard epoch
2000 January $1.^{\mathrm{d}}5$	2 451 545.0	Current standard epoch (J2000.0)

M, day D, and a real value for the time in hours UT. Then

$$\begin{aligned} \mathrm{JD} = {} & \mathrm{INT}[365.25\,y] + \mathrm{INT}[30.6001\,(m+1)] \\ & + D + \mathrm{UT}/24 + 1\,720\,981.5 \end{aligned} \tag{3.23}$$

is the conversion into Julian date where INT denotes the integer part of a real number and y, m are given by

$$\begin{aligned} y &= Y - 1 \quad \text{and} \quad m = M + 12 \qquad \text{if} \quad M \leq 2 \\ y &= Y \qquad \text{and} \quad m = M \qquad\quad \text{if} \quad M > 2\,. \end{aligned}$$

The inverse transformation, that is the conversion from Julian date to civil date, is carried out stepwise. First, the auxiliary numbers

$$\begin{aligned} a &= \mathrm{INT}[\mathrm{JD} + 0.5] \\ b &= a + 1537 \\ c &= \mathrm{INT}[(b - 122.1)/365.25] \\ d &= \mathrm{INT}[365.25\,c] \\ e &= \mathrm{INT}[(b - d)/30.6001] \end{aligned}$$

are calculated. Afterwards, the civil date parameters are obtained from the relations

$$\begin{aligned} D &= b - d - \mathrm{INT}[30.6001\,e] + \mathrm{FRAC}[\mathrm{JD} + 0.5] \\ M &= e - 1 - 12\,\mathrm{INT}[e/14] \\ Y &= c - 4\,715 - \mathrm{INT}[(7 + M)/10] \end{aligned} \tag{3.24}$$

where FRAC denotes the fractional part of a number. As a by-product of date conversion, the day of week can be evaluated by the formula

$$N = \mathrm{modulo}\{\mathrm{INT}[\mathrm{JD} + 0.5],\ 7\} \tag{3.25}$$

where $N = 0$ denotes Monday, $N = 1$ means Tuesday, and so on. A further task is the calculation of the GPS week which is achieved by the relation

$$\mathrm{WEEK} = \mathrm{INT}[(\mathrm{JD} - 2\,444\,244.5)/7]\,. \tag{3.26}$$

The formulas given here can be used to prove the different dates in Table 3.3 or to verify the fact that the epoch J2000.0 corresponds to Saturday in the 1042nd GPS week. Note, however, that at the begin of every 1024th week the week number in the navigation message is set to zero since only 10 bits are reserved for the week number. The first rollover occurred at midnight 21–22 August 1999.

In 2000, "The GPS Toolbox" was opened (Hilla and Jackson 2000). In this toolbox, the source codes for date algorithms according to B. Remondi are available under the Web site http://www.ngs.noaa.gov/gps-toolbox.

4 Satellite orbits

4.1 Introduction

The applications of GPS depend substantially on knowing the satellite orbits. For single receiver positioning, an orbital error is highly correlated with the positional error. In the case of baselines, relative orbital errors are considered to be approximately equal to relative baseline errors.

Orbital information is either transmitted by the satellite as part of the broadcast message or can be obtained in form of precise ephemerides (typically some days after the observation) from several sources.

This chapter provides a review of orbital theory with emphasis on GPS orbits to acquaint the reader with the problems of computing ephemerides.

4.2 Orbit description

4.2.1 Keplerian motion

Orbital parameters

Assume two point masses m_1 and m_2 separated by the distance r. Considering for the moment only the attractive force between the masses and applying Newtonian mechanics, the movement of mass m_2 relative to m_1 is defined by the homogeneous differential equation of second order

$$\ddot{\underline{r}} + \frac{G(m_1 + m_2)}{r^3}\underline{r} = \underline{0} \tag{4.1}$$

where

$\underline{r}$ ... relative position vector with $\|\underline{r}\| = r$

$\ddot{\underline{r}} = \frac{d^2\underline{r}}{dt^2}$ ... relative acceleration vector

G ... universal gravitational constant

and the time parameter t being an inertial time. In fact, the inertial time is provided by the GPS system time.

In the case of motion of an artificial earth satellite, in a first approximation, both bodies can be considered as point masses and the mass of the

B. Hofmann-Wellenhof et al., *Global Positioning System*

satellite can be neglected. The product of G and the earth's mass M_E is denoted as μ and is known as one defining parameter of the WGS-84 reference system (Table 3.1):

$$\mu = G\,M_E = 3\,986\,004.418 \cdot 10^8\,\mathrm{m^3\,s^{-2}}\,.$$

The analytical solution of differential equation (4.1) can be found in textbooks on celestial mechanics (e.g., Brouwer and Clemence 1961, Beutler 1991, 1992), and leads to the well-known Keplerian motion defined by six orbital parameters which correspond to the six integration constants of the second-order vector equation (4.1). Satellite orbits may be restricted to elliptic motion, and the six associated parameters are listed in Table 4.1. The point of closest approach of the satellite with respect to the earth's center of mass is called perigee and the most distant position is the apogee. The intersection between the equatorial and the orbital plane with the unit sphere is termed the nodes, where the ascending node defines the northward crossing of the equator. A graphical representation of the Keplerian orbit is given in Fig. 4.1.

The mean angular satellite velocity n (also known as the mean motion) with revolution period P follows from Kepler's Third Law given by

$$n = \frac{2\pi}{P} = \sqrt{\frac{\mu}{a^3}}\,. \tag{4.2}$$

For GPS orbits, the nominal semimajor axis is $a = 26\,560\,\mathrm{km}$. The substitution of a into Eq. (4.2) yields an orbital period of 12 sidereal hours. The ground track of the satellites, thus, repeats every sidereal day.

The instantaneous position of the satellite within its orbit is described by an angular quantity known as anomaly (where this notation is used for historical reasons).

Table 4.1. Keplerian orbital parameters

Parameter	Notation
Ω	Right ascension of ascending node
i	Inclination of orbital plane
ω	Argument of perigee
a	Semimajor axis of orbital ellipse
e	Numerical eccentricity of ellipse
T_0	Epoch of perigee passage

Table 4.2. Anomalies of the Keplerian orbit

Notation	Anomaly
$M(t)$	Mean anomaly
$E(t)$	Eccentric anomaly
$v(t)$	True anomaly

Table 4.2 lists anomalies commonly used. The mean anomaly $M(t)$ is a mathematical abstraction, while both the eccentric anomaly $E(t)$ and the true anomaly $v(t)$ are geometrically producible (Fig. 4.1). The three

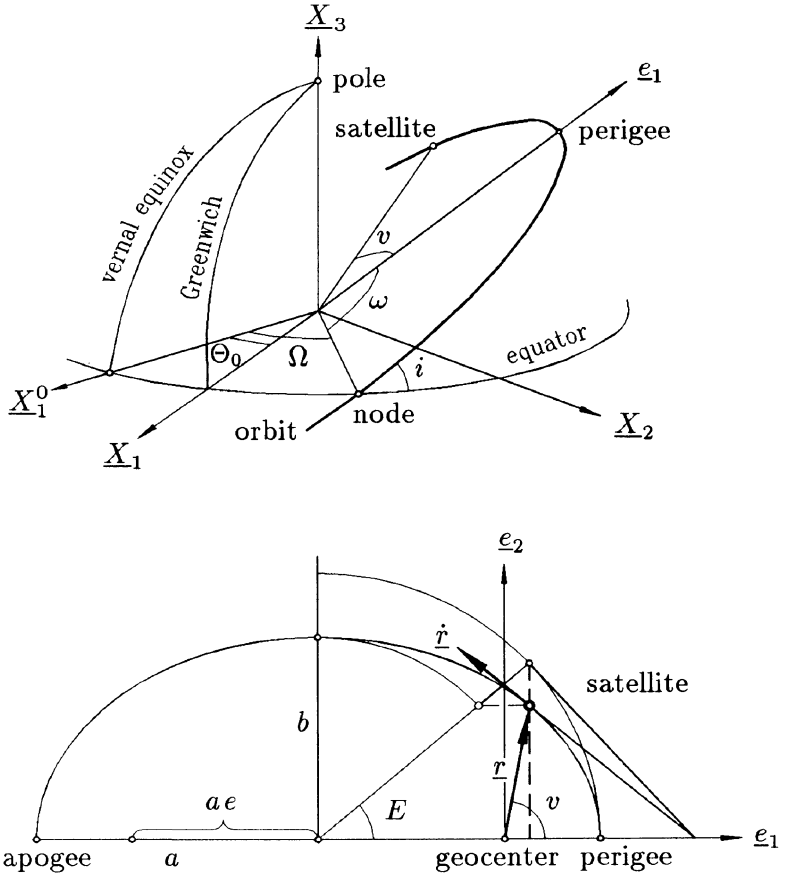


Fig. 4.1. Keplerian orbit

anomalies are related by the formulas

$$M(t) = n\,(t - T_0) \tag{4.3}$$

$$E(t) = M(t) + e \sin E(t) \tag{4.4}$$

$$v(t) = 2 \arctan \left[\sqrt{\frac{1+e}{1-e}} \tan \frac{E(t)}{2} \right] \tag{4.5}$$

where e denotes the eccentricity. Equation (4.3) is valid by definition and shows that the mean anomaly can be used instead of T_0 as a defining parameter. Equation (4.4) is known as Kepler's equation and is obtained in the course of the analytical integration of Eq. (4.1). Finally, Eq. (4.5) follows from purely geometric relations as shown in the following paragraph.

To become more familiar with the various anomalies, assume an orbit with a semidiurnal orbital period and an eccentricity of $e = 0.1$. At an epoch 3 hours after perigee passage, the mean anomaly is $M = 90.^\circ0000$. The calculation of the eccentric anomaly requires iteration and gives the value $E = 95.^\circ7012$. The true anomaly is obtained as $v = 101.^\circ3838$

Orbit representation

The coordinate system $\underline{e}_1$, $\underline{e}_2$ defining the orbital plane is shown in Fig. 4.1. The position vector $\underline{r}$ and the velocity vector $\dot{\underline{r}} = d\underline{r}/dt$ of the satellite can be represented by means of the eccentric as well as the true anomaly:

$$\underline{r} = a \begin{bmatrix} \cos E - e \\ \sqrt{1-e^2} \sin E \end{bmatrix} = r \begin{bmatrix} \cos v \\ \sin v \end{bmatrix} \tag{4.6}$$

$$r = a\,(1 - e \cos E) = \frac{a\,(1-e^2)}{1 + e \cos v} \tag{4.7}$$

$$\dot{\underline{r}} = \frac{n\,a^2}{r} \begin{bmatrix} -\sin E \\ \sqrt{1-e^2} \cos E \end{bmatrix} = \sqrt{\frac{\mu}{a\,(1-e^2)}} \begin{bmatrix} -\sin v \\ \cos v + e \end{bmatrix} \tag{4.8}$$

$$\dot{r} = \frac{n\,a^2}{r} \sqrt{1 - (e \cos E)^2} = \sqrt{\mu \left(\frac{2}{r} - \frac{1}{a} \right)}. \tag{4.9}$$

The components of the vector $\underline{r}$ are evident from the geometry in Fig. 4.1 where the semiminor axis b of the orbital ellipse is replaced by $a\sqrt{1-e^2}$. The geocentric distance $r = r(E)$ corresponds to the norm $\|\underline{r}(E)\|$ and follows

from simple algebra. The representation $r = r(v)$ is known as polar equation of the ellipse (Bronstein and Semendjajew 1996: p. 43).

Equation (4.5) can now be verified. For this purpose, the first components of Eq. (4.6) are equated and $r = r(E)$ of Eq. (4.7) is substituted. This leads to $\cos v = (\cos E - e)/(1 - e\cos E)$. The identity $\tan(\alpha/2) = \sqrt{(1-\cos\alpha)/(1+\cos\alpha)}$ applied twice and using simple algebra yields (4.5).

The derivation of the velocity vector $\dot{\underline{r}} = \dot{\underline{r}}(E)$ requires knowledge of the time derivative $dE/dt = n\,a/r$. This result is obtained by differentiating Kepler's equation and substituting Eq. (4.7). The derivation of $\dot{\underline{r}} = \dot{\underline{r}}(v)$ is much more laborious and the result is given without proof. The norm of $\dot{\underline{r}}$ (i.e., the velocity $\dot{r}$) can be derived from either representation $\dot{\underline{r}}(E)$ or $\dot{\underline{r}}(v)$. The calculation of $\dot{r}(E)$ is straightforward; the proof of $\dot{r}(v)$ is left to the reader. It is worth noting that Eq. (4.9) when squared and divided by two relates kinetic energy on the left side with potential energy on the right side where a by definition is a constant. Hence, Eq. (4.9) can be recognized as the law of energy conservation in the earth-satellite system!

The transformation of $\underline{r}$ and $\dot{\underline{r}}$ into the equatorial system $\underline{X}_i^0$ is performed by a rotation matrix $\underline{R}$ and results in vectors denoted by $\underline{\varrho}$ and $\dot{\underline{\varrho}}$. The superscript "S" generally indicating a satellite is omitted here for simplicity. The vectors expressed in the orbital system must be considered as three-dimensional vectors for the transformation. Therefore, the axes $\underline{e}_1$, $\underline{e}_2$ are supplemented with an $\underline{e}_3$-axis which is orthogonal to the orbital plane. Since $\underline{r}$ and $\dot{\underline{r}}$ are vectors in the orbital plane (represented by $\underline{e}_1$, $\underline{e}_2$), their $\underline{e}_3$-component is zero.

The transformation is defined by

$$\begin{aligned} \underline{\varrho} &= \underline{R}\,\underline{r} \\ \dot{\underline{\varrho}} &= \underline{R}\,\dot{\underline{r}} \end{aligned} \tag{4.10}$$

where the matrix $\underline{R}$ is composed of three successive rotation matrices and is given by

$$\begin{aligned} \underline{R} &= \underline{R}_3\{-\Omega\}\,\underline{R}_1\{-i\}\,\underline{R}_3\{-\omega\} \\ &= \begin{bmatrix} \cos\Omega\cos\omega - \sin\Omega\sin\omega\cos i & -\cos\Omega\sin\omega - \sin\Omega\cos\omega\cos i & \sin\Omega\sin i \\ \sin\Omega\cos\omega + \cos\Omega\sin\omega\cos i & -\sin\Omega\sin\omega + \cos\Omega\cos\omega\cos i & -\cos\Omega\sin i \\ \sin\omega\sin i & \cos\omega\sin i & \cos i \end{bmatrix} \\ &= \begin{bmatrix} \underline{e}_1 & \underline{e}_2 & \underline{e}_3 \end{bmatrix}. \end{aligned} \tag{4.11}$$

The column vectors of the orthonormal matrix $\underline{R}$ are the axes of the orbital coordinate system represented in the equatorial system $\underline{X}_i^0$.

In order to rotate the system $\underline{X}_i^0$ into the terrestrial system $\underline{X}_i$, an additional rotation through the angle Θ_0, the Greenwich sidereal time, is required. The transformation matrix, therefore, becomes

$$\underline{R}' = \underline{R}_3\{\Theta_0\}\,\underline{R}_3\{-\Omega\}\,\underline{R}_1\{-i\}\,\underline{R}_3\{-\omega\}\,. \tag{4.12}$$

The product $\underline{R}_3\{\Theta_0\}\,\underline{R}_3\{-\Omega\}$ can be expressed by a single matrix $\underline{R}_3\{-\ell\}$ where $\ell = \Omega - \Theta_0$. Hence, Eq. (4.12) can be written in the form

$$\underline{R}' = \underline{R}_3\{-\ell\}\,\underline{R}_1\{-i\}\,\underline{R}_3\{-\omega\} \tag{4.13}$$

and matrix $\underline{R}'$ corresponds to matrix $\underline{R}$ if in Eq. (4.11) the parameter Ω is replaced by ℓ.

For a numerical example, assume a satellite orbiting in a Kepler ellipse with the following parameters: $a = 26\,000\,\text{km}$, $e = 0.1$, $\omega = -140°$, $i = 60°$, $\ell = 110°$. To calculate the position and velocity vector in the earth-fixed equatorial system at an epoch where the eccentric anomaly is $E = 45°$, the vectors are first calculated in the orbital plane using Eqs. (4.6) through (4.9). Then, the transformation into the equatorial system is performed by means of Eq. (4.10) using the rotation matrix $\underline{R}'$. The final result is

$$\underline{\varrho} = [11\,465,\ 3\,818,\ -20\,923]^T \qquad [\text{km}]$$

$$\underline{\dot{\varrho}} = [-1.2651,\ 3.9960,\ -0.3081]^T \qquad [\text{km s}^{-1}]$$

where "T" denotes transposition.

In addition to the fixed orbital system $\underline{e}_i$, another orthonormal system $\underline{e}_i^*$ may be defined. This system rotates about the $\underline{e}_3$-axis because the $\underline{e}_1^*$-axis always points towards the instantaneous satellite position. Hence, the unit vectors $\underline{e}_i^*$ can be derived from the position and velocity vectors by

$$\underline{e}_1^* = \frac{\underline{\varrho}}{\|\underline{\varrho}\|} \qquad \underline{e}_3^* = \frac{\underline{\varrho} \times \underline{\dot{\varrho}}}{\|\underline{\varrho} \times \underline{\dot{\varrho}}\|} = \underline{e}_3 \qquad \underline{e}_2^* = \underline{e}_3^* \times \underline{e}_1^*\,. \tag{4.14}$$

Note that the base vectors $\underline{e}_i^*$ correspond to the column vectors of a modified rotation matrix $\underline{R}^*$ if the parameter ω is replaced by $(\omega + v)$ in Eq. (4.11).

The transformation of a change $\Delta\underline{\varrho}$ in the position vector into the orbital system $\underline{e}_i^*$ results in a vector $\Delta\underline{r} = [\Delta r_1,\ \Delta r_2,\ \Delta r_3]^T$ with its components reckoned along the respective axes $\underline{e}_i^*$:

$$\begin{aligned} \Delta r_1 &= \underline{e}_1^* \cdot \Delta\underline{\varrho} && \text{radial component} \\ \Delta r_2 &= \underline{e}_2^* \cdot \Delta\underline{\varrho} && \text{along-track component} \\ \Delta r_3 &= \underline{e}_3^* \cdot \Delta\underline{\varrho} && \text{across-track component.} \end{aligned} \tag{4.15}$$

Inversely, $\Delta\underline{\varrho}$ is calculated if the vector $\Delta\underline{r}$ is given. The solution follows from the inversion of Eq. (4.15) and leads to

$$\Delta\underline{\varrho} = \underline{R}^* \, \Delta\underline{r} \,. \tag{4.16}$$

For a numerical examination assume a change $\Delta\underline{\varrho} = [0.1,\ 1.0,\ -0.5]^T$ [km] in the satellite position of the previous example. Applying Eqs. (4.14) and (4.15) gives $\Delta\underline{r} = [0.638,\ 0.914,\ 0.128]^T$ [km].

Differential relations

The derivatives of $\underline{\varrho}$ and $\dot{\underline{\varrho}}$ with respect to the six Keplerian parameters are required in one of the subsequent sections. The differentiation can be separated into two groups because in Eq. (4.10) the vectors $\underline{r}$ and $\dot{\underline{r}}$ depend only on the parameters a, e, T_0, whereas the matrix $\underline{R}$ is only a function of the remaining parameters ω, i, Ω.

In order to obtain the derivatives of $\underline{r}$ and $\dot{\underline{r}}$, it is advantageous to differentiate the vectors represented by the eccentric anomaly. As a first step, the derivatives of the quantities n, E, r with respect to the parameters a, e, T_0 are calculated. By differentiating Eq. (4.2) one gets

$$\frac{\partial n}{\partial a} = -\frac{3}{2}\frac{n}{a} \qquad \frac{\partial n}{\partial e} = 0 \qquad \frac{\partial n}{\partial T_0} = 0 \,. \tag{4.17}$$

Differentiating the Kepler equation (4.4) and substituting Eq. (4.17) leads to

$$\frac{\partial E}{\partial a} = -\frac{3}{2}\frac{M}{r} \qquad \frac{\partial E}{\partial e} = \frac{a}{r}\sin E \qquad \frac{\partial E}{\partial T_0} = -n\,\frac{a}{r} \,. \tag{4.18}$$

Finally, the differentiation of Eq. (4.7) yields

$$\begin{aligned} \frac{\partial r}{\partial a} &= (1 - e\cos E) + a\,e\,\sin E\,\frac{\partial E}{\partial a} \\ \frac{\partial r}{\partial e} &= -a\,\cos E + a\,e\,\sin E\,\frac{\partial E}{\partial e} \\ \frac{\partial r}{\partial T_0} &= a\,e\,\sin E\,\frac{\partial E}{\partial T_0} \,. \end{aligned} \tag{4.19}$$

Now, the desired result follows after applying the chain rule to Eqs. (4.6) and (4.8), and after algebraic operations where $dm = -n\,dT_0$ is substituted:

$$\frac{\partial \underline{r}}{\partial a} = \begin{bmatrix} \cos E - e + \dfrac{3}{2}\dfrac{a}{r}M\sin E \\ \sqrt{1-e^2}\,(\sin E - \dfrac{3}{2}\dfrac{a}{r}M\cos E) \end{bmatrix} \tag{4.20}$$

$$\frac{\partial \underline{r}}{\partial e} = \begin{bmatrix} -a\left(1 + \frac{a}{r}\sin^2 E\right) \\ \frac{a^2 \sin E}{r\sqrt{1-e^2}}\left(\cos E - e\right) \end{bmatrix} \tag{4.21}$$

$$\frac{\partial \underline{r}}{\partial m} = \begin{bmatrix} -\frac{a^2}{r}\sin E \\ \frac{a^2}{r}\sqrt{1-e^2}\cos E \end{bmatrix} \tag{4.22}$$

$$\frac{\partial \dot{\underline{r}}}{\partial a} = \frac{a\,n}{2r}\begin{bmatrix} \sin E - \frac{3a}{r}M\left(\frac{a\,e}{r}\sin^2 E - \cos E\right) \\ \sqrt{1-e^2}\left(\frac{3a}{r}M\sin E\left(\frac{a\,e}{r}\cos E + 1\right) - \cos E\right) \end{bmatrix} \tag{4.23}$$

$$\frac{\partial \dot{\underline{r}}}{\partial e} = \frac{a^2 n}{r}\begin{bmatrix} \frac{a^2}{2r^2}\left(2e\sin E + \sin(2E)\left(e\cos E - 2\right)\right) \\ -\frac{e}{\sqrt{1-e^2}}\cos E + \frac{a}{r}\sqrt{1-e^2}\left(\frac{a}{r}\sin^2 E - \cos^2 E\right) \end{bmatrix} \tag{4.24}$$

$$\frac{\partial \dot{\underline{r}}}{\partial m} = -\frac{n\,a^4}{r^3}\begin{bmatrix} \cos E - e \\ \sqrt{1-e^2}\sin E \end{bmatrix}. \tag{4.25}$$

Considering Eq. (4.11), the differentiation of the matrix $\underline{R}$ with respect to the parameters ω, i, Ω is simple and does not pose any problem. Hence, the differential relations

$$\begin{aligned} d\underline{\varrho} &= \underline{R}\frac{\partial \underline{r}}{\partial a}da + \underline{R}\frac{\partial \underline{r}}{\partial e}de + \underline{R}\frac{\partial \underline{r}}{\partial m}dm + \frac{\partial \underline{R}}{\partial \omega}\underline{r}\,d\omega + \frac{\partial \underline{R}}{\partial i}\underline{r}\,di + \frac{\partial \underline{R}}{\partial \Omega}\underline{r}\,d\Omega \\ d\dot{\underline{\varrho}} &= \underline{R}\frac{\partial \dot{\underline{r}}}{\partial a}da + \underline{R}\frac{\partial \dot{\underline{r}}}{\partial e}de + \underline{R}\frac{\partial \dot{\underline{r}}}{\partial m}dm + \frac{\partial \underline{R}}{\partial \omega}\dot{\underline{r}}\,d\omega + \frac{\partial \underline{R}}{\partial i}\dot{\underline{r}}\,di + \frac{\partial \underline{R}}{\partial \Omega}\dot{\underline{r}}\,d\Omega \end{aligned} \tag{4.26}$$

are obtained. The mean or eccentric anomaly is contained in Eqs. (4.20) through (4.25). Hence, these derivatives with respect to the parameters a, e, m are time dependent. Because of the appearance of $\underline{r}$ and $\dot{\underline{r}}$, all terms in Eq. (4.26) are time dependent although the derivatives of the matrix $\underline{R}$ with respect to the parameters ω, i, Ω are constant.

4.2.2 Perturbed motion

The Keplerian orbit is a theoretical orbit and does not include actual perturbations. Consequently, disturbing accelerations $d\underline{\ddot{\varrho}}$ must be added to Eq. (4.1) which is now expressed in the equatorial system. The perturbed motion, thus, is based on an inhomogeneous differential equation of second order

$$\underline{\ddot{\varrho}} + \frac{\mu}{\varrho^3}\,\underline{\varrho} = d\underline{\ddot{\varrho}}\,. \tag{4.27}$$

One should note that, for GPS satellites, the acceleration $\|\underline{\ddot{\varrho}}\|$ due to the central attractive force μ/ϱ^2 is at least 10^4 times larger than the disturbing accelerations. Hence, for the analytical solution of Eq. (4.27), perturbation theory may be applied where, initially, only the homogeneous part of the equation is considered. This leads to a Keplerian orbit defined by the six parameters p_{i0}, $i = 1, \ldots, 6$ at the reference epoch t_0. Each disturbing acceleration $d\underline{\ddot{\varrho}}$ causes temporal variations $\dot{p}_{i0} = dp_{i0}/dt$ in the orbital parameters. Therefore, at an arbitrary epoch t, the parameters p_i describing the so-called osculating ellipse are given by

$$p_i = p_{i0} + \dot{p}_{i0}\,(t - t_0)\,. \tag{4.28}$$

In order to obtain time derivatives $\dot{p}_{i0}$, the Keplerian motion is compared to the perturbed motion. In the first case, the parameters p_i are constant, whereas in the second case they are time dependent. Thus, for the position and velocity vector of the perturbed motion one may write

$$\begin{aligned} \underline{\varrho} &= \underline{\varrho}\{t, p_i(t)\} \\ \underline{\dot{\varrho}} &= \underline{\dot{\varrho}}\{t, p_i(t)\}\,. \end{aligned} \tag{4.29}$$

Differentiating the above equations with respect to time and taking into account Eq. (4.27) leads to

$$\underline{\dot{\varrho}} = \frac{\partial \underline{\varrho}}{\partial t} + \sum_{i=1}^{6} \left(\frac{\partial \underline{\varrho}}{\partial p_i}\,\frac{dp_i}{dt} \right) \tag{4.30}$$

$$\underline{\ddot{\varrho}} = \frac{\partial \underline{\dot{\varrho}}}{\partial t} + \sum_{i=1}^{6} \left(\frac{\partial \underline{\dot{\varrho}}}{\partial p_i}\,\frac{dp_i}{dt} \right) = -\frac{\mu}{\varrho^3}\,\underline{\varrho} + d\underline{\ddot{\varrho}}\,. \tag{4.31}$$

Since an (osculating) ellipse is defined for any epoch t, the Eqs. (4.30) and (4.31) must also be valid for Keplerian motion. Evidently, equivalence is

obtained with the following conditions

$$\begin{aligned} \sum_{i=1}^{6} \left(\frac{\partial \underline{\varrho}}{\partial p_i} \frac{dp_i}{dt} \right) &= \underline{0} \\ \sum_{i=1}^{6} \left(\frac{\partial \dot{\underline{\varrho}}}{\partial p_i} \frac{dp_i}{dt} \right) &= d\ddot{\underline{\varrho}} . \end{aligned} \tag{4.32}$$

In the following, for simplicity, only one disturbing acceleration is considered. The two vector equations (4.32) correspond to six linear equations which, in vector notation, are given by

$$\underline{A}\,\underline{u} = \underline{\ell} \tag{4.33}$$

where

$$\begin{aligned} \underline{A} &= \begin{bmatrix} \frac{\partial \underline{\varrho}}{\partial a} & \frac{\partial \underline{\varrho}}{\partial e} & \frac{\partial \underline{\varrho}}{\partial m} & \frac{\partial \underline{\varrho}}{\partial \omega} & \frac{\partial \underline{\varrho}}{\partial i} & \frac{\partial \underline{\varrho}}{\partial \Omega} \\ \frac{\partial \dot{\underline{\varrho}}}{\partial a} & \frac{\partial \dot{\underline{\varrho}}}{\partial e} & \frac{\partial \dot{\underline{\varrho}}}{\partial m} & \frac{\partial \dot{\underline{\varrho}}}{\partial \omega} & \frac{\partial \dot{\underline{\varrho}}}{\partial i} & \frac{\partial \dot{\underline{\varrho}}}{\partial \Omega} \end{bmatrix} \\ &= \begin{bmatrix} \underline{R}\frac{\partial \underline{r}}{\partial a} & \underline{R}\frac{\partial \underline{r}}{\partial e} & \underline{R}\frac{\partial \underline{r}}{\partial m} & \underline{r}\frac{\partial \underline{R}}{\partial \omega} & \underline{r}\frac{\partial \underline{R}}{\partial i} & \underline{r}\frac{\partial \underline{R}}{\partial \Omega} \\ \underline{R}\frac{\partial \dot{\underline{r}}}{\partial a} & \underline{R}\frac{\partial \dot{\underline{r}}}{\partial e} & \underline{R}\frac{\partial \dot{\underline{r}}}{\partial m} & \dot{\underline{r}}\frac{\partial \underline{R}}{\partial \omega} & \dot{\underline{r}}\frac{\partial \underline{R}}{\partial i} & \dot{\underline{r}}\frac{\partial \underline{R}}{\partial \Omega} \end{bmatrix} \\ \underline{u} &= \begin{bmatrix} \frac{da}{dt} & \frac{de}{dt} & \frac{dm}{dt} & \frac{d\omega}{dt} & \frac{di}{dt} & \frac{d\Omega}{dt} \end{bmatrix}^T \\ &= \begin{bmatrix} \dot{a} & \dot{e} & \dot{m} & \dot{\omega} & \dot{i} & \dot{\Omega} \end{bmatrix}^T \\ \underline{\ell} &= \begin{bmatrix} \underline{0} \\ d\ddot{\underline{\varrho}} \end{bmatrix} . \end{aligned}$$

The 6×6 matrix $\underline{A}$ requires the derivatives of $\underline{\varrho}$ and $\dot{\underline{\varrho}}$ with respect to the Keplerian parameters which have been developed in the preceding section, cf. Eq. (4.26). The 6×1 vector $\underline{\ell}$ contains the disturbing acceleration. Finally, the six unknown time derivatives appear in the 6×1 vector $\underline{u}$. The inversion of the system (4.33) leads to Lagrange's equations where the disturbing potential R, associated with the disturbing acceleration by $d\ddot{\underline{\varrho}} = \mathrm{grad} R$, has

been introduced (e.g., Beutler 1992: Eq. (6.96)):

$$\begin{aligned}
\dot{a} &= \frac{2}{n\,a}\,\frac{\partial R}{\partial m} \\
\dot{e} &= \frac{1-e^2}{n\,a^2\,e}\,\frac{\partial R}{\partial m} - \frac{\sqrt{1-e^2}}{n\,a^2\,e}\,\frac{\partial R}{\partial \omega} \\
\dot{m} &= -\frac{2}{n\,a}\,\frac{\partial R}{\partial a} - \frac{1-e^2}{n\,a^2\,e}\,\frac{\partial R}{\partial e} \\
\dot{\omega} &= \frac{\sqrt{1-e^2}}{n\,a^2\,e}\,\frac{\partial R}{\partial e} - \frac{\cos i}{n\,a^2\sqrt{1-e^2}\,\sin i}\,\frac{\partial R}{\partial i} \\
\dot{i} &= \frac{\cos i}{n\,a^2\sqrt{1-e^2}\,\sin i}\,\frac{\partial R}{\partial \omega} - \frac{1}{n\,a^2\sqrt{1-e^2}\,\sin i}\,\frac{\partial R}{\partial \Omega} \\
\dot{\Omega} &= \frac{1}{n\,a^2\sqrt{1-e^2}\,\sin i}\,\frac{\partial R}{\partial i}\,.
\end{aligned} \tag{4.34}$$

Note that the system (4.34) fails for $e = 0$ or $i = 0$. This singularity can be avoided by the substitution of auxiliary parameters (Arnold 1970: p. 28).

The Lagrange equations presuppose that the disturbing potential R is expressed in function of the Keplerian parameters. When the acceleration $d\underline{\ddot{\varrho}}$ is represented by components K_i along the axes $\underline{e}_i^*$, the system (4.34) can be transformed using the identity

$$\frac{\partial R}{\partial p_i} = \mathrm{grad}R \cdot \frac{\partial \underline{\varrho}}{\partial p_i} = (K_1\,\underline{e}_1^* + K_2\,\underline{e}_2^* + K_3\,\underline{e}_3^*) \cdot \frac{\partial \underline{\varrho}}{\partial p_i}\,. \tag{4.35}$$

The simple but cumbersome algebra leads to the Gaussian equations. The result is taken from Seeber (1993), Eq. (3.101), where K_1 and K_3 are interchanged and slight modifications are introduced:

$$\begin{aligned}
\dot{a} &= \frac{2}{n\sqrt{1-e^2}}\,[e\,\sin v\,K_1 + (1+e\,\cos v)\,K_2] \\
\dot{e} &= \frac{\sqrt{1-e^2}}{n\,a}\,[\sin v\,K_1 + (\cos E + \cos v)\,K_2] \\
\dot{m} &= \frac{1-e^2}{n\,a\,e}\left[\left(\frac{-2e}{1+e\,\cos v} + \cos v\right) K_1\right. \\
&\quad \left. - \left(1 + \frac{1}{1+e\,\cos v}\right)\,\sin v\,K_2\right]
\end{aligned}$$

and

$$\begin{aligned}
\dot{\omega} &= \frac{\sqrt{1-e^2}}{n\,a\,e}\left[-\cos v\,K_1 + \left(1+\frac{1}{1+e\cos v}\right)\sin v\,K_2\right. \\
&\quad \left. - \frac{e\,\sin(\omega+v)}{1+e\,\cos v}\cot i\,K_3\right] \\
\dot{i} &= \frac{r\,\cos(\omega+v)}{n\,a^2\sqrt{1-e^2}}K_3 \\
\dot{\Omega} &= \frac{r\,\sin(\omega+v)}{n\,a^2\sqrt{1-e^2}\,\sin i}K_3\,.
\end{aligned} \tag{4.36}$$

Note that in the temporal variations $\dot{i}$ and $\dot{\Omega}$ only the component orthogonal to the orbital plane, K_3, appears, whereas the variations of $\dot{a}$, $\dot{e}$, $\dot{m}$ are affected by both components in the orbital plane, K_1 and K_2.

4.2.3 Disturbing accelerations

In reality, many disturbing accelerations act on a satellite and are responsible for the temporal variations of the Keplerian elements. Roughly speaking, they can be divided into two groups, namely those of gravitational and those of nongravitational origin (Table 4.3). Because the GPS satellites are orbiting at an altitude of approximately 20 200 km, the indirect effect of solar radiation pressure as well as air drag may be neglected. On the other hand, the shape (and, thus, the cross section) of the satellites is irregular which renders the modeling of direct solar radiation pressure more difficult. The variety of materials used for the satellites each has a different heat-absorption which results in additional and complicated perturbing accelerations. Also, accelerations may arise from leaks in the container of the gas-propellant as mentioned by Lichten and Neilan (1990).

To demonstrate the effect of disturbing accelerations, an example is computed by assuming a constant disturbance $d\ddot{\varrho} = 10^{-9}\,\mathrm{m\,s^{-2}}$ acting on a GPS satellite. The associated shift in the position of the satellite results from

Table 4.3. Sources for disturbing accelerations

Gravitational	Nonsphericity of the earth
	Tidal attraction (direct and indirect)
Nongravitational	Solar radiation pressure (direct and indirect)
	Air drag
	Relativistic effects
	Others (e.g., solar wind, magnetic field forces)

double integration over time t and yields $d\varrho = (t^2/2)\, d\ddot{\varrho}$. Substituting the numerical value $t = 12$ hours gives the shift after one revolution which is $d\varrho \approx 1\,\mathrm{m}$. This can be considered as typical value.

Nonsphericity of the earth
The earth's potential V can be represented by a spherical harmonic expansion (Heiskanen and Moritz 1967: p. 342) by

$$V = \frac{\mu}{r}\left[1 - \sum_{n=2}^{\infty}\left(\frac{a_E}{r}\right)^n J_n\, P_n(\sin\varphi) - \sum_{n=2}^{\infty}\sum_{m=1}^{n}\left(\frac{a_E}{r}\right)^n \left[J_{nm}\cos m\lambda + K_{nm}\sin m\lambda\right] P_{nm}(\sin\varphi)\right] \tag{4.37}$$

where a_E is the semimajor axis of the earth, r is the geocentric distance of the satellite, and φ, λ are its latitude and longitude. The J_n, J_{nm}, K_{nm} denote the zonal and tesseral coefficients of the harmonic development known from an earth model. Finally, P_n are the Legendre polynomials and P_{nm} are the associated Legendre functions.

The first term on the right side of Eq. (4.37), μ/r, represents the potential V_0 for a spherical earth, and its gradient, $\mathrm{grad}(\mu/r) = (\mu/r^3)\,\underline{r}$, was considered as the central force for the Keplerian motion. Hence, the disturbing potential R is given by the difference

$$R = V - V_0\,. \tag{4.38}$$

It is shown below that the disturbing acceleration due to J_2, the term representing the oblateness, is smaller by a factor of 10^4 than the acceleration due to V_0. On the other hand, the oblateness term is approximately three orders of magnitude larger than any other coefficient.

A numerical assessment of the central acceleration of GPS satellites gives $\|\ddot{\underline{r}}\| = \mu/r^2 \approx 0.57\,\mathrm{m\,s^{-2}}$. The acceleration corresponding to the oblateness term in the disturbing potential R is given by $\|d\ddot{\underline{r}}\| \approx \|\partial R/\partial r\| = 3\mu\,(a_E/r^2)^2\, J_2\, P_2(\sin\varphi)$. The latitude of a satellite can only reach $\varphi = 55°$, the value of its orbital inclination. The maximum of the function $P_2(\sin\varphi) = \frac{1}{2}\,(3\sin\varphi^2 - 1)$, therefore, becomes 0.5. Finally, with $J_2 \approx 1.1 \cdot 10^{-3}$, the numerical value $\|d\ddot{\underline{r}}\| \approx 5 \cdot 10^{-5}\,\mathrm{m\,s^{-2}}$ is obtained.

In the early days of GPS, a subset of earth model coefficients complete up to degree and order eight was considered sufficient for GPS satellite arcs of a few revolutions. Associated with the revised WGS-84 (G873) is the Earth Gravitational Model 1996 (EGM96) and it is recommended to use

a subset of coefficients up to degree and order 70 for high accuracy orbit determination (National Imagery and Mapping Agency 1997).

Tidal effects

Consider a celestial body with pointmass m_b and the geocentric position vector $\underline{\varrho}_b$ (Fig. 4.2). Note that the geocentric angle z between the celestial body and the satellite can be expressed as a function of $\underline{\varrho}_b$ and $\underline{\varrho}$, the latter denoting the geocentric position vector of the satellite, by

$$\cos z = \frac{\underline{\varrho}_b}{\|\underline{\varrho}_b\|} \cdot \frac{\underline{\varrho}}{\|\underline{\varrho}\|} . \tag{4.39}$$

The additional mass exerts an acceleration with respect to the earth as well as with respect to the satellite. For the perturbed motion of the satellite around the earth, only the difference of the two accelerations is relevant; consequently, the disturbing acceleration is given by

$$d\ddot{\underline{\varrho}} = G\, m_b \left[\frac{\underline{\varrho}_b - \underline{\varrho}}{\|\underline{\varrho}_b - \underline{\varrho}\|^3} - \frac{\underline{\varrho}_b}{\|\underline{\varrho}_b\|^3} \right] . \tag{4.40}$$

Among all the celestial bodies in the solar system, only the sun and the moon must be considered because the effects of the planets are negligible. The geocentric position vector of the sun and the moon are obtained by evaluating known analytical expressions for their motion.

The maximum of the perturbing acceleration is reached when the three bodies in Fig. 4.2 are situated in a straight line. In this case, Eq. (4.40) reduces to $\|d\ddot{\underline{\varrho}}\| = G\, m_b\, (1/\|\underline{\varrho}_b - \underline{\varrho}\|^2 - 1/\|\underline{\varrho}_b\|^2)$. For a numerical assessment, the corresponding numerical values for the sun ($G\, m_b \approx 1.3 \cdot 10^{20}\,\mathrm{m^3\,s^{-2}}$, $\varrho_b \approx 1.5 \cdot 10^{11}\,\mathrm{m}$) and of the moon ($G\, m_b \approx 4.9 \cdot 10^{12}\,\mathrm{m^3\,s^{-2}}$, $\varrho_b \approx 3.8 \cdot 10^{8}\,\mathrm{m}$) are substituted. The resulting numerical values for the perturbing acceleration are $2 \cdot 10^{-6}\,\mathrm{m\,s^{-2}}$ for the sun and $5 \cdot 10^{-6}\,\mathrm{m\,s^{-2}}$ for the moon.

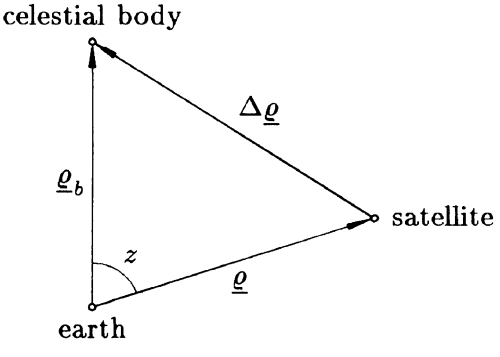


Fig. 4.2. Three-body problem

Apart from the direct effect of the tide-generating bodies, indirect effects due to the tidal deformation of the solid earth and the oceanic tides must be taken into account. Considering only the tidal potential W_2 of second degree, the disturbing potential R due to the tidal deformation of the solid earth (Melchior 1978) is given by

$$R = k \left(\frac{a_E}{\varrho}\right)^3 W_2 = \frac{1}{2} k G m_b \frac{{a_E}^5}{(\varrho\, \varrho_b)^3} (3\cos^2 z - 1) \tag{4.41}$$

with $k \approx 0.3$ being one of the Love numbers. The associated acceleration of the satellite is in the order of $10^{-9}\,\mathrm{m\,s^{-2}}$ as the reader may verify.

The model for the indirect effect due to the oceanic tides is more complicated. Tidal charts with the distribution of the oceanic tides are required. In addition, loading coefficients are needed. These coefficients describe the response of the solid earth to the load of the oceanic water masses. The perturbing acceleration is again in the order of $10^{-9}\,\mathrm{m\,s^{-2}}$.

As a consequence of the tidal deformation and the oceanic loading, the geocentric position vector $\underline{\varrho}_R$ of an observing site varies with time. This variation must be taken into account when modeling receiver dependent biases in the observation equations (e.g., McCarthy 1996).

Solar radiation pressure

Following Fliegel et al. (1985), the perturbing acceleration due to the direct solar radiation pressure has two components. The principal component $d\ddot{\underline{\varrho}}_1$ is directed away from the sun and the smaller component $d\ddot{\underline{\varrho}}_2$ acts along the satellite's y-axis. This is an axis orthogonal to both the vector pointing to the sun and the antenna which is nominally directed towards the center of the earth.

The principal component is usually modeled by

$$d\ddot{\underline{\varrho}}_1 = \nu\, K\, \varrho_S^2 \frac{\underline{\varrho} - \underline{\varrho}_S}{\|\underline{\varrho} - \underline{\varrho}_S\|^3} \tag{4.42}$$

where $\underline{\varrho}_S$ denotes the geocentric position vector of the sun. The factor K depends linearly on the solar radiation term, a factor defining the reflective properties of the satellite, and the area-to-mass ratio of the satellite. The quantity ν is an eclipse factor which is zero when the satellite is in the earth's shadow. This occurs twice a year for each satellite when the sun is in or near the orbital plane. Such an eclipse lasts approximately one hour. The eclipse factor equals one when the satellite is in sunlight and for the penumbra regions the relation $0 < \nu < 1$ holds.

The magnitude of $d\ddot{\underline{\varrho}}_1$ is in the order of $10^{-7}\,\mathrm{m\,s^{-2}}$. Hence, an accurate model for the factors K and ν is required even for short arcs. The modeling

is extremely difficult since the solar radiation term varies unpredictably over the year and a single factor for the reflective properties is not adequate for the satellite. Although the mass in orbit is usually well-known, the irregular shape of the satellites does not allow for an exact determination of the area-to-mass ratio. A further problem is the modeling of the earth's penumbra and the assignment of an eclipse factor, particularly in the transition zone between illumination and shadow. Some models are given in Landau (1988).

The component $d\ddot{\underline{\varrho}}_2$ is often called y-bias and is believed to be caused by a combination of misalignments of the solar panels and thermal radiation along the y-axis. Since the magnitude of this bias can remain constant for several weeks, it is usually introduced as an unknown parameter which is determined in the course of the orbit determination. Note that this bias is two orders of magnitudes smaller than the principal term.

That portion of the solar radiation pressure which is reflected back from the earth's surface causes an effect called albedo. In the case of GPS, the associated perturbing accelerations are smaller than the y-bias and can be neglected.

Relativistic effect
The relativistic effect on the satellite orbit is caused by the gravity field of the earth and gives rise to a perturbing acceleration which is (simplified) given (Beutler 1991: Eq. (2.5)) by

$$d\ddot{\underline{\varrho}} = -\frac{3\mu^2 a(1-e^2)}{c^2}\frac{\underline{\varrho}}{\varrho^5} \tag{4.43}$$

where c denotes the velocity of light. Numerically assessed, the perturbing acceleration results in an order of $3 \cdot 10^{-10}\,\mathrm{m\,s^{-2}}$ (Zhu and Groten 1988). This effect is smaller than the indirect effects by one order of magnitude and is mentioned for the sake of completeness.

4.3 Orbit determination

Here, orbit determination essentially means the determination of orbital parameters and satellite clock biases. In principle, the problem is inverse to the navigational or surveying goal. In the fundamental equation for the range ϱ or the range rate $\dot{\varrho}$ between the observing site R and the satellite S,

$$\varrho = \|\underline{\varrho}^S - \underline{\varrho}_R\| \tag{4.44}$$

$$\dot{\varrho} = \frac{(\underline{\varrho}^S - \underline{\varrho}_R)}{\|\underline{\varrho}^S - \underline{\varrho}_R\|} \cdot \dot{\underline{\varrho}}^S\,, \tag{4.45}$$

the position vector $\underline{\varrho}^S$ and the velocity vector $\dot{\underline{\varrho}}^S$ of the satellite are considered unknown, whereas the position vector $\underline{\varrho}_R$ of the (stationary) observing site is assumed to be known in a geocentric system. Note that the superscript indicates a vector related to a satellite.

The ranges in Eq. (4.44) are obtained with high precision as outlined in Sect. 6.1. This is particularly true for delta range data since biases are eliminated by differencing the ranges. The range rates in Eq. (4.45) are less accurate and are derived from frequency shifts due to the Doppler effect. At present, the observations for the orbit determination are performed at terrestrial sites, but GPS data could also be obtained from orbiting receivers such as the TOPEX/Poseidon mission (Lichten and Neilan 1990).

In the following, the satellite clock biases and other parameters are neglected. The actual orbit determination which is performed in two steps is emphasized. First, a Kepler ellipse is fitted to the observations. In the second step, this ellipse serves as reference for the subsequent improvement of the orbit by taking into account perturbing accelerations.

4.3.1 Keplerian orbit

For the moment, it is assumed that both the position and the velocity vector of the satellite have been derived from observations. Now, the question arises of how to use these data for the derivation of the Keplerian parameters.

The position and velocity vector given at the same epoch t define an initial value problem, and two position vectors at different epochs t_1 and t_2 define a (first) boundary value problem. In principle, a second and a third boundary value problem could also be defined; however, these problems are not of practical importance in the context of GPS and are not treated here.

Initial value problem
As stated previously, the derivation of the Keplerian parameters from position and velocity vectors, both given at the same epoch and expressed in an equatorial system such as $\underline{X}_i$, is an initial value problem for solving the differential equation (4.1). Recall that the two given vectors contain six components which allow for the calculation of the six Keplerian parameters. Since both vectors are given at the same epoch, the time parameter is omitted.

The solution corresponds to a transformation inverse to Eq. (4.10) and makes use of the fact that quantities like distances or angles are invariant

with respect to rotation. Hence, the following equations are obtained:

$$\begin{aligned} \|\underline{\varrho}^S\| &= \|\underline{r}\| \\ \|\dot{\underline{\varrho}}^S\| &= \|\dot{\underline{r}}\| \\ \underline{\varrho}^S \cdot \dot{\underline{\varrho}}^S &= \underline{r} \cdot \dot{\underline{r}} \\ \|\underline{\varrho}^S \times \dot{\underline{\varrho}}^S\| &= \|\underline{r} \times \dot{\underline{r}}\| \,. \end{aligned} \tag{4.46}$$

In addition, by substituting Eqs. (4.6) and (4.8), the relations

$$\underline{\varrho}^S \cdot \dot{\underline{\varrho}}^S = \sqrt{\mu a}\,(e \sin E) \tag{4.47}$$

$$\|\underline{\varrho}^S \times \dot{\underline{\varrho}}^S\| = \sqrt{\mu a\,(1 - e^2)} \tag{4.48}$$

can be derived.

Now, the inverse transformation can be solved as follows. First, the geocentric distance r and the velocity $\dot{r}$ are calculated from the given $\underline{\varrho}^S$ and $\dot{\underline{\varrho}}^S$. Based on these two quantities, the semimajor axis a follows from Eq. (4.9). With a and r determined, $e \cos E$ can be established using Eq. (4.7) and $e \sin E$ can be established using Eq. (4.47). Hence, the eccentricity e and the eccentric anomaly E, and consequently the mean and true anomalies M and v, can be calculated. The vector product of $\underline{\varrho}^S$ and $\dot{\underline{\varrho}}^S$ is equivalent to the vector of angular momentum and is directed orthogonal to the orbital plane. Therefore, this vector is, after normalization, identical to the vector $\underline{e}_3$ in Eq. (4.11) from which the parameters i and ℓ can be deduced. According to Eq. (4.48), the norm of the vector product allows for a check of the previously calculated parameters a and e. For the determination of ω, the unit vector $\underline{k} = [\cos \ell, \sin \ell, 0]^T$ directed to the ascending node is defined. From Fig. 4.1 one can obtain the relations $\underline{\varrho}^S \cdot \underline{k} = r \cos(\omega + v)$ and $\underline{\varrho}^S \cdot \underline{X}_3 = r \sin i \sin(\omega + v)$. The two equations can be uniquely solved for ω, since r, v, i are known.

For a numerical example start with a position vector $\underline{\varrho}^S$ and the corresponding velocity vector $\dot{\underline{\varrho}}^S$

$$\begin{aligned} \underline{\varrho}^S &= [11\,465,\ 3\,818,\ -20\,923]^T \qquad [\mathrm{km}] \\ \dot{\underline{\varrho}}^S &= [-1.2651,\ 3.9960,\ -0.3081]^T \quad [\mathrm{km\,s^{-1}}] \end{aligned}$$

to determine the Keplerian parameters. This is a task inverse to the numerical example in Sect. 4.2.1 and the result can be read from there.

Boundary value problem

Now, it is assumed that two position vectors $\underline{\varrho}^S(t_1)$ and $\underline{\varrho}^S(t_2)$ at epochs t_1 and t_2 are available. Note that position vectors are preferred for orbit determination since they are more accurate than velocity vectors. The given data correspond to boundary values in the solution of the basic second-order differential equation, cf. Eq. (4.1).

An approximate method for the derivation of the Keplerian parameters makes use of initial values defined for an averaged epoch $t = \frac{1}{2}(t_1 + t_2)$:

$$\begin{aligned} \underline{\varrho}^S(t) &= \frac{\underline{\varrho}^S(t_2) + \underline{\varrho}^S(t_1)}{2} \\ \dot{\underline{\varrho}}^S(t) &= \frac{\underline{\varrho}^S(t_2) - \underline{\varrho}^S(t_1)}{t_2 - t_1} \,. \end{aligned} \tag{4.49}$$

The rigorous solution starts with the computation of the geocentric distances

$$\begin{aligned} r_1 &= r(t_1) = \|\underline{\varrho}^S(t_1)\| \\ r_2 &= r(t_2) = \|\underline{\varrho}^S(t_2)\| \,. \end{aligned} \tag{4.50}$$

The unit vector $\underline{e}_3$, orthogonal to the orbital plane, is obtained from a vector product by

$$\underline{e}_3 = \frac{\underline{\varrho}^S(t_1) \times \underline{\varrho}^S(t_2)}{\|\underline{\varrho}^S(t_1) \times \underline{\varrho}^S(t_2)\|} \tag{4.51}$$

and produces the longitude ℓ and the inclination angle i, cf. Eqs. (4.11) and (4.13). As demonstrated earlier, the argument of the latitude $u = \omega + v$ is defined as the angle between the satellite position and the ascending node vector $\underline{k} = [\cos\ell,\ \sin\ell,\ 0]^T$. Hence, the relation

$$r_i \cos u_i = \underline{k} \cdot \underline{\varrho}^S(t_i) \,, \qquad i = 1, 2 \tag{4.52}$$

holds from which the u_i with $u_2 > u_1$ can be deduced uniquely. Now, there are two equations, cf. Eq. (4.7),

$$r_i = \frac{a(1 - e^2)}{1 + e\cos(u_i - \omega)} \,, \qquad i = 1, 2 \tag{4.53}$$

where the parameters a, e, ω are unknown. The system can be solved for a and e after assigning a preliminary value such as the nominal one to ω, the argument of the perigee. Based on the assumed ω and the u_i, the true anomalies v_i and subsequently the mean anomalies M_i are obtained. Therefore, the mean angular velocity n can be calculated twice by the formulas

$$n = \sqrt{\frac{\mu}{a^3}} = \frac{M_2 - M_1}{t_2 - t_1} \,, \tag{4.54}$$

cf. Eqs. (4.2) and (4.3). The equivalence is achieved by varying ω. This iterative procedure is typical for boundary value problems. Finally, the epoch of perigee passage T_0 follows from the relation

$$T_0 = t_i - \frac{M_i}{n} . \tag{4.55}$$

For a numerical solution of the boundary value problem assume two position vectors $\underline{\varrho}^S(t_1)$ and $\underline{\varrho}^S(t_2)$, both represented in the earth-fixed equatorial system $\underline{X}_i$ and with $\Delta t = t_2 - t_1 = 1$ hour:

$$\underline{\varrho}^S(t_1) = [11\,465,\ \ 3\,818, -20\,923]^T \quad [\text{km}]$$

$$\underline{\varrho}^S(t_2) = [\ 5\,220, 16\,754, -18\,421]^T \quad [\text{km}] .$$

The application of Eqs. (4.50) through (4.55) results, apart from rounding errors, in the following set of parameters for the associated Kepler ellipse: $a = 26\,000\,\text{km}$, $e = 0.1$, $\omega = -140°$, $i = 60°$, $\ell = 110°$, and $T_0 = t_1 - 1^{\text{h}}.3183$.

Orbit improvement

If there are redundant observations, the parameters of an instantaneous Kepler ellipse can be improved. The position vector $\underline{\varrho}_0^S$ associated with the reference ellipse can be computed. Each observed range, for example, gives rise to an equation

$$\varrho = \varrho_0 + d\varrho = \|\underline{\varrho}_0^S - \underline{\varrho}_R\| + \frac{\underline{\varrho}_0^S - \underline{\varrho}_R}{\|\underline{\varrho}_0^S - \underline{\varrho}_R\|} \cdot d\underline{\varrho}^S . \tag{4.56}$$

The vector $d\underline{\varrho}^S$ can be expressed as a function of the Keplerian parameters, cf. Eq. (4.26). Thus, Eq. (4.56) actually contains the differential increments for the six orbital parameters.

In the past, orbit improvement was often performed in the course of GPS data processing when, in addition to terrestrial position vectors $\underline{\varrho}_R$, the increments dp_{0i} were determined. The procedure became unstable or even failed for small networks. In the case of orbit relaxation, only three degrees of freedom were assigned to the orbit.

4.3.2 Perturbed orbit

Analytical solution

As known from previous sections, the perturbed motion is characterized by temporal variations of the orbital parameters. The analytical expressions for these variations are given by Eqs. (4.34) or (4.36).

In order to be suitable for Lagrange's equations, the disturbing potential must be expressed as a function of the Keplerian parameters. Kaula (1966)

was the first who has performed this transformation for the earth potential. The resulting relation for the disturbing potential is

$$R = \sum_{n=2}^{\infty} \sum_{m=0}^{n} A_n(a) \sum_{p=0}^{n} F_{nmp}(i) \cdot \sum_{q=-\infty}^{\infty} G_{npq}(e)\, S_{nmpq}(\omega, \Omega, M; \Theta_0, J_{nm}, K_{nm})\,. \tag{4.57}$$

Recall that n denotes the degree and m the order of the spherical harmonics in the disturbing potential development. Each of the functions A_n, F_{nmp}, G_{npq} contains only one parameter of a Kepler ellipse; however, the function S_{nmpq} is composed of several parameters and can be expressed by

$$\begin{aligned} S_{nmpq} &= J_{nm} \cos\psi + K_{nm} \sin\psi \qquad \text{for } n-m \text{ even} \\ S_{nmpq} &= -K_{nm} \cos\psi + J_{nm} \sin\psi \qquad \text{for } n-m \text{ odd}\,. \end{aligned} \tag{4.58}$$

The frequency $\dot{\psi}$ associated to the argument ψ is given by

$$\dot{\psi} = (n-2p)\,\dot{\omega} + (n-2p+q)\,\dot{M} + m\,(\dot{\Omega} - \dot{\Theta}_0) \tag{4.59}$$

and is a measure of the spectrum of the perturbations.

The conditions $(n-2p) = (n-2p+q) = m = 0$ lead to $\dot{\psi} = 0$ and, thus, to secular variations. Because of $m = 0$, they are caused by zonal harmonics. If $(n-2p) \neq 0$, then the variations depend on $\dot{\omega}$ and are, therefore, generally long-periodic. Finally, the conditions $(n-2p+q) \neq 0$ and/or $m \neq 0$ result in short-periodic variations. The integer value $(n-2p+q)$ gives the frequency in cycles per revolution and m the frequency in cycles per (sidereal) day.

A rough overview of the frequency spectrum of the Keplerian parameters due to the gravity field of the earth is given in Table 4.4. Summarizing, one can state that the even degree zonal coefficients produce primarily secular variations and the odd degree zonal coefficients give rise to long-periodic perturbations. The tesseral coefficients are responsible for short-periodic terms. From Table 4.4 one can see that short-periodic variations occur in each parameter. With the exception of the semimajor axis, the parameters are also affected by long-periodic perturbations. Secular effects are only contained in Ω, ω, M. The analytical expression for the secular variations

Table 4.4. Perturbations due to the earth's gravity field

Parameter	Secular	Long-periodic	Short-periodic
a	no	no	yes
e	no	yes	yes
i	no	yes	yes
Ω	yes	yes	yes
ω	yes	yes	yes
M	yes	yes	yes

of these parameters due to the oblateness term J_2 is given as an example:

$$\begin{aligned}
\dot{\Omega} &= -\frac{3}{2}\, n\, a_E^2\, \frac{\cos i}{a^2\,(1-e^2)^2}\, J_2 \\
\dot{\omega} &= \frac{3}{4}\, n\, a_E^2\, \frac{5\cos^2 i - 1}{a^2\,(1-e^2)^2}\, J_2 \\
\dot{m} &= \frac{3}{4}\, n\, a_E^2\, \frac{3\cos^2 i - 1}{a^2\,\sqrt{(1-e^2)^3}}\, J_2 \,.
\end{aligned} \tag{4.60}$$

The first equation describes the regression of the node in the equatorial plane, the second equation expresses the rotation of the perigee, and the third equation contributes to the variation of the mean anomaly by $\dot{M} = n + \dot{m}$. In the case of GPS satellites, the numerical values $\dot{\Omega} \approx -0.^\circ 03$ per day, $\dot{\omega} \approx 0.^\circ 01$ per day, and $\dot{m} \approx 0$ are obtained. The result for $\dot{m}$ is verified immediately since the term $3\cos^2 i - 1$ becomes approximately zero for the nominal inclination $i = 55°$. A graphical representation of the secular perturbations is given in Fig. 4.3. Special attention must be paid to resonance effects which occur when the period of revolution corresponds to a harmonic in the gravity potential. This is why the GPS satellites have been raised into slightly higher orbits to preclude an orbital period very close to half a sidereal day.

The tidal potential also has a harmonic representation and, thus, the tidal perturbations can be analytically modeled. This was performed first by Kozai (1959) and, analogous to the earth's potential effect, has led to analytical expressions for the secular variations of the node's right ascension Ω and of the perigee's argument ω. The reader is referred to the publication of Kozai (1959) for formulas specified.

Numerical solution

If the disturbing acceleration cannot be expressed in analytical form, one has to apply numerical methods for the solution. Therefore, in principle, with

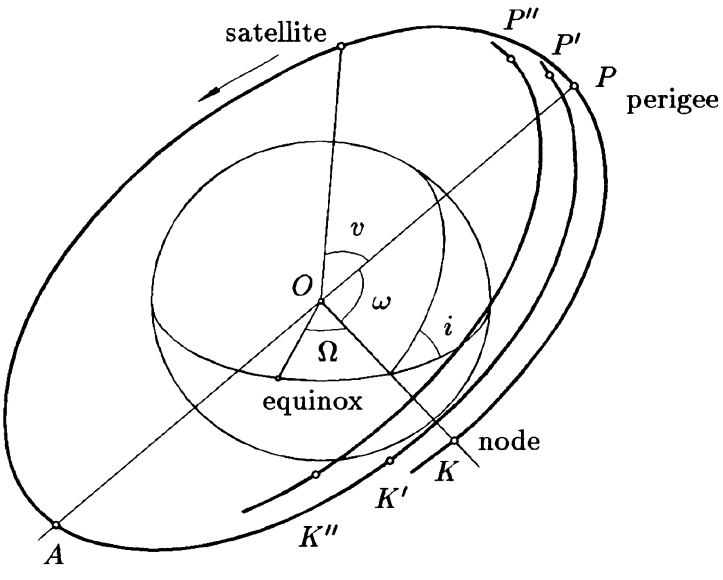


Fig. 4.3. Secular perturbations caused by the oblateness term J_2

initial values such as the position and velocity vectors $\underline{\varrho}(t_0)$ and $\dot{\underline{\varrho}}(t_0)$ at a reference epoch t_0, a numerical integration of Eq. (4.27) could be performed. This simple concept can be improved by the introduction of a Kepler ellipse as a reference. By this means only the smaller difference between the total and the central acceleration must be integrated. The integration results in an increment $\Delta\underline{\varrho}$ which, when added to the position vector computed for the reference ellipse, gives the actual position vector.

The second-order differential equation is usually transformed to a system of two differential equations of the first order. This system is given by

$$\begin{aligned}
\dot{\underline{\varrho}}(t) &= \dot{\underline{\varrho}}(t_0) + \int_{t_0}^{t} \ddot{\underline{\varrho}}(t_0)\, dt \\
&= \dot{\underline{\varrho}}(t_0) + \int_{t_0}^{t} \left[d\ddot{\underline{\varrho}}(t_0) - \frac{\mu}{\varrho^3(t_0)}\, \underline{\varrho}(t_0) \right] dt \qquad (4.61) \\
\underline{\varrho}(t) &= \underline{\varrho}(t_0) + \int_{t_0}^{t} \dot{\underline{\varrho}}(t_0)\, dt\,.
\end{aligned}$$

The numerical integration of this coupled system can be performed by applying the standard Runge–Kutta algorithm (e.g., Kreyszig 1968: p. 89). This

method is briefly outlined. Let $y(x)$ be a function defined in the interval $x_1 \leq x \leq x_2$ and denote by $y' = dy/dx$ its first derivative with respect to the argument x. The general solution of the ordinary differential equation of the first order

$$y' = \frac{dy}{dx} = y'(y, x) \tag{4.62}$$

follows from integration. The particular solution is found after assigning the given numerical initial value $y_1 = y(x_1)$ to the integration constant. For the application of numerical integration, first the integration interval is partitioned into n equal and sufficiently small $\Delta x = (x_2 - x_1)/n$ where n is an arbitrary integer number > 0. Then, the difference between successive functional values is obtained by the weighted mean

$$\begin{aligned} \Delta y &= y(x + \Delta x) - y(x) \\ &= \tfrac{1}{6}\left[\Delta y^{(1)} + 2\left(\Delta y^{(2)} + \Delta y^{(3)}\right) + \Delta y^{(4)}\right] \end{aligned} \tag{4.63}$$

where

$$\begin{aligned} \Delta y^{(1)} &= y'(y,\ x)\,\Delta x \\ \Delta y^{(2)} &= y'(y + \frac{\Delta y^{(1)}}{2},\quad x + \frac{\Delta x}{2})\,\Delta x \\ \Delta y^{(3)} &= y'(y + \frac{\Delta y^{(2)}}{2},\quad x + \frac{\Delta x}{2})\,\Delta x \\ \Delta y^{(4)} &= y'(y + \Delta y^{(3)},\quad x + \Delta x\,)\,\Delta x\,. \end{aligned}$$

Hence, starting with the initial value y_1 for the argument x_1, the function can be calculated for the successive argument $x_1 + \Delta x$ and so on.

Numerical methods can also be applied to the integration of the disturbing equations of Lagrange or Gauss. These equations have the advantage of being differentials of the first order and, therefore, must only be integrated once over time.

This section concludes with an example for numerical integration. Assume the ordinary first-order differential equation $y' = y - x + 1$ with the initial value $y_1 = 1$ for $x_1 = 0$. The differential equation should be solved for the argument $x_2 = 1$ with increments $\Delta x = 0.5$. Starting with the initial values for the first interval, successively the values $\Delta y^{(1)} = 1.000$, $\Delta y^{(2)} = 1.125$, $\Delta y^{(3)} = 1.156$, and $\Delta y^{(4)} = 1.328$ are obtained. The corresponding weighted mean is $\Delta y = 1.148$. Replacing the initial values in the second interval by $x = x_1 + \Delta x = 0.5$ and $y = y_1 + \Delta y = 2.148$ and proceeding as before yields $\Delta y = 1.569$ and, thus, the final result $y_2 = y_1 + \Sigma\,\Delta y = 3.717$. Note that

the function $y = e^x + x$, satisfying the differential equation, gives the true value $y_2 = 3.718$. With an increment $\Delta x = 0.1$ the numerical integration would provide an accuracy in the order of 10^{-6}.

4.4 Orbit dissemination

4.4.1 Tracking networks

Objectives and strategies
The official orbit determination for GPS satellites is based on observations at the five monitor stations of the control segment. The broadcast ephemerides for Block I satellites with cesium clocks were accurate to about 5 m (assuming three uploads per day). For the Block II satellites, the accuracy is in the order of about 1 m (Conley and Lavrakas 1999). However, an orbital accuracy of about 20 cm is required for specific missions such as TOPEX/Poseidon or for investigations which require an accuracy at the level of 10^{-9}. Moreover, the geodetic community should become independent from the vagaries of the DoD. Thus, the need for civil tracking networks for orbit determination is evident.

The minimum number of sites in a global network is six (Lichten and Neilan 1990) if a configuration is desired where at least two satellites can be tracked simultaneously any time from two sites. Global networks result in higher accuracy and reliability of the orbits compared to those determined from regional networks. The tie of the orbital system to terrestrial reference frames is achieved by the collocation of GPS receivers with VLBI and SLR trackers. The distribution of the GPS sites is essential to achieve the highest accuracy. Two different approaches may be compared. In the first case, the sites are regularly distributed around the globe; in the second case, each network site is surrounded by a cluster of additional points to facilitate ambiguity resolution and, thus, strengthen the solution of the orbital parameters by a factor of three to five (Counselman and Abbot 1989).

Examples for global networks
Apart from the GPS control segment, several networks have been established for orbit determination. Some networks are of regional or even continental size such as the Australian GPS orbit determination network. Subsequently, some examples of global networks are given.

The first Global Orbit Tracking Experiment (GOTEX) was performed in the fall of 1988 and aimed at the collocation of GPS data at existing VLBI and SLR sites. The data should permit an accurate tie between the WGS-84 and the VLBI/SLR systems. About 25 sites, distributed worldwide, were

occupied during the three weeks of the campaign.

The Cooperative International GPS Network (CIGNET) was operated by the U.S. National Geodetic Survey (NGS) with tracking stations located at VLBI sites. The service started in 1988 with 8 stations in North America, Europe and Japan. In 1991, already 20 globally distributed stations were participating (Chin 1991). In the following three years, 30 more stations were incorporated into the network. The charter of CIGNET was limited to the collection and distribution of the raw tracking data and there was no intention to provide global orbits.

In 1990, the International Association of Geodesy (IAG) decided to install an International GPS Service for Geodynamics (IGS) (Mueller 1991). After a test campaign, routine activities started on January 1, 1994. The main purpose of this service is dedicated to orbit determination for geodynamic applications which require the highest accuracy. The IGS is headed by the Central Bureau located at the Jet Propulsion Laboratory (JPL) in the U.S.A. In 1999, the tracking network (Fig. 4.4) was based on more than 220 globally distributed GPS tracking sites with coordinates (and velocities) related to the ITRF. The IGS stations collect code ranges and carrier phases from all satellites in view using dual frequency receivers. The data are ana-

Fig. 4.4. IGS tracking network (simplified)

lyzed independently by seven agencies and are daily archived in receiver and software independent exchange format (RINEX and SINEX) by global and regional data centers. Today, the IGS routinely provides high-quality orbits for all GPS satellites. Predicted orbits with an accuracy of about 1 m are available in almost real time, whereas postprocessed orbits have a delay of 2 days (rapid solution) or two weeks (final solution). The accuracy of the final solution is estimated to be ±5 cm. Also, the raw tracking data, the satellite clock parameters, the earth orientation parameters (EOP), and other data like ionospheric and tropospheric information are available through the service. Associated to the IGS is an information service which can be reached under the address http://igscb.jpl.nasa.gov. Further details on the IGS are given in Gurtner (1995), Beutler (1996a), Neilan and Moore (1999).

4.4.2 Ephemerides

Three sets of data are available to determine position and velocity vectors of the satellites in a terrestrial reference frame at any instant: almanac data, broadcast ephemerides, and precise ephemerides. The data differ in accuracy (Table 4.5) and are either available in real time or after the fact.

Almanac data
The purpose of the almanac data is to provide the user with less precise data to facilitate receiver satellite search or for planning tasks such as the computation of visibility charts. The almanac data are updated at least every six days and are broadcast as part of the satellite message (Sect. 5.1.2). The almanac message essentially contains parameters for the orbit and satellite clock correction terms for all satellites (Table 4.6). All angles are expressed in semicircles. The parameter ℓ_0 denotes the difference between the node's right ascension at epoch t_a and the Greenwich sidereal time at t_0, the beginning of the current GPS week. The reduction of the Keplerian parameters

Table 4.5. Uncertainties of ephemerides

Ephemerides	Uncertainty	Remark
Almanac	Some kilometers	Depending on the age of data
Broadcast	1 m	Or even better
Precise	0.05–0.20 m	Depending on the delay

Table 4.6. Almanac data

Parameter	Explanation
ID	Satellite PRN number
WEEK	Current GPS week
t_a	Reference epoch in seconds within the current week
$\sqrt{a}$	Square root of semimajor axis in $\sqrt{\text{meter}}$
e	Eccentricity
M_0	Mean anomaly at reference epoch
ω	Argument of perigee
δi	Inclination offset from 0.3 semicircles ($\hat{=}54°$)
ℓ_0	Longitude of the node at weekly epoch
$\dot{\Omega}$	Drift of node's right ascension per second
a_0	Satellite clock offset in seconds
a_1	Satellite clock drift

to the observation epoch t is obtained by the formulas

$$\begin{aligned} M &= M_0 + n\,(t - t_a) \\ i &= 54° + \delta i \\ \ell &= \ell_0 + \dot{\Omega}\,(t - t_a) - \omega_E\,(t - t_0) \end{aligned} \tag{4.64}$$

where $\omega_E = 7\,292\,115.1467 \cdot 10^{-11}\,\text{rad}\,\text{s}^{-1}$ is the (untruncated) angular velocity of the earth. The other three Keplerian parameters a, e, ω remain unchanged. Note that in the formula for ℓ in Eq. (4.64) the second term on the right side of the equation considers the node's regression and the third term expresses the uniform change in the sidereal time since epoch t_0. An estimate for the satellite clock bias is given by

$$\delta^S = a_0 + a_1\,(t - t_a)\,. \tag{4.65}$$

Almanac data are also available from a variety of information services such as NIS (Table 2.2). Mostly, the data are packed into the YUMA or the SEM format.

Broadcast ephemerides

The broadcast ephemerides are based on observations at the monitor stations of the GPS control segment. The most recent of these data are used

Table 4.7. Broadcast ephemerides

Parameter	Explanation
ID	Satellite PRN number
WEEK	Current GPS week
t_e	Ephemerides reference epoch
$\sqrt{a}$	Square root of semimajor axis in $\sqrt{\text{meter}}$
e	Eccentricity
M_0	Mean anomaly at reference epoch
ω_0	Argument of perigee
i_0	Inclination
ℓ_0	Longitude of the node at weekly epoch
Δn	Mean motion difference
$\dot{i}$	Rate of inclination angle
$\dot{\Omega}$	Rate of node's right ascension
C_{uc}, C_{us}	Correction coefficients (argument of perigee)
C_{rc}, C_{rs}	Correction coefficients (geocentric distance)
C_{ic}, C_{is}	Correction coefficients (inclination)
t_c	Satellite clock reference epoch
a_0	Satellite clock offset
a_1	Satellite clock drift
a_2	Satellite clock frequency drift

to compute a reference orbit for the satellites. Additional tracking data are entered into a Kalman filter and the improved orbits are used for extrapolation. The Master Control Station is responsible for the computation of the ephemerides and the upload to the satellites.

The broadcast ephemerides are part of the satellite message. Essentially, the ephemerides contain records with general information, records with orbital information, and records with information on the satellite clock (Table 4.7). The parameters in the block of orbital information are the reference epoch, six parameters to describe a Kepler ellipse at the reference epoch, three secular correction terms and six periodic correction terms. The correction terms consider perturbation effects due to the nonsphericity of the earth, the direct tidal effect, and the solar radiation pressure. The ephemerides are broadcast (mostly) every hour and should only be used during the prescribed period of approximately four hours to which they refer.

In order to compute the satellite position at the observation epoch t, the following quantities, apart from the parameters a and e, are needed:

$$\begin{aligned} M &= M_0 + \left[\sqrt{\frac{\mu}{a^3}} + \Delta n\right](t - t_e) \\ \ell &= \ell_0 + \dot{\Omega}\,(t - t_e) - \omega_E\,(t - t_0) \\ \omega &= \omega_0 + C_{uc}\cos(2u) + C_{us}\sin(2u) \\ r &= r_0 + C_{rc}\cos(2u) + C_{rs}\sin(2u) \\ i &= i_0 + C_{ic}\cos(2u) + C_{is}\sin(2u) + i\,(t - t_e) \end{aligned} \tag{4.66}$$

where $u = \omega_0 + v$ is the argument of latitude. The geocentric distance r_0 is calculated by Eq. (4.7) using a, e, E at the observation epoch. The vector $\underline{r}$ in the orbital plane follows from the second representation of (4.6). Based on the reference epoch t_e, the computation of ℓ is analogous to that in Eq. (4.64). Note that again the untruncated value for the earth rotation rate must be used. The block with clock parameters enables the computation of the satellite clock error at an observation epoch t by

$$\delta^S = a_0 + a_1\,(t - t_c) + a_2\,(t - t_c)^2\,. \tag{4.67}$$

Precise ephemerides

The official precise orbits are produced by the Naval Surface Warfare Center (NSWC) together with NIMA and are based on observed data in the (extended) tracking network of the control segment. The post-mission orbits are available upon request. The most accurate orbital information is provided by the IGS with a delay of about two weeks. Less accurate information is available about two days after the observations. Currently, IGS data and products are free of charge for all users.

The precise ephemerides consist of satellite positions and velocities at equidistant epochs. Typical spacing of the data is 15 minutes. The format proposed by the NGS is widely used. Since 1985, NGS has been involved in the distribution of precise GPS orbital data. At that time, the data were distributed in the specific ASCII formats SP1 and SP2 and their binary counterparts ECF1 and ECF2. Later, ECF2 was modified to EF13 format. The formats SP1 and ECF1 contain position and velocity data, whereas SP2 and ECF2 contain just positional data. This almost halves the storage amount since the velocity data can be computed from the positional data by numerical differentiation. In 1989, NGS decided to add the GPS satellite

clock offset data into the orbital formats. Furthermore, the second generation formats can handle up to 85 satellites (GPS and others) instead of 35 GPS satellites included in the first generation formats. Apart from the clock corrections, the files normally contain only positional data; however, a header flag allows for the inclusion of velocity data. The corresponding ASCII format is denoted SP3 and the binary counterpart is ECF3 or (in a modified version) EF18.

Each NGS format consists of a header containing general information (epoch interval, orbit type, etc.) followed by the data section for successive epochs. These data are repeated for each satellite. The positions are given in kilometers and the velocities are given in kilometers per second. The NGS formats are described in Remondi (1991b). Also, NGS provides software to translate orbital files from one format to another.

The position and velocity vectors between the given epochs are obtained by interpolation where the Lagrange interpolation based on polynomial base functions is used. Note that Lagrange interpolation is also applicable to variable epoch series and that the coefficients determined can be applied to considerably longer series without updating the coefficients. This interpolation method is a fast procedure and can be easily programmed. Extensive studies by B. Remondi concluded that for GPS satellites a 30-minute epoch interval and a 9th-order interpolator suffices for an accuracy of about 10^{-8}. Another study by Remondi (1991b) using a 17th-order interpolator demonstrates that millimeter-level (10^{-10}) can be achieved based on a 40-minute epoch interval.

For those not familiar with Lagrange interpolation, the principle of this method and a numerical example are given. Assume functional values $f(t_j)$ are given at epochs t_j, $j = 0, \ldots, n$. Then,

$$\ell_j(t) = \frac{(t - t_0)(t - t_1)\cdots(t - t_{j-1})(t - t_{j+1})\cdots(t - t_n)}{(t_j - t_0)(t_j - t_1)\cdots(t_j - t_{j-1})(t_j - t_{j+1})\cdots(t_j - t_n)} \tag{4.68}$$

is the definition of the corresponding base functions $\ell_j(t)$ of degree n related to an arbitrary epoch t. The interpolated functional value at epoch t follows from the summation

$$f(t) = \sum_{j=0}^{n} f(t_j)\,\ell_j(t)\,. \tag{4.69}$$

The following numerical example assumes the functional values $f(t_j)$ given

at the epochs t_j:

$$f(t_1) = f(-3) = 13$$
$$f(t_2) = f(+1) = 17$$
$$f(t_3) = f(+5) = 85$$

The base functions are polynomials of second degree

$$\ell_0(t) = \frac{(t-t_1)(t-t_2)}{(t_0-t_1)(t_0-t_2)} = \frac{1}{32}(t^2 - 6t + 5)$$
$$\ell_1(t) = \frac{(t-t_0)(t-t_2)}{(t_1-t_0)(t_1-t_2)} = -\frac{1}{16}(t^2 - 2t - 15)$$
$$\ell_2(t) = \frac{(t-t_0)(t-t_1)}{(t_2-t_0)(t_2-t_1)} = \frac{1}{32}(t^2 + 2t - 3)$$

and, according to Eq. (4.69), the interpolated value for $t = 4$ is $f(t) = 62$. The result is immediately verified since the given functional values were generated by the polynomial $f(t) = 2t^2 + 5t + 10$.

5 Satellite signal

5.1 Signal structure

5.1.1 Physical fundamentals

Operational satellite geodesy is based upon data transmitted from the satellite to the user by means of electromagnetic waves. Such waves are generated by an oscillating (alternating) electric force. The (sinusoidal) propagation of electromagnetic waves conforms to Maxwell's laws.

During wave propagation, the field strength decreases with increasing distance from the transmitting source. This attenuation is usually expressed in decibels (dB). By definition, an attenuation of n dB means that the original field strength is decreased by a factor of $10^{-0.1\,n}$. Thus, an attenuation of 3 dB reduces the signal strength to about one half of its original strength.

Some of the factors describing the physical behavior of electromagnetic waves are presented in Table 5.1 together with their symbols and dimensions. A fundamental constant is the velocity of light in vacuum and its numerical value is given by

$$c = 299\,792\,458\,\mathrm{m\,s^{-1}}\,. \tag{5.1}$$

Note that integer cycles are equivalent to multiples of 2π radians. Another unit for cycles per second (cps) is Hertz (Hz) or decimal multiples according to the rules of the International System of Units (ISU). Equation (5.2) relates the parameters shown in Table 5.1:

$$f = 2\pi\,\frac{1}{P} = \frac{c}{\lambda}\,. \tag{5.2}$$

Table 5.1. Physical quantities

Quantity	Symbol	Dimension
Circular frequency	f	cycle $\mathrm{s^{-1}}$
Phase	φ	cycle
Wavelength	λ	m $\mathrm{cycle^{-1}}$
Period	P	s
Velocity of light	c	$\mathrm{m\,s^{-1}}$

B. Hofmann-Wellenhof et al., *Global Positioning System*

The instantaneous circular frequency f is also defined by the derivation of the phase φ with respect to time

$$f = \frac{d\varphi}{dt}, \tag{5.3}$$

and, from integrating the frequency between the epochs t_0 and t, the phase

$$\varphi = \int_{t_0}^{t} f \, dt \tag{5.4}$$

results. Assuming a constant frequency, setting the initial phase $\varphi(t_0) = 0$, and taking into account the time span t_ϱ which the signal needs to propagate through the distance ϱ from the emitter to the receiver, yields the phase equation for electromagnetic waves as observed at the receiving site:

$$\varphi = f\,(t - t_\varrho) = f\,(t - \frac{\varrho}{c}). \tag{5.5}$$

For a numerical example consider an electromagnetic wave with frequency $f = 1.5$ GHz and calculate the instantaneous phase at an antenna 20 000 km distant from the transmitting source. Using $c = 3 \cdot 10^5\,\mathrm{km\,s^{-1}}$, a (continuous) phase of exactly 10^8 cycles results. The observable (fractional) phase within one cycle, thus, is zero.

In the case of a moving emitter or a moving receiver, the received frequency is Doppler shifted. This means that the received frequency f_r differs from the emitted frequency f_e by an amount Δf which, apart from relativistic effects, is proportional to the radial velocity or range rate $v_\varrho = d\varrho/dt = \dot{\varrho}$ of the emitter with respect to the receiver:

$$\Delta f = f_r - f_e = -\frac{1}{c}\, v_\varrho\, f_e\,. \tag{5.6}$$

It can be verified by Eq. (4.9) that GPS satellites are orbiting with the mean velocity $\dot{r} = n\,a \approx 3.9\,\mathrm{km\,s^{-1}}$. Assuming a stationary terrestrial receiver and neglecting earth rotation, there is a zero radial velocity and, thus, no Doppler effect at the epoch of closest approach. The maximum radial velocity appears when the satellite crosses the horizon and amounts to $0.9\,\mathrm{km\,s^{-1}}$. Assuming a transmitted frequency $f_e = 1.5\,\mathrm{GHz}$, the Doppler frequency shift is $\Delta f = 4.5 \cdot 10^3\,\mathrm{Hz}$. This frequency shift results in a phase change of 4.5 cycles after 1 millisecond which corresponds to a change in range of 90 cm. The result can be verified by multiplying the radial velocity with the time span of 1 millisecond.

5.1.2 Components of the signal

General remarks
The official description of the GPS signal is given in the GPS Interface Control Document ICD-GPS-200 (Arinc Research Corporation 2000). Details may also be found in Spilker (1996a).

The oscillators on board the satellites generate a fundamental frequency f_0 with a stability in the range of 10^{-13} over one day for the Block II satellites. Two carrier signals in the L-band, denoted L1 and L2, are generated by integer multiplications of f_0. In addition, the carrier L3 is generated for military users only (Langley 1996). For Block IIF satellites, the option of a carrier L5 for civilian use will be implemented. The carriers L1 and L2 are modulated by codes to provide satellite clock readings to the receiver and to transmit information such as the orbital parameters. The codes consist of a sequence with the states $+1$ or -1, corresponding to the binary values 0 or 1. The so-called biphase modulation is performed by a 180° shift in the carrier phase whenever a change in the code state occurs (Fig. 5.1).

The (current) components of the signal and their frequencies are summarized in Table 5.2. Note that the nominal fundamental frequency f_0 is intentionally reduced by about 0.005 Hz to compensate for relativistic effects. More details on this subject are provided in Sect. 6.4.3.

Two codes are used for the satellite clock readings, both characterized by a pseudorandom noise (PRN) sequence. The coarse/acquisition (or clear/access) code (C/A-code) has the frequency $f_0/10$ and is repeated every millisecond. The precision (or protected) code (P-code) has the frequency f_0 and is repeated approximately once every 266.4 days. The W-code is used

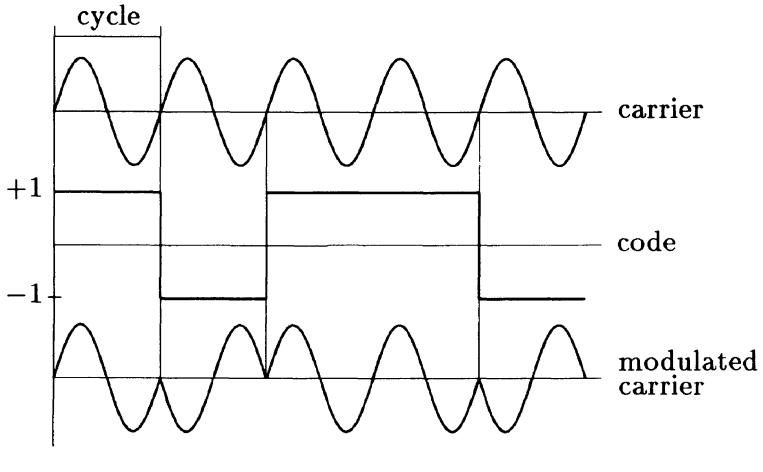


Fig. 5.1. Biphase modulation of carrier wave

Table 5.2. Components of the satellite signal

Component	Frequency (MHz)		
Fundamental frequency	f_0	$= 10.23$	
Carrier L1	$154\, f_0$	$= 1\,575.42$	($\hat{=}19.0$ cm)
Carrier L2	$120\, f_0$	$= 1\,227.60$	($\hat{=}24.4$ cm)
P-code	f_0	$= 10.23$	
C/A-code	$f_0/10$	$= 1.023$	
W-code	$f_0/20$	$= 0.5115$	
Navigation message	$f_0/204\,600$	$= 50 \cdot 10^{-6}$	

to encrypt the P-code to the Y-code when A-S is implemented. The coding of the navigation message requires 1 500 bits and, at the frequency of 50 Hz, is transmitted in 30 seconds.

Both the L1 and L2 carrier are modulated by the P-code (more precisely, by the Y-code). The C/A-code is placed on the L1 carrier in phase quadrature (i.e., 90° offset) with the P-code. Denoting the unmodulated carriers by $\mathrm{L}i(t) = a_i \cos(f_i\, t)$ and the state sequences of the P-code, the C/A-code, the W-code, and the navigation message by $P(t)$, $C/A(t)$, $W(t)$, and $D(t)$ respectively, the modulated carriers are represented by the equations

$$\begin{aligned} \mathrm{L1}(t) &= a_1\, P(t)\, W(t)\, D(t) \cos(f_1\, t) + a_1\, C/A(t)\, D(t) \sin(f_1\, t) \\ \mathrm{L2}(t) &= a_2\, P(t)\, W(t)\, D(t) \cos(f_2\, t)\,. \end{aligned} \tag{5.7}$$

Note that in the frequency domain the broadcast signal is characterized by a spread spectrum. Spread spectrum systems are commonly used for communication, since they are more resistant to jamming. The basic concept is that the information waveform with small bandwidth is converted (or spread) by modulating it with a large-bandwidth waveform. In the case of GPS, the navigation message has a bandwidth of 100 Hz, whereas the P-code bandwidth is about 20 MHz (Wells et al. 1987).

The current signal structure will be augmented in the near future. For some details, the reader is referred to Sect. 13.2.

Pseudorandom noise codes

The generation of the PRN sequences in the codes is based on the use of hardware devices called tapped feedback shift registers. As illustrated in Fig. 5.2, such a device consists of a number of storage cells (here labeled from 1 to 5) constituting one bit. Associated with each clock pulse is a shift of the bits to the right where the content of the rightmost cell is read

Number of cell	1	2	3	4	5
Initial state	1	0	1	1	0
Successive state	1	1	0	1	1

Fig. 5.2. Principle of a tapped feedback shift register

as output. The new value of the leftmost cell is determined by the binary sum of two defined cells where a binary 0 is set if the bits in these two cells are equal. The choice of the defining cells is arbitrary and determines the property of the resulting code. In order to make the procedure more clear consider cells 2 and 3 as defining pair in Fig. 5.2. After one clock pulse, the initial state would deliver a binary 0 as output and a binary 1 would be set in cell 1 of the successive state. Continuing in an analogous way, the bit sequence 1101110010... would be delivered.

The bits of the PRN sequences are often called chips to underscore that these codes do not carry data.

The C/A-code is generated by the combination of two 10-bit tapped feedback shift registers where the output of both registers are added again by binary operation to produce the code sequence. The codes of the two registers are not classified, and the C/A-code is available to civilian users. A unique code is assigned to each satellite depending on the defining cells. The frequency of 1.023 MHz and the repetition rate of 1 millisecond results in a code length of 1 023 chips. Hence, the time interval between two chips is just under 1 microsecond which approximately corresponds to a 300 m chip length.

The P-code is also not classified and originates from a combination of two bit sequences, again each generated by two registers. The first bit sequence repeats every 1.5 second and, because of the frequency f_0, has a length of $1.5345 \cdot 10^7$ bits. The second sequence has 37 more bits. The combination of both sequences, thus, results in a code with approximately $2.3547 \cdot 10^{14}$ bits which corresponds to a time span of approximately 266.4 days as stated earlier. The total code length is partitioned into 37 unique one-week segments and each segment is assigned to a satellite defining its PRN number. The codes are restarted at the beginning of every GPS week at Saturday midnight. The chip length of the P-code is about 30 m. In order to illustrate the P-code generation, an example is given (Fig. 5.3). Consider the two sequences S1 and S2 containing the bit sequences 010 and 10110 respectively. The combination of 5*S1 with 3*S2 results in a code with a length of 15 chips. Since S2 comprises 2 bits more than S1, the combined code can be partitioned into two unique segments.

5*S1	0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
3*S2	1 0 1 1 0 1 0 1 1 0 1 0 1 1 0
Combination	1 1 1 1 1 1 0 0 1 0 0 0 1 0 0

Fig. 5.3. Principle of P-code generation

The PRN code characteristics are summarized in Table 5.3. In order to protect the P-code against spoofing (i.e., the deliberate transmission of incorrect information by an adversary), the P-code is encrypted to the Y-code by A-S (Sect. 2.2.4). Since the Y-code is the modulo two sum of the P-code and the encrypting W-code, access to the P-code is only possible when the secret conversion algorithm is known.

The new military M-code will replace the Y-code in the near future. More details on this subject are given in Sect. 13.2.

Table 5.3. PRN codes characteristics

Parameter	C/A-code	P-code
Chipping rate	$1.023 \cdot 10^6$ bits per second	$10.23 \cdot 10^6$ bits per second
Chip length	≈300 m	≈30 m
Repetition rate	Millisecond	One week
Code type	37 unique codes	37 one-week segments
Properties	Easy to acquire	More accurate

Navigation message

The navigation message essentially contains information about the satellite health status, the satellite clock, the satellite orbit, and various correction data. As shown in Table 5.4, the total message containing 1 500 bits is subdivided into five subframes. One subframe is transmitted in 6 seconds and contains 10 words with 30 bits. The transmission time needed for a word is therefore 0.6 seconds. Note that a receiver requires at least 30 seconds to lock on a satellite in order to receive the complete navigation message.

Each subframe starts with the telemetry word (TLM) containing a synchronization pattern and some diagnostic messages. The second word in each subframe is the hand-over word (HOW). Apart from a subframe identification and some flags, this word contains the time-of-week (TOW) count for the epoch at the start (leading edge) of the next subframe. The TOW count (sometimes called Z-count) is a multiple number of 1.5-second intervals since the beginning of the current GPS week.

Table 5.4. Scheme of navigation message

	Number of bits	Transmission time
Total message	1 500	30 seconds
Subframe (1–5)	300	6 seconds
Word (1–10)	30	0.6 seconds

The first subframe contains the GPS week number, a prediction of the user range accuracy, indicators of the satellite health and of the age of the data, an estimation of the signal group delay, and three coefficients for a quadratic polynomial to model the satellite clock correction.

The second and third subframe transmit the broadcast ephemerides of the satellite as described in Sect. 4.4.2.

The contents of the fourth and the fifth subframe are changed in every message and have a repetition rate of 25. The total information, thus, is packed into 25 pages and requires 12.5 minutes for transmission. Many pages of the fourth subframe are reserved for military use; the remaining contain information on the ionosphere, UTC data, various flags, and the almanac data (i.e., low-accuracy orbital data) for satellites beyond the nominal 24 constellation. The pages of the fifth subframe are mainly dedicated to the almanac data and the health status for the first 24 satellites in orbit. The pages of the fourth and fifth subframe are broadcast by each satellite. Therefore, by tracking only one satellite, the almanac data of all the other satellites in orbit are obtained.

For more information concerning the data contained in the navigation message, the reader is referred to Sect. 4.4 on orbit dissemination. A detailed description of the data format is given in Department of Defense (1995), Arinc Research Corporation (2000).

5.2 Signal processing

The signal emitted from the satellite is represented by Eq. (5.7) and contains three components in the symbolic form (L1, C/A, D), (L1, Y, D), and (L2, Y, D). The received C/A-code signal on L1 is twice as powerful as the Y-code signal on L1. The same ratio exists between the Y-code signals on L1 and L2 (Table 5.5). The present Block II satellites transmit signals with 6 dB more power than the nominal value. The nominal signal strength, however, will be increased in the course of signal modernization.

Table 5.5. Strength reduction of the received signal in dB

Code	Strength reduction	
	at L1	at L2
C/A-code	−156 to −160	-
Y-code	−159 to −163	−162 to −166

The goal of signal processing by the GPS receiver is the recovery of the signal components, including the reconstruction of the carrier wave and the extraction of the codes for the satellite clock readings and the navigation message. The principle is illustrated in Fig. 5.4.

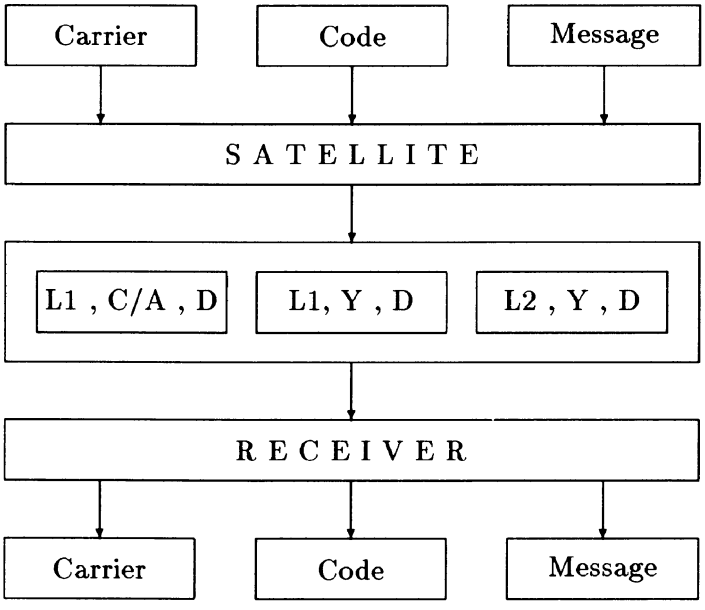


Fig. 5.4. Principle of signal processing

5.2.1 Receiver design

General features

There is a variety of receivers on the market used for different purposes (navigation, surveying, time transfer) and with different features. Despite this variety, all the receivers employ certain common principles which are presented in this section. A more comprehensive tutorial on receiver technology is given by Dierendonck (1999).

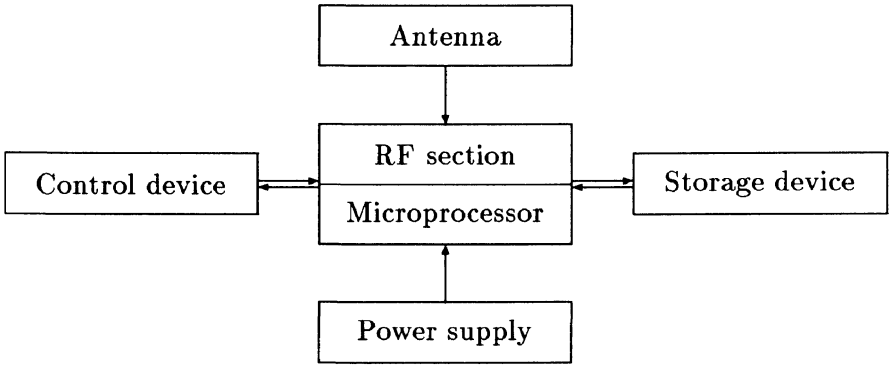


Fig. 5.5. Basic concept of a receiver unit

Following Langley (1991), the receiver unit contains components for signal reception and signal processing. The basic concept is shown in Fig. 5.5.

An omnidirectional antenna receives the signals of all satellites above the horizon and, after preamplification, transmits signals to the radio frequency (RF) section. It is worth noting that the signals are resistant to interference since the PRN codes are unique for each satellite and have very low cross correlations. The antenna may be designed for only the primary carrier L1 or for both L1 and L2 carriers. Most of the antennas sold today are microstrip antennas. One important design criterion of the antenna is the sensitivity of the phase center. The electronic center should be close to its geometric center and should be insensitive to rotation and inclination. This becomes particularly important in kinematic applications where the antenna is moved while in operation. Additionally, the antenna should have a gain pattern which filters low elevation or multipath signals. At present, this is best achieved by choke ring antennas. Additional details on these subjects are provided in Sects. 6.5 and 6.6.

The microprocessor controls the entire system and enables real-time navigation and surveying. The control device provides interactive communication with the receiver. Commands can be keyed in and diagnostic or other messages can be displayed. The control device is, therefore, usually designed as keyboard display unit.

A storage device is necessary for storing the observables and the navigation message so that they are available for later processing. Various media are presently used: PCMCIA cards, microchips, or other nonvolatile storage. Additionally, the receiver can be interfaced to an external computer.

Many receivers have an (optional) internal power supply such as recharge-

able batteries; but, there is always a provision for external batteries or other power supplies.

Radio frequency section
The radio frequency (RF) section is the heart of the receiver. After signal input from the antenna, the discrimination of the signals is achieved by the RF section usually employing the C/A-codes which are, as previously mentioned, unique for each satellite. A second method used to discriminate between satellites is the monitoring of the Doppler shift, which is generally different for each satellite.

The RF section processes the incoming signals employing separate channels. Older receivers used analogue techniques, while today most receivers use digital signal processing. Single frequency units process only the L1 signal, while dual frequency instruments process both the L1 and L2 signals. The data collected by a dual frequency receiver enable a combination where the ionospheric refraction can virtually be eliminated. This subject is treated in Sect. 6.3.2.

An important feature of the RF section is the number of channels and, hence, the number of satellites which can be tracked simultaneously. Older instruments used a limited number of physical channels and alternated satellite tracking by rapidly sequencing (20 milliseconds) satellites in and out of the same channel. Today, most receivers assign one satellite each to a physical channel where the satellites are continuously tracked. Multichannel receivers are more accurate and less sensitive to loss of signal lock but can have interchannel biases. However, with modern receivers these biases can be virtually eliminated by calibration. Receivers with sequencing channels are less expensive but slower, and are seldom used for surveying. Hybrid receivers use a combination of techniques.

The basic elements of the RF section are oscillators used to generate a reference frequency, multipliers to obtain higher frequencies, filters to eliminate undesired frequencies, and mixers. Mixers take two oscillations y_1, y_2 with different amplitudes a_1, a_2 and different frequencies f_1, f_2 and mathematically multiply them. In simplified form, this gives

$$\begin{aligned} y = y_1\, y_2 &= a_1 \cos(f_1\, t)\, a_2 \cos(f_2\, t) \\ &= \frac{a_1\, a_2}{2} \left[\cos\left((f_1 - f_2)\, t \right) + \cos\left((f_1 + f_2)\, t \right) \right] \end{aligned} \tag{5.8}$$

and results in an oscillation y comprised of a low-frequency part and a high-frequency part. After applying a low-pass filter, the high-frequency part is eliminated. The remaining low-frequency signal is processed. The difference

$f_1 - f_2$ between the frequencies is called beat or intermediate frequency (IF) and is more easily managed than the higher frequency signals.

A recently developed silicon germanium powered GPS receiver is able to convert directly RF signals into digital information and does not require IF circuits.

5.2.2 Processing techniques

The actual pseudorange measurements are performed in tracking loop circuits. Due to the motion of the satellites, the received signals are Doppler shifted which means that the tracking loops must be tunable within a frequency range. The code ranges are determined in the delay lock loop (DLL) by using code correlation techniques. After removing the PRN code from the incoming signal and performing some filtering, the unmodulated (Doppler shifted) carrier wave is obtained. This carrier wave is then passed to the phase lock loop (PLL) where the signal is compared with a carrier replica (generated in the receiver) to give the (fractional) phase offset between the two signals.

The code correlation technique provides all components of the satellite signal: the satellite clock reading, the navigation message, and the unmodulated carrier. The drawback is that the procedure requires knowledge of one PRN code. The correlation technique is performed in several steps. First, a reference carrier is generated in the receiver which is then biphase modulated with a replica of the known PRN code. In a second step, the resulting reference signal is correlated with the received satellite signal. The signals are shifted with respect to time so that they optimally match (based on a mathematical correlation). To detect the maximum correlation more accurately, two correlators (called early and late) with one chip spacing are normally used. The accuracy is increased if the spacing of the two correlators is reduced. The narrow correlator spacing technique is standard in high-performance C/A-code receivers. The time shift resulting from the correlation corresponds to the signal travel time from the satellite antenna to the phase center of the receiving antenna where clock biases are neglected. After removal of the PRN code, the received signal still contains the navigation message which can be decoded and eliminated by high-pass filtering. The final result is the Doppler shifted carrier on which a phase measurement can be performed. Since a PRN code is required, the code correlation technique is generally only applicable to the C/A-code, with (presently) only the L1 carrier being reconstructed. If the Y-code is available, both carriers can be reconstructed by the code correlation technique. Normally, the C/A-code is used to lock onto the signal and to initialize the tracking loop. This is

performed in a short time since this code repeats every millisecond and is computationally simple. One result of the C/A-code correlation is the decoded navigation message which contains the HOW in each subframe that tells the receiver where to start the search in the Y-code for signal matching.

Without knowledge of the Y-code, one has to use codeless or quasi-codeless techniques for the reconstruction of the unmodulated carrier wave from which the phase of the base carrier is measured. Most receivers provide a hybrid technique. The L1 carrier is reconstructed by code correlation using the C/A-code, and a (quasi-) codeless technique is applied to reconstruct the L2 carrier. These techniques are subsequently described in more detail since they are essential with A-S being activated.

Squaring technique
This codeless technique was first presented in 1981 by C. Counselman. The procedure is based on autocorrelating the received signal. In other words, the received signal is mixed (i.e., multiplied) with itself and, hence, all modulations are removed. This happens because a 180° phase shift during modulation is equivalent to a change in the sign of the signal. The result is the unmodulated carrier with twice the frequency, cf. Eq. (5.8), and, thus, half the wavelength. In general, it is more difficult to resolve the ambiguities of the squared signals with halved wavelength.

This technique has the advantage of being independent of PRN codes. The drawback of squaring is that the satellite clock and the satellite orbit information are lost in the process. Additionally, the signal-to-noise ratio (SNR) is substantially reduced in the squaring process (Ashjaee 1993). For example, compared to the code correlation technique, the SNR is reduced by 30 dB.

Cross correlation technique
This is another codeless technique which was first described in 1985 by P. MacDoran. The technique is based on the fact that the unknown Y-code is identical on both carriers which enables cross correlation of the L1 and L2 signal. Due to the frequency dependent propagation of an electromagnetic wave through the ionosphere, the Y-code on L2 is slightly slower than the Y-code on L1. The time delay necessary to match the L1 signal with the L2 signal in the receiver is a measure of the travel time difference of the two signals. The delay of the L2 signal is variable and must be adjusted appropriately to achieve maximum correlation between the L1 and the L2 signal. The observables resulting from the correlation process are the range difference between the two signals obtained from the time delay of the Y-code on the two carriers, that is $R_{\mathrm{L2,Y}} - R_{\mathrm{L1,Y}}$, and a phase difference

$\Phi_{L2} - \Phi_{L1}$ obtained from the beat frequency carrier.

The cross correlation outputs may be used to derive the L2 code range and phase by forming

$$R_{L2} = R_{L1,C/A} + (R_{L2,Y} - R_{L1,Y}) \tag{5.9}$$

and

$$\Phi_{L2} = \Phi_{L1,C/A} + (\Phi_{L2} - \Phi_{L1}) \tag{5.10}$$

where the subscripts of $R_{L1,C/A}$ and $\Phi_{L1,C/A}$ indicate that this code range and phase have been derived via the C/A-code on the L1 signal.

Since the power of the L1 ranging signal is twice that of the corresponding L2 signal, the cross correlation of the L1 and the L2 signal leads to an improvement of 3 dB compared to the squaring of the L2 signal (Ashjaee and Lorenz 1992). However, compared to the code correlation technique, an SNR degradation of 27 dB occurs.

Code correlation plus squaring technique

This improved squaring technique was patented by Keegan (1990). The method is also denoted code-aided squaring and involves correlating the received Y-code on the L2 signal with a locally generated replica of the P-code. This correlation is possible because the Y-code originates from a modulo two sum of the P-code and the encrypting W-code. Since the frequency (or chipping rate) of the W-code is about 20 times less the frequency of the Y-code, there always exist Y-code portions which are identical to the original P-code portions (Eissfeller 1993). Therefore, the replica of the P-code is shifted to match with the underlying P-code portion of the satellite Y-code. After the correlation, a low-pass filter is applied by narrowing the bandwidth and, subsequently, the signal is squared to rid the signal of the code. The technique provides code range and, because of squaring, half-wavelength carrier phase. The correlation with the P-code yields a better jamming immunity and an improvement in the multipath performance (Ashjaee and Lorenz 1992).

The subsequent squaring results in 20 times more power (i.e., 13 dB) than by directly squaring the Y-code signal, because the SNR of a squared signal is inversely proportional to its bandwidth (Hatch et al. 1992). In fact, an SNR degradation of 17 dB occurs compared to the code correlation technique.

Z-tracking technique

An improved quasi-codeless technique is denoted Z-tracking^{TM} and has been reported in Ashjaee and Lorenz (1992). The Y-code on both the L1 and

L2 signals is separately correlated with a receiver generated replica of the P-code. Since there is a separate correlation on L1 and L2, the W-code on each frequency is obtained. The subsequent procedure is explained by Ashjaee (1993) in the following way: "The carriers on each frequency also contain the underlying encrypting code. At this time, with enough integration of signal, the encrypting signal bit is estimated for each frequency and is fed to the other frequency. This estimation is used to remove the encrypting code from the signal." This integration is performed by a low-pass filtering where the bandwidth may be reduced to that of the encrypting W-code. There is no need to know the W-code because it is only used for synchronization purposes (Breuer et al. 1993). The removal of the encrypting code leads to the same signals as in the case of no A-S. Thus, the code ranges and full wavelength L1 and L2 carrier phases are obtained. Note that the Y-code pseudoranges are precisely the same observables as obtained by code correlation with the P-code.

Compared to the previous code correlation plus squaring technique, the SNR is improved by 3 dB; however, compared to the code correlation technique, an SNR degradation of 14 dB occurs.

All four approaches to recover the L2 carrier in the presence of A-S suffer from a substantial degradation in the SNR. Without exception, no codeless or quasi-codeless technique recovers GPS signal information as well as the code correlation technique. Also, weaker signals are more sensitive to high ionospheric activities and interfering (jamming) signals which may even cause a loss of lock. Today, most geodetic receivers use techniques resembling Z-tracking.

The numerical values in Table 5.6 summarize the SNR degradation compared to the code correlation technique.

Table 5.7 summarizes the characteristics of the four techniques. The graphical diagrams are adapted from Ashjaee and Lorenz (1992), Eissfeller (1993).

Table 5.6. SNR degradation compared to code correlation technique

Squaring	−30 dB
Cross correlation	−27 dB
Code correlation plus squaring	−17 dB
Z-tracking	−14 dB

Table 5.7. Techniques to recover carrier and code on the L2 signal when A-S is on; the symbol ⊗ indicates the process of correlation

Technique	Input	Operation	Output
Squaring	Y-code on L2	⊗	$\Phi_{L2}(\lambda_{L2}/2)$ no code range
Cross correlation	Y-code on L1; Y-code on L2	Variable delay; ⊗	$\Phi_{L2} - \Phi_{L1}$ $R_{L2,Y} - R_{L1,Y}$
Code correlation plus squaring	Y-code on L2; P-code replica	⊗; Low-pass filter; ⊗	$\Phi_{L2}(\lambda_{L2}/2)$ $R_{L2,P}$
Z-tracking	Y-code on L1; P-code replica; Y-code on L2; P-code replica	⊗; Low-pass filter; ⊗; Low-pass filter	Φ_{L1} $R_{L1,Y}$; Φ_{L2} $R_{L2,Y}$

6 Observables

6.1 Data acquisition

In concept, the GPS observables are ranges which are deduced from measured time or phase differences based on a comparison between received signals and receiver generated signals. Unlike the terrestrial electronic distance measurements, GPS uses the "one way concept" where two clocks are used, namely one in the satellite and the other in the receiver. Thus, the ranges are biased by satellite and receiver clock errors and, consequently, they are denoted as pseudoranges.

6.1.1 Code pseudoranges

Let us denote by t^S the reading of the satellite clock at emission time and by t_R the reading of the receiver clock at signal reception time. Analogously, the delays of the clocks with respect to GPS system time are termed δ^S and δ_R. Recall that the satellite clock reading t^S is transmitted via the PRN code. The difference between the clock readings is equivalent to the time shift Δt which aligns the satellite and reference signal during the code correlation procedure in the receiver. Thus,

$$\begin{aligned} \Delta t = t_R - t^S &= [t_R(\mathrm{GPS}) + \delta_R] - [t^S(\mathrm{GPS}) + \delta^S] \\ &= \Delta t(\mathrm{GPS}) + \Delta\delta\,, \end{aligned} \tag{6.1}$$

where $\Delta t(\mathrm{GPS}) = t_R(\mathrm{GPS}) - t^S(\mathrm{GPS})$ and $\Delta\delta = \delta_R - \delta^S$. The bias δ^S of the satellite clock can be modeled by a polynomial with the coefficients being transmitted in the first subframe of the navigation message. Assuming the δ^S correction is applied, $\Delta\delta$ equals the receiver clock delay. The time interval Δt multiplied by the speed of light c yields the pseudorange R and, hence,

$$R = c\,\Delta t = c\,\Delta t(\mathrm{GPS}) + c\,\Delta\delta = \varrho + c\,\Delta\delta\,. \tag{6.2}$$

Note that the C/A-code repeats every millisecond which corresponds to 300 km in range. Since the satellites are at a distance of about 20 000 km from the earth, C/A-code pseudoranges are ambiguous. However, this ambiguity can be easily resolved during initial satellite acquisition by introducing approximate (within some few hundred kilometers) position coordinates of the receiver (Lachapelle 1991).

The range ϱ is calculated from the true signal travel time. In other words, ϱ corresponds to the distance between the position of the satellite at epoch t^S(GPS) and the position of the antenna of the receiver at epoch t_R(GPS). Since ϱ is a function of two different epochs, it is often expanded into a Taylor series with respect to, e.g., the emission time

$$\begin{aligned} \varrho = \varrho(t^S, t_R) &= \varrho(t^S, (t^S + \Delta t)) \\ &= \varrho(t^S) + \dot{\varrho}(t^S)\,\Delta t \end{aligned} \tag{6.3}$$

where $\dot{\varrho}$ denotes the time derivative of ϱ or the radial velocity of the satellite relative to the receiving antenna. All epochs in Eq. (6.3) are expressed in GPS system time.

The maximum radial velocity for GPS satellites in the case of a stationary receiver is $\dot{\varrho} \approx 0.9\,\mathrm{km\,s^{-1}}$, and the travel time of the satellite signal is about 0.07 s. The correction term in Eq. (6.3), thus, amounts to some 60 m.

The precision of a pseudorange derived from code measurements has been traditionally about 1% of the chip length. Thus, a precision of roughly 3 m and 0.3 m is achieved with C/A-code and P-code pseudoranges respectively. However, recent developments demonstrate that a precision of about 0.1% of the chip length is possible.

6.1.2 Phase pseudoranges

Let us denote by $\varphi^S(t)$ the phase of the received and reconstructed carrier with frequency f^S and by $\varphi_R(t)$ the phase of a reference carrier generated in the receiver with frequency f_R. Here, the parameter t is an epoch in the GPS time system reckoned from an initial epoch $t_0 = 0$. According to Eq. (5.5), the following phase equations are obtained

$$\begin{aligned} \varphi^S(t) &= f^S\,t - f^S\,\frac{\varrho}{c} - \varphi_0^S \\ \varphi_R(t) &= f_R\,t - \varphi_{0R}\,. \end{aligned} \tag{6.4}$$

The initial phases φ_0^S, φ_{0R} are caused by clock errors and are equal to

$$\begin{aligned} \varphi_0^S &= -f^S\,\delta^S \\ \varphi_{0R} &= -f_R\,\delta_R\,. \end{aligned} \tag{6.5}$$

Hence, the beat phase $\varphi_R^S(t)$ is given by

$$\begin{aligned} \varphi_R^S(t) &= \varphi^S(t) - \varphi_R(t) \\ &= -f^S\,\frac{\varrho}{c} + f^S\,\delta^S - f_R\,\delta_R + (f^S - f_R)\,t\,. \end{aligned} \tag{6.6}$$

The deviation of the frequencies f^S, f_R from the nominal frequency f is only in the order of some fractional parts of Hz. This may be verified by considering, e.g., a short-time stability in the frequencies of $df/f = 10^{-12}$. With the nominal carrier frequency $f \approx 1.5\,\mathrm{GHz}$, the frequency error, thus, becomes $df = 1.5 \cdot 10^{-3}\,\mathrm{Hz}$. Such a frequency error may be neglected because during signal propagation (i.e., $t = 0.07$ seconds) a maximum error of 10^{-4} cycles in the beat phase is generated which is below the noise level. The clock errors are in the range of milliseconds and are, thus, less effective. Summarizing, Eq. (6.6) may be written in the simplified form

$$\varphi_R^S(t) = -f\frac{\varrho}{c} - f\,\Delta\delta \tag{6.7}$$

where again $\Delta\delta = \delta_R - \delta^S$ has been introduced. If the assumption of frequency stability is incorrect and the oscillators are unstable, then their behavior has to be modeled by, for example, polynomials where clock and frequency offsets and a frequency drift are determined. A complete carrier phase model which includes the solution of large (e.g., 1 second) receiver clock errors was developed by Remondi (1984). Adequate formulas can also be found in King et al. (1987), p. 55. No further details are given here because in practice eventual residual errors will be eliminated by differencing the measurements.

Switching on a receiver at epoch t_0, the instantaneous fractional beat phase is measured. The initial integer number N of cycles between satellite and receiver is unknown. However, when tracking is continued without loss of lock, the number N, also called integer ambiguity, remains the same and the beat phase at epoch t is given by

$$\varphi_R^S(t) = \Delta\varphi_R^S \Big|_{t_0}^{t} + N \tag{6.8}$$

where $\Delta\varphi_R^S$ denotes the (measurable) fractional phase at epoch t augmented by the number of integer cycles since the initial epoch t_0. A geometrical interpretation of Eq. (6.8) is provided in Fig. 6.1 where $\Delta\varphi_i$ is a shortened notation for $\Delta\varphi_R^S \big|_{t_0}^{t_i}$ and, for simplicity, the initial fractional beat phase $\Delta\varphi_0$ is assumed to be zero. Substituting Eq. (6.8) into Eq. (6.7) and denoting the negative observation quantity by $\Phi = -\Delta\varphi_R^S$ yields the equation for the phase pseudoranges

$$\Phi = \frac{1}{\lambda}\varrho + \frac{c}{\lambda}\Delta\delta + N \tag{6.9}$$

where the wavelength λ has been introduced according to Eq. (5.2). Multiplying the above equation by λ scales the phase expressed in cycles to a range which only differs from the code pseudorange by the integer multiples

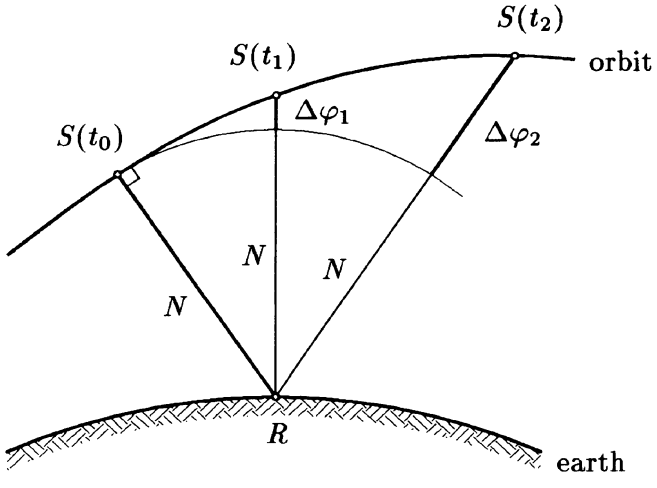


Fig. 6.1. Geometrical interpretation of phase range

of λ. Again, the range ϱ represents the distance between the satellite at emission epoch t and the receiver at reception epoch $t + \Delta t$. The phase of the carrier can be measured to better than 0.01 cycles which corresponds to millimeter precision.

It should be noted that a plus sign convention has been chosen for Eq. (6.9). This choice is somewhat arbitrary since quite often the phase Φ and the distance ϱ have different signs. Actually, the sign is receiver dependent because the beat phase is generated in the receiver and the combination of the satellite and the receiver signal differs for various receiver types.

6.1.3 Doppler data

Some of the first solution models proposed for GPS were to use the Doppler observable as with the TRANSIT system. This system used the integrated Doppler shifts (i.e., phase differences) which were scaled to delta ranges. The raw Doppler shift, cf. Eq. (5.6), being linearly dependent on the radial velocity and, thus, allowing for velocity determination in real time is important for navigation. Considering Eq. (6.9), the equation for the observed Doppler shift scaled to range rate is given by

$$D = \lambda \dot{\Phi} = \dot{\varrho} + c\,\Delta\dot{\delta} \tag{6.10}$$

where the derivatives with respect to time are indicated by a dot. The raw Doppler shift is less accurate than integrated Doppler. An estimate of the

achievable accuracy is 0.001 Hz. This corresponds to $0.3\,\mathrm{m\,s^{-1}}$ if the Doppler shift is measured in the C/A-code tracking loop.

A detailed derivation of Doppler equations within the frame of GPS is given in Remondi (1984) where even relativistic effects are accounted for. It is worth noting here that the raw Doppler shift is also applied to determine integer ambiguities in kinematic surveying (Remondi 1991a) or is used as an additional independent observable for point positioning.

6.1.4 Biases and noise

The code pseudoranges, cf. Eq. (6.2), and phase pseudoranges, cf. Eq. (6.9), are affected by both, systematic errors or biases and random noise. Note that Doppler measurements are affected by the bias rates only. The error sources can be classified into three groups, namely satellite related errors, propagation medium related errors, and receiver related errors. Some range biases are listed in Table 6.1.

Some of the systematic errors can be modeled and give rise to additional terms in the observation equations which will be explained in detail in subsequent sections. As mentioned earlier, systematic effects can also be eliminated by appropriate combinations of the observables. Differencing between receivers eliminates satellite-specific biases, and differencing between satellites eliminates receiver-specific biases. Thus, double-differenced pseudoranges are, to a high degree, free of systematic errors originating from the satellites and from the receivers. With respect to refraction, this is only true for short baselines where the measured ranges at both endpoints are affected equally. In addition, ionospheric refraction can be virtually eliminated by an adequate combination of dual frequency data. Antenna phase center variations are treated in Sect. 6.5. Multipath is caused by multiple reflections

Table 6.1. Range biases

Source	Effect
Satellite	Clock bias
	Orbital errors
Signal propagation	Ionospheric refraction
	Tropospheric refraction
Receiver	Antenna phase center variation
	Clock bias
	Multipath

of the signal (which can also occur at the satellite during signal emission). The interference between the direct and the reflected signal is largely not random; however, it may also appear as a noise. Wells et al. (1987) report a similar effect called imaging where a reflecting obstacle generates an image of the real antenna which distorts the antenna pattern. Both effects, multipath and imaging, can be considerably reduced by selecting sites protected from reflections (buildings, vehicles, trees, etc.) and by an appropriate antenna design. It should be noted that multipath is frequency dependent. Therefore, carrier phases are less affected than code ranges (Lachapelle 1990). More details on the multipath problems are given in Sect. 6.6.

The random noise mainly contains the actual observation noise plus random constituents of multipath (especially for kinematic applications). The pseudorange noise is summarized in Table 6.2.

The measurement noise, an estimation of the satellite biases, and the contributions from the wave propagation are combined in the User Equivalent Range Error (UERE). This UERE is transmitted via the navigation message. In combination with a DOP factor, explained in Sect. 9.6, UERE allows for an estimation of the achievable point positioning accuracy.

Table 6.2. Range noise

Range	Noise
Code range (C/A-code)	10–300 cm
Code range (P-code)	10–30 cm
Phase range	0.2–5 mm

6.2 Data combinations

GPS observables are obtained from the code information or the carrier wave in the broadcast satellite signal. Recall that the P-code is modulated on both carriers L1 and L2, whereas the C/A-code is modulated on L1 only. Consequently, one could measure the code ranges R_{L1}, R_{L2}, the carrier phases Φ_{L1}, Φ_{L2}, and the corresponding Doppler shifts D_{L1}, D_{L2} for a single epoch. Note that the L1 observables may either be derived from the C/A-code or from the P-code. In general, not all observables are available because, for example, a single frequency receiver delivers only data from the L1 frequency. Furthermore, the Doppler observables are not considered here.

The objectives of this section are to show how linear combinations are

developed for dual frequency data, and how code range smoothing by means of carrier phases is performed.

6.2.1 Linear phase combinations

General remarks
The linear combination of two phases φ_1 and φ_2 is defined by

$$\varphi = n_1 \varphi_1 + n_2 \varphi_2 \tag{6.11}$$

where n_1 and n_2 are arbitrary numbers. The substitution of the relations $\varphi_i = f_i t$ for the corresponding frequencies f_1 and f_2 gives

$$\varphi = n_1 f_1 t + n_2 f_2 t = f t . \tag{6.12}$$

Therefore,

$$f = n_1 f_1 + n_2 f_2 \tag{6.13}$$

is the frequency and

$$\lambda = \frac{c}{f} \tag{6.14}$$

is the wavelength of the linear combination.

Considering a certain noise level for the phases, the noise level increases for the linear combination. Applying the error propagation law and assuming the same noise for both phases, the noise of the linear combination is amplified by the factor $\sqrt{n_1^2 + n_2^2}$ compared to the noise of a single phase. To compute this factor correctly, proper noise levels must be taken into account.

Linear combinations with integer numbers
In the case of GPS, the simplest nontrivial linear combinations of L1 and L2 carrier phases Φ_{L1} and Φ_{L2} in Eq. (6.11) are $n_1 = n_2 = 1$, yielding the sum

$$\Phi_{L1+L2} = \Phi_{L1} + \Phi_{L2} \tag{6.15}$$

and $n_1 = 1$, $n_2 = -1$, leading to the difference

$$\Phi_{L1-L2} = \Phi_{L1} - \Phi_{L2} . \tag{6.16}$$

The corresponding wavelengths according to (6.13) and (6.14) are

$$\begin{aligned} \lambda_{L1+L2} &= 10.7 \text{ cm} \\ \lambda_{L1-L2} &= 86.2 \text{ cm} \end{aligned} \tag{6.17}$$

where the numerical values for the carrier frequencies f_{L1} and f_{L2} (Table 5.2) have been substituted. The combination Φ_{L1+L2} is denoted as narrow lane and Φ_{L1-L2} as wide lane. The lane signals are applied for ambiguity resolution (Sect. 9.2).

The advantage of a linear combination with integer numbers is that the integer nature of the ambiguities is preserved.

Linear combinations with real numbers
A slightly more complicated linear combination results from the choice

$$n_1 = 1 \qquad n_2 = -\frac{f_{L2}}{f_{L1}} \tag{6.18}$$

leading to the combination

$$\Phi_{L1,L2} = \Phi_{L1} - \frac{f_{L2}}{f_{L1}} \Phi_{L2} \tag{6.19}$$

which is frequently denoted as geometric residual. This quantity is the kernel in a combination used to reduce ionospheric effects (Sect. 6.3.2).

Another linear combination follows from the reciprocal values of (6.18)

$$n_1 = 1 \qquad n_2 = -\frac{f_{L1}}{f_{L2}} \tag{6.20}$$

leading to the combination

$$\Phi_{L1,L2} = \Phi_{L1} - \frac{f_{L1}}{f_{L2}} \Phi_{L2} \tag{6.21}$$

which is frequently denoted as ionospheric residual. This quantity is used in the context of cycle slip detection (Sect. 9.1.2).

The drawback of a linear combination with real numbers is that the integer nature of the ambiguity is lost.

6.2.2 Code pseudorange smoothing

The objective here is to show the principle of code pseudorange smoothing by means of phase pseudoranges. This is an important issue in accurate real-time positioning.

Assuming dual frequency measurements for epoch t_1, the code pseudoranges $R_{L1}(t_1)$, $R_{L2}(t_1)$ and the carrier phase pseudoranges $\Phi_{L1}(t_1)$, $\Phi_{L2}(t_1)$ are obtained. Further assume, the code pseudoranges are scaled to cycles

(but still denoted as R) by dividing them by the corresponding carrier wavelength. Using the two frequencies f_{L1}, f_{L2}, the combination

$$R(t_1) = \frac{f_{L1}\, R_{L1}(t_1) - f_{L2}\, R_{L2}(t_1)}{f_{L1} + f_{L2}} \tag{6.22}$$

is formed for the code pseudoranges and the wide lane signal

$$\Phi(t_1) = \Phi_{L1}(t_1) - \Phi_{L2}(t_1) \tag{6.23}$$

for the carrier phase pseudoranges. From Eq. (6.22) one can see that the noise of the combined code pseudorange $R(t_1)$ is reduced by a factor of 0.7 compared to the noise of the single code measurement. The increase of the noise in the wide lane signal by a factor of $\sqrt{2}$ has no effect because the noise of the carrier phase pseudoranges is essentially lower than the noise of the code pseudoranges. Note that both signals $R(t_1)$ and $\Phi(t_1)$ have the same frequency and, thus, the same wavelength as may be verified by applying Eq. (6.13).

Combinations (6.22) and (6.23) are formed for each epoch. Additionally, for all epochs t_i after t_1, extrapolated values of the code pseudoranges $R(t_i)_{ex}$ can be calculated from

$$R(t_i)_{ex} = R(t_1) + (\Phi(t_i) - \Phi(t_1))\,. \tag{6.24}$$

The smoothed value $R(t_i)_{sm}$ is finally obtained by the arithmetic mean

$$R(t_i)_{sm} = \tfrac{1}{2}\,(R(t_i) + R(t_i)_{ex})\,. \tag{6.25}$$

Generalizing the above formulas for an arbitrary epoch t_i (with the preceding epoch t_{i-1}), a recursive algorithm is given by

$$\begin{aligned}
R(t_i) &= \frac{f_{L1}\, R_{L1}(t_i) - f_{L2}\, R_{L2}(t_i)}{f_{L1} + f_{L2}} \\
\Phi(t_i) &= \Phi_{L1}(t_i) - \Phi_{L2}(t_i) \\
R(t_i)_{ex} &= R(t_{i-1})_{sm} + (\Phi(t_i) - \Phi(t_{i-1})) \\
R(t_i)_{sm} &= \tfrac{1}{2}\,(R(t_i) + R(t_i)_{ex})
\end{aligned}$$

which works under the initial condition $R(t_1) = R(t_1)_{ex} = R(t_1)_{sm}$ for all $i > 1$.

The above algorithm assumes the data are free of gross errors. However, carrier phase data are sensitive to changes in the integer ambiguity (i.e., cycle slips). To circumvent this problem, a variation of the algorithm is

subsequently given. Using the same notations as before for an epoch t_i, the smoothed code pseudorange is obtained by

$$R(t_i)_{\mathrm{sm}} = w\,R(t_i) + (1-w)\,(R(t_{i-1})_{\mathrm{sm}} + \Phi(t_i) - \Phi(t_{i-1})) \qquad (6.26)$$

where w is a time dependent weight factor. For the first epoch $i = 1$, the weight is set $w = 1$; thus, putting the full weight on the measured code pseudorange. For consecutive epochs, the weight of the code phase is continuously reduced and, thus, emphasizes the influence of the carrier phases. A reduction of the weight by 0.01 from epoch to epoch was tested in a kinematic experiment with a data sampling rate of 1 Hz. After 100 seconds, only the smoothed value of the previous epoch (augmented by the carrier phase difference) is taken into account. Again, in the case of cycle slips, the algorithm would fail. A simple check of the carrier phase difference for two consecutive epochs by the Doppler shift multiplied by the time interval may detect data irregularities such as cycle slips. After the occurrence of a cycle slip, the weight is reset to $w = 1$ which fully eliminates the influence of the erroneous carrier phase data. The key of this approach is that cycle slips must be detected but do not have to be corrected.

Another smoothing algorithm for code pseudoranges uses phase differences $\Delta\Phi(t_i, t_1)$ obtained by the integrated Doppler shift between the current epoch t_i and the starting epoch t_1. Note that the integrated Doppler shifts are insensitive to cycle slips. From each code pseudorange $R(t_i)$ at epoch t_i, an estimate of the code pseudorange at epoch t_1 can be given by

$$R(t_1)_i = R(t_i) - \Delta\Phi(t_i, t_1) \qquad (6.27)$$

where the subscript i on the left side of the equation indicates the epoch that the code pseudorange $R(t_1)$ is computed from. Obtaining an estimate consecutively for each epoch, the arithmetic mean $R(t_1)_{\mathrm{m}}$ of the code pseudorange for n epochs is calculated by

$$R(t_1)_{\mathrm{m}} = \frac{1}{n}\sum_{i=1}^{n} R(t_1)_i \qquad (6.28)$$

and the smoothed code pseudorange for an arbitrary epoch results from

$$R(t_i)_{\mathrm{sm}} = R(t_1)_{\mathrm{m}} + \Delta\Phi(t_i, t_1)\,. \qquad (6.29)$$

The advantage of this procedure lies in the reduction of the noise in the initial code pseudorange by averaging an arbitrary number n of measured code pseudoranges. Note from the three formulas (6.27) through (6.29) that the algorithm may also be applied successively epoch by epoch where the

arithmetic mean must be updated from epoch to epoch. Using the above notations, formula (6.29) also works for epoch t_1, where, of course, $\Delta\Phi(t_1, t_1)$ is zero and there is no smoothing effect.

All the smoothing algorithms are also applicable if only single frequency data are available. In this case, $R(t_i)$, $\Phi(t_i)$, and $\Delta\Phi(t_i, t_1)$ denote the single frequency code pseudorange, carrier phase pseudorange, and phase difference respectively.

6.3 Atmospheric effects

6.3.1 Phase and group velocity

Consider a single electromagnetic wave propagating in space with wavelength λ and frequency f. The velocity of its phase

$$v_{\text{ph}} = \lambda f \tag{6.30}$$

is denoted phase velocity. For GPS, the carrier waves L1 and L2 are propagating with this velocity.

For a group of waves with slightly different frequencies, the propagation of the resultant energy is defined by the group velocity

$$v_{\text{gr}} = -\frac{df}{d\lambda}\lambda^2 \tag{6.31}$$

according to Bauer (1994), p. 96. This velocity has to be considered for GPS code measurements.

A relation between phase and group velocity may be derived by forming the total differential of Eq. (6.30) resulting in

$$dv_{\text{ph}} = f\,d\lambda + \lambda\,df \tag{6.32}$$

which can be rearranged to

$$\frac{df}{d\lambda} = \frac{1}{\lambda}\frac{dv_{\text{ph}}}{d\lambda} - \frac{f}{\lambda}. \tag{6.33}$$

The substitution of (6.33) into (6.31) yields

$$v_{\text{gr}} = -\lambda\frac{dv_{\text{ph}}}{d\lambda} + f\lambda \tag{6.34}$$

or finally the Rayleigh equation

$$v_{\text{gr}} = v_{\text{ph}} - \lambda\frac{dv_{\text{ph}}}{d\lambda}. \tag{6.35}$$

Note that the differentiation (6.32) implicitly contains the dispersion (Joos 1956: p. 57), which is defined as a dependence of the phase velocity on the wavelength or the frequency. Phase and group velocity are equal in nondispersive media and correspond to the speed of light in vacuum.

The wave propagation in a medium depends on the refractive index n. Generally, the propagation velocity is obtained from

$$v = c/n\,. \tag{6.36}$$

Applying this expression to the phase and group velocity, appropriate formulas for the corresponding refractive indices n_{ph} and n_{gr}

$$v_{\mathrm{ph}} = c/n_{\mathrm{ph}} \tag{6.37}$$

$$v_{\mathrm{gr}} = c/n_{\mathrm{gr}} \tag{6.38}$$

are achieved. Differentiation of the phase velocity with respect to λ, that is

$$\frac{dv_{\mathrm{ph}}}{d\lambda} = -\frac{c}{n_{\mathrm{ph}}^2}\,\frac{dn_{\mathrm{ph}}}{d\lambda}\,, \tag{6.39}$$

and substitution of the last three equations into (6.35) yields

$$\frac{c}{n_{\mathrm{gr}}} = \frac{c}{n_{\mathrm{ph}}} + \lambda\,\frac{c}{n_{\mathrm{ph}}^2}\,\frac{dn_{\mathrm{ph}}}{d\lambda} \tag{6.40}$$

or

$$\frac{1}{n_{\mathrm{gr}}} = \frac{1}{n_{\mathrm{ph}}}\left(1 + \lambda\,\frac{1}{n_{\mathrm{ph}}}\,\frac{dn_{\mathrm{ph}}}{d\lambda}\right)\,. \tag{6.41}$$

This equation may be inverted to

$$n_{\mathrm{gr}} = n_{\mathrm{ph}}\left(1 - \lambda\,\frac{1}{n_{\mathrm{ph}}}\,\frac{dn_{\mathrm{ph}}}{d\lambda}\right) \tag{6.42}$$

where the approximation $(1+\varepsilon)^{-1} = 1 - \varepsilon$ has been applied accordingly. Thus,

$$n_{\mathrm{gr}} = n_{\mathrm{ph}} - \lambda\,\frac{dn_{\mathrm{ph}}}{d\lambda} \tag{6.43}$$

is the modified Rayleigh equation. A slightly different form is obtained by differentiating the relation $c = \lambda f$ with respect to λ and f, that is

$$d\lambda/\lambda = -df/f\,, \tag{6.44}$$

and by substituting the result into (6.43):

$$n_{\mathrm{gr}} = n_{\mathrm{ph}} + f\,\frac{dn_{\mathrm{ph}}}{df}\,. \tag{6.45}$$

6.3.2 Ionospheric refraction

The ionosphere, extending in various layers from about 50 km to 1 000 km above earth, is a dispersive medium with respect to the GPS radio signal. Following Seeber (1993), p. 44, the series

$$n_{\text{ph}} = 1 + \frac{c_2}{f^2} + \frac{c_3}{f^3} + \frac{c_4}{f^4} + \dots \tag{6.46}$$

approximates the phase refractive index. The coefficients c_2, c_3, c_4 do not depend on frequency but on the quantity N_e denoting the number of electrons per cubic meter (i.e., the electron density) along the propagation path. Using an approximation by cutting off the series expansion after the quadratic term, that is

$$n_{\text{ph}} = 1 + \frac{c_2}{f^2}\,, \tag{6.47}$$

differentiating this equation

$$dn_{\text{ph}} = -\frac{2c_2}{f^3}\,df\,, \tag{6.48}$$

and substituting (6.47) and (6.48) into (6.45) yields

$$n_{\text{gr}} = 1 + \frac{c_2}{f^2} - f\,\frac{2c_2}{f^3} \tag{6.49}$$

or

$$n_{\text{gr}} = 1 - \frac{c_2}{f^2}\,. \tag{6.50}$$

It can be seen from (6.47) and (6.50) that the group and the phase refractive indices deviate from unity with opposite sign. With an estimate for c_2 (Seeber 1993: p. 44),

$$c_2 = -40.3\,N_e \quad [\text{Hz}^2]\,, \tag{6.51}$$

the relation $n_{\text{gr}} > n_{\text{ph}}$ and, thus, $v_{\text{gr}} < v_{\text{ph}}$ follows because the electron density N_e is always positive. As a consequence of the different velocities, a group delay and a phase advance occur. In other words, GPS code measurements are delayed and the carrier phases are advanced. Therefore, the code pseudoranges are measured too long and the carrier phase pseudoranges are measured too short compared to the geometric range between the satellite and the receiver. The amount of the difference is the same in both cases.

According to Fermat's principle, the measured range s is defined by

$$s = \int n\,ds \tag{6.52}$$

where the integral must be extended along the path of the signal. The geometric range s_0 along the straight line between the satellite and the receiver may be obtained analogously by setting $n = 1$:

$$s_0 = \int ds_0 \,. \tag{6.53}$$

The difference Δ^{Iono} between measured and geometric range is called ionospheric refraction and follows from

$$\Delta^{\text{Iono}} = \int n\, ds - \int ds_0 \tag{6.54}$$

which may be written for a phase refractive index n_{ph} from (6.47) as

$$\Delta_{\text{ph}}^{\text{Iono}} = \int \left(1 + \frac{c_2}{f^2}\right) ds - \int ds_0 \tag{6.55}$$

and for a group refractive index n_{gr} from (6.50) as

$$\Delta_{\text{gr}}^{\text{Iono}} = \int \left(1 - \frac{c_2}{f^2}\right) ds - \int ds_0 \,. \tag{6.56}$$

A simplification is obtained when allowing the integration for the first term in (6.55) and (6.56) along the geometric path. In this case, ds becomes ds_0 and the formulas

$$\Delta_{\text{ph}}^{\text{Iono}} = \int \frac{c_2}{f^2}\, ds_0 \qquad \Delta_{\text{gr}}^{\text{Iono}} = -\int \frac{c_2}{f^2}\, ds_0 \tag{6.57}$$

result which can also be written as

$$\Delta_{\text{ph}}^{\text{Iono}} = -\frac{40.3}{f^2} \int N_e\, ds_0 \qquad \Delta_{\text{gr}}^{\text{Iono}} = \frac{40.3}{f^2} \int N_e\, ds_0 \tag{6.58}$$

where (6.51) has been substituted. Defining the total electron content (TEC) by

$$\text{TEC} = \int N_e\, ds_0 \,, \tag{6.59}$$

and substituting TEC into (6.58) yields

$$\Delta_{\text{ph}}^{\text{Iono}} = -\frac{40.3}{f^2}\,\text{TEC} \qquad \Delta_{\text{gr}}^{\text{Iono}} = \frac{40.3}{f^2}\,\text{TEC} \tag{6.60}$$

as the final result which has the dimension of length. Usually, the TEC is given in TEC units (TECU) where

$$1\,\text{TECU} = 10^{16}\,\text{electrons per m}^2 \,. \tag{6.61}$$

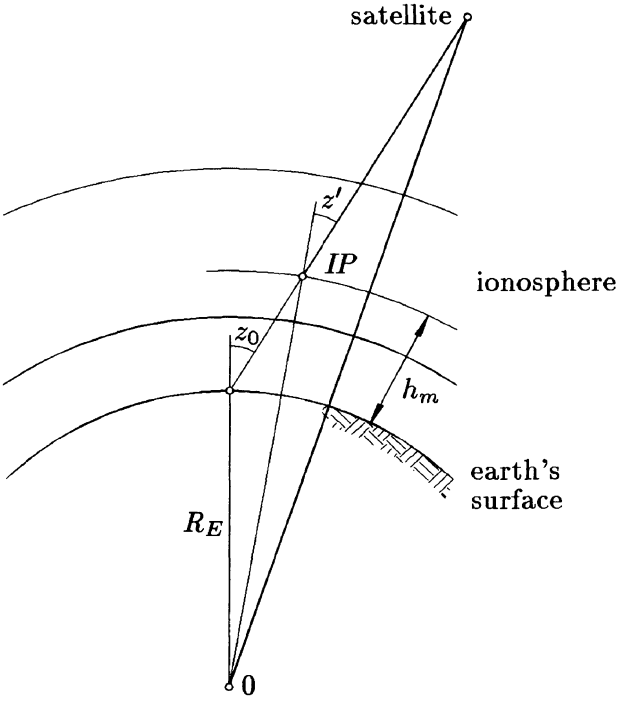


Fig. 6.2. Geometry for the ionospheric path delay

For a numerical example, the C/A-code delay $\Delta_{\mathrm{gr}}^{\mathrm{Iono}} = 0.16\,\mathrm{m}$ is obtained if one TECU is substituted.

Note that TEC as introduced in (6.59) is the total electron content along the signal path between the satellite and the receiver. The integral is assumed to include the electrons in a column with a cross-section of $1\,\mathrm{m}^2$ and extending from the receiver to the satellite. Usually, the total vertical electron content (TVEC) is modeled. More figuratively, this quantity is sometimes denoted as total overhead electron content. If TVEC is introduced in (6.60), the quantities suffice only for satellites at zenith. For arbitrary lines of sight (Fig. 6.2) the zenith angle of the satellite must be taken into account by

$$\Delta_{\mathrm{ph}}^{\mathrm{Iono}} = -\frac{1}{\cos z'}\frac{40.3}{f^2}\,\mathrm{TVEC} \qquad \Delta_{\mathrm{gr}}^{\mathrm{Iono}} = \frac{1}{\cos z'}\frac{40.3}{f^2}\,\mathrm{TVEC} \tag{6.62}$$

since the path length in the ionosphere varies with a changing zenith angle.

These two quantities differ only with respect to the sign. Introducing the notation

$$\Delta^{\mathrm{Iono}} = \frac{1}{\cos z'}\frac{40.3}{f^2}\,\mathrm{TVEC} \tag{6.63}$$

for the (positive) amount of the ionospheric influence on a measured pseudorange allows the omission of the subscripts "ph" or "gr" but requires the consideration of the correct sign for the appropriate models. This means that the ionospheric influence for the code pseudorange is modeled by $+\Delta^{\text{Iono}}$ and for the phase by $-\Delta^{\text{Iono}}$.

Figur 6.2 represents a single-layer model with the assumption that all free electrons are concentrated in an infinitesimally thick spherical shell at the height h_m and containing the ionospheric point. From Fig. 6.2, the relation

$$\sin z' = \frac{R_E}{R_E + h_m} \sin z_0 \tag{6.64}$$

can be seen where R_E is the mean radius of the earth, h_m is a mean value for the height of the ionosphere, and z' and z_0 are the zenith angles at the ionospheric point *IP* and at the observing site. The zenith angle z_0 can be calculated for a known satellite position and approximate coordinates of the observation location. For h_m a value in the range between 300 km and 400 km is typical. The height is only sensitive for low satellite elevations.

As shown by (6.60), the change of range caused by the ionospheric refraction may be restricted to the determination of the TEC. However, the TEC itself is a fairly complicated quantity because it depends on sunspot activities (approximately 11-year cycle), seasonal and diurnal variations, the line of sight which includes elevation and azimuth of the satellite, and the position of the observation site. The TEC may be measured, estimated, its effect computed by models, or eliminated.

Measuring the TEC

Considering as one example Japan, one facility in Tokyo directly measures the TEC. However, since there is a correlation between the TEC and the critical plasma frequency, the TEC may also be calculated by the Japanese ionospheric observatories which make available critical plasma frequency results on an hourly basis. By interpolation, any arbitrary location in Japan can be covered. Many similar experiments are running since "the ionosphere is increasingly used as a laboratory in which active plasma experiments are performed" (Stubbe 1996).

Estimating the TEC

A straightforward estimation of the TEC is to substitute a Taylor series expansion as function of the observation latitude and the local solar time into (6.60). The coefficients in the Taylor series are introduced as unknowns in the pseudorange equations and estimated together with the other unknowns during data processing.

Computing the effect of the TEC
Here, the entire vertical ionospheric refraction is approximated by the model of Klobuchar (1986) and yields the vertical time delay for the code measurements. Although the model is an approximation, it is nevertheless of importance because it uses the ionospheric coefficients broadcast within the fourth subframe of the navigation message (Sect. 5.1.2). The Klobuchar model is

$$\Delta T_v^{\text{Iono}} = A_1 + A_2 \cos\left(\frac{2\pi(t - A_3)}{A_4}\right) \tag{6.65}$$

where

$$\begin{aligned} A_1 &= 5 \cdot 10^{-9} \text{ s} = 5 \text{ ns} \\ A_2 &= \alpha_1 + \alpha_2 \varphi_{IP}^m + \alpha_3 \varphi_{IP}^{m\,2} + \alpha_4 \varphi_{IP}^{m\,3} \\ A_3 &= 14^{\text{h}} \text{ local time} \\ A_4 &= \beta_1 + \beta_2 \varphi_{IP}^m + \beta_3 \varphi_{IP}^{m\,2} + \beta_4 \varphi_{IP}^{m\,3} . \end{aligned} \tag{6.66}$$

The values for A_1 and A_3 are constant, the coefficients α_i, β_i, $i = 1, \ldots, 4$ are uploaded to the satellites and broadcast to the user. The parameter t in (6.65) is the local time of the ionospheric point IP (Fig. 6.2) and may be derived from

$$t = \frac{\lambda_{IP}}{15} + t_{UT} \tag{6.67}$$

where λ_{IP} is the geomagnetic longitude positive to east for the ionospheric point in degrees and t_{UT} is the observation epoch in Universal Time. Finally, φ_{IP}^m in Eq. (6.66) is the spherical distance between the geomagnetic pole and the ionospheric point. Denoting the coordinates of the geomagnetic pole by φ_P, λ_P and those of the ionospheric point by φ_{IP}, λ_{IP}, then φ_{IP}^m is obtained by

$$\cos\varphi_{IP}^m = \sin\varphi_{IP} \sin\varphi_P + \cos\varphi_{IP} \cos\varphi_P \cos(\lambda_{IP} - \lambda_P) \tag{6.68}$$

where the coordinates of the geomagnetic pole are

$$\begin{aligned} \varphi_P &= \ 78.^\circ 3 \text{ N} \\ \lambda_P &= 291.^\circ 0 \text{ E} . \end{aligned} \tag{6.69}$$

Summarizing, the evaluation of the Klobuchar model may be performed by the following steps:

- Compute the azimuth a and the zenith angle z_0 of the satellite for epoch t_{UT}.

- Choose a mean height of the ionosphere and compute the distance s between the observing site and the ionospheric point obtained from the triangle origin – observation site – ionospheric point (Fig. 6.2).
- Compute the coordinates φ_{IP}, λ_{IP} of the ionospheric point by means of the quantities a, z, s.
- Calculate φ_{IP}^m from (6.68).
- Calculate A_2 and A_4 from (6.66) where the coefficients α_i, β_i, $i = 1,\ldots,4$ are received via the satellite navigation message.
- Use (6.66) and (6.67) and compute the vertical (or zenith) delay ΔT_v^{Iono} by (6.65).
- By calculating z' from (6.64) and applying $\Delta T^{\text{Iono}} = \frac{1}{\cos z'} \Delta T_v^{\text{Iono}}$, the transition from the vertical delay to the delay along the wave path is achieved. The result is obtained as a time delay in seconds which must be multiplied by the speed of light to convert it to a change of range.

Eliminating the effect of the TEC

It is difficult to find a satisfying model for the TEC because of the various time dependent influences. The most efficient method, thus, is to eliminate the ionospheric refraction by using two signals with different frequencies. This dual frequency method is the main reason why the GPS signal has two carrier waves L1 and L2.

Starting with the phase pseudorange model (6.9) and adding the frequency dependent ionospheric refraction (6.63) gives

$$\begin{aligned} \lambda_{\text{L1}}\Phi_{\text{L1}} &= \varrho + c\,\Delta\delta + \lambda_{\text{L1}}N_{\text{L1}} - \Delta^{\text{Iono}}(f_{\text{L1}}) \\ \lambda_{\text{L2}}\Phi_{\text{L2}} &= \varrho + c\,\Delta\delta + \lambda_{\text{L2}}N_{\text{L2}} - \Delta^{\text{Iono}}(f_{\text{L2}}) \end{aligned} \tag{6.70}$$

where the frequencies of the two carriers are denoted by f_{L1} and f_{L2}. After dividing by the corresponding wavelengths,

$$\begin{aligned} \Phi_{\text{L1}} &= \frac{1}{\lambda_{\text{L1}}}\varrho + \frac{c}{\lambda_{\text{L1}}}\Delta\delta + N_{\text{L1}} - \frac{1}{\lambda_{\text{L1}}}\Delta^{\text{Iono}}(f_{\text{L1}}) \\ \Phi_{\text{L2}} &= \frac{1}{\lambda_{\text{L2}}}\varrho + \frac{c}{\lambda_{\text{L2}}}\Delta\delta + N_{\text{L2}} - \frac{1}{\lambda_{\text{L2}}}\Delta^{\text{Iono}}(f_{\text{L2}}) \end{aligned} \tag{6.71}$$

are obtained. Using the relation $c = f\,\lambda$ yields

$$\begin{aligned} \Phi_{\text{L1}} &= \frac{f_{\text{L1}}}{c}\varrho + f_{\text{L1}}\,\Delta\delta + N_{\text{L1}} - \frac{f_{\text{L1}}}{c}\Delta^{\text{Iono}}(f_{\text{L1}}) \\ \Phi_{\text{L2}} &= \frac{f_{\text{L2}}}{c}\varrho + f_{\text{L2}}\,\Delta\delta + N_{\text{L2}} - \frac{f_{\text{L2}}}{c}\Delta^{\text{Iono}}(f_{\text{L2}}) \end{aligned} \tag{6.72}$$

which can be written in the form

$$\begin{aligned} \Phi_{L1} &= a f_{L1} + N_{L1} - \frac{b}{f_{L1}} \\ \Phi_{L2} &= a f_{L2} + N_{L2} - \frac{b}{f_{L2}} \end{aligned} \tag{6.73}$$

by introducing

$$\begin{aligned} a &= \frac{\varrho}{c} + \Delta\delta && \text{geometry term} \\ b &= \frac{f^2}{c}\Delta^{\text{Iono}} = \frac{1}{c}\frac{40.3}{\cos z'}\,\text{TVEC} && \text{ionosphere term} \end{aligned} \tag{6.74}$$

where the second expression for b may be verified by substituting Eq. (6.63).

The ionosphere term can be eliminated by the following linear combination. Multiplying the first equation of (6.73) by f_{L1} and the second by f_{L2} and forming the difference yields

$$\Phi_{L1} f_{L1} - \Phi_{L2} f_{L2} = a(f_{L1}^2 - f_{L2}^2) + N_{L1} f_{L1} - N_{L2} f_{L2} \tag{6.75}$$

and, after multiplying the equation by $f_{L1}/(f_{L1}^2 - f_{L2}^2)$ and a slight rearrangement, the so-called ionosphere-free combination

$$\left[\Phi_{L1} - \frac{f_{L2}}{f_{L1}}\Phi_{L2}\right]\frac{f_{L1}^2}{f_{L1}^2 - f_{L2}^2} = a f_{L1} + \left[N_{L1} - \frac{f_{L2}}{f_{L1}}N_{L2}\right]\frac{f_{L1}^2}{f_{L1}^2 - f_{L2}^2} \tag{6.76}$$

is obtained. The significant drawback of the combination is that the integer nature of the ambiguities is lost. Note that on the left side of the equation the geometric residual reappears, cf. Eq. (6.19). Thus, this quantity could also be denoted as reduced ionosphere-free signal.

The derivation of the ionosphere-free combination for code pseudoranges starts with the model equations

$$\begin{aligned} R_{L1} &= \varrho + c\,\Delta\delta + \Delta^{\text{Iono}}(f_{L1}) \\ R_{L2} &= \varrho + c\,\Delta\delta + \Delta^{\text{Iono}}(f_{L2}) \end{aligned} \tag{6.77}$$

where Δ^{Iono} is inversely proportional to the squared respective carrier frequency, cf. Eq. (6.63). Thus, multiplying the first equation of (6.77) by f_{L1}^2 and the second by f_{L2}^2 and then forming the difference yields

$$R_{L1} f_{L1}^2 - R_{L2} f_{L2}^2 = (f_{L1}^2 - f_{L2}^2)(\varrho + c\,\Delta\delta) \tag{6.78}$$

where the ionospheric term is eliminated. After dividing the equation by $(f_{L1}^2 - f_{L2}^2)$ and a slight rearrangement, the ionosphere-free combination

$$\left[R_{L1} - \frac{f_{L2}^2}{f_{L1}^2} R_{L2}\right] \frac{f_{L1}^2}{f_{L1}^2 - f_{L2}^2} = \varrho + c\,\Delta\delta \tag{6.79}$$

is obtained.

The advantage of the ionosphere-free combination is the elimination (or more precisely, the reduction) of ionospheric effects. Remembering the derivation, it should be clear that the term "ionosphere-free" is not fully correct because there are some approximations involved, for instance Eq. (6.47) or the integration is not carried out along the true signal path in (6.57). Brunner and Gu (1991) propose an improved model to account for the higher-order terms arising from the series expansion of the refractive index, the geomagnetic field effect, and the bending effects of the ray paths.

6.3.3 Tropospheric refraction

The effect of the neutral atmosphere (i.e., the nonionized part) is denoted as tropospheric refraction, tropospheric path delay, or simply tropospheric delay. The naming is slightly incorrect because it excludes the stratosphere which is another constituent of the neutral atmosphere. However, the dominant contribution of the troposphere explains the choice of the name.

The neutral atmosphere is a nondispersive medium with respect to radio waves up to frequencies of 15 GHz. Thus, the propagation is frequency independent. Consequently, a distinction between carrier phases and code ranges derived from different carriers L1 or L2 is not necessary. The disadvantage is that an elimination of the tropospheric refraction by dual frequency methods is not possible.

The tropospheric path delay is defined by

$$\Delta^{\text{Trop}} = \int (n-1)\, ds \tag{6.80}$$

which is analogous to the ionospheric formula (6.54). Again an approximation is introduced so that the integration is performed along the geometric path of the signal. Usually, instead of the refractive index n the refractivity

$$N^{\text{Trop}} = 10^6 (n-1) \tag{6.81}$$

is used so that Eq. (6.80) becomes

$$\Delta^{\text{Trop}} = 10^{-6} \int N^{\text{Trop}}\, ds\,. \tag{6.82}$$

Hopfield (1969) shows the possibility of separating N^{Trop} into a dry and a wet component

$$N^{\text{Trop}} = N_d^{\text{Trop}} + N_w^{\text{Trop}} \tag{6.83}$$

where the dry part results from the dry atmosphere and the wet part from the water vapor. Correspondingly, the relations

$$\Delta_d^{\text{Trop}} = 10^{-6} \int N_d^{\text{Trop}}\, ds \tag{6.84}$$

$$\Delta_w^{\text{Trop}} = 10^{-6} \int N_w^{\text{Trop}}\, ds \tag{6.85}$$

and

$$\begin{aligned} \Delta^{\text{Trop}} &= \Delta_d^{\text{Trop}} + \Delta_w^{\text{Trop}} \\ &= 10^{-6} \int N_d^{\text{Trop}}\, ds + 10^{-6} \int N_w^{\text{Trop}}\, ds \end{aligned} \tag{6.86}$$

are obtained. About 90% of the tropospheric refraction arise from the dry and about 10% from the wet component. In practice, models for the refractivities are introduced in Eq. (6.86) and the integration is performed by numerical methods or analytically after, e.g., series expansions of the integrand. Models for the dry and wet refractivity at the surface of the earth have been known for some time (e.g., Essen and Froome 1951). The corresponding dry component is

$$N_{d,0}^{\text{Trop}} = \bar{c}_1 \frac{p}{T}, \qquad \bar{c}_1 = 77.64 \ \text{K mb}^{-1} \tag{6.87}$$

where p is the atmospheric pressure in millibar (mb) and T is the temperature in Kelvin (K). The wet component was found to be

$$\begin{aligned} N_{w,0}^{\text{Trop}} &= \bar{c}_2 \frac{e}{T} + \bar{c}_3 \frac{e}{T^2}, \qquad \bar{c}_2 = -12.96 \ \text{K mb}^{-1} \\ &\qquad\qquad\qquad\qquad\quad \bar{c}_3 = 3.718 \cdot 10^5 \ \text{K}^2\,\text{mb}^{-1} \end{aligned} \tag{6.88}$$

where e is the partial pressure of water vapor in mb and T again the temperature in K. The overbar in the coefficients only stresses that there is absolutely no relationship to the coefficients for the ionosphere in, e.g., (6.57).

The values for $\bar{c}_1$, $\bar{c}_2$, and $\bar{c}_3$ are empirically determined and, certainly, cannot fully describe the local situation. An improvement is obtained by measuring meteorological data at the observation site. The following paragraphs present several models where meteorological surface data are taken into account.

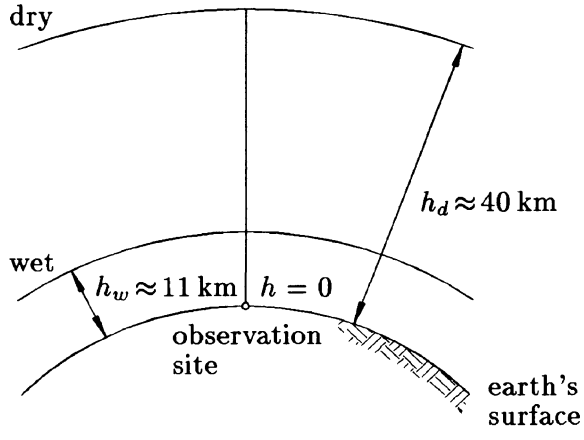


Fig. 6.3. Thickness of polytropic layers for the troposphere

Hopfield model
Using real data covering the whole earth, Hopfield (1969) has empirically found a representation of the dry refractivity as a function of the height h above the surface by

$$N_d^{\text{Trop}}(h) = N_{d,0}^{\text{Trop}} \left[\frac{h_d - h}{h_d}\right]^4 \tag{6.89}$$

under the assumption of a polytropic layer with thickness

$$h_d = 40\,136 + 148.72\,(T - 273.16) \quad [\text{m}]\,, \tag{6.90}$$

as shown in Fig. 6.3. Substitution of (6.89) into (6.84) yields (for the dry part) the tropospheric path delay

$$\Delta_d^{\text{Trop}} = 10^{-6}\, N_{d,0}^{\text{Trop}} \int \left[\frac{h_d - h}{h_d}\right]^4 ds\,. \tag{6.91}$$

The integral can be solved if the delay is calculated along the vertical direction and if the curvature of the signal path is neglected. Thus, for an observation site on the surface of the earth (i.e., $h = 0$), Eq. (6.91) becomes

$$\Delta_d^{\text{Trop}} = 10^{-6}\, N_{d,0}^{\text{Trop}} \frac{1}{{h_d}^4} \int_{h=0}^{h=h_d} (h_d - h)^4\, dh \tag{6.92}$$

where the constant denominator has been extracted. After integration,

$$\Delta_d^{\text{Trop}} = 10^{-6}\, N_{d,0}^{\text{Trop}} \frac{1}{{h_d}^4} \left[-\frac{1}{5}(h_d - h)^5 \Big|_{h=0}^{h=h_d}\right] \tag{6.93}$$

is obtained. The evaluation of the expression between the brackets gives $h_d^5/5$ so that

$$\Delta_d^{\text{Trop}} = \frac{10^{-6}}{5} N_{d,0}^{\text{Trop}} h_d \tag{6.94}$$

is the dry portion of the tropospheric zenith delay.

The wet portion is much more difficult to model because of the strong variations of the water vapor with respect to time and space. Nevertheless, due to lack of an appropriate alternative, the Hopfield model assumes the same functional model for both the wet and dry components. Thus,

$$N_w^{\text{Trop}}(h) = N_{w,0}^{\text{Trop}} \left[\frac{h_w - h}{h_w}\right]^4 \tag{6.95}$$

where the mean value

$$h_w = 11\,000 \text{ m} \tag{6.96}$$

is used. Sometimes other values such as $h_w = 12\,000\,\text{m}$ have been proposed. Unique values for h_d and h_w cannot be given because of their dependence on location and temperature. In Germany, a local model for estimating the tropospheric path delay at microwave frequencies using radio sonde data over 4.5 years yielded for the region of the observation site $h_d = 41.6\,\text{km}$ and $h_w = 11.5\,\text{km}$. The effective troposphere heights are given as $40\,\text{km} \leq h_d \leq 45\,\text{km}$ and $10\,\text{km} \leq h_w \leq 13\,\text{km}$.

The integration of (6.95) is completely analogous to (6.91) and results in

$$\Delta_w^{\text{Trop}} = \frac{10^{-6}}{5} N_{w,0}^{\text{Trop}} h_w \,. \tag{6.97}$$

Therefore, the total tropospheric zenith delay is

$$\Delta^{\text{Trop}} = \frac{10^{-6}}{5} \left[N_{d,0}^{\text{Trop}} h_d + N_{w,0}^{\text{Trop}} h_w\right] \tag{6.98}$$

with the dimension meters. The model in its present form does not account for an arbitrary zenith angle of the signal. Considering the line of sight, an obliquity factor must be applied which, in its simplest form, is the projection from the zenith onto the line of sight. Frequently, the transition of the zenith delay to a delay with arbitrary zenith angle is denoted as the application of a mapping function.

Introducing the mapping function, Eq. (6.98) becomes

$$\Delta^{\text{Trop}} = \frac{10^{-6}}{5} \left[N_{d,0}^{\text{Trop}} h_d \, m_d(E) + N_{w,0}^{\text{Trop}} h_w \, m_w(E)\right] \tag{6.99}$$

where $m_d(E)$ and $m_w(E)$ are the mapping functions for the dry and the wet part and E (expressed in degrees) indicates the elevation at the observing site (where the line of sight is simplified as straight line). Explicitly,

$$\begin{aligned} m_d(E) &= \frac{1}{\sin\sqrt{E^2+6.25}} \\ m_w(E) &= \frac{1}{\sin\sqrt{E^2+2.25}} \end{aligned} \tag{6.100}$$

are the mapping functions for the Hopfield model. In more compact form, Eq. (6.99) can be represented as

$$\Delta^{\text{Trop}}(E) = \Delta_d^{\text{Trop}}(E) + \Delta_w^{\text{Trop}}(E) \tag{6.101}$$

where the terms on the right side of the equation are given by

$$\begin{aligned} \Delta_d^{\text{Trop}}(E) &= \frac{10^{-6}}{5} \frac{N_{d,0}^{\text{Trop}}\, h_d}{\sin\sqrt{E^2+6.25}} \\ \Delta_w^{\text{Trop}}(E) &= \frac{10^{-6}}{5} \frac{N_{w,0}^{\text{Trop}}\, h_w}{\sin\sqrt{E^2+2.25}} \end{aligned} \tag{6.102}$$

or, after substituting (6.87), (6.90) and (6.88), (6.96) respectively, by

$$\begin{aligned} \Delta_d^{\text{Trop}}(E) &= \frac{10^{-6}}{5} \frac{77.64\,\frac{p}{T}}{\sin\sqrt{E^2+6.25}} \left[40\,136 + 148.72\,(T-273.16)\right] \\ \Delta_w^{\text{Trop}}(E) &= \frac{10^{-6}}{5} \frac{-12.96\,T + 3.718\cdot 10^5}{\sin\sqrt{E^2+2.25}} \frac{e}{T^2}\, 11\,000\,. \end{aligned} \tag{6.103}$$

Measuring p, T, e at the observation location and calculating the elevation angle E, the total tropospheric path delay is obtained in meters by (6.101) after evaluating (6.103).

Modified Hopfield models

The empirical function (6.89) is now rewritten by introducing lengths of position vectors instead of heights. Denoting the radius of the earth by R_E, the corresponding lengths are $r_d = R_E + h_d$ and $r = R_E + h$ (Fig. 6.4). Thus, the dry refractivity in the form

$$N_d^{\text{Trop}}(r) = N_{d,0}^{\text{Trop}} \left[\frac{r_d - r}{r_d - R_E}\right]^4 \tag{6.104}$$

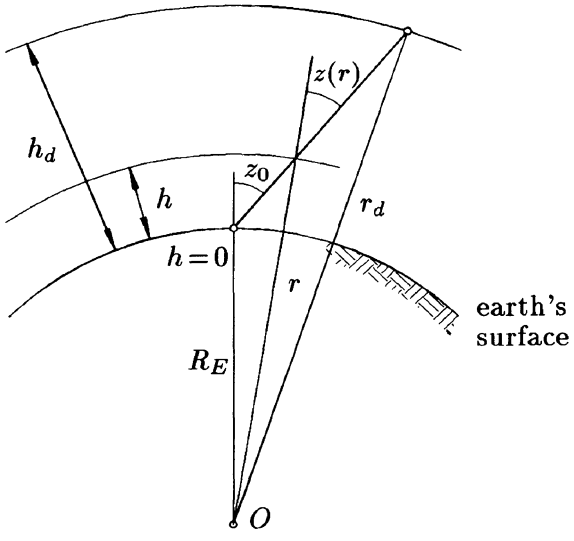


Fig. 6.4. Geometry for the tropospheric path delay

is equivalent to (6.89). Applying Eq. (6.84) and introducing a mapping function gives

$$\Delta_d^{\text{Trop}}(z) = 10^{-6} \int_{r=R_E}^{r=r_d} N_d^{\text{Trop}}(r) \frac{1}{\cos z(r)}\, dr \tag{6.105}$$

for the dry path delay. Note that the zenith angle $z(r)$ is variable. Denoting the zenith angle at the observation site by z_0, the sine-law

$$\sin z(r) = \frac{R_E}{r} \sin z_0 \tag{6.106}$$

can be applied (Fig. 6.4). From Eq. (6.106) follows

$$\cos z(r) = \sqrt{1 - \frac{R_E^2}{r^2} \sin^2 z_0} \tag{6.107}$$

which is equivalent to

$$\cos z(r) = \frac{1}{r} \sqrt{r^2 - R_E^2 \sin^2 z_0}\,. \tag{6.108}$$

Substituting (6.108) and (6.104) into (6.105) yields

$$\Delta_d^{\text{Trop}}(z) = \frac{10^{-6} N_{d,0}^{\text{Trop}}}{(r_d - R_E)^4} \int_{r=R_E}^{r=r_d} \frac{r\,(r_d - r)^4}{\sqrt{r^2 - R_E^2 \sin^2 z_0}}\, dr \tag{6.109}$$

where the terms being constant with respect to the integration variable r have been extracted from the integral. Assuming the same model for the wet portion, the corresponding formula is given by

$$\Delta_w^{\text{Trop}}(z) = \frac{10^{-6}\, N_{w,0}^{\text{Trop}}}{(r_w - R_E)^4} \int_{r=R_E}^{r=r_w} \frac{r\,(r_w - r)^4}{\sqrt{r^2 - R_E^2 \sin^2 z_0}}\, dr\,. \tag{6.110}$$

Instead of the zenith angle z the elevation angle $E = 90° - z$ could also be used. Many modified Hopfield models have been derived, depending solely on the method to solve the integral. Here, one model is presented based on a series expansion of the integrand. The resulting formulas can be found, e.g., in Remondi (1984) where a subscript i is introduced which reflects either the dry component (replace i by d) or the wet component (replace i by w). With

$$r_i = \sqrt{(R_E + h_i)^2 - (R_E \cos E)^2} - R_E \sin E \tag{6.111}$$

the tropospheric delay in meters is

$$\Delta_i^{\text{Trop}}(E) = 10^{-12}\, N_{i,0}^{\text{Trop}} \left[\sum_{k=1}^{9} \frac{\alpha_{k,i}}{k}\, r_i^k \right] \tag{6.112}$$

where

$$\begin{aligned}
\alpha_{1,i} &= 1 & \alpha_{6,i} &= 4a_i\, b_i\, (a_i^2 + 3b_i) \\
\alpha_{2,i} &= 4a_i & \alpha_{7,i} &= b_i^2\, (6a_i^2 + 4b_i) \\
\alpha_{3,i} &= 6a_i^2 + 4b_i & \alpha_{8,i} &= 4a_i\, b_i^3 \\
\alpha_{4,i} &= 4a_i\, (a_i^2 + 3b_i) & \alpha_{9,i} &= b_i^4 \\
\alpha_{5,i} &= a_i^4 + 12a_i^2\, b_i + 6b_i^2
\end{aligned} \tag{6.113}$$

and

$$a_i = -\frac{\sin E}{h_i}\,, \qquad b_i = -\frac{cos^2 E}{2h_i\, R_E}\,. \tag{6.114}$$

Substituting $i = d$, the dry part results where in (6.112) for $N_{d,0}^{\text{Trop}}$ Eq. (6.87) and for h_d Eq. (6.90) must be introduced. Analogously, Eqs. (6.88) and (6.96) must be used for $N_{w,0}^{\text{Trop}}$ and for h_w.

Saastamoinen model

The refractivity can alternatively be deduced from gas laws. The Saastamoinen model is based on this approach where again some approximations have been employed (Saastamoinen 1973). Here, any theoretical derivation is omitted. Saastamoinen models the tropospheric delay, expressed in meters,

$$\Delta^{\text{Trop}} = \frac{0.002277}{\cos z}\left[p + \left(\frac{1255}{T} + 0.05\right)e - \tan^2 z\right] \tag{6.115}$$

as a function of z, p, T and e. As before, z denotes the zenith angle of the satellite, p the atmospheric pressure in millibar, T the temperature in Kelvin, and e the partial pressure of water vapor in millibar. A numerical assessment using parameters of a standard atmosphere at sea level results in a tropospheric zenith delay of about 2.3 m.

Saastamoinen has refined this model by adding two correction terms, one being dependent on the height of the observing site and the other on the height and on the zenith angle. Another refined formula is

$$\Delta^{\text{Trop}} = \frac{0.002277}{\cos z}\left[p + \left(\frac{1255}{T} + 0.05\right)e - B\,\tan^2 z\right] + \delta R \tag{6.116}$$

where the correction terms B, δR are interpolated from Tables 6.3 and 6.4.

Table 6.3. Correction term B for the refined Saastamoinen model

Height [km]	B [mb]
0.0	1.156
0.5	1.079
1.0	1.006
1.5	0.938
2.0	0.874
2.5	0.813
3.0	0.757
4.0	0.654
5.0	0.563

Table 6.4. Correction term δR [m] for refined Saastamoinen model

Zenith angle	Station height above sea level [km]							
	0	0.5	1.0	1.5	2.0	3.0	4.0	5.0
60°00′	0.003	0.003	0.002	0.002	0.002	0.002	0.001	0.001
66°00′	0.006	0.006	0.005	0.005	0.004	0.003	0.003	0.002
70°00′	0.012	0.011	0.010	0.009	0.008	0.006	0.005	0.004
73°00′	0.020	0.018	0.017	0.015	0.013	0.011	0.009	0.007
75°00′	0.031	0.028	0.025	0.023	0.021	0.017	0.014	0.011
76°00′	0.039	0.035	0.032	0.029	0.026	0.021	0.017	0.014
77°00′	0.050	0.045	0.041	0.037	0.033	0.027	0.022	0.018
78°00′	0.065	0.059	0.054	0.049	0.044	0.036	0.030	0.024
78°30′	0.075	0.068	0.062	0.056	0.051	0.042	0.034	0.028
79°00′	0.087	0.079	0.072	0.065	0.059	0.049	0.040	0.033
79°30′	0.102	0.093	0.085	0.077	0.070	0.058	0.047	0.039
79°45′	0.111	0.101	0.092	0.083	0.076	0.063	0.052	0.043
80°00′	0.121	0.110	0.100	0.091	0.083	0.068	0.056	0.047

Models using the mapping function of Marini

In 1972, Marini developed a continued fraction of the mapping function. Herring (1992) specified this function with three constants and normalized to unity at the zenith. For the dry component, the mapping function

$$m_d(E) = \frac{1 + \dfrac{a_d}{1 + \dfrac{b_d}{1 + c_d}}}{\sin E + \dfrac{a_d}{\sin E + \dfrac{b_d}{\sin E + c_d}}} \tag{6.117}$$

is used where the coefficients are defined as

$$\begin{aligned} a_d &= [1.2320 + 0.0139 \cos\varphi - 0.0209\, h + 0.00215\,(T - 283)] \cdot 10^{-3} \\ b_d &= [3.1612 - 0.1600 \cos\varphi - 0.0331\, h + 0.00206\,(T - 283)] \cdot 10^{-3} \\ c_d &= [71.244 - 4.293 \cos\varphi - 0.149\, h - 0.0021\,(T - 283)] \cdot 10^{-3} \end{aligned} \tag{6.118}$$

depending on the latitude φ and height h of the observing site and on the temperature T in Kelvin.

For the wet part, the mapping function is the same as in (6.117) but the subscript d must be replaced by w. The corresponding coefficients are obtained as

$$\begin{aligned} a_w &= [0.583 - 0.011\cos\varphi - 0.052\,h + 0.0014\,(T - 283)] \cdot 10^{-3} \\ b_w &= [1.402 - 0.102\cos\varphi - 0.101\,h + 0.0020\,(T - 283)] \cdot 10^{-3} \\ c_w &= [45.85 - 1.91\cos\varphi - 1.29\,h + 0.015\,(T - 283)] \cdot 10^{-3}\,. \end{aligned} \tag{6.119}$$

Niell (1996) uses the same type of mapping function as Herring, i.e., the continued fraction of the Marini mapping function is restricted to three coefficients. The coefficients for the dry part depend on the latitude and the height at the observing site and on the day of the year, whereas the coefficients for the wet part depend only on the site latitude. Numerical values of the coefficients are given for some specific latitudes in Niell (1996). Interpolation must be used to obtain the coefficients for arbitrary latitudes and days.

The transition to tropospheric models is achieved by substituting the mapping functions given in this paragraph into models for the zenith delay, e.g., Eq. (6.99).

Tropospheric problems

There are many other tropospheric models which are similar to the models given here. Janes et al. (1991) and Spilker (1996b) analyze several other tropospheric models. The question arises why there are so many different approaches. One reason is the difficulty in modeling the water vapor. The simple use of surface measurements cannot give the utmost accuracy so that water vapor radiometers have been developed. These instruments measure the sky brightness temperature by radiometric microwave observations along the signal path enabling the calculation of the wet path delay. Accurate water vapor radiometers are expensive and experience problems at low elevation angles since the tropospheric zenith delay is amplified by the mapping function.

The difficulty in modeling the tropospheric effect will require continuation of research and development for some years. One solution is to combine surface and radio sonde meteorological data, water vapor radiometer measurements and statistics. This is a major task and an appropriate model has not yet been found.

Any standard model suffers from the estimation of the zenith delay from measured ground parameters. Another approach is to estimate the zenith

delay in the least squares adjustment of the phase observations. Some processing software offers this option. Usually one zenith delay for each site and session is estimated; however, it is good practice to estimate more than one zenith delay per session (Brunner and Welsch 1993).

6.3.4 Atmospheric monitoring

Ionospheric tomography
Tomography has developed from a medical diagnostics tool, commonly denoted as computer tomography, to become an imaging technique for many applications, including geodesy and geophysics (Leitinger 1996). Referring to ionospheric tomography, the line integral of electron density, i.e., the TEC, is measured over a large number of ray paths transitioning the ionosphere. This dataset is inverted to produce an image of electron density in ionosphere maps.

More GPS-specifically, TEC monitoring using satellite positioning systems is possible (Jakowski 1996). As one very representative example, the Center for Orbit Determination in Europe (CODE) estimates Global Ionosphere Maps (GIMs) as an additional product since January 1, 1996. The main idea is to analyze the ionospheric residuals of dual frequency phases which contain the information on ionospheric refraction. Following closely Schaer (1997, 1999), the TEC is developed into a series of spherical harmonics adopting a single-layer model in a sun-fixed reference frame. For each day, a set of TEC coefficients is determined which approximates an average distribution of the vertical TEC on a global scale. The GIMs produced may contribute to improve the ambiguity resolution, as demonstrated in the CODE processing. Also spaceborne applications, e.g., altimetry, may benefit from the TEC maps. For ionosphere physicists, these maps are an alternative source of information about the deterministic behavior of the ionosphere that may be correlated with solar and geomagnetic parameters and compared to theoretical ionosphere models. All details on how to use the maps are given under http://www.cx.unibe.ch/aiub/igs.html.

The GIMs are based on the single-layer model in Fig. 6.2. The electron density of the surface is modeled by

$$\mathrm{TVEC}(\beta, \Delta\lambda) = \sum_{n=0}^{n_{\max}} \sum_{m=0}^{n} [a_{nm} \cos m\Delta\lambda + b_{nm} \sin m\Delta\lambda] \bar{P}_{nm}(\sin\beta) \tag{6.120}$$

yielding the total vertical electron content as a function of the ionospheric point expressed by the geocentric latitude β and the longitude difference

$\Delta\lambda = \lambda - \lambda_0$ between the earth-fixed longitude λ and the longitude of the sun λ_0. The coefficients a_{nm} and b_{nm} are the coefficients to be determined representing the parameters of the GIM. Finally, $\bar{P}_{nm}(\sin\beta)$ are the fully normalized associated Legendre functions of degree n and order m.

If the solar-geographical reference frame refers to the mean sun, the geographic longitude of the sun may be written in function of the Universal Time (UT)

$$\lambda_0 = 12^{\mathrm{h}} - \mathrm{UT} \tag{6.121}$$

and the geographic latitude of the sun is set to zero. Note that $\mathrm{TVEC}(\beta, \Delta\lambda)$ may equivalently be expressed in the solar-geomagnetic frame.

The global ionosphere maps are generated on a daily basis by CODE. The TEC (more precisely TVEC) is modeled with a spherical harmonic expansion up to degree $n = 12$ and order $m = 8$ referring to a solar-geomagnetic reference frame. Per day twelve 2-hour sets are derived from GPS data of the global IGS network. From Schaer (1997) some statistical values are given: the maximum and minimum values in TEC units (TECU) for day 73 of 1996 are

$$\mathrm{TVEC}_{\mathbf{max}}(\beta, \Delta\lambda) = \mathrm{TVEC}(-7.60°, 45.37°) \quad = 35.79\,\mathrm{TECU}$$

$$\mathrm{TVEC}_{\mathbf{min}}(\beta, \Delta\lambda) = \mathrm{TVEC}(60.91°, -106.64°) = \ 0.34\,\mathrm{TECU}$$

and the mean TECU, averaged from the one-day GIM, roughly describes the evolution of the ionospheric activity in a global sense and varies for a 28-month time span starting with January 1, 1995 between about 6 and 18 TECU.

Apart from the global ionosphere maps, for Europe also regional ionosphere maps based on some 30 European IGS stations are provided. The application is restricted to the corresponding definition area.

Troposphere sounding

Reliable information on global climate change processes over future decades and better weather forecasting on near- and medium-term time scales are only possible on the basis of global and regional data records. These data are used to accurately model atmospheric state parameters with high spatial and temporal resolution.

Water vapor is one of the most significant constituents of the troposphere. This parameter plays a fundamental role with regard to weather and climate since it has the capability to transport moisture and heat through the atmosphere. Meteorologists have started to use GPS as a low-cost tool for measuring the water vapor. Thereby, tropospheric refraction (in the past considered as nuisance parameter) has become a well appreciated signal.

The tropospheric zenith delay is estimated during data processing. The dry component can be computed with high accuracy, the remaining wet component is a function of the water vapor in the atmosphere. Short-periodic variations of the Integrated Water Vapor (IWV) improve numerical weather prediction whereas long-periodic variations have impact on climate research. More details on the subject are found in Bevis et al. (1992), Gendt et al. (1999).

CHAMP mission
The Challenging Mini-satellite Payload (CHAMP) is used for geophysical research and application. The mission started in 2000 and is scheduled to last five years in order to provide a sufficiently long observation time to resolve long-term temporal variations in the magnetic field, in the gravity field, and within the atmosphere. Some specifications of the satellite of this mission: altitude 300–470 km, inclination 87.3 degree, eccentricity 0.001.

Among other equipment, the satellite has a dual frequency GPS receiver on board. The measurable refractional effects on GPS signals propagating through the atmospheric limb may be used to derive profiles for a variety of atmospheric parameters. Ionospheric refraction is used for the derivation of electron density in profiles between 60 km and the CHAMP orbital height. In conjunction with TEC measurements from a network of terrestrial stations, a comprehensive model of the ionosphere is possible with high resolution in space and time. Refractional effects in the atmosphere ranging from the earth surface up to about 60 km altitude (bending of signal path, tropospheric path delay) give rise to profiles for meteorological parameters (density, pressure, temperature, water vapor). The final objective of the CHAMP mission is the determination of all these parameters in (near) real time.

More on CHAMP may be found under http://op.gfz-potsdam.de/champ of the GeoForschungsZentrum Potsdam, Germany, where the information of this section has been extracted from.

6.4 Relativistic effects

6.4.1 Special relativity

Lorentz transformation
Consider two four-dimensional systems $S(x, y, z, t)$ and $S'(x', y', z', t')$ where the union of space coordinates x, y, z and the time coordinate t is characterized by space-time coordinates. The system S is at rest and, relative to S, the system S' is uniformly translating with velocity v. For simplicity, it

is assumed that both systems coincide at an initial epoch $t = 0$ and that the translation takes place along the x-axis.

The transformation of the space-time coordinates is given by

$$\begin{aligned} x' &= \frac{x - v\,t}{\sqrt{1 - \frac{v^2}{c^2}}} \\ y' &= y \\ z' &= z \\ t' &= \frac{t - \frac{v}{c^2}\,x}{\sqrt{1 - \frac{v^2}{c^2}}} \end{aligned} \tag{6.122}$$

where c is the speed of light. Note that the equations above describe the moving system S' with respect to the system S at rest (more figuratively: as viewed from the moving system). Equivalently, the system S at rest may be described with respect to the moving system S' (more figuratively: as viewed from the system at rest). The corresponding formulas follow by solving Eq. (6.122) for the space-time coordinates in the system S at rest or simply by interchanging the role of the primed and the unprimed coordinates and reversing the sign of the velocity v. Thus, the relations

$$\begin{aligned} x &= \frac{x' + v\,t'}{\sqrt{1 - \frac{v^2}{c^2}}} \\ y &= y' \\ z &= z' \\ t &= \frac{t' + \frac{v}{c^2}\,x'}{\sqrt{1 - \frac{v^2}{c^2}}} \end{aligned} \tag{6.123}$$

are obtained. Equations (6.122) and (6.123) are known as Lorentz transformation. An elegant and simple derivation of these formulas can be found in Joos (1956), p. 217. Using Eqs. (6.122) or (6.123), the relation

$$x^2 + y^2 + z^2 - c^2t^2 = x'^2 + y'^2 + z'^2 - c^2t'^2 \tag{6.124}$$

may be verified. This means that the norm of a vector in space-time coordinates is invariant with respect to the choice of its reference system. Note

that in the case of $c = \infty$, the Lorentz transformation (6.122) converts to the Galilei transformation

$$\begin{aligned} x' &= x - v\,t \\ y' &= y \\ z' &= z \\ t' &= t \end{aligned} \tag{6.125}$$

which is fundamental in classical Newtonian mechanics.

The theory of special relativity is, by definition, restricted to inertial systems. The application of the Lorentz transformation reveals some features of that theory.

Time dilation

Consider an observer moving with the system S'. At a specific location x' the time events t'_1 and t'_2 are recorded. The corresponding time events t_1 and t_2 in the system S at rest, according to Lorentz transformation (6.123), are

$$t_1 = \frac{t'_1 + \frac{v}{c^2}\,x'}{\sqrt{1 - \frac{v^2}{c^2}}} \qquad t_2 = \frac{t'_2 + \frac{v}{c^2}\,x'}{\sqrt{1 - \frac{v^2}{c^2}}}\,. \tag{6.126}$$

The time interval $\Delta t' = t'_2 - t'_1$ in the moving system is called proper time and the time interval $\Delta t = t_2 - t_1$ in the system at rest is called coordinate time. The relation between proper and coordinate time is found by the difference of the two expressions in (6.126)

$$\Delta t = \frac{\Delta t'}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{6.127}$$

which means that as viewed at the system at rest, the time interval recorded by the moving observer is lengthened or dilated. The same holds for the inverse situation: a time interval recorded by an observer in the system at rest is dilated for an observer in the moving system. The result $\Delta t' = \Delta t/\sqrt{1 - v^2/c^2}$ may be verified by the reader by using Eqs. (6.122) or is simply obtained by interchanging the role of the primed and the unprimed coordinates in (6.127) (reversing the sign of v has here no effect because this quantity is squared). The time dilation is the reason why moving clocks run slower than clocks at rest.

Lorentz contraction

The derivation of the Lorentz contraction is similar to that of time dilation.

Consider now two locations x_1' and x_2' in the moving system S' at a specific epoch t'. The corresponding locations x_1 and x_2 in the system S at rest, according to the Lorentz transformation (6.123), are

$$x_1 = \frac{x_1' + v\,t'}{\sqrt{1 - \frac{v^2}{c^2}}} \qquad x_2 = \frac{x_2' + v\,t'}{\sqrt{1 - \frac{v^2}{c^2}}}\,. \tag{6.128}$$

Using the abbreviations $\Delta x = x_2 - x_1$ and $\Delta x' = x_2' - x_1'$, the difference of the two expressions in (6.128) gives

$$\Delta x = \frac{\Delta x'}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{6.129}$$

which means that as viewed from the system S at rest, $\Delta x'$ is lengthened to Δx. Expressing it in another way, the dimension of a body moving with the observer in the system S' seems to be contracted.

Second-order Doppler effect
Since frequency is inversely proportional to time, one can deduce immediately from the considerations on time dilation the formula

$$f = f' \sqrt{1 - \frac{v^2}{c^2}} \tag{6.130}$$

which means that the frequency f' of a moving emitter would be reduced to f. This is the second-order Doppler effect.

Mass relation
Special relativity also affects masses. Denoting the masses in the two reference frames S and S' by m and m' respectively, then

$$m = \frac{m'}{\sqrt{1 - \frac{v^2}{c^2}}} \tag{6.131}$$

is the corresponding mass relation.

The previous formulas comprise the same square root. Expansion into binomial series yields

$$\begin{aligned} \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} &= 1 + \frac{1}{2}\left(\frac{v}{c}\right)^2 \ldots \\ \sqrt{1 - \frac{v^2}{c^2}} &= 1 - \frac{1}{2}\left(\frac{v}{c}\right)^2 \ldots \end{aligned} \tag{6.132}$$

which may be substituted accordingly into Eqs. (6.127), (6.129) (6.130), and (6.131). Related to an observer at rest,

$$\frac{\Delta t' - \Delta t}{\Delta t} = \frac{\Delta x' - \Delta x}{\Delta x} = -\frac{f' - f}{f} = \frac{m' - m}{m} = -\frac{1}{2}\left(\frac{v}{c}\right)^2 \quad (6.133)$$

accounts for the mentioned effects of the special relativity in one formula.

6.4.2 General relativity

The theory of general relativity includes accelerated reference systems too, where the gravitational field plays the key role. Formulas analogous to (6.133) may be derived if the kinetic energy $\frac{1}{2}v^2$ in special relativity is replaced by the potential energy ΔU. Thus,

$$\frac{\Delta t' - \Delta t}{\Delta t} = \frac{\Delta x' - \Delta x}{\Delta x} = -\frac{f' - f}{f} = \frac{m' - m}{m} = -\frac{\Delta U}{c^2} \quad (6.134)$$

represents the relations in general relativity where ΔU is the difference of the gravitational potential in the two reference frames under consideration.

6.4.3 Relevant relativistic effects for GPS

The reference frame (relatively) at rest is located in the center of the earth and an accelerated reference frame is attached to each GPS satellite. Therefore, the theory of special and general relativity must be taken into account. Relativistic effects are relevant for the satellite orbit, the satellite signal propagation, and both the satellite and receiver clock. An overview of all these effects is given for example in Zhu and Groten (1988). The relativistic effects on rotating and gravitating clocks is also treated in Grafarend and Schwarze (1991). With respect to general relativity, Ashby (1987) shows that only the gravitational field of the earth must be considered. Sun and moon and consequently all other masses in the solar system are negligible. Deines (1992) investigates the uncompensated effects if the noninertial GPS observations are not transformed to an inertial frame.

Relativity affecting the satellite orbit
The gravitational field of the earth causes relativistic perturbations in the satellite orbits. An approximate formula for the disturbing acceleration is given by Eq. (4.43). For more details see Zhu and Groten (1988).

Relativity affecting the satellite signal
The gravitational field gives rise to a space-time curvature of the satellite

signal. Therefore, a propagation correction must be applied to get the Euclidean range for instance. The range correction may be represented in the form

$$\delta^{\text{rel}} = \frac{2\mu}{c^2} \ln \frac{\varrho^j + \varrho_i + \varrho_i^j}{\varrho^j + \varrho_i - \varrho_i^j} \tag{6.135}$$

where μ is the earth's gravitational constant. The geocentric distances of satellite j and observing site i are denoted ϱ^j and ϱ_i, and ϱ_i^j is the distance between the satellite and the observing site. In order to estimate the maximum effect for a point on the surface of the earth take the mean radius $R_E = 6\,370\,\text{km}$ and a mean altitude of $h = 20\,200\,\text{km}$ for the satellites. The maximum distance ϱ_i^j results from the Pythagorean theorem and is about 25 800 km. Substituting these values, the maximum range error $\delta^{\text{rel}} = 18.7\,\text{mm}$ results from (6.135). Note that this maximum value only applies to point positioning. In relative positioning the effect is much smaller and amounts to 0.001 ppm (Zhu and Groten 1988).

Relativity affecting the satellite clock

The fundamental frequency f_0 of the satellite clock is 10.23 MHz. All the signals are based on this frequency which is influenced by the motion of the satellite and by the difference of the gravitational field at the satellite and the observing site. The corresponding effects of special and general relativity are small and may be linearly superposed. Thus,

$$\delta^{\text{rel}} \equiv \frac{f_0' - f_0}{f_0} = \underbrace{\frac{1}{2}\left(\frac{v}{c}\right)^2}_{\text{special}} + \underbrace{\frac{\Delta U}{c^2}}_{\text{general relativity}} \tag{6.136}$$

is the effect on the frequency of the satellite clock where Eqs. (6.133) and (6.134) have been used. To get a numerical value, circular orbits and a spherical earth with the observing site on its surface are assumed. Backing on these simplifications, (6.136) takes the form

$$\delta^{\text{rel}} \equiv \frac{f_0' - f_0}{f_0} = \frac{1}{2}\left(\frac{v}{c}\right)^2 + \frac{\mu}{c^2}\left[\frac{1}{R_E + h} - \frac{1}{R_E}\right] \tag{6.137}$$

with v being the mean velocity of the satellite. Substituting numerical values yields

$$\frac{f_0' - f_0}{f_0} = -4.464 \cdot 10^{-10} \tag{6.138}$$

which, despite the simplifications, is sufficiently accurate. Ashby (1987) for instance takes into account the J_2-term for the potential and the centrifugal forces and gets the only slightly different result $-4.465 \cdot 10^{-10}$. Recall that f_0' is the emitted frequency and f_0 is the frequency received at the observation site. Thus, it can be seen that the satellite transmitted nominal frequency would be increased by $df = 4.464 \cdot 10^{-10} f_0 = 4.57 \cdot 10^{-3}$ Hz. However, it is desired to receive the nominal frequency. This is achieved by an offset df in the satellite clock frequency, so that 10.22999999543 MHz are emitted.

Another periodic effect arises due to the assumption of a circular orbit (which is almost true for the Block II satellites). An adequate correction formula is given as

$$\delta^{\text{rel}} = \frac{2}{c} \sqrt{\mu a}\, e \sin E \tag{6.139}$$

where e denotes the eccentricity, a the semimajor axis, and E the eccentric anomaly. This relativistic effect is included in the clock polynomial broadcast via the navigation message, where the time dependent eccentric anomaly E is expanded into a Taylor series. Thus, if the more accurate Eq. (6.139) is used, the relativistic effects must be subtracted from the clock polynomial coefficients. In the case of relative positioning, the effect cancels (Zhu and Groten 1988).

Relativity affecting the receiver clock
A receiver clock located at the surface of the earth is rotating with respect to the resting reference frame at the geocenter. The associated linear velocity at the equator is approximately 0.5 km s^{-1} and, thus, roughly one tenth of the velocity of the satellite. Substituting this value into the special relativistic part of Eq. (6.136) yields a relative frequency shift in the order of 10^{-12} which after 3 hours corresponds to a clock error of 10 nanoseconds (1 ns = 10^{-9} s $\doteq$ 30 cm). Ashby (1987) presented the correction formula for this Sagnac effect; however, since the correction is usually performed by the receiver software the formula is not explicitly given here.

6.5 Antenna phase center offset and variation

The phase center of the antennas is the point to which the radio signal measurement is referred and generally is not identical with the geometric antenna center. The offset depends on the elevation, the azimuth, and the intensity of the satellite signal and is different for L1 and L2. Two effects must be distinguished: the offset and the variation of the antenna phase center. The precision of an antenna should be based on the antenna phase

center variation and not on the offset. A constant offset could be determined and taken into account.

The true antenna phase center may be different from the center indicated by the manufacturer. This antenna offset may simply arise from inaccurate production series. An investigation to determine this offset was based on test measurements, e.g., by rotating the antenna in a laboratory environment.

The antenna phase center varies with respect to the incoming satellite signals. The variation is systematic and may be investigated by test series where variations experienced amount to 1–2 cm. However, it is fairly difficult to model the antenna phase center variation because it is different for each antenna and also for various types. Geiger (1988) shows the different characteristics of conical spiral antennas, microstrip antennas, dipole antennas (crossed pair of horizontal, half-wavelength dipole), and helices. As a consequence, the direct computation of the antenna effects on the distance measurements with respect to azimuth and elevation was proposed. Simple functions for an appropriate modeling may also be found by laboratory tests (Schupler and Clark 1991).

The geometric effect of the antenna orientation on the carrier phase is investigated by Wu et al. (1993). The observed carrier phase depends on the orientation of the antennas of the transmitter and the receiver as well as the direction of the line of sight. Wu et al. (1993) demonstrate that the effect does not cancel for double-differences and may amount to 1 part in 10^9.

6.6 Multipath

6.6.1 General remarks

The effect is well described by its name: a satellite emitted signal arrives at the receiver via more than one path. Multipath is mainly caused by reflecting surfaces near the receiver (Fig. 6.5). Secondary effects are reflections at the satellite during signal transmission.

Referring to Fig. 6.5, the satellite signal arrives at the receiver on three different paths, one direct and two indirect ones. As a consequence, the received signals have relative phase offsets and the phase differences are proportional to the differences of the path lengths. There is no general model of the multipath effect because of the arbitrarily different geometric situations. The influence of the multipath, however, can be estimated by using a combination of L1 and L2 code and carrier phase measurements. The principle is based on the fact that the troposphere, clock errors, and relativistic effects influence code and carrier phases by the same amount. This is not true for ionospheric refraction and multipath which are frequency dependent. Taking

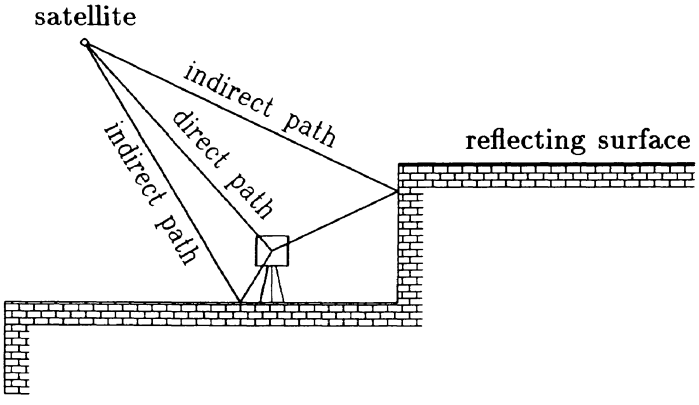


Fig. 6.5. Multipath effect

ionosphere-free code ranges and carrier phases, and forming corresponding differences, all mentioned effects except for multipath are canceled. The residuals, apart from the noise level, reflect the multipath effect.

Tranquilla and Carr (1990/91) group the multipath errors of pseudoranges into three classes: (1) diffuse forward scattering from a widely distributed area (e.g., the signal passes through a cluttered metallic environment), (2) specular reflection from well-defined objects or reflective surfaces in the vicinity of the antenna, and (3) fluctuations of very low frequency, usually associated with reflection from the surface of water.

Purely from geometry it is clear that signals received from low satellite elevations are more susceptible to multipath than signals from high elevations. Note also that code ranges are more highly affected by multipath than carrier phases. Comparing single epochs, the multipath effect may amount to 10–20 m for code pseudoranges (Wells et al. 1987). Under certain circumstances, the error resulting from multipath may grow to about 100 m in the vicinity of buildings (Nee 1992). In severe cases of multipath, loss of lock may even occur.

The multipath effects on carrier phases for relative positioning with short baselines, should, generally, not be greater than 1 cm (good satellite geometry and a reasonably long observation interval). But even in those cases, a simple change of the height of the receiver may increase the multipath and, thus, deteriorate the results. When performing static surveys where the observation times tend to be relatively long, intermittent periods of multipath contamination are not a problem. Such situations occur when the receiver is set up in the center of a highway and large metal trucks continually pass by the antenna. Rapid static surveys may be more contaminated in such cases, and longer observation times would be appropriate.

6.6.2 Mathematical model

The effect of multipath on carrier phases may be estimated by the following considerations (Fig. 6.5). The direct and indirect signals interfere at the antenna center and may be represented by

$$\begin{array}{ll} a\cos\varphi & \ldots \text{ direct signal} \\ \beta\, a\cos(\varphi+\Delta\varphi) & \ldots \text{ indirect signal} \end{array} \tag{6.140}$$

where a and φ denote the amplitude and the phase of the direct signal. The amplitude of the indirect signal is reduced by the damping factor β because of the reflection at a surface (Seeber 1993: p. 308). The phase of the indirect signal is delayed by the phase shift $\Delta\varphi$ which is a function of the geometric configuration. The superposition of the signals in (6.140) is represented by

$$a\cos\varphi+\beta\, a\cos(\varphi+\Delta\varphi)\,. \tag{6.141}$$

Applying the cosine-theorem yields

$$a\cos\varphi+\beta\, a\cos\varphi\cos\Delta\varphi-\beta\, a\sin\varphi\sin\Delta\varphi \tag{6.142}$$

which is slightly rearranged to

$$(1+\beta\cos\Delta\varphi)\, a\cos\varphi-(\beta\sin\Delta\varphi)\, a\sin\varphi. \tag{6.143}$$

This resultant signal may be represented (Joos 1956: p. 44) in the form

$$\beta_M\, a\cos(\varphi+\Delta\varphi_M) \tag{6.144}$$

where the subscript M indicates multipath. The cosine-theorem gives

$$(\beta_M\cos\Delta\varphi_M)\, a\cos\varphi-(\beta_M\sin\Delta\varphi_M)\, a\sin\varphi\,. \tag{6.145}$$

Comparing the coefficients for $a\sin\varphi$ and $a\cos\varphi$ of Eqs. (6.143) and (6.145) leads to the relations

$$\begin{aligned} \beta_M\sin\Delta\varphi_M &= \beta\sin\Delta\varphi \\ \beta_M\cos\Delta\varphi_M &= 1+\beta\cos\Delta\varphi \end{aligned} \tag{6.146}$$

which represent two equations for the desired quantities β_M and $\Delta\varphi_M$. An explicit expression for β_M follows by squaring and adding the two equations. Thus,

$$\beta_M=\sqrt{1+\beta^2+2\beta\cos\Delta\varphi} \tag{6.147}$$

is obtained. An explicit expression for $\Delta\varphi_M$ follows by dividing the two equations in (6.146). Thus,

$$\tan \Delta\varphi_M = \frac{\beta \sin \Delta\varphi}{1 + \beta \cos \Delta\varphi} \tag{6.148}$$

is the solution.

The damping factor β may vary between 0 and 1. The substitution of $\beta = 0$ (i.e., there is no reflected signal and no multipath) into (6.147) and (6.148) gives $\beta_M = 1$ and $\Delta\varphi_M = 0$. This means that the "resultant" signal is identical to the direct signal. The strongest possible reflection is defined by $\beta = 1$. The substitution of this value into (6.147) and (6.148) leads to

$$\beta_M = \sqrt{2(1 + \cos \Delta\varphi)} = 2 \cos \frac{\Delta\varphi}{2} \tag{6.149}$$

and

$$\tan \Delta\varphi_M = \frac{\sin \Delta\varphi}{1 + \cos \Delta\varphi} = \tan \frac{\Delta\varphi}{2} \tag{6.150}$$

yielding

$$\Delta\varphi_M = \tfrac{1}{2} \Delta\varphi \,. \tag{6.151}$$

Examples for numerical values for β_M and $\Delta\varphi_M$ as a function of $\Delta\varphi$ are

$\Delta\varphi$	β_M	$\Delta\varphi_M$
0°	2	0°
90°	$\sqrt{2}$	45°
180°	0	90°

which shows that the maximum effect of multipath on phase measurements occurs for $\Delta\varphi_M = 90° = 1/4$ cycle. Converting this phase shift to range gives $\lambda/4$ or, with $\lambda = 20\,\text{cm}$, the maximum change in range of about 5 cm. However, it should be noted that this value may increase if linear phase combinations are used.

The phase shift $\Delta\varphi$ can be expressed as a function of the extra pathlength Δs. In the case of a horizontal reflector (ground),

$$\Delta\varphi = \frac{1}{\lambda} \Delta s = \frac{2h}{\lambda} \sin E \tag{6.152}$$

is obtained where the phase shift is expressed in cycles. The parameter h denotes the vertical distance between the antenna and the ground and E is the elevation of the satellite (Fig. 6.6). Multipath is periodic because E

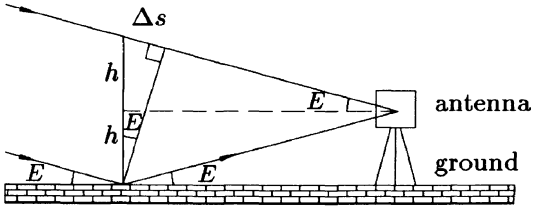


Fig. 6.6. Geometry of multipath

varies with time. The frequency of multipath is

$$f = \frac{d(\Delta\varphi)}{dt} = \frac{2h}{\lambda} \cos E \frac{dE}{dt} . \tag{6.153}$$

Substituting typical values like $E = 45°$ and $dE/dt = 0.07$ mrad per second leads to the approximation for the L1 carrier

$$f = 0.521 \cdot 10^{-3} h \tag{6.154}$$

where the frequency is obtained in hertz if h is given in meters (Wei and Schwarz 1995a). Thus, an antenna height of 2 m leads to an approximate period of 16 minutes for the multipath error.

6.6.3 Multipath reduction

To reduce or estimate the multipath effects, various methods were developed that are classified by Ray et al. (1999) as (1) antenna-based mitigation, (2) improved receiver technology, and (3) signal and data processing.

Among the antenna-based mitigation methods, improving the antenna gain pattern by choke rings, creating special antenna designs and arrays are very effective (Moelker 1997, Bartone and Graas 1998). The elimination of multipath signals is possible by selecting an antenna that takes advantage of the signal polarization. The transmitted GPS signals are right-handed circularly polarized, whereas the reflected signals are left-handed polarized. A reduction of the multipath effect may also be achieved by digital filtering, wideband antennas, radio frequency absorbent antenna ground planes, choke ring antennas including the advanced dual frequency choke ring design providing optimal multipath rejection for both L1 and L2 band, while the previous choke ring designs yield optimal multipath rejection only in L2 band (Philippov et al. 1999). A refined version of the choke ring idea uses spiral arms. The advantage of the more recent developments over the choke rings is the sharper radiation pattern roll-off (which reduces the multipath susceptibility), no phase center offset between the L1 and L2 carriers, no

necessity for any alignment, e.g., to the north direction, due to its symmetry, and its planar structure (Kunysz 2000). The absorbent antenna ground plane reduces the interference of satellite signals with low or even negative elevation angles which occur in case of multipath.

Improving the receiver technology for multipath reduction includes narrow correlator spacing, extending the multipath estimation delay lock loop, enhancing the strobe correlator multipath rejection; more details in Dierendonck and Braasch (1997), Garin and Rousseau (1997). One example for the ongoing research for multipath reduction is the continuous improvement based on Multipath Estimating Delay Lock Loop (MEDLL) (Townsend et al. 1995, 2000). This technique separates the incoming signal into the direct line-of-sight component and the indirect signal by using an array of correlators and measuring the received correlation function. Tests with MEDLL have shown an error reduction up to 90% (Fenton and Townsend 1994).

Numerous methods investigate multipath mitigation by signal and data processing: exploring the signal-to-noise ratio, using multiple reference stations, smoothing carrier phases, using data combinations.

The detection and reduction of multipath in the spectral domain is proposed by Li et al. (1993). The measured data are transformed by the fast Fourier transform into the spectral domain. The detection and reduction of multipath is carried out by amplitude filtering. Afterwards, the inverse fast Fourier transform outputs the filtered data.

The relation (6.148) expressing the carrier phase error due to multipath is investigated by Ray et al. (1999) to mitigate multipath effects using multiple closely spaced antennas for static applications. The nearby placed antennas cause a strong correlation of the reflected signals. The phase of the reflected signal at each antenna phase center depends on the signal direction which may be expressed by azimuth and elevation. Also the geometry of the antennas must be taken into account. Ray et al. (1999) introduce a reference antenna and five antennas assembled around it. For each satellite, a Kalman filter is implemented. The four-element state vector of the estimator comprises the damping factor β, the reflected signal phase at the antenna, and azimuth and elevation of the reflected signal. By individually combining the data of the reference antenna with the others, profiting from the known geometry of the antennas and using a single external stable clock to get a negligible receiver clock bias, the model for the measurements may strongly be simplified so that it mainly reflects the oscillatory multipath error and the random carrier phase noise. The results of the Kalman filter estimation may be adapted to finally apply to (6.148) allowing for the multipath error determination in the carrier phase at each antenna. Neglecting the filter convergence period, test measurements show an about 70% improvement.

The most effective counter-measure to multipath is to avoid sites where it could be a problem (e.g., near chain link fence). Considering Fig. 6.5, placing the antenna directly on the reflecting ground without tripod would eliminate one of the two indirect paths; however, the vertical reflecting surface would still contaminate the results. The general recommendation is, therefore, to avoid, as far as possible, reflecting surfaces in the neighborhood of the receivers.

Nowadays, multipath analysis and mitigation is no longer restricted to high precision (static) applications. Car navigation is one example with feasibility studies using multiple antennas to isolate and detect multipath on code measurements (Nayak et al. 2000). Multipath on code measurements remains the most significant error source for differential GPS vehicle navigation. Compared to static applications, the positions of various reflectors are changing rapidly, increasing the difficulty of a proper model. For the multipath affecting code measurements, the residuals of code and phase may be analyzed since the carrier receiver noise and the multipath affecting phases are very small compared to the corresponding code values. Successful identification and elimination of the multipath corrupted measurement is the final objective being demonstrated in some experiments by Nayak et al. (2000). The gain in position accuracy improvement depends on the size of the multipath errors. Even with high-performance correlator receivers, multipath errors of several meters frequently occur.

7 Surveying with GPS

7.1 Introduction

This chapter is mainly concerned with the practical aspects of GPS surveying and addresses planning, performance, and in situ data processing. Overlapping with other chapters is intentional to provide complete information in a single chapter for readers more interested in practical considerations.

7.1.1 Terminology definitions

The enormous interest in GPS is reflected by the numerous papers published today. Unfortunately, a standard terminology is missing, although several authors have attempted to provide a list of terms used. To avoid confusion in the following sections, some of the more important definitions are given here and are used throughout the entire text.

Code range versus carrier phase
Typically, GPS observables are pseudoranges derived from code or carrier phase measurements. Generally speaking, the accuracy of code ranges is at the meter level, whereas the accuracy of carrier phases is in the millimeter range. The accuracy of code ranges can be improved, however, by the narrow correlator spacing technology or by smoothing techniques. Unlike the carrier phases, the code ranges are virtually unambiguous. This makes code ranges immune from cycle slips (i.e., changes of the phase ambiguities) and to some extent from site obstructions. The determination of the phase ambiguities is a critical issue in GPS surveying.

Real-time processing versus postprocessing
To qualify as real-time GPS, the position results must be available in the field immediately or while still on the station. The results are denoted as "instantaneous" if the observables of a single epoch are used for the position computation and the processing time is negligible. The original concept of GPS aimed at instantaneous navigation of moving vehicles (i.e., ships, cars, aircraft) by unsmoothed code pseudoranges. A different and less stringent definition is "real-time" which includes computing results with a slight delay. Precisely speaking, these are quasi (or near) real-time results. Today, radio data links allow the combination of measurements from different sites in real time. Postprocessing refers to when data are processed after the fact.

B. Hofmann-Wellenhof et al., *Global Positioning System*

Point positioning versus relative positioning
The coordinates of a single point are determined by point positioning when using a single receiver which measures pseudoranges to (normally four or more) satellites. Instead of "point positioning" the term "single-point positioning" or the term "absolute point positioning" is used. Here, the term "absolute" reflects the opposite of "relative".

Instead of "relative positioning" the term "differential positioning" is often used. Note, however, that the two methods are (at least theoretically) different. Differential positioning is rather an improved single-point positioning technique and is based on applying (predicted) corrections to pseudoranges at an unknown site. The technique provides instantaneous solutions (usually denoted as real-time solutions) where improved accuracies with respect to a reference station are achieved.

Relative positioning is possible if (as in the case of differential positioning) two receivers are used and (code or carrier phase) measurements, to the same satellites, are simultaneously made at two sites. The measurements at both sites are (in contrast to differential positioning) directly combined. This direct combination further improves the positional accuracy but prevents instantaneous solutions in the strict sense. Normally, the coordinates of one site are known and the position of the other site is to be determined relative to the known site (i.e., the vector between the two sites is determined). In general, the receiver placed on the known site is stationary while observing.

In the past, point positioning was associated with navigation and relative positioning with surveying. Also, the term "relative" was used for carrier phase observations, whereas the term "differential" was used for code range observations.

Static versus kinematic
Static denotes a stationary observation location, while kinematic implies motion. A temporary loss of signal lock in static mode is not as critical as in kinematic mode.

Attention should be paid to the difference between the terms "kinematic" and "dynamic". A very intuitive example given in Schwarz et al. (1987) points up the difference: "Modeling the movement of a vehicle in three-dimensional space requires either the knowledge of the forces causing the motion or the measurement of the vehicle motion in a given three-dimensional coordinate system. The first type of modeling will be called dynamic, the second kinematic." The modeling of the orbit for GPS satellites is a dynamic procedure. As soon as the positions of the satellites are known, positioning of a moving vehicle is regarded as kinematic procedure.

The terms "static" and "kinematic" must be considered in the context of point or relative positioning. Typical examples of these modes are given to acquaint the reader with these terms.

Static point positioning is useful if points are needed with moderate accuracy. Today (i.e., without SA), positional accuracies of about 10 m can be achieved in real time using code ranges.

Kinematic point positioning can be used to determine a vehicle's three-dimensional trajectory as a function of time. Therefore, a typical example for kinematic point positioning is vehicle navigation.

Static relative positioning by carrier phases is the most accurate positioning technique and the one most frequently used by surveyors. The method is also called *static surveying.* This technique aims at the determination of the vector (often called baseline) between two stationary receivers. One has to distinguish between single or multipoint baseline determination. Obviously, the multipoint solution concerns more than two sites. In static surveying, 1 ppm to 0.1 ppm (or even better) accuracies are achievable. This is equivalent to millimeter accuracy for baselines up to some kilometers.

Kinematic relative positioning involves one stationary and one moving receiver. The two receivers perform the observations simultaneously. The possible applications are basically the same as for kinematic point positioning, but better accuracies at the centimeter level are achievable. The use of radio links between the baseline stations leads to the real-time kinematic (RTK) technique. Depending on the data (phases or phase corrections), which are transmitted in real time to the roving receiver, the technique belongs to relative or differential positioning.

7.1.2 Observation techniques

The selection of the observation technique in a GPS survey depends upon the particular requirements of the project; the desired accuracy especially plays a dominant role.

Point positioning

When using a single receiver, only point positioning with code pseudoranges makes sense. The concept of point positioning is simple, it is trilateration in space. For point positioning, GPS provides two levels of service: (1) the Standard Positioning Service (SPS) with access for civilian users and (2) the Precise Positioning Service (PPS) with access for authorized users.

For the SPS, only the C/A-code is available. With SA turned off, the achievable real-time accuracies are about 10 m at the 95% probability level.

The PPS has access to both codes and accuracies down to the meter level

can be obtained. The availability of PPS is primarily designed for U.S. military use; however, it is also available to U.S. Federal and Allied Government (civil and military) users through special agreements with DoD. Furthermore, limited nongovernmental civil use of GPS, domestic and foreign, will be considered upon request (Olsen 1992).

Differential GPS

Basic concept

The degradation of the point positioning accuracy by SA has led to the development of Differential GPS (DGPS). This technique is based on the use of two (or more) receivers, where one (stationary) reference or base receiver is located at a known point and the position of the (mostly moving) remote receiver is to be determined (Fig. 7.1). At least four common satellites must be tracked simultaneously at both sites. The known position of the reference receiver is used to calculate corrections to the GPS derived position or to the observed pseudoranges. These corrections are then transmitted via telemetry (i.e., controlled radio link) to the roving receiver and allow the computation of the rover position with far more accuracy than for the single-point positioning mode.

An alternative to the navigation mode is the surveillance mode, where the remote receiver broadcasts the raw observation data to the (fixed) monitor station where the correct position of the rover is computed. The surveillance mode has the advantage that the roving receiver does not have to perform a large number of computations (Kasties and Harrer 1991).

As mentioned previously, two correction methods are in use. In the first method, the reference receiver at a known location calculates its position

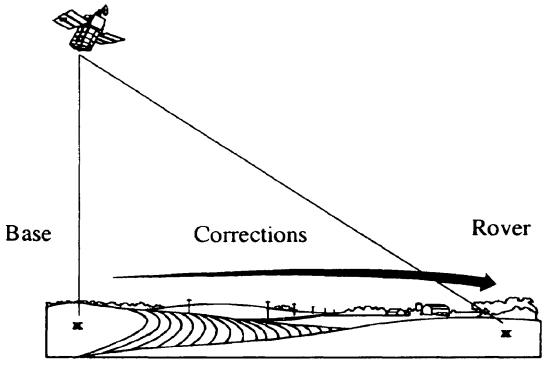


Fig. 7.1. Concept of DGPS (courtesy W. Oberegger, Wels, Austria)

using the same set of satellites as the roving receiver. The difference (this explains the notation "differential") of the known from the calculated position yields position corrections. These values are then applied to the roving receiver to obtain an improved position. This method is conceptually simple but requires more complex satellite selection. The second method is based on pseudorange corrections which are derived from the difference between calculated ranges and observed (code or phase) pseudoranges at the reference site. Apart from range corrections, range rate corrections are also computed at the reference site. The observed pseudoranges at the roving site may be corrected by applying pseudorange corrections of the reference station. This method is more flexible, gives higher accuracy, and is the one in general use. The disadvantage is that more computing effort is required for the algorithms. The higher accuracy is based on the fact that GPS error sources are very similar over a distance of about 500 km and are, therefore, virtually eliminated by the differential technique.

The accuracy requirements of GPS users are very different and vary between several hundred meters and centimeter level. A very large group of users is interested in a real-time accuracy at the meter level. This accuracy cannot be obtained by point positioning with SPS but can be achieved by DGPS. Using C/A-code ranges, accuracies at the 1–5 m level can be routinely achieved. To obtain the submeter level, phase smoothed code ranges or high performance C/A-code receivers must be used (Lachapelle et al. 1992). An even higher accuracy level can be reached by the use of carrier phases. For ranges up to some 20 km, accuracies at the subdecimeter level can be obtained in real time (DeLoach and Remondi 1991). To achieve this accuracy, the ambiguities must be resolved on-the-fly and therefore (generally) dual frequency receivers are required.

Wide Area DGPS

An extension of DGPS (in the sense of Local Area DGPS) is Wide Area DGPS (WADGPS) which uses a network of GPS reference stations. As the name implies, WADGPS covers a larger territory than can be reasonably accommodated by a single reference station. One of the main advantages of WADGPS is that a more consistent accuracy can be achieved throughout the region supported by the network. In the case of DGPS with a single reference station, the accuracy decreases as a function of distance from the reference station at a rate of approximately 1 cm per 1 km. Other advantages of WADGPS are that inaccessible regions can be covered, e.g., large bodies of water, and that in case of a failure in one of the reference stations, the network will still maintain a relatively high level of integrity and reliability compared to a collection of individual DGPS reference stations.

Apart from the monitor stations, the WADGPS network includes (at least) one master station. This station collects the range corrections from the monitor stations, processes these data to form WADGPS corrections which are transmitted to the user community as well as to the monitor stations (Mueller 1994). The networks may cause slight additional delay beyond regular DGPS due to the additional communication required between the monitor stations and the master station.

WADGPS corrections are generated using two main approaches, namely the measurement domain approach and the state-space approach. In the measurement domain technique, the differential corrections of the individual monitor stations are weighted (Szabo and Tubman 1994) to form one set of corrections. This simple concept has the disadvantage that the accuracy depends on the distance of the user from the nearest monitor station. A more complex technique is the state-space approach where individual (orbital, tropospheric, ionospheric) errors are modeled and evaluated in the network solution. Based on these errors, the pseudorange corrections related to each reference station are computed. Thus, the accuracy is consistent over the entire network. Algorithms for the state-space approach can be found in Mueller et al. (1994), Kee (1996).

Since the reference stations of the WADGPS network may be very distant from the user location, the Virtual Reference Station (VRS) concept has been developed (Wanninger 1999). Here, the user gets range corrections or even the observables of a nonexistent (i.e., virtual) reference station at a user-specified position. This concept is a prerequisite mainly for RTK applications which require short distances to reference stations to facilitate ambiguity resolution.

Data link

An essential component of DGPS is the data communication link. This link is usually a radio channel of some sort. The radio link requires compatible hardware at the monitor and remote sites. For DGPS using code ranges, the required data transfer rates are typically at 200 bits per second (bps). Data rates of at least 2 400 bps up to 19 200 bps (depending on the update rate of the data) are required in the case of RTK applications.

There are many (ground-based and space-based) communication systems capable of transmitting DGPS corrections. For global data transfer, space-based radio links are required. One example for communication satellites is the International Maritime Satellite (INMARSAT) system consisting of several geostationary (GEO) satellites. Most recently, DGPS makes use of the worldwide telecommunication satellite program. This program uses numerous low earth orbit (LEO) satellites that provide mobile telephone

service anywhere in the world. Many of the proposed WADGPS approaches rely on the use of satellite communication for the transmission of differential data which is beneficial, e.g., to users in remote mountainous regions and avoids difficulties with receiving ground-based radio transmissions.

In comparison to satellite-based links, the establishment and operation of terrestrial links is much cheaper. The data can be transmitted via conventional data radios, cellular radiophones, or by modulating the data onto the sidebands of existing FM radio broadcast as, e.g., in the Radio Data System (RDS) (Weber and Tiwari 1994). This standardized method has been defined by the European Broadcasting Union and consists of transmitting digital data on a subcarrier superimposed upon the normal radio (audio) broadcast. An improved technique is the Data Radio Channel (DARC) which was developed in Japan and supports higher data rates than RDS. An interesting future possibility is the use of ground-based pseudolites. The name is a contraction of the words "pseudo" and "satellite" and refers to a transmitter that emits radio signals like a GPS satellite. The differential corrections could be modulated onto the signal.

Apart from radiation power, the frequency of the radio link is the critical parameter for transmission performance. The higher the frequency the more data can be transferred in the time unit. Thus, VHF (very high frequency) and UHF (ultra high frequency) radio links are mostly used for terrestrial data links. The drawback is that the range of high-frequency waves emitted from terrestrial stations is limited to the line of sight between transmitter and receiver. The maximum range d in kilometers is given by

$$d = 3.57\sqrt{k}\,(\sqrt{h_t} + \sqrt{h_r}) \tag{7.1}$$

where h_t and h_r are the heights in meters of the transmitting and receiving antennas above mean terrain. The factor k depends on the vertical gradient of refractivity and varies from about 1.2 to 1.6. Satellite communication allows high frequencies in the gigahertz range and enables data rates up to 19 200 bps (and even much higher in the future) over long distances. An overview of data rates is given in Table 7.1 (Lichtenegger 1998). For further details on data links see Langley (1993), Scott (1993).

RTCM format

Although some receiver manufacturers have devised their own proprietary formats, the transmission of correction data between the reference receiver and the remote receivers has been standardized since 1985 according to the proposals of the Radio Technical Commission for Maritime Services, Special Committee 104. This standard is shortly named RTCM format and

Table 7.1. Transmission rates of radio links

Medium	Baudrate
Very Low Frequency (VLF)	50
Low Frequency (LF)	300
Radio Data System (RDS)	1 000
Ultra High Frequency (UHF)	2 400
Cell phones	9 600
INMARSAT	19 200

the current version is 2.2 (RTCM 1998). There are 64 message types available (some of them are listed in Table 7.2), but some of them are presently still undefined. Note that the messages concern both, GPS and its Russian counterpart GLONASS.

The format of the messages is (almost) identical to that of the GPS navigation message and consists of a sequence of 30-bit words. The last six bits in each word are parity bits. Each message starts with a header. The first word contains a fixed preamble, the message type identification, and the reference station identifier. The second word contains the frame time tag in the form of the Z-count, the sequence number, the message length, and a reference station health indicator. In some messages a third word is added to the header.

Table 7.2. Some message types of RTCM version 2.2

Type	Meaning
1	GPS differential corrections
3	GPS reference station parameters
6	Null frame (filler)
9	Partial set of type 1 messages
16	Special GPS message
18	Raw carrier phase measurements
19	Raw code range measurements
20	Carrier phase corrections
21	Code range corrections
31	GLONASS differential corrections
32	GLONASS reference station parameters
34	Partial set of type 31 messages
36	Special GLONASS message

The messages 1–17 were available already in the former version 2.0, while the messages 18–21, each with a three-word header, were added starting with version 2.1. The messages related to GLONASS are available since version 2.2. The messages 18 and 19 contain the raw carrier phase and code range observed at the reference station with GPS and GLONASS, and support relative positioning. The messages 20 and 21 contain corrections to the corresponding measurements and, thus, support differential positioning. A detailed description of all message types can be found in RTCM (1998).

Relative positioning

At present, highest accuracies are achieved in the relative positioning mode with observed carrier phases. Processing a baseline vector requires the phases be simultaneously observed at both baseline endpoints (Fig. 7.2). Originally, relative positioning was only possible by postprocessing data. Today, real-time data transfer over short baselines is routinely possible, which enables real-time computation of baseline vectors, and has led to the real-time kinematic (RTK) technique.

Static relative positioning

The static surveying method is the most commonly used since the only basic requirement is a relatively unobstructed view of the sky for the occupied points. Conventional static surveys require observation periods depending on the baseline length, the number of visible satellites, the geometric configuration, and the method used. The accuracy is correlated with the baseline length and amounts to $1-0.1$ ppm for baselines up to some 100 km and even better for longer baselines.

Static relative positioning also includes the rapid static technique based on fast ambiguity resolution techniques. These techniques generally use

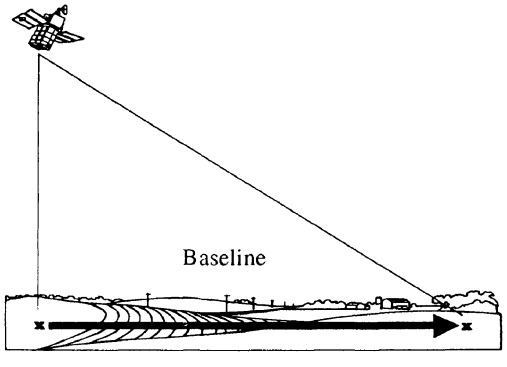


Fig. 7.2. Concept of relative positioning (courtesy W. Oberegger, Wels, Austria)

Table 7.3. Session lengths for static surveys

Receiver	Conventional static	Rapid static
L1	30 minutes + 3 minutes/km	20 minutes + 2 minutes/km
L1+L2	20 minutes + 2 minutes/km	10 minutes + 1 minute/km

code and carrier phase combinations on both frequencies. Thus, dual frequency receivers and optimum satellite geometry are required. Restricting the method to 20 km baselines, accuracies at the subcentimeter level can be achieved. This was first demonstrated in a P-code receiver test by the U.S. Federal Geodetic Control Subcommittee (FGCS), formerly called U.S. Federal Geodetic Control Committee (FGCC), in 1991.

Standard values for the session lengths of static observations (particularly for baseline up to some 20 km) are listed in Table 7.3. These values are based on the visibility of four satellites, good geometry, and normal atmospheric conditions. Note that an additional satellite may reduce the session lengths by 20%. The numbers may be regarded as too conservative; however, they assure correct ambiguity resolution and, thus, high accuracies.

Typical uses of static surveying include: state, county, and local control surveys, photo-control surveys, boundary surveys, and deformation surveys.

Kinematic relative positioning

Kinematic surveys are the most productive in that the greatest number of points can be determined in the least time. The drawback is that after initialization of the survey a continuous lock on at least four satellites must be maintained.

The semikinematic or stop-and-go technique is characterized by alternatively stopping and moving one receiver to determine the positions of fixed points along the trajectory. The most important feature of this method is the increase in accuracy when several measurement epochs at the stop locations are accumulated and averaged. This technique is often referred to as simply kinematic survey. Relative positional accuracies at the centimeter level can be achieved for baselines up to some 20 km.

The kinematic technique requires the resolution of the phase ambiguities before starting the survey. The initialization can be performed by static or kinematic techniques. Currently available commercial software (for dual frequency receivers) only requires 1–2 minutes of observation for baselines up to 20 km to resolve the ambiguities kinematically. The required lock on at least four satellites throughout the entire survey means that kinematic surveys require considerable reconnaissance since not only the occupied sites but also the route taken to travel between sites should be free of obstructions.

The kinematic method is best suited for wide open areas (such as cleared construction sites) where there are few obstructions. Areas such as suburban subdivisions can also be surveyed by the kinematic technique if there are not too many large trees overhanging the roads. A practical use of the kinematic technique is to determine positions of a receiver mounted on an all-terrain vehicle that travels across a given area in a series of cross section lines. The three-dimensional coordinates of this vehicle-mounted receiver can be determined to high accuracy (few centimeters) so that an accurate topographic map of the area can be rapidly prepared.

Pseudokinematic relative positioning
The technique most nearly like the static method is the pseudokinematic survey. This technique, developed by B. Remondi, is also named intermittent static or reoccupation method. These surveys require less occupation time but one must occupy the "point-pair" two times. A typical scenario would involve occupying a pair of points for five minutes, moving to other points, and finally returning to the first point pair about one hour after the initial occupation for a second 5-minute occupation. The advantage of the (total) 10-minute occupation versus, e.g., a 60-minute occupation for conventional static surveying is somewhat offset by the time lost in traveling to the points a second time. The best use of the pseudokinematic method is when the points to be occupied are along a road where the observers can move quickly between setups. Subcentimeter accuracy can be obtained with the pseudokinematic method. This high accuracy is achievable because the integer biases can be fixed due to the fact that the satellite geometry changes in the hour between occupations. The main advantage of the pseudokinematic method is that for a given observation time more sites can be occupied than with conventional static surveying. Compared to the kinematic method, a loss of lock may occur and the number of satellites does not play that essential role. There is no requirement for maintaining signal lock between the reoccupation of points. The receiver may even be turned off while moving. The main weakness is the necessity of revisiting the site. This restricts the method to local applications. The main competitor to the pseudokinematic method is the rapid static positioning approach where each site has to be occupied only once. Typical pseudokinematic applications include: photocontrol, lower-order control surveys, and mining surveys.

In practice, it is best to use a mixture of the three methods when using single frequency receivers. For example, static and pseudokinematic methods can be used to establish a broad framework of control and to set points on either side of obstructions such as bridges. Kinematic surveys can then be

employed to determine the coordinates of the major portion of points, using the static points as control and check points. A thorough reconnaissance is required for these mixed surveys.

The reduction in the cost of dual frequency receivers will eventually obviate the use of some methods described in the preceding section. With these instruments, occupation times of a few minutes yield subcentimeter accuracy, and sophisticated software virtually eliminates the loss-of-lock problem for kinematic surveys.

7.1.3 Field equipment

The field equipment includes receiver units and auxiliary devices such as meteorological sensors, tribrachs, tripods, bipods, and other ancillary equipment. Geodetic receivers which perform precise baseline vector measurements are mainly considered in this section. The selection of an appropriate receiver depends on the special project requirements. Therefore, only some general considerations will be given in this section.

For moderate length baselines up to about 20 km, single frequency receivers provide equivalent results to the dual frequency receivers because the ionospheric refraction (mostly) cancels by differencing the phase measurements between the baseline sites. During periods of moderate solar activity, lines of up to 100 km have been accurately measured with single frequency receivers by observing for several hours. The ionospheric aberrations mostly have a highly central tendency and tend to average out. Baseline distances must be reduced in periods of high sunspot activity. This activity cycle has a repetition rate of about 11 years with a maximum in 2002. Dual frequency receivers compensate (and virtually eliminate) ionospheric refraction by the ionosphere-free combination of the two carrier phases (Sect. 6.3.2).

Receivers with Y-code capability enable the reconstruction of both carriers by the code correlation technique. Thus, they offer the best signal strength and yield the best results. Unfortunately, the DoD, to date, refuses to consider eliminating A-S so the number of civilian users is very limited. Civilian users achieve the most accurate observables with P-code survey receivers. Using this type of receivers, baseline vectors can be determined more accurately and rapidly than with C/A-code receivers. One reason that P-code receivers (generally) provide a better code range resolution is that they are less affected by multipath and imaging. Another reason is that the P-code enables the reconstruction of both carriers by quasi-codeless techniques. Accurate phase measurements on (full wavelength) L1 and L2 can be combined to produce the wide lane signal with a wavelength of about 86 cm. The integer ambiguity of this combined wavelength is more easily re-

solved than the shorter 19 cm base carrier wavelength. Aside from the loss of some signal strength, P-code receivers provide virtually the same capability as Y-code receivers when used for surveying.

State-of-the-art receivers provide options to process DGPS corrections. Most manufacturers offer receivers with RTK capability. In both cases, data links are required. Today, the data links are still a critical factor in high-accuracy real-time positioning.

An important aspect of receiver design is the data sampling rate. A fast rate produces a large volume of data and requires a significant amount of storage. This fast rate is necessary for kinematic applications and facilitates cycle slip detection and repair in static techniques. Also, the use of receivers with more than the minimum four channels is appropriate, since the additional satellites provide redundant information.

Another feature for receiver selection (particularly for kinematic surveys) is its capability of bandwidth selection in the tracking loops. The bandwidth should be wide enough to prevent loss of signal, but narrow enough to provide a high SNR. Therefore, receivers that are able to adapt the bandwidth depending on the dynamics will provide optimum results.

Different receiver types involved in a survey may cause problems due to different numbers of channels, different signal processing techniques, and different time tagging.

The phase center of the antenna should be stable and repeatable. Microstrip antennas, to date, have provided the most economical solution and have provided reasonable results. When different antenna types are used, the phase center of each type must be calibrated. It should be noted that the phase center depends on frequency and is, therefore, different for L1 and L2 carrier observations. The choke ring antenna appears to provide the best multipath immunity at a reasonable size. Mixing antennas during a survey can be a problem, especially with the heights. A convenient calibration would be to measure a short (100 m) base using different antenna combinations to determine the antenna offset for a given pair. Most of the GPS receiver manufacturers offer coaxial antenna cables with different lengths up to 60 m where 10 m is standard. A long cable provides more versatility for site access; however, the recommendations of manufacturers on cable size (type) should be followed to avoid signal loss. Large-diameter low-loss cable is available for tracking sites where the cable length exceeds 60 m.

Today, weather-proof, lightweight (<1 kg), small ($<1\,000$ cm^3), and low power consuming (<5 W) dual frequency receivers capable of tracking "all-in-view" satellites (6–12 channels) are available for a fraction of the cost of earlier receivers. New technologies will further miniaturize the receivers and lower the cost.

All comparable GPS equipment being marketed today provides excellent results so the decision of which equipment to use is usually made based upon ease of use and monetary considerations. The cost of GPS receivers is still decreasing and their capability is constantly increasing. The TI-4100 produced in 1984 was priced at $ 170 000. Today, a simple geodetic C/A-code receiver costs less than $ 10 000. Some years ago, Henstridge (1991) estimated the initial capital and start-up costs for three GPS receivers and related items such as software and computers to amount $ 100 000; today, however, this amount is much lower.

The final factor that should be considered is the primary use of the equipment. For example, if one plans to use the equipment for kinematic surveys or in heavily treed areas, a receiver should be chosen that has a separate antenna that can be mounted on a prism pole or survey mast. Equipment with built-in antennas have optional antennas that can be purchased. Another important feature is the weight of the receiver and its power requirements if kinematic surveys on foot are planned. A heavy, very power consuming receiver is not the type of instrument to carry for extended periods.

The number of surveying receiver manufacturers is constantly increasing and their products are continuously being improved. The manufacturers should be contacted directly to obtain actual information on a particular product. Addresses are provided, e.g., yearly in the January issue ("Receiver Survey") or in the June issue ("Buyers Guide") of *GPS World.*

7.2 Planning a GPS survey

7.2.1 General remarks

Designing a GPS network gives rise to several practical questions which are just as important as the theoretical aspects of GPS as stated by Bevis (1991). The issues of equipment, observation technique, and organization are all important. In this section the last item will be addressed.

GPS surveying differs essentially from classical surveying because it is weather independent and there is no need for intervisibility between the sites. Because of these differences, GPS surveys require different planning, execution, and processing techniques. At present, only a few organizations have published standards and specifications for GPS surveys.

Planning a GPS survey may or may not be appropriate for a particular application or project. Before deciding whether or not to expend resources on planning, one must decide the ultimate purpose of the survey and the accuracy of the desired results (coordinates). The FGCS classifies surveys by designating orders for the various proportional accuracies desired. Generally,

surveys of the highest order (A and B) are reserved for special purposes such as state and county control networks and scientific studies. The remaining first-, second-, and third-order surveys are the classifications normally specified for mapping, surveying, and engineering projects. There is no question that a substantial planning effort is required to successfully complete a high-order survey; however, first-order and lower-order surveys may not require extensive planning except in heavily treed or obstructed areas.

The optimum planning of a GPS survey has to consider several parameters such as site or satellite configurations, the number and the type of receivers to be used, and economic aspects. Contrary to the design of triangulation or trilateration networks which involves considerable effort to maintain the geometric strength, geometry and line length are not so critical for GPS networks. The planning phase should also include some data processing considerations; for instance, whether the available software allows for the computation of single baseline vectors or for multipoint solutions.

For large projects with many sites and many receivers, planning a GPS survey could be aided by the use of computer programs to save time and resources. For instance, the Geodetic Survey of Canada has developed an appropriate software package (Klees 1990). This program plans vehicle routes, the selection of satellites, and the delineation of the network design.

7.2.2 Presurvey planning

Point selection

The first step in planning a GPS survey is to obtain the largest scale map of the area upon which the desired points can be plotted. Topographic maps from 1:25 000 to 1:100 000 scale are excellent for this purpose; also, county road maps are quite useful. All desired survey points are plotted on the map along with the known control points. In most instances it is not worth using control points that are not included in the national system. Using the coordinates of points of unknown accuracy can create many problems so that it is better to choose a national network point even if it is at a greater distance from the project site than other control monuments.

In planning a GPS survey, there are only three basic considerations in choosing a point:

1. No obstructions above 20° elevation to avoid satellite signal blockage.
2. No reflecting surfaces (e.g., metal structures, fences, water surfaces) in the vicinity of the antenna to avoid multipath.
3. No nearby electrical installations (e.g., transmitters) to avoid signal disturbances.

The requirements above are of primary importance; however, the proximity to a road is a convenience that increases production. Further desirable GPS site characteristics are: mark on publicly owned land and not likely to be disturbed, clear site for visible azimuth mark, and space for parked vehicle.

The complete tie of a GPS network to the national datum requires the occupation of three or more control points. Many areas now have supernets which provide convenient reference monuments for GPS surveys. If supernet points are available, one may not even have to visit them prior to beginning field work since the supernet points have been selected for GPS occupation. A reasonable horizontal tie would consist of measurements to national control points on either side of the project area. Ties to the vertical datum normally require more planning. Three benchmarks at the corners of the project area would provide the minimum acceptable vertical datum reference.

Observation window

The second step of presurvey planning is to determine the optimum daily observation period and to decide how it should be subdivided into sessions (i.e., periods where two or more receivers simultaneously track the same satellites).

The optimum window of satellite availability is the period when a maximum of satellites can be observed simultaneously. The windows of satellite availability are shifted by four minutes each day, due to the difference between sidereal time and Universal Time (UT). For example, if some of the satellites appeared in a given geometric configuration at 9:00 UT today, they would be roughly in the same position in the sky at 8:56 UT the following day. The length of the window is a function of the location. The optimum window is found by inspecting azimuth-elevation charts which are produced by the software of GPS equipment manufacturers.

The calculation of azimuth and elevation is based on the projection of the unit vector $\underline{\varrho}$, pointing from the observing site to the instantaneous satellite position, onto the orthogonal axes of the local coordinate frame. The unit vector $\underline{\varrho}$ is defined by

$$\underline{\varrho} = \frac{\underline{\varrho}^S - \underline{\varrho}_R}{\|\underline{\varrho}^S - \underline{\varrho}_R\|} \tag{7.2}$$

with $\underline{\varrho}^S$ being the geocentric position vector of the satellite and $\underline{\varrho}_R$ being the geocentric position vector of the observing site which is needed only approximately (i.e., coordinates from a map are sufficient). The vector $\underline{\varrho}^S$ can be calculated by procedures described in Chap. 4 and the components of vector $\underline{\varrho}_R$ are defined in Eq. (10.1).

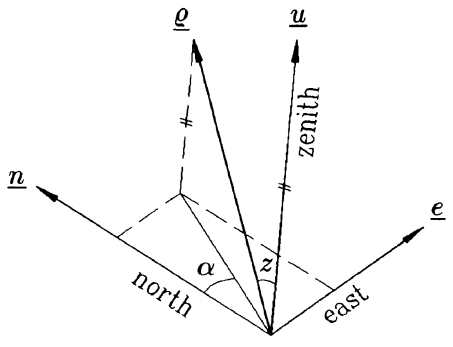


Fig. 7.3. Satellite's zenith angle z and azimuth α

With φ and λ being the ellipsoidal latitude and longitude of the observing site, the axes $\underline{n}$, $\underline{e}$, $\underline{u}$ of the local coordinate frame, representing the north, east, up directions, are given by

$$\begin{aligned} \underline{n} &= \begin{bmatrix} -\sin\varphi\cos\lambda \\ -\sin\varphi\sin\lambda \\ \cos\varphi \end{bmatrix} = \frac{\partial \underline{u}}{\partial \varphi} \\ \underline{e} &= \begin{bmatrix} -\sin\lambda \\ \cos\lambda \\ 0 \end{bmatrix} \qquad = \frac{1}{\cos\varphi}\frac{\partial \underline{u}}{\partial \lambda} \\ \underline{u} &= \begin{bmatrix} \cos\varphi\cos\lambda \\ \cos\varphi\sin\lambda \\ \sin\varphi \end{bmatrix} . \end{aligned} \tag{7.3}$$

Hence, adequate equations for the zenith angle z and the azimuth α, reckoned from north towards east positively, follow from the inner products (Fig. 7.3)

$$\begin{aligned} \underline{\varrho} \cdot \underline{n} &= \sin z \cos\alpha \\ \underline{\varrho} \cdot \underline{e} &= \sin z \sin\alpha \\ \underline{\varrho} \cdot \underline{u} &= \cos z . \end{aligned} \tag{7.4}$$

Table 7.4 is part of a typical listing of satellite locations. The leftmost column shows the Universal Time (UT) and each subsequent column shows the vertical (elevation) angle $E = 90° - z$ and azimuth α for several satellites.

A graphical representation of the complete elevation-azimuth list, based on an elevation cutoff angle of 15°, is given in Fig. 7.4. This figure indicates

Table 7.4. Part of elevation E and azimuth α list (both in degrees) for Graz, Austria, on September 2, 1995 (cutoff elevation: 15°)

SV	02		12		16		18		19		24		26		27	
UT	E	α	E	α	E	α	E	α	E	α	E	α	E	α	E	α
18:30					42	305	41	85	77	91	29	245			67	207
18:45					48	302	35	90	72	72	24	239			75	206
19:00	19	169			54	296	30	95	65	64	19	234			82	201
19:15	25	167			59	287	24	99	59	61					87	126
19:30	32	165			62	274	19	103	52	60					82	60
19:45	39	162	16	218	63	259	15	106	46	61			17	292	75	55
20:00	46	159	22	221	62	243			40	63			22	295	68	56
20:15	53	154	28	226	58	229			33	65			28	297	61	58
20:30	59	146	35	231	53	220			28	67			34	299	55	61

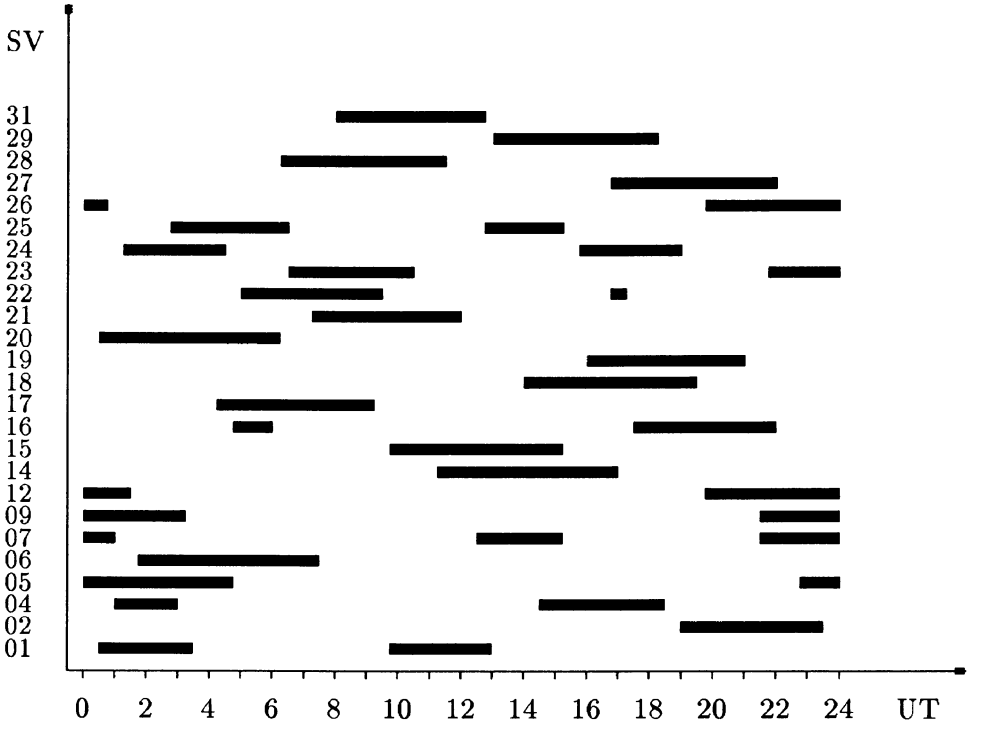


Fig. 7.4. Satellite visibility for Graz, Austria, on September 2, 1995 (cutoff elevation: 15°)

the satellites in view shown on the ordinate of the coordinate system and indicated as SV (space vehicle) with the corresponding PRN number for a certain epoch. No matter what observation technique is used or the precision

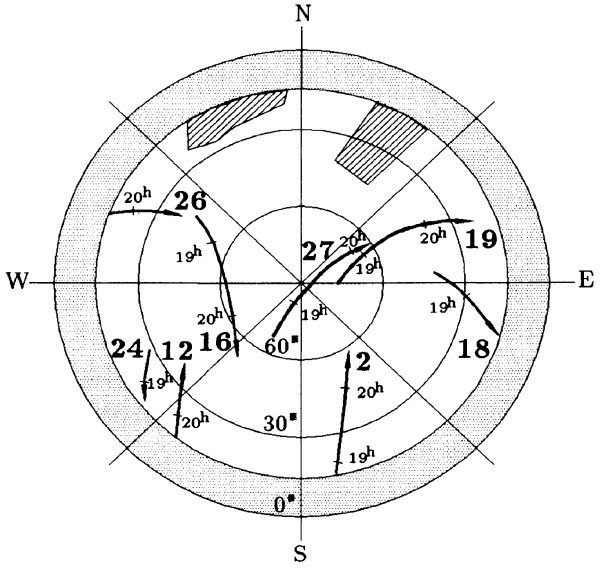


Fig. 7.5. Polar sky plot for Table 7.4

desired, only periods with four or more satellites in view are considered to provide viable satellite coverage.

An improved representation of visibility is given by the sky plots. These polar or orthogonal plots show the satellite pathes as a function of elevation angle and azimuth (Fig. 7.5). Such sky plots are often supplemented by time tags and by the image of the local horizon. The latter may be obtained from field reconnaissance (e.g., fish-eye photos) or can be generated by the computer from a digital terrain model.

Apart from visibility, the tracked satellites should be geometrically well distributed with (ideally) one in each of the four quadrants. This is only a guideline, however, not a firm requirement. In static surveys, poor satellite geometry and even lack of the fourth satellite can often be offset by observing for a longer period of time. The movement of the satellites with respect to each other improves the geometry and, thus, the solution. The surest method of selecting periods of adequate satellite coverage is to make test observations over known (or determined) baseline vectors for the full period being considered. This large data set can then be divided into smaller subsets for processing during periods being considered for observations. This method allows for testing the acceptability of both the specific period of coverage and the chosen session length. A measure for satellite geometry is the Geometric Dilution of Precision (GDOP) factor. Normally, GDOPs under six are considered good and those above six are considered as being

too high. The GDOPs reflect only the instantaneous geometry related to a single point. Therefore, factors for baseline vectors, accumulated over the time span of a session, are more appropriate precision indicators (Merminod et al. 1990). It has been proposed that quality factors can be computed by the receiver itself, and the selection of the optimum configuration could be automated. Appropriate formulas for GDOP calculation are provided in Sect. 9.6 where details on the decomposition of GDOP into several components such as Position Dilution of Precision (PDOP) are also given.

Another aspect for the selection of the window concerns the ionospheric refraction. Observations during night hours may be appropriate because the ionospheric effect is usually quieter during this time. Normally, however, daylight hours are preferred for organizational reasons.

Sessions

The specific time period chosen for an observation is called a session. The first observation of the day would be designated session "a", the second "b", and so on. Some manufacturers use number designation but numbers have the disadvantage of requiring two digits for sessions in excess of 10. Normally, the session designator begins with "a" again each day, and days are expressed as the consecutive calendar day (1 to 365 or, in leap years, 366). For example, session 105c means the third session of day 105.

A good time to begin the first session of static surveys is when four or more satellites are above the 15 to 20 degree elevation angle, and the last observation of this session should, generally, end when the fourth satellite drops below 15 to 20 degrees. This is only a general rule since three-satellite time prior to rise of the fourth satellite and after the fourth satellite has set is useful. There are the following factors that determine the length of a particular observation:

- the length of the baseline,
- the number of visible satellites (affects geometry),
- the relative geometry of the satellites and the change in geometry,
- the SNR of the received satellite signal.

In general, the more satellites that are available, the better the geometry, and the shorter the length of observation required. The length of a session may also be reduced in the case of shorter baselines. For example, sessions with lines 1–2 km in length could be as short as 20 minutes with six satellites and using single frequency receivers. Longer lines between control points might, on the other hand, require 90 minutes of data to achieve good results. For conventional static surveying, Table 7.5 (based on Table 7.3) may be used as a general guide to plan the session length when four to six satellites are available and the ionospheric conditions are normal.

Table 7.5. Session length in function of baseline length for conventional static surveying and single frequency receivers

Baseline [km]	Session [min]
1	20–35
5	25–45
10	35–60
20	55–90

What is the reason for needing these relatively long sessions? Chapter 6 discusses the observables and it is seen that at each observation epoch the carrier phase is measured to millimeter accuracy or even better. In effect, a single observation would be sufficient to provide the precision required for a geodetic survey. The difficulty is that one can measure the decimeters, centimeters, and millimeters precisely; however, the observation must last long enough to determine the meters by resolving the integer number of cycles. On short (less than 1 km) baselines, the integer cycles can often be resolved in 5–10 minutes using L1 only phase. With dual frequency receivers using the wide lane technique, long (15 km) lines can be accurately measured with as little as two minutes data.

The best method of determining the optimum session lengths for large projects is to make longer than normal observations on the first day to obtain typical data sets. For example, observations lasting 60 minutes for short (1–5 km) lines and 120 minutes for longer (5–20 km) lines would be made. These data sets when processed would yield excellent results. The observations could then be reprocessed using portions of the data set to determine the point where good results can no longer be obtained. For example, consecutive 30-minute data sets could be processed and compared with the full data set to determine if the shorter observation times were sufficient to achieve good results.

As discussed previously, a session should be long enough to guarantee the required accuracy; but, one should also consider that longer sessions cost more. In any case, the time between the sessions should be long enough to transport the equipment to another site and allow for accurate setup. Older receivers may also require oscillator warm-up.

In order to reference the single sessions to a common datum, at least one site of the network must be occupied during the entire project (radial surveying), or subsequent sessions must contain at least one reoccupied site (leap frogging). The reoccupation of more than one site improves the precision and reliability of the network. When planning sessions for kinematic

surveys, there are two factors to consider. Normally, times are selected when five or more satellites are above 20 degrees and when the satellites have a GDOP of less than six. For most locations, the GDOP condition is satisfied when five or more satellites are available. Problems occur when one of the satellites is obstructed during the survey and the remaining four satellites have a high GDOP. This problem is solved by keeping the roving receiver stationary until the fifth satellite is reacquired.

Nonplanned surveys

Before proceeding with descriptions of the planning steps, a brief discussion of nonplanned surveys will be given. Not all surveys require extensive planning. Some surveyors are now using GPS as they would use other survey equipment and are not necessarily planning a geodetic "campaign". A good example of a survey requiring a minimum of planning is a photo-control survey. Since many types of GPS surveys require that points be placed in locations where there may be substantial obstructions, the major planning effort is to layout a scheme and select sites that are relatively free from obstructions. Some types of photo-control surveys do not present the typical obstruction planning problem.

If the area being mapped is a suburban or residential area with clearings well-scattered throughout the area, one can perform the GPS (and photo) point selection during the actual survey. In this type of area, the only reconnaissance or presurvey activity required is to locate both horizontal and vertical control used to reference or tie the survey to the national datum.

Assuming that a photo-control project is performed in a fairly obstruction free area and that acceptable horizontal and vertical control points are nearby, the field crew can be sent to the area with the approximate desired photo-control sites plotted on a county road map or some other relatively small-scale maps. A good plan is to begin the survey in a portion of the project area where the sites are close together so the project manager can quickly coordinate the start of the survey. During the first session, the observers set up on their assigned points, while the project manager selects and "monuments" the next set of points. The monumentation may be an iron pin driven into the ground or it may be a nail driven into the pavement. Following selection of the second set of points and completion of the first session, the GPS observers are instructed to proceed to the second session, and the project manager selects the third set of points. Upon conclusion of the local scheme, the ties to existing control can be made to complete the survey.

A second example of nonplanned GPS surveys would be the establishment of control at a construction site. Many times, survey crews are sent

to construction sites with plans, coordinate lists, and other supporting data, and the crew chief makes decisions and plans the field work onsite. The same procedure could be followed when using GPS equipment. The crew chief would design the network onsite and establish points in areas where they were needed. Immediately following the observations, the data can be transferred from the GPS receivers to a laptop computer (for example in the survey vehicle). Modern laptop computers are able to process GPS data in a matter of seconds per line so that the data from several sessions (occupations) could be completed in less than one hour. The crew chief would then have accurate (first-order) coordinates of control onsite which could be used to layout desired construction stakes.

7.2.3 Field reconnaissance

After the GPS points have been plotted on a map and descriptions of how to reach the existing control have been obtained, one is ready to perform a field reconnaissance. This is also a good time to assign each point a unique identifier. The most obvious method is to consecutively assign each point a number. Points can have more descriptive (full name) identifiers as well, but a simple consecutive number facilitates future reference to each point. The reconnaissance surveyor visits each site to check its suitability based on the factors listed in the previous section.

First of all, static GPS surveys need an unobstructed view of the sky above an elevation of 15 to 20 degrees and a nonreflective environment. This is a critical requirement for kinematic applications where the path of the roving antenna should be selected carefully in advance. Easy access of the site is desired to save time between the sessions. This may be less important when cross-country vehicles are used. Eccentric occupations may often become necessary in the case of forests or urban areas. Sites that have many obstructions require additional consideration. At these sites the reconnaissance surveyor should prepare a polar plot showing the vertical angle and azimuth to obstructions over 20 degrees. This plot is then overlayed on the polar sky plot of the satellites (Fig. 7.5). The obstruction problem is solved in two ways. The first is to place the antenna on top of a survey mast so that the desired 20 degree visibility angle is obtained. The Geodetic Survey of Sweden, for instance, has devised 30 meter (guyed) survey masts that are quickly erected and plumbed over the mark by two offset theodolites. Several manufacturers produce prism poles that extend to 10 meters which also can be used for this purpose. The second technique for overcoming the problem of obstructions is to choose a time when a sufficient number of satellites is electronically visible at the site. The example in Fig. 7.5 shows

hatched portions that depict areas obstructed by trees, buildings, hills, etc. It is seen that these obstructions do not affect the observations at this one site. The data used to process a baseline vector consist of the common observations between two points; so the same obstruction check must be made for both ends of a line. Blockage of a satellite at one end of a line effectively eliminates that satellite from the solution; so care must be used in performing the analysis. The manual method of making this analysis is to produce satellite sky plots for every hour of the useable satellite span and then visually compare the site obstruction plots with the various one-hour plots. Another more automated method is to use the software produced by the manufacturers to perform this analysis on the computer.

Field reconnaissance is a must prior to conducting a kinematic survey (except when using receivers with OTF capability). Each site must be checked for sky visibility and the route taken to travel between points also must have good sky visibility. Since the kinematic method requires that lock be continuously maintained on four or more satellites, good sky visibility practically means obstruction free (above 20 degree elevation angle) situations. When obstructions on the route of travel (such as bridges) occur, static points can be placed on either side of the obstruction so that the roving receiver can be reinitialized. The path that the surveyor is to follow between points should be clearly marked on a large-scale map to make sure that unwanted cycle slips do not occur.

Reconnaissance for pseudokinematic surveys is not as important because this technique only requires the revisited sites be unobstructed. In the case of differential (navigation) surveys where code ranges are measured, reconnaissance is not critical at all because the receivers can be simply turned on and measurements made when desired. This mode would be used more as a precise navigation system than a survey system.

Apart from obstructions, it is important to consider the multipath problem. Multipath (more fully discussed in Sect. 6.6) is the effect of unwanted reflected satellite signals that are received by the antenna. This problem is most severe when the antenna is placed near a chain link fence or another metal structure. The satellite signals are reflected by the metal structure and corrupt the direct signals causing phase errors. In the case of chain link fences, the antenna can be elevated above the fence to eliminate the problem. When the point is close to a metal building, the only practical solution is to move the point to another location. Multipath does not appear to be a problem for points located in the median of highways where large trucks pass by at high speed. The multipath caused when a truck is near the antenna is of too brief duration to cause significant problems. However, one should avoid having a large metal transport truck parked next to the antenna.

When the site meets all requirements, the point can be marked for the marksetters to set the monument. The reconnaissance surveyor should plot the location of the chosen site on the largest scale map of the area and prepare a preliminary "to reach" description that describes how to reach the point from a known prominent location (e.g., local post office). This description will save hours in wasted time in the future since it will allow the marksetters and observers to quickly find the point. In cases where a different crew sets the actual monument, this mark setting crew completes the "to reach" description adding the final specific location of the mark and ties to nearby prominent objects. Surveys performed for inclusion in national networks require that a "to reach" description be prepared, and in some cases that a sketch of the site be made. The final site documentation should also include photos of the station surroundings, name and address of the owner, preliminary coordinates, power supply, etc.

7.2.4 Monumentation

The monumentation normally set eventually will become unnecessary when active control networks have been established. For the present, monumentation is still specified for projects where the sites are planned to be reoccupied for, e.g., geodynamical studies (Avdis et al. 1990). Note that monuments set other than in solid rock or by deep concrete pillars are not adequate for high precision surveys.

Monumentation is a general term to describe any object used to mark a point. Land surveyors and others commonly use sections of steel reinforcing bar upon which a cap is crimped to monument a point. Each surveyor must decide which particular type of mark is appropriate for the project. A steel rod may be appropriate to mark a photo-panel point; whereas, a massive concrete monument would be more appropriate for a county geodetic survey. The main consideration is the mark should be easily found, at least for the duration of the survey.

7.2.5 Organizational design

The planning phase for static surveys ends with the layout of an organizational design for the project. First, each field crew is allocated a number of personnel and appropriate vehicle and survey equipment. Each crew is then assigned sites to occupy during specific sessions. Field crews should be fully acquainted with the area and should be able to move quickly between points. Today, the operation of GPS receivers no longer requires highly qualified survey personnel. Still, malfunctions are better solved by trained crews.

The minimum number n of sessions in a network with s sites and using r receivers is given by

$$n = \frac{s - o}{r - o} \tag{7.5}$$

where o denotes the number of overlapping sites between the sessions. Equation (7.5) only makes sense for $o \geq 1$ and $r > o$. In the case of a real number, n must be rounded to the next higher integer.

Another approach for the design implies that each network site should be occupied m times. In this case, the minimum number of sessions equals

$$n = \frac{m\,s}{r} \tag{7.6}$$

where n again must be rounded to the next higher integer.

The number s_r of redundant occupied sites with respect to the minimum overlapping $o = 1$ is given by

$$s_r = n\,r - [s + (n - 1)]\,. \tag{7.7}$$

Consider a squared network consisting of nine sites P1, ..., P9 equally spaced and assume that three receivers A, B, C are available. If the "leap frogging" method with one-point overlapping between the sessions is the observation technique selected, Eq. (7.5) gives $n = 4$ as minimum number of sessions. The first four sessions in Table 7.6 represent one possible solution for the corresponding organizational design where the sites P4, P5, and P6 are reoccupied. When each point should be occupied twice then $m = 2$ and Eq. (7.6) gives $n = 6$ which is reflected by two more sessions in Table 7.6. The chosen design has the property that all short baseline vectors between adjacent sites are observed which provides a homogeneous accuracy in the network. For the elimination of receiver or antenna biases, an interchange of the equipment between the sessions is recommended.

The organizational design also depends upon the type of the network, i.e., the distribution of the sites (Unguendoli 1990). There are two basic types of GPS networks: (1) radial and (2) closed geometric figures.

Table 7.6. Principles of organizational design

Receiver	Session	a	b	c	d	e	f
A		P2	P4	P4	P6	P7	P1
B		P5	P5	P1	P3	P8	P2
C		P8	P6	P7	P9	P9	P3

Radial surveys

Radial (or cartwheel) surveys are performed by placing one receiver at a fixed site, and measuring lines from this fixed site to receivers placed at other locations. A typical radial survey configuration is shown in Fig. 7.6. There is no geometric consideration for planning this type of survey except that points in close proximity should be connected by direct observation. Consider in Fig. 7.6 the points 2 and 3 located 10 km from fixed point 1 and 100 m distant from one another. If they were surveyed at different times, the error between the two points could be considerable. A relative baseline error of 10 ppm, for example, produces a 0.1 m error in the points 2 and 3. Therefore, the expected error between the two points would be 0.14 m (as obtained from $\sqrt{0.1^2 + 0.1^2}$) which corresponds to a relative error of 1:714. This large relative error would not be tolerated by most surveyors since it would give a bad impression on the accuracy of the overall survey.

In general, kinematic surveys are radial surveys, and many pseudokinematic surveys are performed in the radial mode. Each point established by the radial method is a "no check" position since there is only one determination of the coordinates and there is no geometric check on the position. An appropriate use of radial surveys might be to establish photo-control since the photogrammetrist is able to make an independent check on the coordinates using analytical bridging. Other uses would be to provide positions for wells or geological features where precise coordinates may generally not be needed.

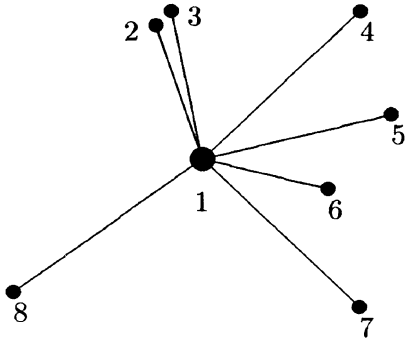


Fig. 7.6. Radial survey

Network survey

GPS surveys performed by static (and pseudokinematic) methods where accuracy is a primary consideration require that observations be performed in

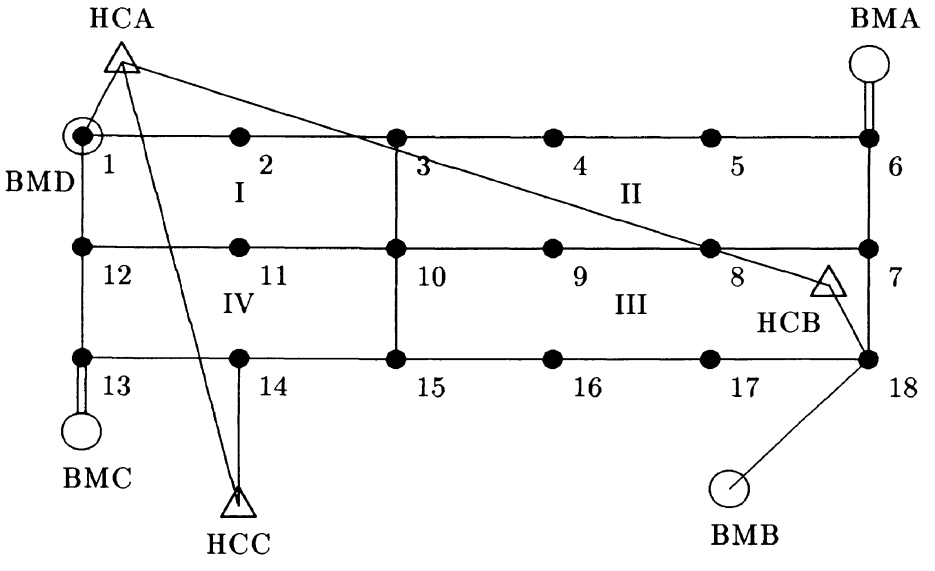


Fig. 7.7. Static network design

a systematic manner and that closed geometric figures be formed to provide closed loops. Figure 7.7 shows a typical scheme consisting of 18 points to be determined. The preferred observation scheme is to occupy adjacent points consecutively and traverse around the figure. For example, the scheme shown in Fig. 7.7 would be approached in the following manner if three receivers were used. Receiver A would be placed on point 1, receiver B on point 2, and receiver C on point 3 for the first session. Data would be collected from the three receivers for the chosen session length. The receivers would then be turned off and receiver A would be moved to point 11, and receiver B to point 10, while receiver C remains on point 3 to overlap between the first two sessions. Following the period during which the receivers were being moved, the three receivers would be turned on for the second session and data collection would be continued. An alternative to this approach is to move all three receivers so that receiver A would occupy point 3, receiver B point 11, and receiver C point 10 for the second session. Either plan works equally well; however, the second approach would eliminate eventual receiver biases, since the overlapping point 3 is occupied by different receivers in consecutive sessions. The leapfrog traversing technique is the preferred method because it provides the required loop checks and gives the maximum productivity. In general, the goal in devising a measuring scheme is to directly connect as many adjacent points as possible.

When national datum coordinates and elevations are desired for points in a scheme, ties to existing control must be made. In Fig. 7.7, horizontal

control points are marked by triangles and denoted HCA, HCB, HCC. Proper connection to control points requires that direct measurements be made between the existing control. The purpose of this measurement is to both verify the accuracy of the existing control and to determine scale, shift, and rotation between the control and the new GPS network. After the horizontal control has been occupied, measurements are made between the control and the nearest network point. For example, the points HCA, HCB, and HCC in Fig. 7.7 would be occupied during one session and the tie between HCA and point 1 would be made during another session.

Ties to vertical control points are performed somewhat differently than ties to horizontal points. There is no geometric reason for measuring between existing benchmarks (vertical control) since the error in the geoidal height masks any possible GPS error. Vertical ties are normally made from the benchmark to the closest GPS point. In Fig. 7.7, the benchmarks are shown as BMA, BMB, BMC, and BMD. In this figure, BMD was selected to be point 1. Thus, the need of a tie for this point is eliminated. Whenever possible, existing monuments should be chosen as GPS points to save placing a mark in the ground. In the figure, BMA and BMC were assumed to be close to points 6 and 13, respectively, so that a conventional level (loop) tie was made between benchmark and GPS point. Benchmark BMB was assumed to be far enough away from point 18 so that a GPS measurement was used to connect the two points. In the case of this benchmark, FGCS specifications call for two determinations of the vector between point 18 and point BMB.

Figure 7.7 is an idealized scheme; however, the basic principles apply for all control network schemes. In summary, the following are the major points of network design:

1. The network should consist of closed loops or other geometric figures.
2. Ties should be made to at least three horizontal control points which should also be directly occupied.
3. Ties should be made to at least four vertical control points (benchmarks) by the most direct means.

In addition to these general specifications, other measurements may be prescribed by various agencies, such as making a certain percentage of repeat measurements and setups.

In the case of one-point overlapping per session and using three receivers, a network consisting of 22 sites (i.e., 18 GPS points, 3 horizontal control points, 1 benchmark) would require 11 sessions where one baseline is measured twice, cf. Eqs. (7.5) and (7.7). With 15 sessions, nine loop closures can

Table 7.7. Organizational design for the configuration in Fig. 7.5

Session	Receiver		
	A	B	C
a	HCA	HCB	HCC
b	HCA	1	2
c	3	10	2
d	11	10	12
e	1	13	12
f	HCC	13	14
g	15	10	14
h	3	4	5
i	6	7	5
j	9	7	8
k	9	10	15
l	16	17	15
m	BMB	17	18
n	BMB	HCB	18
o	8	7	18

be formed. One possible occupation schedule for determining the coordinates of the 22 points is listed in Table 7.7.

Under the assumption of 60-minute sessions and 30 minutes for site change, the 15 sessions would require two days under ideal conditions. A more realistic schedule would be to take three days to make the sessions longer than one hour.

7.3 Surveying procedure

The actual execution of GPS surveys is greatly facilitated by good planning. A well-planned survey normally progresses quite smoothly. This section will discuss the various aspects of conducting a GPS survey and provide information to make the effort more productive.

7.3.1 Preobservation

Antenna setup

To avoid multipath or imaging effects, it is recommended that any vehicle be parked as far away as possible from the antenna. Parking the survey vehicle the full 10 m (antenna cable length) away from the point should be adequate in most cases.

The antennas are mounted on pillars, tripods (with tribrach), or on a range (or prism) pole equipped with steadying legs. This range pole with legs is called a bipod or tripod pole and its use can materially speed up the conduct of a survey. When performing kinematic or pseudokinematic surveys, the bipod pole is essential to speed up the setup time because the antenna is maintained at a fixed height. The use of a bipod pole for static surveys is not critical; however, its use reduces the possibility of an undetected blunder in measuring the antenna height (since the height of the rod is fixed).

A rather minor problem using GPS should not be overlooked: the calibration of tribrachs. A survey will only be as accurate as the surveyor's ability to center the antenna over the survey point. The best way to avoid this problem is to use two-piece collimating tribrachs that can be rotated to check the centering. A more cost effective approach is to hang a plumb bob from the tribrach at each setup to check the optical centering device.

The antenna phase center offset discussed in Sect. 6.5 is due to the fact that the geometric center of the antenna is not at the electronic center. This problem is virtually eliminated by using only one type of antenna and by pointing the antennas in the same direction. In the surveying mode, differences between points are measured so that any systematic offset will be canceled by uniform antenna orientation.

The measurement of the antenna phase center above the reference point is an important aspect which is often overlooked. Experience has shown that the mismeasured antenna height is the single most vexing problem in conducting GPS surveys and causes the greatest number of errors. The best way to avoid this problem (when using tripods) is to measure the antenna height twice, once at the beginning and once after completion of the survey. Some manufacturers are providing special rods to facilitate antenna height measurements. When using tripods, the setup at "swing" or repeat points should be broken during the move and the tripod reset over the point at a different height.

Receiver calibration

In general, GPS receivers are considered to be self-calibrating and users do not normally perform equipment calibration. One simple test that can be performed, however, is a zero baseline measurement. This measurement is made by connecting two or more receivers to one antenna. Care should be taken in doing this to use a special device that blocks the voltage being fed to the antenna from all but one receiver. Also, a signal splitter must be used to divide the incoming signal to the multiple receivers.

A normal session (e.g., 60 minutes) is observed and the baseline is computed in the normal manner. Since a single antenna is used, the baseline components should all be zero. This measurement essentially checks the functioning of the receiver circuits and electronics and is a convenient method of trouble shooting receiver problems independent from antenna biases. The zero baseline test is also one way to satisfy specifications which call for equipment calibration.

Initialization

In static surveying, the initialization of some receivers requires the preprogramming or the on-site input of parameters. Some of these are: the selection of the sampling rate, the bandwidth, the minimum number of satellites to track, the start and stop time for the session, the cutoff elevation angle, and the assignment of a data file name. Most of the modern receivers have several channels and track all satellites in view. A preselection is, thus, only necessary to disregard a satellite. The ephemerides are generally gathered and stored automatically by the receiver. Many receivers have been designed to require the minimum of operator interaction. These receivers are operated by simply turning them on.

In the kinematic mode, the phase ambiguities are determined during initialization by static or kinematic techniques. One static technique is based on the occupation of a short, known baseline which allows ambiguity resolution after a few observation epochs. Another method is to perform a (rapid) static survey to determine the vector between the fixed point and the unknown starting point for the kinematic survey. A third static method is to perform an antenna swap between the fixed point and the starting point (Hofmann-Wellenhof and Remondi 1988). This method is accomplished by placing receiver A at the fixed point and receiver B on the starting point. After a few epochs of observation, receiver A is moved to the starting point and receiver B is moved to the fixed point. During this move, both receivers must continuously track a minimum of four satellites. The antenna swap is completed by moving receiver A back to the fixed point and receiver B back to the starting point. By this antenna swap, the vector between the two points is determined to millimeter accuracy for short lines.

The kinematic initialization on-the-fly (OTF) is the most advanced technique to resolve phase ambiguities. This technique is nearly identical to the rapid static technique. The method theoretically can use single frequency receivers if the code range error is sufficiently small, e.g., 10 cm (this is achievable with the narrow correlator spacing technique which reduces both noise and multipath). The method works best with dual frequency full wavelength receivers where the carrier phase and code phase are measured

on both frequencies. Tests with this procedure show that 20 km baselines can be determined to centimeter accuracy in as little as two minutes. The technique is described in more detail in Sect. 9.2.

Logistically, if the fixed receiver antenna were at a remote location such as the roof of the surveyor's office, the survey starting point at a given site would be located by a static survey. On the other hand, if the fixed point were at the project site, the starting point could be placed near (i.e., 10 m) the fixed site and surveyed by the antenna swap method for single frequency, or by OTF for dual frequency receivers.

7.3.2 Observation

Communications between survey crews are desirable and generally increase efficiency. Today, the major portion of geodetic surveys, where lines average 20 km in length, are performed with communication between observers using portable cell phones. In cases when an observer is late reaching the point, this possibly prevents a resurvey of the lines to that point since the schedule could be revised on the spot. Land (as opposed to geodetic) GPS surveys normally measure lines averaging 5 km so that normal frequency modulated (FM) transceivers (hand-held radios) can be used to communicate between observers. The need for communications is most critical when conducting pseudokinematic surveys because it is important that all receivers collect data during the same time span (except a fixed full time running reference receiver is used).

Most static observations can be performed in an automated mode so that an operator is not required. However, it is good practice to perform data checks during the session, and any irregularity should be noted in the field log. Also, an operator is often needed to keep the receiver from harm or loss. When conducting geodetic surveys (campaigns) of highest accuracy, meteorological data (wet and dry temperature, air pressure) should be collected during the session at each site as far above the ground as possible. In extreme cases, backscatter radiometers may be used to measure the atmospheric water vapor. The inclusion of observed meteorological data does not improve the results for short baselines as opposed to using standard tropospheric refraction models. The data may possibly be used for future studies.

In kinematic surveying applications, after initialization, the fixed receiver and the roving receiver are placed on the fixed and (now known) starting point for a few epochs of observation. Then the roving receiver proceeds to the points for which coordinates are desired. As long as four or more satellites (with low PDOP) are continuously tracked by both receivers, vectors from

the fixed point can be measured to a high degree of accuracy. If (in the case of four satellites) the signal lock is lost or a cycle slip occurs, the initialization must be repeated. This may happen due to shadowing (i.e., obstruction) of the satellite signal by buildings, bridges, trees, or other objects. In practice, points surveyed should be visited twice so that a check on the determination can be made. Also, several points whose coordinates are known (i.e., from static surveys) should be included in the survey, when possible, to provide additional checks.

The roving receiver normally remains at a point for a few epochs so that observations can be averaged to obtain a more accurate position. A second way to perform kinematic surveys is to alternate the fixed and roving receivers by leapfrogging units. For example, consider consecutive points $1, 2, 3, \ldots, n$ along an open highway. Receiver A would be placed at point 1 and receiver B at starting point 2 near point 1. The two receivers could perform an antenna swap to obtain the starting coordinates of point 2. Receiver B would then move to point 3, with receiver A remaining at point 1. Receiver B would then remain fixed at point 3, while receiver A was moved to point 4. In this way, the vectors 2–3, and 3–4 will be measured. Continuing in this manner, a geometric traverse of kinematically surveyed points is obtained.

Precaution is recommended during thunderstorms. Lightning strikes may destroy the instrument. Therefore, the receivers should be switched off and the antenna disconnected.

7.3.3 Postobservation

At the completion of a session, a check of the antenna position including a remeasurement of its height is recommended. Older instruments with external recording devices require the labeling of the data storage media.

A final recommendation is to prepare a site occupation sheet that should be completed at the end of the session. This sheet should contain at least the following information:

- project and station name,
- date and session number,
- start and stop times,
- station identifier used for file name,
- name of observer,
- receiver and antenna serial number,
- height of antenna and eccentricities in position,
- meteorological data,
- problems experienced.

In addition to this minimum amount of information, it is good practice to make a "rubbing" of the top of the survey mark or make a photograph of the mark and to prepare a site sketch of the mark's location (or check existing sketch).

7.3.4 Ties to control monuments

The practice of connecting a GPS network to a horizontal control point is most simple when the control point is in a clear location so that the GPS antenna is placed directly upon the mark. In many cases, however, this is not possible because the control monument is located in an area with many obstructions or where multipath is a problem. In these instances, the point must be occupied eccentrically.

Figure 7.8 shows a typical eccentric occupation. In this figure, it is assumed that the control point C is obstructed and that the GPS receiver can be placed on A, a nearby eccentric point. A conventional azimuth mark and second auxiliary point B are also shown. When an azimuth mark exists at the control point, the problem is simplified because the azimuth to the eccentric point can be determined by measuring the angle at the control point between the azimuth mark and the eccentric point. An alternative method would be to measure the astronomic (third-order) azimuth to the eccentric point. A second auxiliary point such as point B should be used as a check on the measured distance from the control point to the eccentric point, and all angles and distances should be measured in the triangle ABC. A second technique can be used when there is no azimuth mark at the control point. Referring to Fig. 7.8, one GPS antenna would be placed on the eccentric mark A and a second receiver placed upon point B. In this case, the distance between the eccentric and the auxiliary point should exceed the distance from the eccentric point to the control station. From the GPS baseline vector, the azimuth of the line from A to B can be derived and

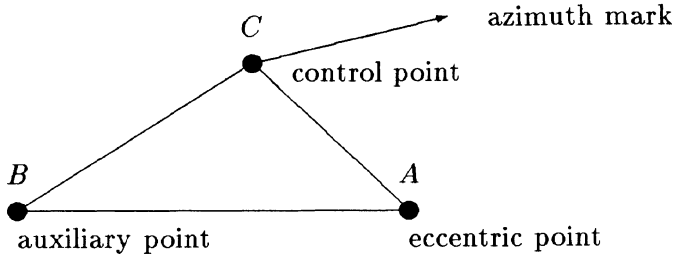


Fig. 7.8. Eccentric occupation

the azimuth from A to the control point C can be computed by the angle measured at A. The azimuth from the control point to the eccentric point is determined in one iteration by subtracting the convergence from the reverse azimuth.

Occupying a benchmark eccentrically is accomplished by setting an eccentric point in a clear area near the benchmark and double running levels between the two marks. It should be remembered when occupying horizontal control points eccentrically, where the control point has an accurate elevation, to level between the marks.

7.4 In situ data processing

7.4.1 Data transfer

Modern equipment stores the GPS observations internally. The first step in processing is to transfer the data from the receiver to a computer hard disk. This transfer is accomplished using software provided by the manufacturer. Observation files for a given session contain the phase and other observables. In addition, the broadcast ephemerides and site data consisting of station identifier, antenna height, and possibly navigation position may be contained in other files. The main task in transferring files is to make sure the files are named correctly and that the antenna height is correct.

A good procedure to ensure that the filenames and antenna heights are correct is to prepare an abstract form (Table 7.8) containing the assignment of sites to sessions, the measured antenna heights, and the start and stop times of the sessions.

Many times, an observer will enter the incorrect site identifier. This identifier must be immediately corrected before proceeding with the processing. The controlling document should be the site occupation sheets from which the information in Table 7.8 was abstracted. With this information, one knows for instance that point 001 was occupied by receiver B from 02:29 to 03:30 UT. If this point were misidentified, one could look at the start and stop times in the given data file also noting the navigation latitude and longitude. This start and stop time and position contained in each data file pinpoints that particular file to a given time and location which can then be matched with the data in Table 7.8. Using this technique, misnamed files can be identified and renamed.

Most of the batch file processing software automatically extracts the antenna height from the site file data stored in the receiver. After correcting the names of the various files, one should then check and correct all antenna

Table 7.8. Control abstract

Session	Receiver						Time (UT)	
	A		B		C		Start	Stop
	Site	Height	Site	Height	Site	Height		
a	HCA	1.234	HCB	1.574	HCC	1.342	01:00	02:00
b	HCA	0.987	001	1.782	002	1.543	02:29	03:30
c	003	1.344	010	1.328	002	1.452	04:01	05:00
d	011	1.324	010	1.563	012	1.437	05:31	06:30
e	001	1.564	013	1.453	012	1.455	06:59	08:00

heights. Again a form such as Table 7.8 will be helpful in keeping track of the various antenna heights. Also, a table listing the various occupations and antenna heights is useful for inclusion in the project report.

As soon as all files have been corrected, the observed data should be backed up on at least two sets of storage medium (e.g., floppies). A good practice is to keep one set of original data in a different secure location. Today, many organizations keep a complete set of storage medium in a vault.

7.4.2 Data processing

In the case of remote and extended GPS surveys, a quality control check of the data at least once per day should be made. This check could also include preliminary computations of baseline vectors in the field before leaving the survey area. The on-site vector processing on a daily basis helps ensure that adequate measurements are being made.

Today, most of the routine processing is performed by batch file processing. All batch files are generated using the three- or four-digit site identification so that the first task in processing GPS data is to ensure that all sites are properly named. A good practice is to first number the control points and then consecutively assign numbers to adjacent unknown points. This numbering should have been accomplished during the planning and reconnaissance phase to facilitate record keeping.

Processing of static surveys

Modern processing software uses batch processing to compute baseline vectors. Usually, the data for a given day are loaded into a subdirectory on a hard disk. Processing software is normally in another directory and the "path" of the computer has to be set to access the programs. Once the software has been initiated (generally by menu commands), the lines are computed in order, automatically. There are two types of processing software: (1) vector by vector and (2) multipoint solutions.

The vector-by-vector or single-baseline solution type is presently the most common, and in any case should be used prior to processing with the multi-point software. In some instances, one of the points in an observation session will be corrupted and if all points are processed together, the errors from the bad point are distributed among the vectors and the error is masked. The single vector software provides a better check on bad lines or points. The bad point can be more easily isolated by noting that the statistics (i.e., root mean square error, standard error) to lines leading to this one point are worse than the statistics for the other lines. Additionally, the vectors can be summed for the lines in the session, and if the sum around the perimeter is not a small value (e.g., 1 ppm), this indicates that one of the points in the session is bad.

In general, the individual vector processing software performs the following steps:

1. Generation of orbit files.
2. Computation of the best fit value for point positions from code pseudoranges.
3. Creation of undifferenced phase data from receiver carrier phase readings and satellite orbit data. Time tags may also be corrected.
4. Creation of differenced phase data and of computation of their correlations.
5. Computation of an estimate of the vector using triple-difference processing. This method is insensitive to cycle slips but provides least accurate results.
6. Computation of the double-difference solution solving for vector and (floating point or real) values of phase ambiguities.
7. Estimation of integer values for the phase ambiguities computed in the previous step, and decision whether to continue with fixed ambiguities.
8. Computation of the fixed bias solution based upon best ambiguity estimates computed in the previous step.
9. Computation of several other fixed bias solutions using integer values differing slightly (e.g., by 1) from selected values.
10. Computation of the ratio of statistical fit between chosen fixed solution and the next best solution. This ratio should be at least two to three indicating that the chosen solution is at least two to three times better than the next most likely solution.

Processing of kinematic surveys
Today, many of the kinematic surveys are performed in the RTK mode where the observations at the reference site are transmitted via radio link to the rover. The simultaneous observations of both the reference and the rover site are then combined and processed by receiver internal software in real time. In the postprocessing mode, the basic steps are similar for processing static and kinematic surveys. The data files are downloaded from the receiver to the computer and the filenames and antenna heights are checked. The actual processing differs depending on the software used; however, much of the newer software is automated so that hands-on interaction is not needed. The main check for kinematic vectors is to compute positions of the roving receiver and check that similar values are obtained on separate visits to the same point. Also, it is good survey practice to visit points whose coordinates are known during the survey as a further check of the method.

7.4.3 Trouble shooting and quality control

This subsection mainly considers postprocessing. In real-time applications, trouble shooting and quality control is preferably performed automatically by the receiver internal software.

Single baseline vectors
There are various quality numbers that indicate how well the survey was performed. The first analysis involves inspecting the statistics of individual vectors. The key for determining which vectors are bad is to compare the statistics of a good line with those of a bad line. The two baselines shown in Table 7.9 are actual data collected by single frequency receivers. Both lines of similar length were of 30 minutes duration, and five satellites were in view for nearly the entire period. The sampling rate was 20 seconds.

The results were abstracted from the output or solution files. The first column in Table 7.9 indicates the type of solution: triple-difference (TRP),

Table 7.9. Baseline statistics

1	2 ΔX [m]	3 ΔY [m]	4 ΔZ [m]	5 $\sigma_{\Delta X}$ [m]	6 $\sigma_{\Delta Y}$ [m]	7 $\sigma_{\Delta Z}$ [m]	8 Ratio	9 rms [m]
TRP	-303.457	135.317	158.292	1.419	1.011	0.520		0.003
FLT	-303.431	135.314	158.284	0.068	0.062	0.027		0.003
FIX	-303.437	135.327	158.263	0.003	0.006	0.004	30.1	0.004
TRP	-191.888	-343.451	-546.721	6.686	0.699	1.488		0.004
FLT	-192.221	-343.366	-546.721	0.538	0.069	0.102		0.007
FIX	-192.217	-343.192	-546.689	0.027	0.107	0.051	1.7	0.052

double-difference with real valued (float) ambiguities (FLT), and double-difference with the ambiguity biases fixed to integers (FIX). The next three columns are the components of the baseline vector (i.e., differences of ECEF coordinates defined in Chap. 3). The fifth through the seventh column are the standard errors of the three coordinate differences listed in columns two through four. Column eight (single number) is the ratio between the chosen fixed bias solution and the next best solution. The last column is the root mean square (rms) error of fit in meters.

The basic line statistics in Table 7.9 indicate both a good baseline and a problem baseline. The single most important number in the solution output is the ratio shown in column eight. This number is a strong indicator of both good and bad lines, and should be greater than three, especially for short baselines up to about 5 km. When the ratio is a large number as in the first solution, the line is an accurate measurement. In this case, the difference between the fixed and float solutions is small and the change in the rms between the fixed and float solutions is not dramatic. The first baseline is included in Table 7.9 as an example of a good line with "normal" statistics. The second baseline in Table 7.9 is a bad line which is clearly shown by several indicators. First, note that the ratio (column 8) is 1.7 compared to a similar line with a ratio of 30.1. The standard error of the vector components (columns 5 to 7) are substantially greater for the second line although the number of observations was actually greater for this baseline. The critical numbers in Table 7.9 are the differences between the fixed and float solutions. Note that the difference in the ΔY-value for the second baseline is 0.174 m. In the computation of this line, the software default was chosen to compute the fixed bias solution even though the ratio did not indicate the fixed solution would be a valid one. This "forced" computation of the fixed bias often yields valid solutions, but particularly on short lines with brief observations, the forced solution yields bad results. Another indicator that the observations do not fit the solution model is the large jump in the rms shown in the last column. The rms of the second baseline jumped from 0.007 to 0.052, a sevenfold increase. Compare this change to the rms of the first line which shows the normal slight increase of the three types of solutions.

The statistics of the second line clearly indicate that the fixed bias solution is invalid. This fact is further demonstrated by the inconclusive bias values. The computed ambiguities are compared in Table 7.10. Satellite 24 and satellite 16 are the reference satellites for the double-differences of the first and the second baseline respectively. It is clearly seen that the ambiguities of the first baseline are obvious by inspection since the values are close to integers. On the other hand, the integer biases of the second line are not obvious. The bias 17-16 could be rounded either up or down, and the

Table 7.10. Float ambiguity solution

First baseline		Second baseline	
Satellites	Ambiguities	Satellites	Ambiguities
12-24	-970431.089	03-16	-184540.781
13-24	-832899.977	17-16	-155781.542
16-24	2235113.884	20-16	29969.388
20-24	61256.060	24-16	12931.837

biases 03-16, and 20-16 are too close to 0.5 to be clearly defined. It is a good practice, when in doubt, to inspect the ambiguities to determine if there is an obvious problem. If the measured line is short and sufficient observations are available, one can try reprocessing the line using a different reference satellite, or try eliminating satellites with troublesome bias values. In the case of the second line, there was simply an insufficient amount of data.

Networks

For networks the best single tool for finding problem lines is to use loop closure software that sums the vector components around a loop to determine misclosures. Most GPS software packages have this type of software.

The first step in checking loops is to prepare a simple sketch showing the measured lines. Depending upon the software, the lines should then be numbered to correspond with the data file used by the software, or, in some cases, the path through the network can be chosen by referring to certain node points in the network. Bad lines in a loop are found by computing different combinations that both include and exclude the bad line. For example, if the line from point 10 to 11 (Fig. 7.7) were a bad line, it would be discovered by a high misclosure in the loops I and IV. When the larger loop around both small loops is computed (i.e., 1–3–15–13–1), the loop would have an acceptable closure. Therefore, the bad line could be narrowed to either line 10–11, or line 11–12. Normally, one would then inspect the output files for these lines more closely to determine if one or the other was more suspect. At this point, either the suspect line would be remeasured or point 11 would be dropped or downweighted.

The second quality control procedure in a network is to compute a minimally constrained least squares adjustment using one of the numerous adjustment programs available today. This adjustment should normally be performed only after bad lines have been eliminated using the loop closure programs. Chapter 9 gives a more complete description of least squares adjustment and only some practical uses will be discussed here. Each software package contains programs that prepare the necessary input files automati-

cally from the various output files.

Least squares adjustment programs perform three basic tasks: (1) they first shift all vectors so that they are connected in a contiguous network, (2) they add small corrections to each vector component to obtain a "flat" geometric figure closure, and (3) they compute coordinates and elevations of all points. The network design is important in obtaining useful information from an adjustment. A single loop would provide the minimum required information, while a network such as the one in Fig. 7.7 would provide a high degree of redundancy needed for network analysis.

In the batch processing of GPS data, the navigation position is often used as the starting coordinate for the vector computation and the solution is linearized around this value. Note that an error of 20 m in the starting position introduces a baseline error of about 1 ppm. When performing a minimally constrained adjustment, the datum coordinates of one point are normally entered, and the coordinates of all vectors are shifted to agree with the chosen coordinates of the one point. Therefore, a listing of the amounts each point was shifted to be consistent with the fixed point is available and can be printed. It is good practice to scan this list of shifts to determine if the starting coordinates for any of the vectors was in error by more than a reasonable amount. For example, if one of the shifts were 200 meters, you would probably wish to recompute that particular vector using a better starting coordinate (e.g., obtained from preliminary adjustment). Also, these shifts give a good indication of how well the C/A-code pseudorange positions agree with the datum position.

The proper weighting of observations (vectors) is always a problem. Many of the adjustment programs build a weight matrix based upon the vector correlation matrix and standard errors in the GPS output file. The formal standard errors of the vector computation are generally optimistic by a factor of three to ten times. Therefore, the weight matrix composed of these optimistic values must be scaled to arrive at a true estimate of the network errors. The problem is to determine what scale factor to use. Experience with the various software packages will show how large a scale factor is appropriate. Also, the adjustment software or the GPS software documentation should indicate the various scaling factors to use.

Adjustment software that does not use the vector correlation matrix requires the a priori estimates of the vector errors. The allowable error for first-order surveys, that is 1 : 100 000, according to the specifications of the FCGS at the 2σ confidence level is $\pm(10\,\text{mm} + 10\,\text{ppm})$ of the distance. This error can be chosen as the a priori value for the adjustment, but first must be halved to express the value in terms of the 1σ or standard error level. When this $\pm(5\,\text{mm} + 5\,\text{ppm})$ a priori error is used, and the adjustment yields a

standard error of unit weight of one or less, one knows that the work meets first-order specifications.

Another method for checking the quality of the GPS network is to inspect the residuals which result from the adjustment. Normally, there are two residual lists. One comprises the actual amounts that vector components have been corrected to achieve (exact) closure of the network. These residuals should consist of small values for short lines and larger values for longer lines. Blunders or large errors will be distributed throughout a general area and may be difficult to isolate. A better way to isolate large errors is to use the loop closure software in combination with adjustment programs.

The second list of residuals provided by most programs comprises the normalized or standardized residuals. These values are unitless and are the actual residuals scaled. A value of 1.0 for a normalized residual would indicate that the residual was as large as expected using the a priori weight model. A value less than one indicates the residual is less than expected and a value greater than one indicates a larger than expected value. In an adjustment, a few normalized residuals as high as 2.0 are to be expected. When more than 5% of the residuals are greater than 2.0 and for any value over 3.0, further investigations should be performed, and some lines possibly eliminated.

A typical case where redundant lines should be eliminated from a least squares adjustment is shown in Fig. 7.7. Suppose the receivers were placed on points 1, 2, 3, and 4. The vectors 1–2, 2–3, and 3–4 could have accurate fixed bias solutions, while the vector 1–4 could be a float bias solution. This difference would result from the fact that integer biases are more easily fixed on short lines but are difficult to fix on long lines. In this case, it would be appropriate to eliminate vector 1–4 from the adjustment.

When a kinematic survey has been performed using the leapfrog traversing technique, the traverse or geometric figure can be adjusted in a manner similar to a static survey. The same procedure should be followed to isolate bad lines, or bad elevations, or horizontal points.

7.4.4 Datum transformations

To this point only the computation of vectors and the analysis of their accuracy have been dealt with. After having successfully isolated and eliminated bad lines, the coordinates of points in the network can be computed. The first such computation is the final minimally constrained adjustment using all good vectors.

The tie of GPS surveys to existing triangulation monuments and vertical benchmarks enables the transformation of GPS results into the national

datum. These passive control networks have the disadvantage that many sites have to be occupied and maintained. This system is appropriate where dense national triangulation networks exist and when the control network serves other purposes such as geodynamical investigations. Such a dense passive control network has been established, e.g., in Austria as reported by Stangl et al. (1991).

The network shown in Fig. 7.7 has three horizontal control points. One of these points would be designated as fixed and one of the vertical points would be fixed in the free adjustment. If the point HCA were held fixed, the horizontal coordinates of the other two points would be determined. These (free adjusted) values should agree with the known values within a reasonable amount. If, for example, one point agreed well with the adjusted value while the other point disagreed, one could isolate the bad control point.

In most cases, the fixed control agrees quite well and the final adjustment is then performed. The constrained (fixed) adjustment results when two or more horizontal control points and three or more vertical control points are held fixed.

Horizontal datum

In Fig. 7.7, the horizontal coordinates of HCA, HCB, and HCC would be fixed and the GPS network would be rotated and scaled to fit these values. Rotation values of 1 or 2 arcseconds and scales of up to 10 ppm are common when the fixed control has been established by triangulation, and less than 1 arcsecond and a few ppm are expected when the fixed control is established by dual frequency GPS (e.g., supernets).

The coordinates of points in a GPS network are computed in the same manner as coordinates of triangulation or traverse points. Network coordinates are, essentially, interpolated between the fixed control points using the measured vectors to apportion the values. Geodetic coordinates are, thus, referenced to the values of the fixed control and are, therefore, on the same datum as the fixed control. Proper ellipsoidal parameters should be used when performing the least squares adjustment. For example, when North American Datum 1983 (NAD-83) coordinates are used in an adjustment, the Geodetic Reference System 1980 (GRS-80) ellipsoid should be used. When the computation of coordinates for a survey on a different datum is desired, the coordinates of the fixed control and the proper ellipsoidal values must be used for the new adjustment.

Vertical datum

The computation of elevations with GPS is complicated by the fact that GPS heights are purely geometric and referenced to a surface known as the

ellipsoid, while elevations measured with leveling equipment are referenced to a surface known as the geoid (Sect. 10.2.4). Geodesists at various universities and national geodetic organizations have developed a mathematical model that approximates the shape of the geoid (which is quite irregular). These models are based on data from satellites, astronomic observations, and gravity measurements. These geoid models have been used to compute geoid-ellipsoid separation values (i.e., geoidal heights) on regular grids. For example, the NGS distributes a file containing these geoidal heights for every three minutes of arc in latitude and longitude for the conterminous United States. These theoretical geoidal heights may be used to provide (theoretically) more accurate adjusted elevations.

In general, the ellipsoid and geoid are separated from one another (e.g., some 30 m in eastern U.S.) and are also tilted to one another. Over a moderate size area (10 km $\times$ 10 km) with smooth topography, it can be assumed that the geoid has a similar curvature like the ellipsoid and compute the rotation angles and translation distance between the two surfaces (Sect. 10.3.3).

In practice, more than the minimum three required elevations may be available for points in a GPS network, and a least squares adjustment could be performed. Referring to Fig. 7.7, the elevations of points 1, 6, and 13 have been determined and a vector has been measured (twice) to BMB, so that there are four known elevations connected by a network. The transformation of ellipsoidal heights to elevations is accomplished by holding the vertical component of the four points fixed in the adjustment. The adjustment will rotate and translate the entire network so that the elevations of all points are quite close to the true values. This technique works well where the geoid is similar in shape to the ellipsoid; however, in mountainous areas, this assumption is not valid and the difference must be accounted for.

In certain cases, many more than the minimum required elevations may be available for points in a GPS network. For example, a 100-point network covering a county size area may have ten elevations scattered throughout the area. Out of the ten known elevations, there may be as many as five or six combinations of three points that satisfy the "well distributed geometrically" criterion. Therefore, to check for badly known elevations, five to six preliminary adjustments of the 100-station network would have to be performed. Another way of isolating bad elevations can be found in Collins (1989) where an example is given. This method first involves performing a minimally constrained or "free" adjustment holding one of the vertical points fixed (and at least one horizontal point). The geoidal heights are included in this free adjustment so that all points in the network now have so-called pseudoelevations. For example, in Fig. 7.7, the elevation of BMD could be fixed and the pseudoelevations of the remaining points computed

by using the geoidal height differences (derived from the geoid model) and the ellipsoidal height differences (determined from the adjustment). The pseudoelevations approximate the true elevations although they contain a systematic error. In a moderate size nonmountainous area (10 km × 10 km), the pseudoelevations are normally within 10 to 20 cm of the true elevations. Many times, the pseudoelevations will be within a few centimeters of the true elevations. Standard regression or least squares adjustment can be employed to compute the 3 × 3 normal equation matrix used to determine the transformation parameters (i.e., one shift parameter and two rotation angles). Bad elevations can be quickly isolated in this way since the rotated elevations for the redundant points should agree within a few centimeters of the fixed elevation. After all given elevations have been checked for blunders, the final elevations of points in the network are computed. This is easily accomplished by fixing a minimum of three (geometrically diverse) elevations and performing a fixed adjustment. For example, the elevations at points 1, 6, and 18 in Fig. 7.7 would be fixed in the adjustment, and the elevation of point 13 would serve as a check. When three elevations are fixed in the adjustment, the rotation angles about the north-south and east-west axes are computed. These rotation angles serve as a further check since they should (normally) not exceed a few arcseconds.

The differences of theoretical geoidal heights are also useful in determining elevations in a small area where only one benchmark is available. The pseudoelevations for the area will be closer to the true values than if no geoid model were used and a single point were held fixed. This method should only be used when approximate elevations are desired and the technique should be used with caution since all elevations depend on the single fixed point.

No definite answer can be given to the question of the accuracy of GPS derived elevations since the accuracy depends upon the shape of the geoid in the particular area. Normally, in flat continental areas, the geoid is relatively smooth and elevations within a 10 km × 10 km area can be determined to 3 cm or better. In mountainous areas, the interpolated (holding three elevations fixed) elevations can be in error by several decimeters.

7.4.5 Computation of plane coordinates

GPS surveys performed for local engineering projects or for mapping normally require that plane coordinates be computed for all points. There are several conformal projections used for this purpose. The two most popular in the U.S. are the Transverse Mercator projection and the Lambert projection. In the U.S., each state has a different projection system (in some cases several) and programs to transform latitude and longitude to and from these

projections can be purchased from the NGS. Also, most GPS software includes these transformation programs. In Europe, transformation programs are readily available from the various mapping and surveying organizations.

It should be remembered that geodetic latitude and longitude are specifically referenced to a given datum. In the U.S., the older North American Datum 1927 (NAD-27) has a specific set of projections to use, while the NAD-83 uses a completely different set of projections. The projections between datums should never be mixed.

A number of commercially available transformation programs transforms both ellipsoidal and plane coordinates between datums and projections. It should be noted that these transformation programs are not intended for geodetic use; they only approximate the coordinates of the new datum. Such programs can produce coordinates with errors of up to a meter.

Special local transformations can provide quite accurate coordinates of points that do not have published values in the new system. For example, not all geodetic points have published NAD-83 values. Appropriate software uses the old and new geodetic coordinates of points surrounding the old (without new values) point to compute local transformation parameters used to compute new coordinates for the old point. This type of program works quite well and provides geodetic quality coordinates.

7.5 Survey report

A final survey report is helpful to others in analyzing the conduct of a survey. In effect, a project report should document the survey so that another competent individual could recreate all computations arriving at similar answers to those obtained. This report serves as an "audit trail" for the survey. A project report should address the following topics:

1. Location of the survey and a description of the project area. A general map showing the locality is recommended.
2. Purpose of the survey and the extent that the requirements were satisfied. Also, the specifications followed should be mentioned (e.g., FGCS GPS Specifications Order A).
3. A description of the monumentation used. It should be specifically noted if underground as well as surface marks were used. A section of the report should explain which existing monuments were searched for and which ones were found. For monuments not found, an estimate of the time spent searching for the mark should be given to aid others who wish to recover that mark. A list of all control searched for, control found, and control used is helpful in analyzing the project.

4. A description of the instrumentation used should include both the GPS equipment and conventional equipment along with serial numbers. An explanation of how the tribrachs or bipods were tested for plumb should be given. If survey towers or special range poles were used, their use should be described with an explanation of how the antenna was collimated.
5. The computation scheme for the project should be described including which version of the processing software was used and which least squares adjustment was applied. An exhibit such as Table 7.8 is particularly useful in visualizing how the survey was conducted. The satellites tracked during each session should be included in this listing or in a separate exhibit. Also, an exhibit such as Table 7.9 should be included to show the quality factors for each line. The inclusion of abstracts of each vector would also be helpful.
6. The computation of coordinates for all eccentric points should be included. In the case of horizontal points, sketches and the direct computation results should be shown. For vertical points, a copy of the level records should be included in an appendix.
7. All problems encountered should be discussed and equipment failures listed. Unusual solar activity should be mentioned as well as multipath problems and other factors affecting the survey.
8. The following lists should be included in the report:
   - list of loop closures,
   - occupation schedule (e.g., Table 7.8),
   - vector statistics (e.g., Table 7.9),
   - free least squares adjustment,
   - portion of fixed adjustment showing rotation angles and statistics,
   - list of adjusted positions and plane coordinates,
   - project statistics,
   - copies of "original" site occupation logs,
   - equipment malfunction log,
   - project sketch showing all points and control with a title box, scale, and projection tic marks.
9. Finally, a copy of the original observations (or translated RINEX) should be transmitted as part of the survey. Copies of the adjustment input file and vector output files would be advantageous.

When surveys are properly performed and documented, they provide a lasting contribution to the profession. Often, data can be used in later years by others to study a particular phenomenon or the work may be included in a larger project. Proper use of measurements can only be made when the survey is thoroughly documented for posterity.

8 Mathematical models for positioning

8.1 Point positioning

8.1.1 Point positioning with code ranges

Code range model

The code pseudorange at an epoch t can be modeled, cf. Eq. (6.2), by

$$R_i^j(t) = \varrho_i^j(t) + c\,\Delta\delta_i^j(t)\,. \tag{8.1}$$

Here, $R_i^j(t)$ is the measured code pseudorange between the observing site i and the satellite j, $\varrho_i^j(t)$ is the geometric distance between the satellite and the observing point, and c is the speed of light. The last item to be explained is $\Delta\delta_i^j(t)$. This clock bias represents the combined clock offsets of the receiver and the satellite clock with respect to GPS time, cf. Eq. (6.1).

Examining Eq. (8.1), the desired point coordinates to be determined are implicit in the distance $\varrho_i^j(t)$, which can explicitly be written as

$$\varrho_i^j(t) = \sqrt{(X^j(t) - X_i)^2 + (Y^j(t) - Y_i)^2 + (Z^j(t) - Z_i)^2} \tag{8.2}$$

where $X^j(t)$, $Y^j(t)$, $Z^j(t)$ are the components of the geocentric position vector of the satellite at epoch t, and X_i, Y_i, Z_i are the three unknown ECEF coordinates of the observing site. Now, the clock bias $\Delta\delta_i^j(t)$ must be investigated in more detail. For the moment consider a single epoch; a single position i is automatically implied. Each satellite contributes one unknown clock bias which can be recognized from the superscript j at the clock term. Neglecting, for the present, the site i clock bias, the pseudorange equation for the first satellite would have four unknowns. These are the three site coordinates and one clock bias of this satellite. Each additional satellite adds one equation with the same site coordinates but with a new satellite clock bias. Thus, there would always be more unknowns than measurements. Even when an additional epoch is considered, new satellite clock biases must be modeled due to clock drift. Fortunately, the satellite clock information is known and transmitted via the broadcast navigation message in the form of three polynomial coefficients a_0, a_1, a_2 with a reference time t_c. Therefore, the equation

$$\delta^j(t) = a_0 + a_1(t - t_c) + a_2(t - t_c)^2 \tag{8.3}$$

B. Hofmann-Wellenhof et al., *Global Positioning System*

enables the calculation of the satellite clock bias at epoch t. It should be noted that the polynomial (8.3) removes a great deal of the satellite clock bias, but a small amount of error remains.

The combined bias term $\Delta\delta_i^j(t)$ is split into two parts by

$$\Delta\delta_i^j(t) = \delta_i(t) - \delta^j(t) \tag{8.4}$$

where the satellite related part is known by (8.3) and the receiver related term $\delta_i(t)$ remains unknown. Substituting (8.4) into (8.1) and shifting the satellite clock bias to the left side of the equation yields

$$R_i^j(t) + c\,\delta^j(t) = \varrho_i^j(t) + c\,\delta_i(t)\,. \tag{8.5}$$

Note that the left side of the equation contains observed or known quantities, while the terms on the right side are unknown.

Basic configurations

Basic configurations are defined by the condition that the number of observations must be equal to or greater than the number of unknowns. This condition is sufficient but does not necessarily give a solution. The reason for this is that inherent rank deficiencies may prevent a numerical solution because of a singularity. More explanations are given later when the rank deficiency becomes an issue.

The number of observations is $n_j\,n_t$ where n_j denotes the number of satellites and n_t the number of epochs.

For static point positioning, the three coordinates of the observing site and the receiver clock bias for each observation epoch are unknown. Thus, the number of unknowns is $3 + n_t$. The basic configuration is defined by

$$n_j\,n_t \geq 3 + n_t \tag{8.6}$$

which yields the explicit relation

$$n_t \geq \frac{3}{n_j - 1}\,. \tag{8.7}$$

The minimum number of satellites to get a solution is $n_j = 2$ leading to $n_t \geq 3$ observation epochs. For $n_j = 4$, the solution $n_t \geq 1$ is obtained. This solution reflects the instantaneous positioning capability of GPS, where the four unknowns at any epoch are solved if at least four satellites can be tracked.

For kinematic point positioning, the basic configuration can be directly derived from the following consideration. Due to the motion of the receiver,

the number of the unknown station coordinates is $3n_t$. Adding the n_t unknown receiver clock biases, the total number of unknowns is $4n_t$. Hence, the basic configuration is defined by Eq. (8.6),

$$n_j\, n_t \geq 4n_t \tag{8.8}$$

yielding $n_j \geq 4$. In other words, the position (and velocity) of a moving receiver can be determined at any instant as long as at least four satellites are tracked. Geometrically, the solution is represented by the intersection of four pseudoranges. For the rigorous analytical solution see Kleusberg (1994), Lichtenegger (1995).

The basic configurations must be considered from a theoretical point of view. The solution $n_j = 2$, $n_t \geq 3$ for static point positioning, for example, means that simultaneous observations of two satellites over three epochs would theoretically suffice. In practice, however, this situation would yield unacceptable results or the computation would fail because of an ill-conditioned system of observation equations unless the epochs were widely spaced (e.g., hours). A solution is also possible if observations of three epochs for two satellites are made, followed by three additional epochs (e.g., seconds apart) for two other satellites. Such an application will be rare but is imaginable under special circumstances (e.g., in urban areas).

8.1.2 Point positioning with carrier phases

Phase range model

Pseudoranges can also be obtained from carrier phase measurements. The mathematical model for these measurements, cf. Eq. (6.9), is given by

$$\Phi_i^j(t) = \frac{1}{\lambda}\,\varrho_i^j(t) + N_i^j + f^j\,\Delta\delta_i^j(t) \tag{8.9}$$

where $\Phi_i^j(t)$ is the measured carrier phase expressed in cycles, λ is the wavelength, and $\varrho_i^j(t)$ is the same as for the code range model. The time independent phase ambiguity N_i^j is an integer number and, therefore, often called integer ambiguity or integer unknown or simply ambiguity. The term f^j denotes the frequency of the satellite signal, and $\Delta\delta_i^j(t)$ is the combined receiver and satellite clock bias.

Substituting Eq. (8.4) into Eq. (8.9) and shifting the satellite clock bias, known from Eq. (8.3), to the left side of the equation yields

$$\Phi_i^j(t) + f^j\,\delta^j(t) = \frac{1}{\lambda}\,\varrho_i^j(t) + N_i^j + f^j\,\delta_i(t)\,. \tag{8.10}$$

Basic configurations

Using the same notations as before, the number of observations is again

$n_j\, n_t$. The number of unknowns, however, is increased by the number n_j because of the ambiguities.

For static point positioning, the number of unknowns is composed of 3 coordinates of the observing station, n_j unknown ambiguities, and n_t unknown receiver clock biases. Referring to (8.10), the problem of the rank deficiency is encountered. Mathematically less interested readers may skip the next paragraph.

A few basics on rank and rank deficiency are given here. Deeper insight may be obtained from Koch (1987), Sects. 132, 333. Assume a large number of equations of type (8.10) being prepared to be solved for the unknowns. This implies a matrix-vector representation where the right side is composed of a product of a design matrix $\underline{A}$ and a vector comprising the unknowns in linear form. The rank of the design matrix is equal to the order of the largest nonsingular matrix that can be formed inside $\underline{A}$. Formulated differently: the maximum number of the linearly independent rows of matrix $\underline{A}$ is called the rank r of the matrix and is denoted by $r = \operatorname{rank} \underline{A}$. Linear dependence of two rows means that their linear combination yields a zero vector. The word rows in this definition may also be replaced by the word columns. For a simpler discussion, assume a quadratic matrix with $m \times m$ rows and columns. Thus, if the largest nonsingular matrix is the matrix $\underline{A}$ itself, the rank equals $r = \operatorname{rank} \underline{A} = m$ and the matrix is regular, i.e., it may be inverted without troubles. On the other hand, if the largest nonsingular matrix inside $\underline{A}$ is a matrix with, e.g., $(m-2) \times (m-2)$ rows and columns, the rank would be $r = m - 2$ and implies a rank deficiency of $m - r$ which turns out to be $m - (m-2)$ which amounts to 2. As a consequence, the singular system becomes regularly solvable if two unknowns (also denoted as parameters) are arbitrarily chosen. This equals the "fixing" of two parameters. Figuratively speaking, two of the parameters may be transferred to the left side of the matrix-vector system comprising the measurements. This transfer reduces on the other hand the columns of the matrix on the right side by the amount of the rank deficiency, i.e., in the example discussed by two. This concludes the short discussion on rank and rank deficiency.

The model in the form (8.10) comprises a rank deficiency of 1, this means that one of the unknown parameters may (and must) be arbitrarily chosen. Suppose that a receiver clock bias at one epoch is chosen, then, instead of n_t unknown receiver clock biases, only $n_t - 1$ clock biases remain. Therefore, the basic configuration for static point positioning without rank deficiency is defined by the relation

$$n_j\, n_t \geq 3 + n_j + (n_t - 1) \tag{8.11}$$

which yields explicitly the required number of epochs as

$$n_t \geq \frac{n_j + 2}{n_j - 1}. \tag{8.12}$$

The minimum number of satellites to get a solution is $n_j = 2$ leading to $n_t \geq 4$ observation epochs. Another integer solution pair is $n_j = 4$, $n_t \geq 2$.

For kinematic point positioning with phases, $3n_t$ unknown station coordinates must be considered because of the roving receiver compared to the 3 unknowns in (8.11). The other considerations including the discussion on the rank deficiency remain unchanged. Therefore, the basic configuration is defined by

$$n_j\, n_t \geq 3n_t + n_j + (n_t - 1) \tag{8.13}$$

yielding the explicit relation

$$n_t \geq \frac{n_j - 1}{n_j - 4}. \tag{8.14}$$

The minimum number of satellites to get a solution is $n_j = 5$ which have to be tracked for $n_t \geq 4$ epochs. Another integer solution pair is $n_j = 7$, $n_t \geq 2$.

Note that solutions for a single epoch (i.e., $n_t = 1$) do not exist for point positioning with carrier phases. As a consequence, kinematic point positioning with phases is only possible if the n_j phase ambiguities are known from some initialization. In this case, the phase range model converts to the code range model.

8.1.3 Point positioning with Doppler data

The mathematical model for Doppler data, cf. Eq. (6.10), is

$$D_i^j(t) = \dot{\varrho}_i^j(t) + c\,\Delta\dot{\delta}_i^j(t) \tag{8.15}$$

and may be considered as time derivative of a code or phase pseudorange. In this equation, $D_i^j(t)$ denotes the observed Doppler shift scaled to range rate, $\dot{\varrho}_i^j(t)$ is the instantaneous radial velocity between the satellite and the receiver, and $\Delta\dot{\delta}_i^j(t)$ is the time derivative of the combined clock bias term.

The radial velocity for a stationary receiver, cf. Eq. (4.45),

$$\dot{\varrho}_i^j(t) = \frac{\underline{\varrho}^j(t) - \underline{\varrho}_i}{\|\underline{\varrho}^j(t) - \underline{\varrho}_i\|} \cdot \dot{\underline{\varrho}}^j(t) \tag{8.16}$$

relates the unknown position vector $\underline{\varrho}_i$ of the receiver to the instantaneous position vector $\underline{\varrho}^j(t)$ and velocity vector $\dot{\underline{\varrho}}^j(t)$ of the satellite. These vectors

can be calculated from the satellite ephemerides. The contribution of the satellite clock to $\Delta\dot{\delta}_i^j(t)$ is given by, cf. Eq. (8.3),

$$\dot{\delta}^j(t) = a_1 + 2a_2(t - t_c) \tag{8.17}$$

and is known. Summarizing, the observation equation (8.15) contains four unknowns. These unknowns are the three receiver coordinates $\underline{\varrho}_i$ and the receiver clock drift $\dot{\delta}_i(t)$. Hence, compared to the code range model, the Doppler equation contains the receiver clock drift instead of the receiver clock offset.

The concept of combined code pseudorange and Doppler data processing leads to a total of five unknowns. These unknowns are the three point coordinates, the receiver clock offset, and the receiver clock drift. Each satellite contributes two equations, one code pseudorange and one Doppler equation. Therefore, three satellites are sufficient to solve for the five unknowns.

The similarity of the pseudorange and the Doppler equation gives rise to the question of a linear dependence of the equations. However, it can be shown that the surfaces of constant pseudoranges and the surfaces of constant Doppler are orthogonal and hence independent (Levanon 1999).

8.2 Differential positioning

8.2.1 Basic concept

As shown in Sect. 7.1.2, differential positioning with GPS, abbreviated by DGPS, is a real-time positioning technique where two or more receivers are used. One receiver, usually at rest, is located at the reference or base station with known coordinates and the remote receivers are usually roving and their coordinates are to be determined. The reference station calculates pseudorange corrections (PRC) and range rate corrections (RRC) which are transmitted to the remote receiver in real time. The remote receiver applies the corrections to the measured pseudoranges and performs point positioning with the corrected pseudoranges. The use of the corrected pseudoranges improves the positional accuracy with respect to the base station.

8.2.2 DGPS with code ranges

Generalizing (8.5) and following Lichtenegger (1998), the code range at base station A to satellite j measured at epoch t_0 may be modeled by

$$R_A^j(t_0) = \varrho_A^j(t_0) + \Delta\varrho_A^j(t_0) + \Delta\varrho^j(t_0) + \Delta\varrho_A(t_0) \tag{8.18}$$

where $\varrho_A^j(t_0)$ is the geometric range, the term $\Delta\varrho_A^j(t_0)$ denotes range biases depending on the terrestrial base position and satellite position as well (e.g.,

radial orbital error, refraction effects), the range bias $\Delta\varrho^j(t_0)$ is purely satellite dependent (e.g., effect of satellite clock error), and the range bias $\Delta\varrho_A(t_0)$ is purely receiver dependent (e.g., effect of receiver clock error, multipath). Note that noise has been neglected in (8.18).

The pseudorange correction for satellite j at reference epoch t_0 is defined by the relation

$$\begin{aligned} \mathrm{PRC}^j(t_0) &= \varrho_A^j(t_0) - R_A^j(t_0) \\ &= -\Delta\varrho_A^j(t_0) - \Delta\varrho^j(t_0) - \Delta\varrho_A(t_0) \end{aligned} \tag{8.19}$$

and can be calculated since the geometric range $\varrho_A^j(t_0)$ is obtained from the known position of the reference station and the broadcast ephemerides and $R_A^j(t_0)$ is the measured quantity. In addition to the pseudorange correction $\mathrm{PRC}^j(t_0)$, the time derivative or range rate correction $\mathrm{RRC}^j(t_0)$ is determined at the base station.

Range and range rate corrections referring to the reference epoch t_0 are transmitted to the rover site B in real time. At B the pseudorange corrections are predicted for the observation epoch t using the relation

$$\mathrm{PRC}^j(t) = \mathrm{PRC}^j(t_0) + \mathrm{RRC}^j(t_0)\,(t - t_0) \tag{8.20}$$

where $t - t_0$ is defined as latency. The achievable accuracy increases for smaller variations of the pseudorange corrections and for smaller latencies.

Adapting (8.18) to the rover site B and epoch t, the code pseudorange measured at the rover can be modeled by

$$R_B^j(t) = \varrho_B^j(t) + \Delta\varrho_B^j(t) + \Delta\varrho^j(t) + \Delta\varrho_B(t)\,. \tag{8.21}$$

Applying the predicted pseudorange correction $\mathrm{PRC}^j(t)$, cf. Eq. (8.20), to the measured pseudorange $R_B^j(t)$ yields

$$R_B^j(t)_{\mathrm{corr}} = R_B^j(t) + \mathrm{PRC}^j(t) \tag{8.22}$$

or, after substitution of (8.21) and the pseudorange correction according to (8.19) and (8.20), respectively,

$$R_B^j(t)_{\mathrm{corr}} = \varrho_B^j(t) + [\Delta\varrho_B^j(t) - \Delta\varrho_A^j(t)] + [\Delta\varrho_B(t) - \Delta\varrho_A(t)] \tag{8.23}$$

where the satellite dependent bias has canceled out. For moderate distances between the base and the rover site, the satellite-receiver specific biases are highly correlated. Therefore, the influence of radial orbital errors and

of refraction is significantly reduced. Neglecting these biases, Eq. (8.23) simplifies to

$$R_B^j(t)_{\text{corr}} = \varrho_B^j(t) + \Delta\varrho_{AB}(t) \tag{8.24}$$

where $\Delta\varrho_{AB}(t) = \Delta\varrho_B(t) - \Delta\varrho_A(t)$. If multipath is neglected, this term converts to the combined receiver clock bias scaled to range, i.e., $\Delta\varrho_{AB}(t) = c\,\delta_{AB}(t) = c\,\delta_B(t) - c\,\delta_A(t)$. If no latency exists, the equation above is identical with the between-receiver single-difference of code ranges measured at A and B, and differential positioning converts to relative positioning (Sect. 8.3).

Positioning at the rover site B is performed with the corrected code pseudoranges $R_B^j(t)_{\text{corr}}$ leading to improved positional accuracies. The basic configuration for DGPS with code ranges is identical with that for kinematic point positioning with code ranges, cf. Eq. (8.8).

8.2.3 DGPS with phase ranges

Generalizing (8.9) and following Lichtenegger (1998), the phase pseudorange measured at the base station A at epoch t_0 can be modeled by

$$\lambda\,\Phi_A^j(t_0) = \varrho_A^j(t_0) + \Delta\varrho_A^j(t_0) + \Delta\varrho^j(t_0) + \Delta\varrho_A(t_0) + \lambda\,N_A^j \tag{8.25}$$

where again, in analogy to the code range model, $\varrho_A^j(t_0)$ is the geometric range, $\Delta\varrho_A^j(t_0)$ is the satellite-receiver dependent bias, $\Delta\varrho^j(t_0)$ is purely satellite dependent, $\Delta\varrho_A(t_0)$ is purely receiver dependent. Finally, N_A^j is the phase ambiguity. Consequently, the phase range correction at reference epoch t_0 is given by

$$\begin{aligned} \text{PRC}^j(t_0) &= \varrho_A^j(t_0) - \lambda\,\Phi_A^j(t_0) \\ &= -\Delta\varrho_A^j(t_0) - \Delta\varrho^j(t_0) - \Delta\varrho_A(t_0) - \lambda\,N_A^j\,. \end{aligned} \tag{8.26}$$

The formulation of range rate corrections in the base station A as well as the application of predicted range corrections to the observed phase ranges in the rover site B is carried out in full analogy to the previously described code range procedure. Therefore,

$$\lambda\,\Phi_B^j(t)_{\text{corr}} = \varrho_B^j(t) + \Delta\varrho_{AB}(t) + \lambda\,N_{AB}^j \tag{8.27}$$

results for the corrected phase ranges where $\Delta\varrho_{AB}(t) = \Delta\varrho_B(t) - \Delta\varrho_A(t)$ and $N_{AB}^j = N_B^j - N_A^j$ is the (single-) difference of the phase ambiguities. As in the code range model, if multipath is neglected, the term $\Delta\varrho_{AB}(t)$ converts to the combined receiver clock bias scaled to range, i.e., $\Delta\varrho_{AB}(t) = c\,\delta_{AB}(t) = c\,\delta_B(t) - c\,\delta_A(t)$.

Point positioning at the rover site B is performed with the corrected phase pseudoranges $\lambda \Phi_B^j(t)_{\text{corr}}$. The basic configuration for DGPS with phase ranges is identical with that for kinematic point positioning with phase ranges, cf. Eq. (8.14).

DGPS with phase ranges, sometimes denoted as carrier phase differential technique, is used for most precise real-time kinematic applications. For this mode of operation, OTF techniques are required to resolve the ambiguities. Note that for these techniques at least five common satellites must be observed in the base and in the rover station. Note also that differential DGPS with phases converts to RTK if the latency becomes zero.

8.3 Relative positioning

The objective of relative positioning is to determine the coordinates of an unknown point with respect to a known point which, for most applications, is stationary. In other words, relative positioning aims at the determination of the vector between the two points which is often called the baseline vector or simply baseline (Fig. 7.2). Let A denote the (known) reference point, B the unknown point, and $\underline{b}_{AB}$ the baseline vector. Introducing the corresponding position vectors $\underline{X}_A$, $\underline{X}_B$, the relation

$$\underline{X}_B = \underline{X}_A + \underline{b}_{AB} \tag{8.28}$$

may be formulated, and the components of the baseline vector $\underline{b}_{AB}$ are

$$\underline{b}_{AB} = \begin{bmatrix} X_B - X_A \\ Y_B - Y_A \\ Z_B - Z_A \end{bmatrix} = \begin{bmatrix} \Delta X_{AB} \\ \Delta Y_{AB} \\ \Delta Z_{AB} \end{bmatrix} . \tag{8.29}$$

The coordinates of the reference point must be given in the WGS-84 system and are usually approximated by a code range solution.

Relative positioning can be performed with code ranges, cf. Eq. 8.5), or with phase ranges, cf. Eq. (8.10). Subsequently, only phase ranges are explicitly considered. Relative positioning requires simultaneous observations at both the reference and the unknown point. This means that the observation time tags for the two points must be the same. Assuming such simultaneous observations at the two points A and B to satellites j and k, linear combinations can be formed leading to single-differences, double-differences, and triple-differences. Differencing can basically be accomplished in three different ways: across receivers, across satellites, across time (Logsdon 1992: p. 96). Instead of "across" frequently "between" is used. In order to avoid overburdened expressions, shorthand notations

will be used throughout the textbook with the following meanings: single-difference corresponds to across-receiver difference (or between-receiver difference), double-difference corresponds to across-receiver and across-satellite difference, and triple-difference corresponds to across-receiver and across-satellite and across-time difference. Most postprocessing software uses these three difference techniques, so their basic mathematical modeling is shown in the following sections.

8.3.1 Phase differences

Single-differences
Two points and one satellite are involved. Denoting the points by A and B and the satellite by j and using Eq. (8.10), the phase equations for the two points are

$$\begin{aligned} \Phi_A^j(t) + f^j\,\delta^j(t) &= \frac{1}{\lambda}\,\varrho_A^j(t) + N_A^j + f^j\,\delta_A(t) \\ \Phi_B^j(t) + f^j\,\delta^j(t) &= \frac{1}{\lambda}\,\varrho_B^j(t) + N_B^j + f^j\,\delta_B(t) \end{aligned} \tag{8.30}$$

and the difference of the two equations is

$$\begin{aligned} \Phi_B^j(t) - \Phi_A^j(t) &= \frac{1}{\lambda}\left[\varrho_B^j(t) - \varrho_A^j(t)\right] + N_B^j - N_A^j \\ &\quad + f^j\left[\delta_B(t) - \delta_A(t)\right]. \end{aligned} \tag{8.31}$$

Equation (8.31) is referred to as the single-difference equation. This equation stresses one aspect of the solution for the unknowns on the right side. A system of such equations would lead to a rank deficiency even in the case of an arbitrarily large redundancy. This means that the design matrix of the adjustment has linearly dependent columns and a rank deficiency exists. Therefore, the relative quantities

$$\begin{aligned} N_{AB}^j &= N_B^j - N_A^j \\ \delta_{AB}(t) &= \delta_B(t) - \delta_A(t) \end{aligned} \tag{8.32}$$

are introduced. Using additionally the shorthand notations

$$\begin{aligned} \Phi_{AB}^j(t) &= \Phi_B^j(t) - \Phi_A^j(t) \\ \varrho_{AB}^j(t) &= \varrho_B^j(t) - \varrho_A^j(t), \end{aligned} \tag{8.33}$$

and substituting (8.32) and (8.33) into (8.31) gives

$$\Phi_{AB}^j(t) = \frac{1}{\lambda}\,\varrho_{AB}^j(t) + N_{AB}^j + f^j\,\delta_{AB}(t) \tag{8.34}$$

which is the final form of the single-difference equation. Note that the satellite clock bias has canceled, compared to the phase equation (8.10).

Double-differences
Assuming the two points A, B, and the two satellites j, k, two single-differences according to Eq. (8.34) may be formed:

$$\begin{aligned}\Phi^j_{AB}(t) &= \frac{1}{\lambda}\varrho^j_{AB}(t) + N^j_{AB} + f^j\,\delta_{AB}(t)\\ \Phi^k_{AB}(t) &= \frac{1}{\lambda}\varrho^k_{AB}(t) + N^k_{AB} + f^k\,\delta_{AB}(t)\,.\end{aligned} \tag{8.35}$$

To obtain a double-difference, these single-differences are subtracted. Assuming equal frequencies $f^j = f^k$ for the satellite signals, the result is

$$\Phi^k_{AB}(t) - \Phi^j_{AB}(t) = \frac{1}{\lambda}\left[\varrho^k_{AB}(t) - \varrho^j_{AB}(t)\right] + N^k_{AB} - N^j_{AB}\,. \tag{8.36}$$

Using shorthand notations for the satellites j and k analogously to (8.33), the final form of the double-difference equation is

$$\Phi^{jk}_{AB}(t) = \frac{1}{\lambda}\varrho^{jk}_{AB}(t) + N^{jk}_{AB}\,. \tag{8.37}$$

The canceling effect of the receiver clock biases is the reason why double-differences are preferably used. This cancellation resulted from the assumptions of simultaneous observations and equal frequencies of the satellite signals.

Symbolically, the convention

$$*^{jk}_{AB} = *^k_{AB} - *^j_{AB} \tag{8.38}$$

has been introduced where the asterisk may be replaced by Φ, ϱ, or N. Note that these terms comprising two subscripts and two superscripts are actually composed of four terms. The symbolic notation

$$*^{jk}_{AB} = *^k_B - *^j_B - *^k_A + *^j_A \tag{8.39}$$

characterizes, in detail, the terms in the double-difference equation:

$$\begin{aligned}\Phi^{jk}_{AB}(t) &= \Phi^k_B(t) - \Phi^j_B(t) - \Phi^k_A(t) + \Phi^j_A(t)\\ \varrho^{jk}_{AB}(t) &= \varrho^k_B(t) - \varrho^j_B(t) - \varrho^k_A(t) + \varrho^j_A(t)\\ N^{jk}_{AB} &= N^k_B - N^j_B - N^k_A + N^j_A\,.\end{aligned} \tag{8.40}$$

Triple-differences
So far only one epoch t has been considered. To eliminate the time independent ambiguities, Remondi (1984) has suggested differencing double-differences between two epochs. Denoting the two epochs by t_1 and t_2, then

$$\begin{aligned}\Phi_{AB}^{jk}(t_1) &= \frac{1}{\lambda}\varrho_{AB}^{jk}(t_1) + N_{AB}^{jk} \\ \Phi_{AB}^{jk}(t_2) &= \frac{1}{\lambda}\varrho_{AB}^{jk}(t_2) + N_{AB}^{jk}\end{aligned} \tag{8.41}$$

are the two double-differences, and

$$\Phi_{AB}^{jk}(t_2) - \Phi_{AB}^{jk}(t_1) = \frac{1}{\lambda}\left[\varrho_{AB}^{jk}(t_2) - \varrho_{AB}^{jk}(t_1)\right] \tag{8.42}$$

is the triple-difference which may be written in the simplified form

$$\Phi_{AB}^{jk}(t_{12}) = \frac{1}{\lambda}\varrho_{AB}^{jk}(t_{12}) \tag{8.43}$$

if the symbolic formula

$$*(t_{12}) = *(t_2) - *(t_1) \tag{8.44}$$

is applied to the terms Φ and ϱ. It should be noted that both $\Phi_{AB}^{jk}(t_{12})$ and $\varrho_{AB}^{jk}(t_{12})$ are actually composed of eight terms each. Resubstituting (8.42) and either (8.39) or (8.40) yields

$$\begin{aligned}\Phi_{AB}^{jk}(t_{12}) \;=\; &+ \Phi_B^k(t_2) - \Phi_B^j(t_2) - \Phi_A^k(t_2) + \Phi_A^j(t_2) \\ &- \Phi_B^k(t_1) + \Phi_B^j(t_1) + \Phi_A^k(t_1) - \Phi_A^j(t_1)\end{aligned} \tag{8.45}$$

and

$$\begin{aligned}\varrho_{AB}^{jk}(t_{12}) \;=\; &+ \varrho_B^k(t_2) - \varrho_B^j(t_2) - \varrho_A^k(t_2) + \varrho_A^j(t_2) \\ &- \varrho_B^k(t_1) + \varrho_B^j(t_1) + \varrho_A^k(t_1) - \varrho_A^j(t_1)\,.\end{aligned} \tag{8.46}$$

The advantage of triple-differences is the canceling effect for the ambiguities and, thus, the immunity from changes in the ambiguities. Such changes are called cycle slips and are treated in more detail in Sect. 9.1.2.

8.3.2 Correlations of the phase combinations

In general, there are two groups of correlations, (1) the physical and (2) the mathematical correlations. The phases from one satellite received at two points, for example $\Phi_A^j(t)$ and $\Phi_B^j(t)$, are physically correlated since they refer to the same satellite. Usually, the physical correlation is not

taken into account. Therefore, main interest is directed to the mathematical correlations introduced by differencing.

The assumption may be made that the phase errors show a random behavior resulting in a normal distribution with expectation value zero and variance σ^2. Measured (or raw) phases are, therefore, linearly independent or uncorrelated. Introducing a vector $\underline{\Phi}$ containing the phases, then

$$\mathrm{cov}(\underline{\Phi}) = \sigma^2 \underline{I} \tag{8.47}$$

is the covariance matrix for the phases where $\underline{I}$ is the unit matrix.

Single-differences

Considering the two points A, B and the satellite j at epoch t gives

$$\Phi^j_{AB}(t) = \Phi^j_B(t) - \Phi^j_A(t) \tag{8.48}$$

as the corresponding single-difference. Forming a second single-difference for the same two points but with another satellite k at the same epoch yields

$$\Phi^k_{AB}(t) = \Phi^k_B(t) - \Phi^k_A(t)\,. \tag{8.49}$$

The two single-differences may be computed from the matrix–vector relation

$$\underline{SD} = \underline{C}\,\underline{\Phi} \tag{8.50}$$

where

$$\underline{SD} = \begin{bmatrix} \Phi^j_{AB}(t) \\ \Phi^k_{AB}(t) \end{bmatrix}$$

$$\underline{C} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix} \qquad \underline{\Phi} = \begin{bmatrix} \Phi^j_A(t) \\ \Phi^j_B(t) \\ \Phi^k_A(t) \\ \Phi^k_B(t) \end{bmatrix}. \tag{8.51}$$

The covariance law applied to Eq. (8.50) gives

$$\mathrm{cov}(\underline{SD}) = \underline{C}\,\mathrm{cov}(\underline{\Phi})\,\underline{C}^T \tag{8.52}$$

and, by substituting Eq. (8.47),

$$\mathrm{cov}(\underline{SD}) = \underline{C}\,\sigma^2\,\underline{I}\,\underline{C}^T = \sigma^2\,\underline{C}\,\underline{C}^T \tag{8.53}$$

is obtained. Taking $\underline{C}$ from (8.51), the matrix product

$$\underline{C}\,\underline{C}^T = 2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 2\,\underline{I} \tag{8.54}$$

substituted into (8.53) leads to the covariance of the single-differences

$$\mathrm{cov}(\underline{SD}) = 2\sigma^2\, \underline{I}\,. \tag{8.55}$$

This shows that single-differences are uncorrelated. Note that the dimension of the unit matrix in (8.55) corresponds to the number of single-differences at epoch t, whereas the factor 2 does not depend on the number of single-differences. Considering more than one epoch, the covariance matrix is again a unit matrix with the dimension equivalent to the total number of single-differences.

Double-differences
Now, three satellites j, k, ℓ with j as reference satellite are considered. For the two points A, B and epoch t, the double-differences

$$\begin{aligned} \Phi^{jk}_{AB}(t) &= \Phi^{k}_{AB}(t) - \Phi^{j}_{AB}(t) \\ \Phi^{j\ell}_{AB}(t) &= \Phi^{\ell}_{AB}(t) - \Phi^{j}_{AB}(t) \end{aligned} \tag{8.56}$$

can be derived from the single-differences. These two equations can be written in the matrix–vector form

$$\underline{DD} = \underline{C}\,\underline{SD} \tag{8.57}$$

where

$$\underline{DD} = \begin{bmatrix} \Phi^{jk}_{AB}(t) \\ \Phi^{j\ell}_{AB}(t) \end{bmatrix}$$

$$\underline{C} = \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \qquad \underline{SD} = \begin{bmatrix} \Phi^{j}_{AB}(t) \\ \Phi^{k}_{AB}(t) \\ \Phi^{\ell}_{AB}(t) \end{bmatrix} \tag{8.58}$$

have been introduced. The covariance matrix for the double-differences is given by

$$\mathrm{cov}(\underline{DD}) = \underline{C}\,\mathrm{cov}(\underline{SD})\,\underline{C}^T \tag{8.59}$$

and substituting (8.55) leads to

$$\mathrm{cov}(\underline{DD}) = 2\sigma^2\, \underline{C}\,\underline{C}^T \tag{8.60}$$

or, explicitly, using $\underline{C}$ from (8.58)

$$\mathrm{cov}(\underline{DD}) = 2\sigma^2 \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}. \tag{8.61}$$

This shows that double-differences are correlated. The weight or correlation matrix $\underline{P}(t)$ is obtained from the inverse of the covariance matrix

$$\underline{P}(t) = [\mathrm{cov}(\underline{DD})]^{-1} = \frac{1}{2\sigma^2}\frac{1}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \tag{8.62}$$

where two double-differences at one epoch were used. Generally, with n_{DD} being the number of double-differences at epoch t, the correlation matrix is given by

$$\underline{P}(t) = \frac{1}{2\sigma^2}\frac{1}{n_{DD}+1}\begin{bmatrix} n_{DD} & -1 & -1 & \dots \\ -1 & n_{DD} & -1 & \dots \\ -1 & & & \\ \vdots & \dots & & n_{DD} \end{bmatrix} \tag{8.63}$$

where the dimension of the matrix is $n_{DD} \times n_{DD}$. For a better illustration, assume four double-differences. In this case, the 4×4 matrix

$$\underline{P}(t) = \frac{1}{2\sigma^2}\frac{1}{5}\begin{bmatrix} 4 & -1 & -1 & -1 \\ -1 & 4 & -1 & -1 \\ -1 & -1 & 4 & -1 \\ -1 & -1 & -1 & 4 \end{bmatrix} \tag{8.64}$$

is the correlation matrix. So far only one epoch has been considered. For epochs t_1, t_2, t_3, ..., the correlation matrix becomes a block-diagonal matrix

$$\underline{P}(t) = \begin{bmatrix} \underline{P}(t_1) & & & \\ & \underline{P}(t_2) & & \\ & & \underline{P}(t_3) & \\ & & & \ddots \end{bmatrix} \tag{8.65}$$

where each "element" of the matrix is itself a matrix. The matrices $\underline{P}(t_1)$, $\underline{P}(t_2)$, $\underline{P}(t_3)$, ... do not necessarily have to be of the same dimension because there may be different numbers of double-differences at different epochs.

Triple-differences

The triple-difference equations are slightly more complicated because several different cases must be considered. The covariance of a single triple-difference is computed by applying the covariance propagation law to the relation, cf. Eqs. (8.45) and (8.48),

$$\Phi^{jk}_{AB}(t_{12}) = \Phi^{k}_{AB}(t_2) - \Phi^{j}_{AB}(t_2) - \Phi^{k}_{AB}(t_1) + \Phi^{j}_{AB}(t_1)\,. \tag{8.66}$$

Now, two triple-differences with the same epochs and sharing one satellite are considered. The first triple-difference using the satellites j, k is given by Eq. (8.66). The second triple-difference corresponds to the satellites j, ℓ:

$$\begin{aligned} \Phi_{AB}^{jk}(t_{12}) &= \Phi_{AB}^{k}(t_2) - \Phi_{AB}^{j}(t_2) - \Phi_{AB}^{k}(t_1) + \Phi_{AB}^{j}(t_1) \\ \Phi_{AB}^{j\ell}(t_{12}) &= \Phi_{AB}^{\ell}(t_2) - \Phi_{AB}^{j}(t_2) - \Phi_{AB}^{\ell}(t_1) + \Phi_{AB}^{j}(t_1) \,. \end{aligned} \tag{8.67}$$

By introducing

$$\underline{TD} = \begin{bmatrix} \Phi_{AB}^{jk}(t_{12}) \\ \Phi_{AB}^{j\ell}(t_{12}) \end{bmatrix}$$

$$\underline{C} = \begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 \end{bmatrix} \qquad \underline{SD} = \begin{bmatrix} \Phi_{AB}^{j}(t_1) \\ \Phi_{AB}^{k}(t_1) \\ \Phi_{AB}^{\ell}(t_1) \\ \Phi_{AB}^{j}(t_2) \\ \Phi_{AB}^{k}(t_2) \\ \Phi_{AB}^{\ell}(t_2) \end{bmatrix} \tag{8.68}$$

the matrix–vector relation

$$\underline{TD} = \underline{C}\,\underline{SD} \tag{8.69}$$

can be formed, and the covariance for the triple-difference follows from

$$\mathrm{cov}(\underline{TD}) = \underline{C}\,\mathrm{cov}(\underline{SD})\,\underline{C}^T \tag{8.70}$$

or, by substituting (8.55),

$$\mathrm{cov}(\underline{TD}) = 2\sigma^2\,\underline{C}\,\underline{C}^T \tag{8.71}$$

is obtained which, using (8.68), yields

$$\mathrm{cov}(\underline{TD}) = 2\sigma^2 \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix} \tag{8.72}$$

for the two triple-differences (8.67). The tedious derivation may be abbreviated by setting up Table 8.1 where the point names A, B have been omitted.

It can be seen that the triple-difference $TD^{jk}(t_{12})$, for example, is composed of the two single-differences (with the signs according to the table) for the satellites j and k at epoch t_1 and of the two single-differences for the same satellites but epoch t_2. Accordingly, the same applies for the other triple-difference $TD^{j\ell}(t_{12})$. Thus, the coefficients of Table 8.1 are the same as those of matrix $\underline{C}$ in Eq. (8.68). Finally, the product $\underline{C}\,\underline{C}^T$, appearing in Eq. (8.71), is also aided by referring to Table 8.1. All combinations of inner

Table 8.1. Symbolic composition of triple-differences

Epoch	t_1			t_2		
Satellite	j	k	ℓ	j	k	ℓ
$TD^{jk}(t_{12})$	1	−1	0	−1	1	0
$TD^{j\ell}(t_{12})$	1	0	−1	−1	0	1

products of the two rows (one row represents one triple-difference) must be taken. The inner product (row 1 · row 1) yields the first-row, first-column element of $\underline{C}\,\underline{C}^T$, the inner product (row 1 · row 2) yields the first-row, second-column element of $\underline{C}\,\underline{C}^T$, etc. Based on the general formula (8.66) and Table 8.1, arbitrary cases may be derived systematically. Table 8.2 shows the second group of triple-difference correlations if adjacent epochs t_1, t_2, t_3 are taken. Two cases are considered. It can be seen from Table 8.2 that an exchange of the satellites for one triple-difference causes a change of the sign in the off-diagonal elements of the matrix $\underline{C}\,\underline{C}^T$. Therefore, the correlation of $TD^{kj}(t_{12})$ and $TD^{j\ell}(t_{23})$ produces +1 as off-diagonal element. Based on a table such as Table 8.2, each case may be handled with ease. According to Remondi (1984), p. 142, computer program adaptations require only a few simple rules. These are the basic mathematical correlations for single-, double-, and triple-differences.

More sophisticated models are investigated in Euler and Goad (1991), Gerdan (1995), Jin and Jong (1996) by taking into account the elevation dependence of the observation variances. Gianniou (1996) introduces variable weights by forming differences, applying polynomial fitting, and by using the signal-to-noise ratio for code ranges as well as for phases. Jonkman (1998)

Table 8.2. Triple-difference correlations

Epoch	t_1			t_2			t_3			$\underline{C}\,\underline{C}^T$	
Satellite	j	k	ℓ	j	k	ℓ	j	k	ℓ		
$TD^{jk}(t_{12})$	1	−1	0	−1	1	0	0	0	0	4	−2
$TD^{jk}(t_{23})$	0	0	0	1	−1	0	−1	1	0	−2	4
$TD^{jk}(t_{12})$	1	−1	0	−1	1	0	0	0	0	4	−1
$TD^{j\ell}(t_{23})$	0	0	0	1	0	−1	−1	0	1	−1	4

and Tiberius (1998) consider time correlation and cross correlation of the code ranges and the phases.

8.3.3 Static relative positioning

In a static survey of a single baseline vector between points A and B, the two receivers must stay stationary during the entire observation session. In the following, the single-, double-, and triple-differencing are investigated with respect to the number of observation equations and unknowns. It is assumed that the two sites A and B are able to observe the same satellites at the same epochs. The practical problem of satellite blockage is not considered here. The number of epochs is again denoted by n_t, and n_j denotes the number of satellites.

The undifferenced phase as shown in Eq. (8.10) (where the satellite clock is assumed to be known) is not included here, because there would be no connection (no common unknown) between point A and point B. The two data sets could be solved separately, which would be equivalent to point positioning.

A single-difference may be expressed for each satellite and for each epoch. The number of measurements is, therefore, $n_j\, n_t$. The number of unknowns is written below the corresponding terms of the single-difference equation, cf. Eq. (8.34):

$$\begin{aligned} \Phi^j_{AB}(t) &= \frac{1}{\lambda}\,\varrho^j_{AB}(t) + N^j_{AB} + f^j\,\delta_{AB}(t) \\ n_j\, n_t \;&\geq\; 3 \;\; + \;\; n_j \;\; + (n_t - 1)\,. \end{aligned} \tag{8.73}$$

The $n_t - 1$ unknown clock biases indicate a rank deficiency of 1. The explanation is the same as for static point positioning, cf. Eq. (8.11). From above, the relation

$$n_t \geq \frac{n_j + 2}{n_j - 1} \tag{8.74}$$

may be derived. Although this equation is equivalent to Eq. (8.12), it is useful to repeat the (theoretically) minimum requirements for a solution. A single satellite does not provide a solution because the denominator of (8.74) becomes zero. With two satellites, there results $n_t \geq 4$, and for the normal case of four satellites, $n_t \geq 2$ is obtained.

For double-differences, the relationship of measurements and unknowns is achieved using the same logic. Note that for one double-difference two satellites are necessary. For n_j satellites, therefore, $n_j - 1$ double-differences are obtained at each epoch so that the total number of double-differences

is $(n_j - 1)\, n_t$. The number of unknowns is written below the corresponding terms of the double-difference equation, cf. Eq. (8.37):

$$\begin{aligned} \Phi_{AB}^{jk}(t) \quad &= \frac{1}{\lambda}\varrho_{AB}^{jk}(t) + \quad N_{AB}^{jk} \\ (n_j - 1)\, n_t \geq \quad & 3 \qquad + (n_j - 1)\,. \end{aligned} \tag{8.75}$$

From above, the relation

$$n_t \geq \frac{n_j + 2}{n_j - 1} \tag{8.76}$$

is obtained which is identical with Eq. (8.74) and, therefore, the basic configurations are again given by the pairs $n_j = 2, n_t \geq 4$ and $n_j = 4, n_t \geq 2$. To avoid linearly dependent equations when forming double-differences, a reference satellite is used, against which the measurements of the other satellites are differenced. For example, take the case where measurements are made to the satellites 6, 9, 11, 12, and 6 is used as reference satellite. Then, at each epoch the following double-differences can be formed: (9-6), (11-6), and (12-6). Other double-differences are linear combinations and, thus, linearly dependent. For instance, the double-difference (11-9) can be formed by subtracting (11-6) and (9-6).

The triple-difference model includes only the three unknown point coordinates. For a single triple-difference, two epochs are necessary. Consequently, in the case of n_t epochs, $n_t - 1$ linearly independent epoch combinations are possible. Thus,

$$\begin{aligned} \Phi_{AB}^{jk}(t_{12}) \quad &= \frac{1}{\lambda}\varrho_{AB}^{jk}(t_{12}) \\ (n_j - 1)\,(n_t - 1) \geq \quad & 3 \end{aligned} \tag{8.77}$$

are the resulting equations. From above, the relation

$$n_t \geq \frac{n_j + 2}{n_j - 1} \tag{8.78}$$

is obtained. This equation is identical with Eq. (8.74) and, hence, the basic configurations are again given by the pairs $n_j = 2, n_t \geq 4$ and $n_j = 4, n_t \geq 2$.

This completes the discussion on static relative positioning. As shown, each of the mathematical models: single-difference, double-difference, triple-difference, may be used. The relationships between the number of observation equations and the number of unknowns will be referred to again in the discussion of the kinematic case.

8.3.4 Kinematic relative positioning

In kinematic relative positioning, the receiver on the known point A of the baseline vector remains fixed. The second receiver moves, and its position is to be determined for arbitrary epochs. The models for single-, double-, and triple-difference implicitly contain the motion in the geometric distance. Considering point B and satellite j, the geometric distance in the static case is given by, cf. Eq. (8.2),

$$\varrho_B^j(t) = \sqrt{(X^j(t) - X_B)^2 + (Y^j(t) - Y_B)^2 + (Z^j(t) - Z_B)^2} \quad (8.79)$$

and in the kinematic case by

$$\varrho_B^j(t) = \sqrt{(X^j(t) - X_B(t))^2 + (Y^j(t) - Y_B(t))^2 + (Z^j(t) - Z_B(t))^2} \quad (8.80)$$

where the time dependence for point B appears. In this mathematical model, three coordinates are unknown at each epoch. Thus, the total number of unknown site coordinates is $3\,n_t$ for n_t epochs. The relations between the number of observations and the number of unknowns for the kinematic case follow from the static single- and double-difference models, cf. Eqs. (8.73), (8.75):

$$\begin{aligned} &\text{single-difference:} && n_j\,n_t && \geq 3\,n_t + n_j + (n_t - 1) \\ &\text{double-difference:} && (n_j - 1)\,n_t && \geq 3\,n_t + (n_j - 1) \end{aligned} \quad (8.81)$$

For example, the relation

$$n_t = \frac{n_j - 1}{n_j - 4} \quad (8.82)$$

is the basic configuration for the single-differences which is equivalent to Eq. (8.14).

The continuous motion of the roving receiver restricts the available data for the determination of its position to one epoch. But none of the above two models provides a useful solution for $n_t = 1$. Thus, these models are modified: the number of unknowns is reduced by omitting the ambiguity unknowns, i.e., the ambiguities are assumed to be known. For the single-difference case, this has a twofold effect: first, the n_j ambiguities may be omitted and, second, the rank deficiency vanishes because of the known ambiguities so that n_t unknown clock biases have to be determined. The modified observation requirement for the single-difference is therefore $n_j\,n_t \geq 4\,n_t$ and reduces to $n_j \geq 4$ for a single epoch. Similarly, for the double-difference

$n_j - 1$ ambiguities are omitted in (8.81) so that $(n_j - 1)\, n_t \geq 3\, n_t$ results which reduces to $n_j \geq 4$ for a single epoch. Hence, the single-difference and the double-difference models end up again with the fundamental requirement of four simultaneously observable satellites.

The use of triple-differences for kinematic cases is strongly restricted. In principle, the definition of triple-differences with two satellites at two epochs and two stations at – with respect to the two epochs – fixed positions exclude any application since the rover position changes epoch by epoch. However, triple-differences could be used if the coordinates of the roving receiver were known at the reference epoch. In this case, adapting (8.77) to the kinematic case with $3\, n_t$ unknowns and reducing the number of unknown rover positions by 3 because of the known rover position at the reference epoch, the relationship obtained would be $(n_j - 1)\,(n_t - 1) \geq 3\,(n_t - 1)$. This leads to $n_j \geq 4$ which is the same requirement as for the ambiguity-reduced single- and double-differences.

Omitting the ambiguities for single- and double-difference means that they must be known. The corresponding equations are simply obtained by rewriting (8.73) and (8.75) with the ambiguities shifted to the left side of the equations. The single-differences become

$$\Phi^j_{AB}(t) - N^j_{AB} = \frac{1}{\lambda}\, \varrho^j_{AB}(t) + f^j\, \delta_{AB}(t) \tag{8.83}$$

and the double-differences

$$\Phi^{jk}_{AB}(t) - N^{jk}_{AB} = \frac{1}{\lambda}\, \varrho^{jk}_{AB}(t) \tag{8.84}$$

where the unknowns now appear only on the right sides.

Thus, all of the equations can be solved if one position of the moving receiver is known. Preferably (but not necessarily), this will be the starting point of the moving receiver. The baseline related to this starting point is denoted as the starting vector. With a known starting vector, the ambiguities are determined and are known for all subsequent positions of the roving receiver as long as no loss of signal lock occurs and a minimum of four satellites is in view.

Static initialization

Three methods are available for the static determination of the starting vector. In the first method, the moving receiver is initially placed at a known point, creating a known starting vector. The ambiguities can then be calculated from the double-difference model (8.75) as real values and are then fixed to integers. A second method is to perform a static determination of the starting vector. The third initialization technique is the antenna swap

method according to B. Remondi. The antenna swap is performed as follows: denoting the reference mark as A and the starting position of the moving receiver as B, a few measurements are taken in this configuration, and with continuous tracking, the receiver at A is moved to B, while the receiver at B is moved to A, where again a few measurements are taken. This is sufficient to precisely determine the starting vector in a very short time (e.g., 30 seconds). Usually, the antenna swap is completed by moving the receivers to their starting positions.

Kinematic initialization
Special applications require kinematic GPS without static initialization since the moving object whose position is to be calculated is in a permanent motion (e.g., an airplane while flying). Translated to model equations, this means that the most challenging case is the determination of the ambiguities on-the-fly (OTF). The solution requires an instantaneous ambiguity resolution or an instantaneous positioning (i.e., for a single epoch). This strategy sounds very simple and it actually is. The main problem is to find the position as fast and as accurately as possible. This is achieved by starting with approximations for the position and improving them by least squares adjustments or search techniques.

8.3.5 Pseudokinematic relative positioning

The pseudokinematic method can be identified with static surveying with large data gaps (Kleusberg 1990a). The mathematical model, e.g., for double-differences, corresponds to Eq. (8.75) where generally two sets of phase ambiguities must be resolved since the point is occupied at different times. Processing of the data could start with a triple-difference solution for the few minutes of data collected during the two occupations of a site. Based on this solution, the connection between the two ambiguity sets is computed (Remondi 1990b). After the successful ambiguity connection, the normal double-difference solutions are performed.

The time span between the two occupations is an important factor affecting accuracy. Willis and Boucher (1990) investigate the accuracy improvements by an increasing time span between the two occupations. As a rule of thumb, the minimum time span should be one hour.

9 Data processing

9.1 Data preprocessing

9.1.1 Data handling

Downloading
The observables as well as the navigation message and additional information are generally stored in a binary (and receiver dependent) format. The downloading of the data from the receiver is necessary before postprocessing can begin.

Most GPS manufacturers have designed a data management system which they recommend for processing data. Individual software is fully documented in the manuals of the manufacturers and will not be covered here. In the following sections, a general universal processing scheme is provided which describes the basic principles used.

Data management
During a multisession and multibaseline GPS survey, a large amount of data (in the Gigabyte range) may be collected. In order to archive and to process these data in a reasonable time, an appropriate data structure has to be used. Among different concepts, one based on projects is briefly described here.

Usually, the project management information is stored in one project file comprising the observation files of all the sites involved. The observation files can be stored at any arbitrary computer location. An appropriate pointer in the project file forms the necessary link to the directory where the data are stored after download. This means that there is no need to move data in the computer once they have been stored.

A flexible software allows limited editing of the observation files during download and import. Typical examples of editing refer to receiver or antenna type names.

Data exchange
Although the binary receiver data may have been converted into computer independent ASCII format during downloading, the data are still receiver dependent. Also, each GPS processing software has its own format which necessitates the conversion of specific data into a software independent format when they are processed with a different type of program.

B. Hofmann-Wellenhof et al., *Global Positioning System*

From the preceding, one may conclude that a receiver independent format of GPS data promotes data exchange. This has been realized by the Receiver Independent Exchange (RINEX) format. This format was first defined in 1989 and has been published in a second version by Gurtner and Mader (1990). Later, several minor changes were adopted and in 1997 the format was extended to also account for GLONASS.

Currently, the format consists of four ASCII file types: (1) the observation data file, (2) the navigation message file, (3) the meteorological data file, and (4) the GLONASS navigation message file. The records of the files have variable lengths with a maximum of 80 characters per line. Each file is composed of a header section and a data section. The header section contains generic file information, and the data section contains the actual data.

Basically, the observation and meteorological data files must be created for each site of the session. The RINEX (version 2) also allows to include observation data from more than one site subsequently occupied by a roving receiver in rapid static or kinematic applications.

The navigation message file is more or less site independent. In order to avoid the collection of identical satellite navigation messages from different receivers, one navigation message file only is created containing non-redundant information and being possibly composite from several receivers.

RINEX uses the filename convention "ssssdddf.yyt". The first four characters of the sequence (ssss) are the site identifier, the next three (ddd) indicate the day of year, and the eighth character (f) is the session indicator. The first two file extension characters (yy) denote the last two digits of the current year, and the file type (t) is given by the last character. The satellite designation is defined in the form "snn". The first character (s) is an identifier of the satellite system, and the remaining two digits denote the satellite number (e.g., the PRN number). Thus, the RINEX format enables the combination of observations of different satellite types.

At present, RINEX is the most favored format. As a consequence, receiver manufacturers produce software for the conversion of their receiver dependent format into RINEX. The U.S. National Geodetic Survey (NGS) has acted as a coordinator for these efforts. A detailed description of RINEX (version 2) is given in Gurtner (1998).

The Software Independent Exchange (SINEX) format is mentioned for the sake of completeness. This format enables the exchange of processing results and is used, for example, by the IGS (Mervart 1999).

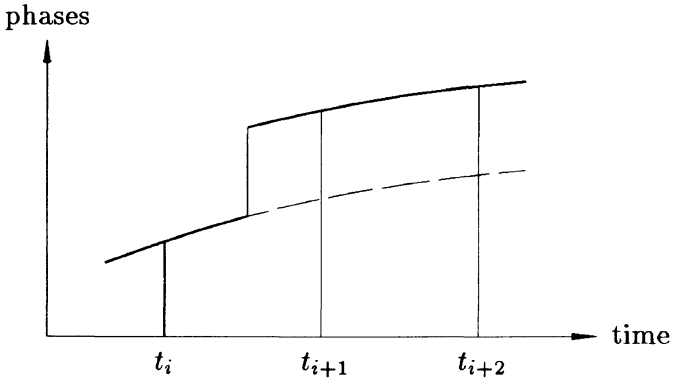


Fig. 9.1. Graphical representation of a cycle slip

9.1.2 Cycle slip detection and repair

Definition of cycle slips
When a receiver is turned on, the fractional part of the beat phase (i.e., the difference between the satellite transmitted carrier and a receiver generated replica signal) is observed and an integer counter is initialized. During tracking, the counter is incremented by one cycle whenever the fractional phase changes from 2π to 0. Thus, at a given epoch, the observed accumulated phase $\Delta\varphi$ is the sum of the fractional phase φ and the integer count n. The initial integer number N of cycles between the satellite and the receiver is unknown. This phase ambiguity N remains constant as long as no loss of the signal lock occurs. In this event, the integer counter is reinitialized which causes a jump in the instantaneous accumulated phase by an integer number of cycles. This jump is called cycle slip which, of course, is restricted to phase measurements.

A graphical representation of a cycle slip is given in Fig. 9.1. When the measured phases are plotted versus time, a fairly smooth curve should be obtained. In the case of a cycle slip, a sudden jump appears in the plotted curve (Hofmann-Wellenhof and Lichtenegger 1988: p. 98).

Three sources for cycle slips can be distinguished. First, cycle slips are caused by obstructions of the satellite signal due to trees, buildings, bridges, mountains, etc. This source is the most frequent one. The second source for cycle slips is a low SNR due to bad ionospheric conditions, multipath, high receiver dynamics, or low satellite elevation. A third source is a failure in the receiver software (Hein 1990b), which leads to incorrect signal processing. Cycle slips could also be caused by malfunctioning satellite oscillators, but these cases are rare.

As seen from Fig. 9.1, cycle slip detection and repair requires the location

of the jump (i.e., cycle slip) and the determination of its size. Detection is accomplished by a testing quantity. In the example given, this is the measured phase. Repairs of cycle slips are made by correcting all subsequent phase observations for this satellite and this carrier by a fixed amount. The determination of the cycle slip size and the correction of the phase data is often denoted as cycle slip fixing.

Testing quantities

The formulation of testing quantities is based on measured carrier phases and code ranges. For a single site, the testing quantities are raw phases, phase combinations, combinations of phases and code ranges, or combinations of phases and integrated Doppler frequencies. Table 9.1 summarizes a number of candidate testing quantities for a single site. Single receiver tests are important because they enable in situ cycle slip detection and repair by the internal software of the receiver. When two sites are involved, single-, double-, and triple-differences provide testing quantities.

The measured raw phase $\Phi_i^j(t)$ can be modeled by

$$\lambda \Phi_i^j(t) = \varrho_i^j(t) + \lambda N_i^j + c\,\Delta\delta_i^j(t) - \Delta^{\mathrm{Iono}}(t) + \ldots \tag{9.1}$$

where i and j denote the site and the satellite respectively. Note that the phase model contains a number of time dependent terms on the right side of (9.1) which may prevent cycle slip detection.

The model for the dual frequency phase combination is developed considering a single site and a single satellite. Thus, the sub- and superscripts and even the time dependency in Eq. (9.1) may be omitted. According to

Table 9.1. Testing quantities to detect cycle slips for a single site

Required data	Testing quantity
Single frequency phase (L1 or L2)	Raw phase
Dual frequency phases (L1 and L2)	Phase combinations
Single frequency phase and code range	Phase and code range combination
Single frequency phase and Doppler	Phase and integrated Doppler combination

Equation (6.73), the phases are modeled by

$$\begin{aligned}\Phi_{L1} &= a\, f_{L1} + N_{L1} - \frac{b}{f_{L1}} \\ \Phi_{L2} &= a\, f_{L2} + N_{L2} - \frac{b}{f_{L2}}\end{aligned} \tag{9.2}$$

where the frequency dependency is shown explicitly by the subscripts L1 and L2.

In order to eliminate the geometry term a, the first equation of (9.2) is multiplied by f_{L2} and the second by f_{L1}. Subtracting the resulting equations yields

$$f_{L2}\Phi_{L1} - f_{L1}\Phi_{L2} = f_{L2}N_{L1} - f_{L1}N_{L2} - b\left(\frac{f_{L2}}{f_{L1}} - \frac{f_{L1}}{f_{L2}}\right) \tag{9.3}$$

and dividing the difference by f_{L2} gives

$$\Phi_{L1} - \frac{f_{L1}}{f_{L2}}\Phi_{L2} = N_{L1} - \frac{f_{L1}}{f_{L2}}N_{L2} - \frac{b}{f_{L2}}\left(\frac{f_{L2}}{f_{L1}} - \frac{f_{L1}}{f_{L2}}\right) \tag{9.4}$$

or, by extracting f_{L2}/f_{L1} from the term in parentheses on the right side of the equation, the final form of the geometry-free phase combination

$$\Phi_{L1} - \frac{f_{L1}}{f_{L2}}\Phi_{L2} = N_{L1} - \frac{f_{L1}}{f_{L2}}N_{L2} - \frac{b}{f_{L1}}\left(1 - \frac{f_{L1}^2}{f_{L2}^2}\right) \tag{9.5}$$

is obtained. The left side of Eq. (9.5) is identical with the ionospheric residual, cf. Eq. (6.21). The right side shows that the only time-varying quantity is the ionosphere term b. In comparison to the influence on the raw phases in Eq. (9.1), the influence of the ionosphere on the dual frequency combination is reduced by a factor $(1 - f_{L1}^2/f_{L2}^2)$. Substituting the appropriate values for f_{L1} and f_{L2} yields a reduction of 65%.

If there are no cycle slips, the temporal variations of the ionospheric residual would be small for normal ionospheric conditions and for short baselines. Indicators of cycle slips are sudden jumps in successive values of the ionospheric residual. The remaining problem is to determine whether the cycle slip was on L1, L2, or both. This will be investigated in the next paragraph.

Another testing quantity follows from a phase and code range combination. Modeling the carrier phase and the code pseudoranges by

$$\begin{aligned}\lambda\,\Phi_i^j(t) &= \varrho_i^j(t) + \lambda\, N_i^j + c\,\Delta\delta_i^j(t) - \Delta^{\mathrm{Iono}}(t) + \Delta^{\mathrm{Trop}}(t) \\ R_i^j(t) &= \varrho_i^j(t) \qquad\qquad + c\,\Delta\delta_i^j(t) + \Delta^{\mathrm{Iono}}(t) + \Delta^{\mathrm{Trop}}(t)\end{aligned} \tag{9.6}$$

and forming the difference

$$\lambda \Phi_i^j(t) - R_i^j(t) = \lambda N_i^j - 2\Delta^{\text{Iono}}(t) \tag{9.7}$$

provides a formula where the time dependent terms (except the ionospheric refraction) vanish from the right side of the equation. Thus, the phase and code range combination could also be used as testing quantity. The ionospheric influence may either be modeled or neglected. The change of $\Delta^{\text{Iono}}(t)$ will be fairly small between closely spaced epochs; this might justify neglecting the ionospheric term. It may also be neglected when using double-differences.

The simple testing quantity (9.7) has a shortcoming which is related to the noise level. The phase and code range combinations have a noise level in the range of ± 5 cycles. This noise is mainly caused by the noise level of the code measurements and to a minor extent by the ionosphere. The noise of code measurements is larger than the noise for phase measurements because resolution and multipath are proportional to the wavelength. Traditionally, the measurement resolution was $\lambda/100$; today, receiver hardware is achieving improved measurement resolutions approaching $\lambda/1000$. In other words, this leads to code range noise levels of a few centimeters. Hence, the phase and code range combination could be an ideal testing quantity for cycle slip detection.

Comparing differences of measured phases with phase differences derived from integrated Doppler which has the advantage of being immune from cycle slips is a further possibilty.

When the data of two sites are combined, single-, double-, or triple-differences can be used for cycle slip detection. This means that, in a first step, unrepaired phase combinations are used to process an approximate baseline vector. The corresponding residuals are then tested. Quite often several iterations are necessary to improve the baseline solution. Note that triple-differences can achieve convergence and rather high accuracy without fixing cycle slips.

Detection and repair

Each of the described testing quantities allows the location of cycle slips by checking the difference of two consecutive epoch values. This also yields an approximate size of the cycle slip. To find the correct size, the time series of the testing quantity must be investigated in more detail. Note that for all previously mentioned testing quantities, except the ionospheric residual, the detected cycle slip must be an integer.

One of the methods for cycle slip detection is the scheme of differences. The very elegant principle can be seen from the following example. Assume

Table 9.2. Scheme of differences

t_i	$y(t_i)$	y^1	y^2	y^3	y^4
t_1	0				
		0			
t_2	0		0		
		0		ε	
t_3	0		ε		-3ε
		ε		-2ε	
t_4	ε		$-\varepsilon$		3ε
		0		ε	
t_5	ε		0		$-\varepsilon$
		0		0	
t_6	ε		0		
		0			
t_7	ε				

$y(t_i)$, $i = 1, 2, \ldots, 7$ as a time series for a signal which contains a jump of ε at epoch t_4 (Table 9.2). In this scheme, y^1, y^2, y^3, y^4 denote the first-order, second-order, third-order, and fourth-order differences. The important property in the context of data irregularities is the amplification of a jump in higher-order differences and, thus, the improved possibility of detecting the jump. The theoretical reason implied is the fact that differences are generated by subtractive filters. These are high-pass filters damping low frequencies and eliminating constant parts. High frequency constituents such as a jump are amplified. Replacing the signal $y(t_i)$ for example by the phase and assuming ε to be a cycle slip, the effect of the scheme of differences becomes evident. Any of the described testing quantities may be used as signal for the scheme of differences.

A method to determine the size of a cycle slip is to fit a curve through the testing quantities before and after the cycle slip. The size of the cycle slip is found from the shift between the two curves. The fits may be obtained from a simple linear regression or from more realistic least squares models (Beutler et al. 1984). These methods are generally called interpolation techniques. Other possibilities are prediction methods such as Kalman filtering. At a certain epoch, the function value (i.e., one of the testing quantities) for the next epoch is predicted based on the information obtained from preceding function values. The predicted value is then compared with the observed value to detect a cycle slip. The application of Kalman filtering for cycle

slip detection is demonstrated in detail by, e.g., Landau (1988). More details on Kalman filtering are provided in Sect. 9.3.2.

When a cycle slip has been detected (by one of the methods previously discussed), the testing quantities can be corrected by adding the size of the cycle slip to each of the subsequent quantities. The assignment of the detected cycle slip to a single phase observation is ambiguous if the testing quantities were phase combinations. An exception is the ionospheric residual. Under special circumstances, this testing quantity permits a unique separation. Consider Eq. (9.5) and assume ambiguity changes ΔN_{L1} and ΔN_{L2} caused by cycle slips. Consequently, a jump ΔN in the ionospheric residual would be detected. This jump is equivalent to

$$\Delta N = \Delta N_{L1} - \frac{f_{L1}}{f_{L2}} \Delta N_{L2} \tag{9.8}$$

where ΔN is no longer an integer. Equation (9.8) represents a diophantine equation for the two integer unknowns ΔN_{L1} and ΔN_{L2}. One equation and two unknowns; hence, there is no unique solution. This can be seen by solving for integer values ΔN_{L1} and ΔN_{L2} such that ΔN becomes zero. To get $\Delta N = 0$, the condition

$$\Delta N_{L1} = \frac{f_{L1}}{f_{L2}} \Delta N_{L2} = \frac{154}{120} \Delta N_{L2} \tag{9.9}$$

must be fulfilled which may be rewritten as

$$\Delta N_{L1} = \frac{77}{60} \Delta N_{L2} \,. \tag{9.10}$$

As an example, this means that $\Delta N_{L1} = 77$ and $\Delta N_{L2} = 60$ cannot be distinguished from $\Delta N_{L1} = 154$ and $\Delta N_{L2} = 120$ since both solutions satisfy Eq. (9.10). However, the solution would be unambiguous if ΔN_{L1} is less than 77 cycles. The consideration so far assumed error free measurements. To be more realistic, the effect of measurement noise must be taken into account. A simple model for the phase measurement noise is

$$m_\Phi = \pm 0.01 \text{ cycles} \tag{9.11}$$

which corresponds to a resolution of $\lambda/100$. The same model is applied to both carriers L1 and L2 and, thus, frequency dependent noise such as multipath is neglected. The assumption is not correct for codeless or quasi-codeless receivers since additional noise is introduced during signal processing (Table 5.6).

The value ΔN, in principle, is derived from two consecutive ionospheric residuals. Hence,

$$\Delta N = \Phi_{L1}(t+\Delta t) - \frac{f_{L1}}{f_{L2}}\Phi_{L2}(t+\Delta t) - \left[\Phi_{L1}(t) - \frac{f_{L1}}{f_{L2}}\Phi_{L2}(t)\right] \tag{9.12}$$

and applying to this equation the error propagation law gives

$$m_{\Delta N} = \pm 2.3\, m_{\Phi} = \pm 0.023 \text{ cycles}. \tag{9.13}$$

The 3σ error yields approximately ± 0.07 cycles. This may be interpreted as the resolution of ΔN. The conclusion is that two ΔN calculated by (9.8) and using arbitrary integers ΔN_{L1} and ΔN_{L2} must differ by at least 0.07 cycles in order to be uniquely separable. A systematic investigation of the lowest values for ΔN_{L1}, ΔN_{L2} is given in Table 9.3. For ΔN_{L1} and ΔN_{L2}, the values $0, \pm 1, \pm 2, \ldots, \pm 5$ have been permutated and ΔN calculated by (9.8). Table 9.3 is sorted with increasing ΔN in the first column. In the second column, the first-order differences of the function values ΔN are given. To shorten the length of the table, only the negative function values ΔN and zero are displayed. For supplementing with positive function values, the signs in the first, third, and fourth column must be reversed.

Those lines in Table 9.3 being marked with an asterisk do not fulfill the criterion of an at least 0.07 cycle difference. For these values, an unambiguous separation is not possible because the measurement noise is larger than the separation value. Consider the next to the last line in Table 9.3. A jump in the ionospheric residual of about 0.14 cycle could result from the pair of cycle slips $\Delta N_{L1} = -4$, $\Delta N_{L2} = -3$ or $\Delta N_{L1} = 5$, $\Delta N_{L2} = 4$; however, notice that for the marked lines either ΔN_{L1} or ΔN_{L2} equals 5 (plus or minus). Therefore, omitting the values for $\Delta N_{L1} = \pm 5$ and $\Delta N_{L2} = \pm 5$ creates uniqueness in the sense of separability. Up to ± 4 cycles the function values ΔN are discernible by 0.12 cycles.

The conclusions for cycle slip repair using the ionospheric residual are as follows. Based on the measurement noise assumption in (9.11), the separation of the cycle slips is unambiguously possible for up to ± 4 cycles. A smaller measurement noise increases the separability. For larger cycle slips, another method should be used in order to avoid wrong choices in ambiguous situations.

Most often, there will be more than one cycle slip. In these cases, each cycle slip must be detected and corrected in turn. The corrected phases, single-, double- or triple-differences are then used to process the baseline.

Table 9.3. Resulting ΔN by permutating ambiguity changes of L1 and L2

ΔN	Diff.	ΔN_{L1}	ΔN_{L2}
-11.42		-5	5
	1.00		
-10.42		-4	5
	0.29		
-10.13		-5	4
	0.71		
-9.42		-3	5
	0.29		
-9.13		-4	4
	0.28		
-8.85		-5	3
	0.43		
-8.42		-2	5
	0.29		
-8.13		-3	4
	0.28		
-7.85		-4	3
	0.29		
-7.56		-5	2
	0.14		
-7.42		-1	5
	0.29		
-7.13		-2	4
	0.28		
-6.85		-3	3
	0.29		
-6.56		-4	2
	0.14		
-6.42		0	5
	0.14		
-6.28		-5	1
	0.15		
-6.13		-1	4
	0.28		
-5.85		-2	-3
	0.29		
-5.56		-3	2
	0.14		
-5.42		1	5
	0.14		
-5.28		-4	1
	0.15		
-5.13		0	4
	0.13		
-5.00		-5	0
	0.15		
-4.85		-1	3
	0.29		
-4.56		-2	2
	0.14		
-4.42		2	5
	0.14		
-4.28		-3	1
	0.15		
-4.13		1	4
	0.13		
-4.00		-4	0
	0.15		
-3.85		0	3
	0.13		
-3.72		-5	-1

ΔN	Diff.	ΔN_{L1}	ΔN_{L2}	
-3.72		-5	-1	
	0.16			
-3.56		-1	2	
	0.14			
-3.42		3	5	
	0.14			
-3.28		-2	1	
	0.15			
-3.13		2	4	
	0.13			
-3.00		-3	0	
	0.15			
-2.85		1	3	
	0.13			
-2.72		-4	-1	
	0.16			
-2.56		0	2	
	0.12			
-2.44		-5	-2	
	0.02			
-2.42		4	5	*
	0.14			
-2.28		-1	1	
	0.15			
-2.13		3	4	
	0.13			
-2.00		-2	0	
	0.15			
-1.85		2	3	
	0.13			
-1.52		-3	-1	
	0.16			
-1.56		1	2	
	0.12			
-1.44		-4	-2	
	0.02			
-1.42		5	5	*
	0.14			
-1.28		0	1	
	0.13			
-1.15		-5	-3	*
	0.02			
-1.13		4	4	
	0.13			
-1.00		-1	0	
	0.15			
-0.85		3	3	
	0.13			
-0.72		-2	-1	
	0.16			
-0.56		2	2	
	0.12			
-0.44		-3	-2	
	0.16			
-0.28		1	1	
	0.13			
-0.15		-4	-3	
	0.02			
-0.13		5	4	*
	0.13			
0.00		0	0	

In recent years, fixing cycle slips by combining GPS data with data of other sensors, mainly Inertial Navigation Systems (INS), succeeded to some extent. As Colombo et al. (1999) demonstrate, even a moderately accurate (and low-cost) INS, which is also small, lightweight, and portable, can substantially enhance the ability to detect and correct cycle slips. If the INS data must bridge GPS data gaps in a stand-alone mode, this bridging time is the critical factor for keeping the desired high accuracy. It depends on several factors as, e.g., the type of application, the baseline length, the accuracy of the INS. Accordingly, the bridging time may be limited from a few seconds only to a few minutes. Details on the GPS/INS data modeling and on tests are given in Schwarz et al. (1994), Colombo et al. (1999), Altmayer (2000), El-Sheimy (2000).

9.2 Ambiguity resolution

9.2.1 General aspects

The ambiguity inherent with phase measurements depends upon both the receiver and the satellite. There is no time dependency as long as tracking is maintained without interruption. In the model for the phase,

$$\Phi = \frac{1}{\lambda}\varrho + f\,\Delta\delta + N - \frac{1}{\lambda}\Delta^{\text{Iono}}, \tag{9.14}$$

the ambiguity is denoted by N. As soon as the ambiguity is determined as an integer value, the ambiguity is said to be resolved or fixed. In general, ambiguity fixing strengthens the baseline solution. Joosten and Tiberius (2000) give an illustrative example. First, a short baseline is computed conventionally and the ambiguities are resolved. Then, introducing the float and the fixed ambiguities, respectively, as given quantities, single point positions on an epoch-by-epoch basis show a strong difference in the precision: with real-valued ambiguities, the solutions are scattering in the meter range in the coordinate components north, east, and up. For the integer-fixed solution, the precision of the coordinates is below the 1 cm level. But sometimes solutions with fixed ambiguities (i.e., integer values) and float ambiguities (i.e., real values) may agree within a few millimeters.

The use of double-differences instead of single-differences for carrier phase processing is important. The reason is that in the case of single-differences an additional unknown parameter for the receiver clock offset must be considered which prevents an effective separation of the integer ambiguities from the clock offset. In the case of double-differences, the clock terms have been eliminated and the isolation of the ambiguities is possible.

In order to fully exploit the high accuracy of the carrier phase observable, the ambiguities must be resolved to their correct integer value since one cycle on the L1 carrier may translate, in the maximum, to a 19 cm position error. It should be stressed here that integer ambiguity resolution may not always be possible. One of the reasons is the baseline length. When considering short baselines (less than 20 km), the model for double-difference phases may be simplified to

$$\lambda\,\Phi_{AB}^{jk}(t) = \varrho_{AB}^{jk}(t) + \lambda\,N_{AB}^{jk} + \text{noise} \tag{9.15}$$

since the effects of the ionosphere, the troposphere, and other minor effects may in general be neglected. As mentioned previously, any significant residual error from these neglected terms will spill over into the unknown parameters, namely station coordinates and ambiguities, and has the effect of degrading the positional accuracy as well as the integer nature of the

ambiguities. Thus, if applications require a long range from the reference station, there may be a need to install several reference stations for integer ambiguity resolution.

Another important aspect of ambiguity resolution is the satellite geometry which can be viewed from two points. First, the number of satellites tracked at any instant in general translates into a better dilution of precision value. Thus, all-in-view receivers with the ability to track seven or eight satellites are preferable since redundant satellites aid in the efficiency and reliability of ambiguity resolution. The second point with respect to geometry is the length of time required to resolve ambiguities. The information content of the carrier phase is a function of time which is directly correlated to the movement of the satellite. This last point can be illustrated through an example. Suppose two datasets. The first one is comprised of observations collected every 15 seconds for one hour, for a total of 240 measurements per satellite. Measurements for the second dataset are collected every second for four minutes, for a total of 240 measurements per satellite. Although the number of measurements is the same, the information content clearly is not. The first dataset has a higher probability of correct ambiguity resolution since the elapsed time is longer. The time is a critical component of ambiguity resolution even under good geometric conditions.

Multipath is also a critical factor for ambiguity resolution. Since multipath is station dependent, it may be significant for even short baselines. As in the case of atmospheric and orbital errors for long baselines, multipath has the effect of both contaminating the station coordinates and ambiguities.

Ambiguity resolution involves three major steps. The first step is the generation of potential integer ambiguity combinations that should be considered by the algorithm. A combination is comprised of an integer ambiguity for, e.g., each of the double-difference satellite pairs. In order to determine these combinations, a search space must be constructed. The search space is the volume of uncertainty which surrounds the approximate coordinates of the unknown antenna location. Since the search space dictates which integer ambiguities will be considered, it should be conservatively selected since it must contain the true antenna location. In the case of static positioning, this search space can be realized from the so-called float ambiguity solution, while for kinematic positioning it is realized from a code range solution. An important aspect of this first step in ambiguity resolution is that the size of the search space will affect the efficiency, i.e., computational speed, of the process. A larger search space gives a higher number of potential integer ambiguity combinations to assess, which in turn increases the computational burden. This is typically important for kinematic applications where a real-time implementation may be sought. It is, therefore, necessary to balance

computational load with a conservative search space size.

The second major step in the ambiguity resolution process is the identification of the correct integer ambiguity combination. The criterion used by many ambiguity resolution techniques is the selection of the integer combination which minimizes the sum of squared residuals (SSR) in the sense of least squares adjustment. The reasoning here comes from the argument that the combination which best fits the data should be the correct result.

The third step in the ambiguity resolution process should be a validation (or verification) of the ambiguities. The assessment of the correctness of the integer numbers obtained should gain more attention. The ambiguity success rate as defined in Joosten and Tiberius (2000) may be used as a tool for determining the probability of correct integer estimation. The ambiguity success rate depends on three factors: the observation equations (i.e., the functional model), the precision of the observables (i.e., the stochastic model), and the method of integer ambiguity estimation.

Although the above criterion based on residual analysis is rather straightforward, a few remarks should be made with respect to some of the potential difficulties of this approach. The first issue is the basic assumption in least squares theory that the residuals should be normally distributed. In many cases, this assumption is not fulfilled due to systematic effects from multipath, the orbit, and the atmosphere. This is the reason why ambiguity resolution generally fails for long baselines; however, if strong multipath exists it may even fail for short baselines. A second related issue is the need for statistical significance when the integer ambiguity decision is made. This means that the integer ambiguity combination which best fits the measurements should do so significantly better than all the other combinations. Statistical criteria can be used in this decision as will be discussed in some of the following subsections. Remaining systematic effects mentioned above play a role here as well as the aspect of time, i.e., ambiguity resolution is more difficult for shorter time intervals.

This three-step approach, (1) generation of potential integer ambiguity combinations, (2) identification of the optimum integer ambiguity combination, and (3) validation of the ambiguities, may also be refined and expanded. Han and Rizos (1997) propose six general classes and include the ambiguity recovery techniques (to reestimate ambiguities when cycle slips occur) as well as integrated models using GPS measurements and data from other sensors.

Subsequently, a few key principles from the numerous kinds of ambiguity resolution techniques will be demonstrated. Many variations may be derived from these basic methods (e.g., Mervart 1995).

9.2.2 Basic approaches

Resolving ambiguities with single frequency phase data
When phase measurements for only one frequency (L1 or L2) are available, the most direct approach is as follows. The measurements are modeled by Eq. (9.14), and the linearized equations are processed. Depending on the model chosen, a number of unknowns (e.g., point coordinates, clock parameters, etc.) is estimated along with N in a common adjustment. In this geometric approach, the unmodeled errors affect all estimated parameters. Therefore, the integer nature of the ambiguities is lost and they are estimated as real values. To fix ambiguities to integer values, a sequential adjustment could be performed. After an initial adjustment, the ambiguity with a computed value closest to an integer and with minimum standard error is considered to be determined most reliably. This bias is then fixed, and the adjustment is repeated (with one less unknown) to fix another ambiguity and so on. When using double-differences over short baselines, this approach is usually successful. The critical factor is the ionospheric refraction which must be modeled and which may prevent a correct resolution of all ambiguities.

Resolving ambiguities with dual frequency phase data
The situation for the ambiguity resolution changes significantly when using dual frequency phase data. There are many advantages implied in dual frequency data because of the various possible linear combinations that can be formed like the wide lane and narrow lane techniques. Denoting the L1 and L2 phase data Φ_{L1} and Φ_{L2}, then, according to Eq. (6.16),

$$\Phi_{L1-L2} = \Phi_{L1} - \Phi_{L2} \tag{9.16}$$

is the wide lane signal. The frequency of this signal is $f_{L1-L2} = 347.82\,\text{MHz}$ and the corresponding wavelength $\lambda_{L1-L2} = 86.2\,\text{cm}$. This is a significant increase compared to the original wavelengths of 19.0 and 24.4 cm. The increased wide lane wavelength λ_{L1-L2} provides an increased ambiguity spacing. This is the key to easier resolution of the integer ambiguities. To show the principle, consider the phase models for the carriers L1 and L2 in the modified form, cf. Eq. (9.2):

$$\begin{aligned} \Phi_{L1} &= a\, f_{L1} + N_{L1} - \frac{b}{f_{L1}} \\ \Phi_{L2} &= a\, f_{L2} + N_{L2} - \frac{b}{f_{L2}} \end{aligned} \tag{9.17}$$

with the geometry term a and the ionosphere term b. The difference of the two equations gives

$$\Phi_{L1-L2} = a\, f_{L1-L2} + N_{L1-L2} - b\left(\frac{1}{f_{L1}} - \frac{1}{f_{L2}}\right) \tag{9.18}$$

with the wide lane quantities

$$\begin{aligned}\Phi_{L1-L2} &= \Phi_{L1} - \Phi_{L2}\\ f_{L1-L2} &= f_{L1} - f_{L2}\\ N_{L1-L2} &= N_{L1} - N_{L2}\,.\end{aligned}$$

The adjustment based on the wide lane model gives wide lane ambiguities N_{L1-L2} which are more easily resolved than the base carrier ambiguities.

To compute the ambiguities for the measured phases (e.g., for L1), divide the first equation of (9.17) by f_{L1} and (9.18) by f_{L1-L2}:

$$\begin{aligned}\frac{\Phi_{L1}}{f_{L1}} &= a + \frac{N_{L1}}{f_{L1}} - \frac{b}{f_{L1}^2}\\ \frac{\Phi_{L1-L2}}{f_{L1-L2}} &= a + \frac{N_{L1-L2}}{f_{L1-L2}} - \frac{b}{f_{L1-L2}}\left(\frac{1}{f_{L1}} - \frac{1}{f_{L2}}\right),\end{aligned} \tag{9.19}$$

and the difference of the two equations gives

$$\frac{\Phi_{L1}}{f_{L1}} - \frac{\Phi_{L1-L2}}{f_{L1-L2}} = \frac{N_{L1}}{f_{L1}} - \frac{N_{L1-L2}}{f_{L1-L2}} - \frac{b}{f_{L1}^2} + \frac{b}{f_{L1-L2}}\left(\frac{1}{f_{L1}} - \frac{1}{f_{L2}}\right). \tag{9.20}$$

The desired ambiguity N_{L1} follows explicitly after rearranging and multiplying the equation above by f_{L1}:

$$N_{L1} = \Phi_{L1} - \frac{f_{L1}}{f_{L1-L2}}(\Phi_{L1-L2} - N_{L1-L2}) + \frac{b}{f_{L1}} - \frac{b}{f_{L1-L2}}\left(1 - \frac{f_{L1}}{f_{L2}}\right). \tag{9.21}$$

The terms reflecting the ionospheric influence may be treated as follows:

$$\begin{aligned}\frac{b}{f_{L1}} - \frac{b}{f_{L1-L2}}\left(1 - \frac{f_{L1}}{f_{L2}}\right) &= b\,\frac{f_{L1-L2}\,f_{L2} - f_{L1}\,f_{L2} + f_{L1}^2}{f_{L1}\,f_{L1-L2}\,f_{L2}}\\ &= b\,\frac{f_{L1-L2}\,f_{L2} + f_{L1}\,(f_{L1} - f_{L2})}{f_{L1}\,f_{L1-L2}\,f_{L2}}\\ &= b\,\frac{f_{L2} + f_{L1}}{f_{L1}\,f_{L2}}\end{aligned} \tag{9.22}$$

where on the right side the term in parentheses was replaced by the wide lane frequency f_{L1-L2} which then canceled. Therefore, the phase ambiguity N_{L1} in (9.21) can be calculated from the wide lane ambiguity by

$$N_{L1} = \Phi_{L1} - \frac{f_{L1}}{f_{L1-L2}} (\Phi_{L1-L2} - N_{L1-L2}) + b \frac{f_{L1} + f_{L2}}{f_{L1} f_{L2}} \tag{9.23}$$

and, in an analogous way, for N_{L2} by exchanging the roles of L1 by L2 in the equation above accordingly. Equation (9.23) represents the so-called geometry-free linear phase combination since the geometric distance ϱ and the clock bias term $\Delta\delta$ do not appear explicitly. Note, however, that these terms are implicitly contained in N_{L1-L2}, cf. Eq. (9.18). The ionospheric term is most annoying. This term will cancel for short baselines with similar ionospheric refraction at both sites (using differenced phases). For long baselines or irregular ionospheric conditions, however, the ionospheric term may cause problems.

To eliminate the ionosphere dependent term b in the computation of the ambiguities for the measured phases (e.g., for L1), one could proceed as follows. Start again with the phase equations (9.17) and multiply the first equation by f_{L1} and the second by f_{L2}. Form the differences of the resulting equations and, thus,

$$f_{L2} \Phi_{L2} - f_{L1} \Phi_{L1} = a (f_{L2}^2 - f_{L1}^2) + f_{L2} N_{L2} - f_{L1} N_{L1} \tag{9.24}$$

is obtained. Eliminating N_{L2} via the relation $N_{L2} = N_{L1} - N_{L1-L2}$ leads to

$$f_{L2} \Phi_{L2} - f_{L1} \Phi_{L1} = a (f_{L2}^2 - f_{L1}^2) - f_{L2} N_{L1-L2} + N_{L1} (f_{L2} - f_{L1}) \tag{9.25}$$

or, introducing $f_{L1-L2} = f_{L1} - f_{L2}$ and dividing the equation by this relation,

$$N_{L1} = \frac{f_{L1}}{f_{L1-L2}} \Phi_{L1} - \frac{f_{L2}}{f_{L1-L2}} (\Phi_{L2} + N_{L1-L2}) - a (f_{L1} + f_{L2}) \tag{9.26}$$

results. By simple linear algebra it may be verified that Eq. (9.26) is another representation of the ionosphere-free phase combination, cf. Eq. (6.76)

A final remark concerning the ambiguities is appropriate. Combining the terms containing N_{L1} and N_{L2} into a single term in the geometry-free or ionosphere-free combination destroys the integer nature of the term. This integer nature can be preserved by separately calculating the ambiguities, first N_{L1-L2} and then N_{L1} by (9.23) or (9.26).

Linear combinations other than wide lane have been considered such as the ionosphere-free linear combination. The disadvantage of this combination is that the corresponding ambiguity is no longer an integer. This is

a kind of circulus vitiosus: either the ambiguities may be resolved where the ionosphere is a problem or the ionospheric influence is eliminated which destroys the integer nature of the ambiguities.

Resolving ambiguities by combining dual frequency carrier phase and code data

The most unreliable factor of the wide lane technique described in the previous paragraph is the influence of the ionosphere which increases with baseline length. This drawback can be eliminated by a combination of phase and code data. The models for dual frequency carrier phases and code ranges, both expressed in cycles of the corresponding carrier, can be written in the form

$$\begin{aligned} \Phi_{L1} &= a\, f_{L1} - \frac{b}{f_{L1}} + N_{L1} \\ \Phi_{L2} &= a\, f_{L2} - \frac{b}{f_{L2}} + N_{L2} \\ R_{L1} &= a\, f_{L1} + \frac{b}{f_{L1}} \\ R_{L2} &= a\, f_{L2} + \frac{b}{f_{L2}} \end{aligned} \tag{9.27}$$

with the geometry term a and the ionosphere term b as known from (6.74). Note that four equations are available with four unknowns for each epoch. The unknowns are a, b, and the ambiguities N_{L1}, N_{L2} and may be expressed explicitly as a function of the measured quantities by inverting the system represented by (9.27).

Multiplying the third equation of (9.27) by f_{L1} and the fourth by f_{L2} and differencing the resulting equations yields the geometry term

$$a = \frac{1}{f_{L2}^2 - f_{L1}^2}\,(R_{L2} f_{L2} - R_{L1} f_{L1})\,. \tag{9.28}$$

Multiplying now the third equation of (9.27) by f_{L2} and the fourth by f_{L1} and differencing the resulting equations yields the ionosphere term

$$b = \frac{f_{L1} f_{L2}}{f_{L2}^2 - f_{L1}^2}\,(R_{L1} f_{L2} - R_{L2} f_{L1})\,. \tag{9.29}$$

Substituting (9.28) and (9.29) into the first two equations of (9.27) leads to explicit expressions for the phase ambiguities

$$\begin{aligned} N_{L1} &= \Phi_{L1} + \frac{f_{L2}^2 + f_{L1}^2}{f_{L2}^2 - f_{L1}^2} R_{L1} - \frac{2 f_{L1} f_{L2}}{f_{L2}^2 - f_{L1}^2} R_{L2} \\ N_{L2} &= \Phi_{L2} + \frac{2 f_{L1} f_{L2}}{f_{L2}^2 - f_{L1}^2} R_{L1} - \frac{f_{L2}^2 + f_{L1}^2}{f_{L2}^2 - f_{L1}^2} R_{L2} \end{aligned} \tag{9.30}$$

and by forming the difference $N_{L1-L2} = N_{L1} - N_{L2}$ one finally obtains

$$N_{L1-L2} = \Phi_{L1-L2} - \frac{f_{L1} - f_{L2}}{f_{L1} + f_{L2}} (R_{L1} + R_{L2}) . \tag{9.31}$$

where the short-hand notation $\Phi_{L1-L2} = \Phi_{L1} - \Phi_{L2}$ has been introduced accordingly. This rather elegant equation allows for the determination of the wide lane ambiguity N_{L1-L2} for each epoch and each site. It is independent of the baseline length and of the ionospheric effects. Even if all modeled systematic effects cancel out in (9.31), the multipath effect remains and affects phase and code differently. Multipath is almost exclusively responsible for a variation of N_{L1-L2} by several cycles from epoch to epoch. These variations may be overcome by averaging over a longer period.

According to Euler and Goad (1991) and Euler and Landau (1992), the ambiguity resolution for the combination of dual frequency code data with a reasonably low noise level and phase data will be possible "under all circumstances" with a few epochs of data. The approach described is even appropriate for instantaneous ambiguity resolution in kinematic applications. Hatch (1990) mentions that a single epoch solution is usually possible for short baselines if seven or more satellites can be tracked. Note that several variations of the technique are known.

Resolving ambiguities by combining triple frequency carrier phase and code data

Upon completion of the GPS modernization, a third civil frequency will be available. The technique for ambiguity resolution based on three carriers is denoted as Three-Carrier Ambiguity Resolution (TCAR) (Vollath et al. 1999). Before pointing out the model equations, a few remarks are appropriate when comparing this TCAR with the previously described dual frequency carrier phase and code data ambiguity resolution.

Theoretically, the four unknowns a, b, N_{L1}, N_{L2} of (9.27) can be determined instantaneously by solving the four equations. This means that in principle the unknowns can be determined epoch by epoch. In reality, fixing the ambiguities N_{L1}, N_{L2} to their correct values will be very unlikely even for short baselines. Therefore, the deviation tour via the wide lane ambiguities is taken.

Similarly, it may be expected that an instantaneous TCAR solution is also possible. This expectation becomes true as it may be seen immediately

from the triple frequency carrier phase and code data model

$$\begin{aligned}
\Phi_{L1} &= a\, f_{L1} - \frac{b}{f_{L1}} + N_{L1} \\
\Phi_{L2} &= a\, f_{L2} - \frac{b}{f_{L2}} + N_{L2} \\
\Phi_{L5} &= a\, f_{L5} - \frac{b}{f_{L5}} + N_{L5} \\
R_{L1} &= a\, f_{L1} + \frac{b}{f_{L1}} \\
R_{L2} &= a\, f_{L2} + \frac{b}{f_{L2}} \\
R_{L5} &= a\, f_{L5} + \frac{b}{f_{L5}}
\end{aligned} \tag{9.32}$$

where, apart from the two carriers L1, L2, the third carrier is designated as L5. This system of six equations contains five unknowns: the geometry term a, the ionosphere term b, and the ambiguities N_{L1}, N_{L2}, N_{L5}. Therefore, the system has the redundancy 1 and could be solved by least squares adjustment. Note, however, referring to the estimated ambiguities, Sjöberg (1997, 1998) indicates "that these estimates are too poor to be useful". By contrast, it is possible to determine a wide lane ambiguity accurately. In Table 9.4, the three carriers and wide lane combinations are given.

From the dual frequency approach, the result

$$N_{L1-L2} = \Phi_{L1} - \Phi_{L2} - \frac{f_{L1} - f_{L2}}{f_{L1} + f_{L2}} (R_{L1} + R_{L2}) \tag{9.33}$$

of Eq. (9.31) is obtained for the L1–L2 combination. Replacing now L2 by L5

Table 9.4. Carriers and wide lane combinations

	Frequency (MHz)	Wavelength (m)
L1	1575.42	0.1903
L2	1227.60	0.2442
L5	1176.45	0.2548
L1–L5	398.97	0.7514
L1–L2	347.82	0.8619
L2–L5	51.15	5.8610

leads to the wide lane combination

$$N_{\mathrm{L1-L5}} = \Phi_{\mathrm{L1}} - \Phi_{\mathrm{L5}} - \frac{f_{\mathrm{L1}} - f_{\mathrm{L5}}}{f_{\mathrm{L1}} + f_{\mathrm{L5}}} (R_{\mathrm{L1}} + R_{\mathrm{L5}}) . \tag{9.34}$$

Following from the wide lane definitions $N_{\mathrm{L1-L2}} = N_{\mathrm{L1}} - N_{\mathrm{L2}}$ and $N_{\mathrm{L1-L5}} = N_{\mathrm{L1}} - N_{\mathrm{L5}}$, the individual ambiguities are

$$\begin{aligned} N_{\mathrm{L2}} &= N_{\mathrm{L1}} - N_{\mathrm{L1-L2}} \\ N_{\mathrm{L5}} &= N_{\mathrm{L1}} - N_{\mathrm{L1-L5}} \end{aligned} \tag{9.35}$$

where N_{L1} is still unknown and to be determined. These equations are resubstituted into (9.32), the initial set of model equations:

$$\begin{aligned} \Phi_{\mathrm{L1}} &= a\, f_{\mathrm{L1}} - \frac{b}{f_{\mathrm{L1}}} + N_{\mathrm{L1}} \\ \Phi_{\mathrm{L2}} + N_{\mathrm{L1-L2}} &= a\, f_{\mathrm{L2}} - \frac{b}{f_{\mathrm{L2}}} + N_{\mathrm{L1}} \\ \Phi_{\mathrm{L5}} + N_{\mathrm{L1-L5}} &= a\, f_{\mathrm{L5}} - \frac{b}{f_{\mathrm{L5}}} + N_{\mathrm{L1}} \\ R_{\mathrm{L1}} &= a\, f_{\mathrm{L1}} + \frac{b}{f_{\mathrm{L1}}} \\ R_{\mathrm{L2}} &= a\, f_{\mathrm{L2}} + \frac{b}{f_{\mathrm{L2}}} \\ R_{\mathrm{L5}} &= a\, f_{\mathrm{L5}} + \frac{b}{f_{\mathrm{L5}}} \end{aligned} \tag{9.36}$$

where the known wide lane ambiguities $N_{\mathrm{L1-L2}}$ and $N_{\mathrm{L1-L5}}$ have been shifted to the left side of the equations. This system of six equations comprises only three unknowns: a, b and N_{L1}, thus the redundancy amounts to 3. Inherently, this combined data set of code and phase measurements reflects two accuracy classes because the last three code range equations are much less accurate compared to the first three mainly phase derived equations. Sjöberg (1999) neglects the three code range equations by arguing that they contribute little to the least squares solution. With the remaining phase equations, the calculation of the three unknowns is still possible for a single epoch.

After the successful computation of the N_{L1} ambiguity, the same procedure may be applied accordingly to get the other two carrier ambiguities N_{L2} and N_{L5}.

Vollath et al. (1999) use the same set of equations as given in (9.36) but with an extended modeling of the ionospheric influence and apply a recursive least squares adjustment.

Several other procedures exist. Hatch et al. (2000), e.g., conclude that over short baselines the ambiguities may be resolved much more quickly (often in a single epoch), whereas for longer baselines there is limited gain from the third frequency. Vollath et al. (1999) conclude similarly, that the TCAR procedure will generally not suffice to resolve the ambiguities instantaneously, i.e., using data of a single epoch unless very short baselines are considered. Accumulating several epochs will, on the one hand, reduce the noise but, on the other hand, the main error components caused by the ionosphere and multipath remain because of their long correlation times. Therefore, a search for the optimal solution along with a validation is still required. However, the number of possible candidates for this optimal result is substantially reduced.

9.2.3 Search techniques

A standard approach

When processing the data based on double-differences by least squares adjustment, the ambiguities are estimated as real or floating point numbers, hence the first double-difference solution is called the float ambiguity solution. The output is the best estimate of the station coordinates as well as double-difference ambiguities. If the baseline is relatively short, say five kilometers, and the observation span relatively long, say one hour, these float ambiguities would typically be very close to integers. Ambiguity resolution in this case is merely used to refine the achievable positioning accuracy. The change in the station coordinates from the float solution to the fixed ambiguity solution should not be large and in the case when ambiguity resolution fails, the float solution is generally a very good alternative.

As the observation span becomes smaller, the float solution will weaken due to loss of information. Ambiguity resolution will then play a more important role, since its effect on the station coordinates will now be significant. If the observation span is further reduced, the success of ambiguity resolution may determine whether or not the user's positioning specifications are met. As this discussion implies, there is a risk associated with a reduction in the observation span.

The search space concept can be generated for the static case by considering the positional accuracy of the float ambiguity solution. A conceptually simpler approach, however, is to directly use the estimated accuracies of the float ambiguities to set their search range. For example, if an ambiguity is

estimated to be −87457341.88 cycles with a standard deviation of 0.30 cycles, all the integer ambiguities that fall within ± 3 standard deviations of that value (for a high statistical probability) should be searched. This would give potential integer ambiguities of −87457340.0 to −87457343.0 by being conservative. This procedure can be repeated for each of the double-difference ambiguities and the result is a set of potential integer ambiguity combinations.

The number of ambiguity sets to be considered is dependent on the number of satellites tracked and the search range of the double-difference ambiguities. For example, there are five ambiguities if six satellites are tracked and if the range for each ambiguity is three cycles, the number of combinations to test is $3^5 = 243$. If the search range is increased to five cycles, the total number of combinations is 3 125. Larger combinations can typically occur when the observation time is decreased since the geometry weakens, thus increasing the estimated standard deviation used to derive the range.

Once all the potential ambiguity combinations are identified, each one is tested by constraining (fixing) the ambiguities to the selected integer combination and then computing the measurement residuals. The total redundancy is increased in the fixed ambiguity adjustment since only the station coordinates are estimated. However, the residuals are larger than for the float ambiguity solution (see, e.g., Table 7.9). King et al. (1987) present a technique by which the influence of various integer ambiguity combinations can be computed from the float ambiguity solution, rather than initiating a new least squares adjustment for each of the potential ambiguity combinations.

The sum of squared residuals is used as the final measure of the fit of the ambiguity combination. The integer ambiguity solution corresponding to the smallest sum of squared residuals should be the candidate which is selected. Due to reasons stated earlier, however, no candidate may be significantly better than the other to warrant selection. A ratio test is often used to make this decision. For example, if the ratio of the second smallest sum of squared residuals to the smallest sum of squared residuals is 2 or 3 (depending on the algorithm), then a decision to select the smallest sum of squared residuals as the true solution can be made. Otherwise, no integer ambiguity solution can be determined and then the best estimate for the station coordinates is the float ambiguity solution.

An example given in Cannon and Lachapelle (1993) will illustrate this concept. On a 720 m baseline, six satellites were tracked for 10 minutes. Using double-differences with satellite 19 as reference satellite, the least squares approach yielded for the ambiguities the following values in cycles:

DD SV#	Float ambiguity
2 – 19	17 329 426.278
6 – 19	14 178 677.032
11 – 19	11 027 757.713
16 – 19	−1 575 518.876
18 – 19	−15 754 175.795

The abbreviation DD SV# indicates double-differences (DD) for the specified space vehicle numbers (SV#). To get integer values, the float solution is simply rounded to the nearest integer values. To check this solution, possible other ambiguity sets are established by varying each ambiguity in a certain range, say by ±2 cycles, so that, apart from the integer solution obtained from the table above, each ambiguity is varied by −2, −1 and +1, +2 cycles. This means that for each ambiguity five cases are checked. Considering the five double-differences, in total $5^5 = 3\,125$ possible integer sets arise which are to be compared with respect to the sum of the squared residuals. Subsequently, the results for the three smallest sums of squared residuals (SSR) are given:

	1st smallest SSR = 0.044	2nd smallest SSR = 0.386	3rd smallest SSR = 0.453
DD SV#	Ambiguity	Ambiguity	Ambiguity
2 – 19	17 329 426	17 329 426	17 329 426
6 – 19	14 178 677	14 178 676	14 178 678
11 – 19	11 027 758	11 027 757	11 027 759
16 – 19	−1 575 519	−1 575 518	−1 575 520
18 – 19	−15 754 176	−15 754 176	−15 754 176

The ambiguity set with the smallest sum of squared residuals is likely to represent the correct integers only if its SSR compared to the 2nd smallest SSR is significantly smaller. The ratio, which amounts to 0.386/0.044 = 8.8 in the example above, should be >3 (which has empirically been established). If this ratio test fails, then it is safer to use the coordinates from the double-difference float ambiguity solution as the best estimate for a particular station.

To demonstrate a failing of the ratio test, the same example is taken but the data set is reduced to 5 minutes instead of the original 10 minutes. The

results for the double-difference float solution are:

DD SV#	Float ambiguity
2 – 19	17 329 426.455
6 – 19	14 178 677.192
11 – 19	11 027 757.762
16 – 19	−1 575 518.471
18 – 19	−15 754 175.411

and, when checking again 3 125 possible integer ambiguity sets, i.e., five cases for each ambiguity, namely the solution rounded to the nearest integer value and ±2 cycles, the following ambiguity sets represent the best solutions in the sense of minimal sum of squared residuals:

	1st smallest SSR = 0.137	2nd smallest SSR = 0.155	3rd smallest SSR = 0.230
DD SV#	Ambiguity	Ambiguity	Ambiguity
2 – 19	17 329 425	17 329 426	17 329 426
6 – 19	14 178 675	14 178 677	14 178 675
11 – 19	11 027 757	11 027 758	11 027 756
16 – 19	−1 575 516	−1 575 519	−1 575 518
18 – 19	−15 754 175	−15 754 176	−15 754 175

The ratio test for the smallest and the second smallest yields 0.155/0.137 = 1.1 and, thus, fails. This means that from the statistical point of view with regard to the squared sum of residuals, the correct solution cannot be extracted safely. Note, however, that the solution of the second smallest sum of squared residuals gives the correct integer ambiguities (as compared to the solution of the full 10-minute data set), but from the chosen criterion of the ratio this is not recognized. This shows that the technique of comparing the sum of squared residuals is certainly not the most advanced technique.

Ambiguity resolution on-the-fly

The notation "on-the-fly" reflects any type of rover motion. The terms AROF (ambiguity resolution on the fly), OTF (on-the-fly), and sometimes OTR (on-the-run) are different abbreviations with the same meaning, namely the development of ambiguity resolution techniques for the kinematic case. Numerous techniques have been developed to deal with the kinematic case.

Code ranges are generally used to define the search space for the kinematic case. A relative code range position is used as the best estimate of antenna location, and the associated standard deviations are used to define the size of the search space. This space can be determined in several ways, for example, it can be a cube, a cylinder, or an ellipsoid.

In order to reduce the number of integer ambiguity combinations to be tested, the code solution should be as accurate as possible which means that receiver selection becomes important. The availablility of low noise, narrow correlator-type code ranges is advantageous since they have a resolution in the order of 10 cm as well as improved multipath reduction compared with standard C/A-code receivers.

An example is used to show the direct correlation between the code accuracy and the size of the potential ambiguities to be searched. Suppose a standard C/A-code receiver is used to define the search cube. The accuracy of the resulting position is approximately $\pm$ 2 m to give a cube size of 4 m on a side. If six satellites are tracked, there are five double-difference ambiguities to consider. The search range for each ambiguity is approximately 4 m/0.2 m = 20 cycles (for L1 processing) to give 3.2 million total combinations. If, in contrast, a narrow correlator-type receiver is used, the accuracy of the resulting position is approximately $\pm$1 m to give a cube of 2 m on a side and a search range of 2 m/0.2 m = 10 cycles. Under the same six satellite geometry, the total combinations are reduced to 100 000 which is a significant difference.

The importance of the carrier phase wide lane should be mentioned here in the context of the number of potential ambiguity combinations. If the above example is repeated using the wide lane which has a wavelength of 86 cm, the number of potential ambiguities for narrow correlator-type C/A-code case would be about 35. The advantage of using this observable instead of the L1 carrier phase is clear as it tremendously reduces the search time. The only disadvantage of using the wide lane is that the measurement is significantly noisier than L1. Many OTF implementations use the wide lane to resolve integer ambiguities and then use the resulting position to directly compute the ambiguities on the L1 carrier phase data, or at least to significantly limit the number of L1 ambiguities to be considered. The wide lane is also used extensively for fast static applications where the station occupation time is limited.

The OTF techniques have common features like, e.g., the determination of an initial solution; they differ only in how these features are carried out. A summary of the main features is given in Table 9.5 which is closely related to Erickson (1992b). A feature unique to the ambiguity function method and not mentioned in the table is the insensitivity with respect to cycle slips. As far as the search technique (domain, space, reduction of trials) is concerned, there are also combinations of several listed characteristics (e.g., Abidin et al. 1992). Illustrative graphic representations of search spaces lead to an easier understanding of the reduction of trials (Hatch 1991, Erickson 1992a, Frei and Schubernigg 1992, Abidin 1993).

Table 9.5. Characteristics and options for OTF ambiguity resolution techniques

Initial solution	• Code solution for position X, Y, Z and its accuracy $\sigma_X, \sigma_Y, \sigma_Z$ • Carrier solution for X, Y, Z and N_j and accuracies $\sigma_X, \sigma_Y, \sigma_Z, \sigma_{N_j}$
Search domain	• Test points (three-dimensional space) • Ambiguity sets (n-dimensional integer space where n is the number of ambiguities)
Search space	• $k\,\sigma_X, k\,\sigma_Y, k\,\sigma_Z$ • $k\,\sigma_{N_j}$
Determination of k	• Empirically • Statistically
Reduction of trials	• Grid search (fine, coarse) • Double-difference plane intersection • Statistically (correlation of ambiguities for example)
Selection criterion	• Maximum ambiguity function • Minimum variance σ_0^2
Acceptance criterion	• Ratio of largest and second largest ambiguity function • Ratio of smallest and second smallest variance σ_0^2
Observation period	• Instantaneous potential • Some minutes
Data required	• Single or dual frequency • Phase only or phase and code

The double-difference plane intersection to reduce the number of trials requires a brief explanation. Positions are derived from three double-differences with sets of possible ambiguities. Geometrically, each (linearized) double-difference with its trial ambiguity defines a plane in three-dimensional space (Hatch 1990). Thus, the intersection of three planes yields a possible solution position. The grid spacing is the wavelength of the carrier and is equivalent to the grid spacing in the ambiguity search domain.

Minimizing the variance σ_0^2 as selection criterion is in principle the same as minimizing the sum of the squared residuals. If the position of the receiver is eliminated by a mapping function, as proposed by Walsh (1992), the residuals reflect the ambiguities only.

The subsequent paragraphs explain some out of the many OTF techniques that can be used. Examples are: the ambiguity function method, the least squares ambiguity search, the fast ambiguity resolution approach, the fast ambiguity search filter, least squares ambiguity decorrelation adjustment method, and ambiguity determination with special constraints.

Numerous approaches may be found in publications, e.g., the fast ambiguity resolution using an integer nonlinear programming method (Wei and Schwarz 1995b); a maximum likelihood method based on undifferenced phases (Knight 1994); the fitting of individual epoch residuals for potential ambiguity candidates to low-order polynomials (Borge and Forssell 1994). Additional methods may be found in the review papers by Chen and Lachapelle (1994), Hatch (1994), Hein (1995).

Ambiguity function method

Counselman and Gourevitch (1981) proposed the principle of the ambiguity function, Remondi (1984, 1990a) and Mader (1990) further investigated this method. The concept should become clear from the following description. Assume the model (8.34) for the single-difference phase represented by

$$\Phi^j_{AB}(t) = \frac{1}{\lambda}\varrho^j_{AB}(t) + N^j_{AB} + f\,\delta_{AB}(t) \tag{9.37}$$

for the points A and B, and the satellite j. If point A is assumed known and B is a selected candidate from the gridded cube, then the term $\varrho^j_{AB}(t)$ is known and may be shifted to the left side of the equation:

$$\Phi^j_{AB}(t) - \frac{1}{\lambda}\varrho^j_{AB}(t) = N^j_{AB} + f\,\delta_{AB}(t)\,. \tag{9.38}$$

The key is to circumvent the ambiguities N^j_{AB}. A special effect occurs if the term $2\pi\,N^j_{AB}$ is used as the argument of a cosine or sine function because N^j_{AB} is an integer. Therefore, the whole expression (9.38) is multiplied by 2π and placed into the complex plane by raising both the left and right side to the power of $e^i = \exp\{i\}$ where $i = \sqrt{-1}$ is the imaginary unit. In detail,

$$\exp\left\{i[2\pi\,\Phi^j_{AB}(t) - \frac{2\pi}{\lambda}\varrho^j_{AB}(t)]\right\} = \exp\{i[2\pi\,N^j_{AB} + 2\pi\,f\,\delta_{AB}(t)]\} \tag{9.39}$$

where the right side may also be written as

$$\exp\{i\,2\pi\,N^j_{AB}\}\exp\{i\,2\pi\,f\,\delta_{AB}(t)\}\,. \tag{9.40}$$

It is illustrative to consider this situation in the complex plane (Fig. 9.2). Note the equivalence

$$\exp\{i\alpha\} = \cos\alpha + i\,\sin\alpha \tag{9.41}$$

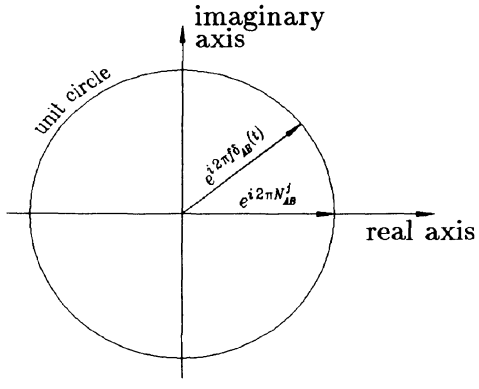


Fig. 9.2. Vector representation in the complex plane

which may be represented as a unit vector with the components $\cos\alpha$ and $\sin\alpha$ if a real axis and an imaginary axis are used. Therefore,

$$\exp\{i\,2\pi\,N_{AB}^{j}\} = \cos(2\pi\,N_{AB}^{j}) + i\,\sin(2\pi\,N_{AB}^{j}) = 1 + i\cdot 0 \qquad (9.42)$$

results because of the integer nature of N_{AB}^{j}. Hence, for one epoch and one satellite, (9.39) reduces to

$$\exp\left\{i[2\pi\,\Phi_{AB}^{j}(t) - \frac{2\pi}{\lambda}\,\varrho_{AB}^{j}(t)]\right\} = \exp\{i\,2\pi\,f\,\delta_{AB}(t)\} \qquad (9.43)$$

by applying (9.40) and (9.42). Considering n_j satellites and forming the sum over these satellites for the epoch t leads to

$$\sum_{j=1}^{n_j}\exp\left\{i[2\pi\,\Phi_{AB}^{j}(t) - \frac{2\pi}{\lambda}\,\varrho_{AB}^{j}(t)]\right\} = n_j\exp\{i\,2\pi\,f\,\delta_{AB}(t)\}\,. \qquad (9.44)$$

Considering more than one epoch, the fact that the clock error $\delta_{AB}(t)$ varies with time must be taken into account. Recall that $\exp\{i\,2\pi\,f\,\delta_{AB}(t)\}$ is a unit vector as indicated in Fig. 9.2. Thus, when $\|\exp\{i\,2\pi\,f\,\delta_{AB}(t)\}\| = 1$ is applied to (9.44), the relation

$$\left\|\sum_{j=1}^{n_j}\exp\left\{i[2\pi\,\Phi_{AB}^{j}(t) - \frac{2\pi}{\lambda}\,\varrho_{AB}^{j}(t)]\right\}\right\| = n_j\cdot 1 \qquad (9.45)$$

is obtained where the clock error has now vanished.

Take for example four satellites and an error free situation (i.e., neither measurement errors nor model errors, and correct coordinates for the points A and B). In this case, the evaluation of the left side of (9.45) should yield

4 where $\Phi^j_{AB}(t)$ are the single-differences of measured phases and $\varrho^j_{AB}(t)$ can be calculated from the known points and satellite positions. However, if point B was chosen incorrectly then the result must be less than 4. In reality, this maximum can probably never be achieved precisely because of measurement errors and incomplete modeling. Thus, the task is restricted to obtaining the maximum of (9.45) by varying B.

With highly stable receiver clocks and close epoch spacing it is theoretically possible to include more than one epoch within the absolute value. Using n_t epochs, the contribution of all epochs may be summed up by

$$\sum_{t=1}^{n_t} \left\| \sum_{j=1}^{n_j} \exp\left\{ i[2\pi\, \Phi^j_{AB}(t) - \frac{2\pi}{\lambda}\, \varrho^j_{AB}(t)] \right\} \right\| = n_t\, n_j \tag{9.46}$$

where for simplicity the same number of satellites at all epochs is assumed. Following Remondi (1984, 1990a), the left side of (9.46), i.e., the double sum, is denoted as an ambiguity function. Analogous to the case with one epoch, the maximum of the ambiguity function must be found. In general it will, as before, be less than the theoretical value $n_t\, n_j$.

The ambiguity function procedure is simple. Assume an approximate solution for point B, e.g., by triple-differences. Then, place this solution into the center of a cube (Fig. 9.3) and partition the cube into grid points. Each grid point is a candidate for the final solution, and the ambiguity function (9.46) is calculated for all single-differences. The grid point yielding the maximum ambiguity function value, which should theoretically be equal to the total number of single-differences (i.e., $n_t\, n_j$), is the desired solution. Having found this solution, the ambiguities could be computed using double-differences. Also, an adjustment using double-differences might be performed to verify the position of B and the ambiguities. The computation of point B with fixed ambiguities is the final step.

It is worth noting that the ambiguity function method is completely insensitive to cycle slips. The reason can easily be seen from Eq. (9.42). Even if the ambiguity changes by an arbitrary integer amount ΔN^j_{AB}, then $\exp\{i\, 2\pi\, (N^j_{AB} + \Delta N^j_{AB})\}$ is still a unit vector and the subsequent equations,

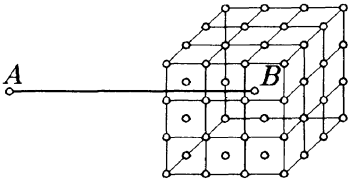


Fig. 9.3. Search technique

therefore, remain unchanged. Other methods require that cycle slips be repaired before computing the ambiguity.

Remondi (1984) shows detailed examples of how to speed up the procedure, how to choose the density of the grid points within the cube, and how to find the correct maximum if there are many relative maxima for the ambiguity function. These considerations are significant, since the computational burden could, otherwise, become overwhelming. For illustrative purposes, assume a 6 m × 6 m × 6 m cube with a one centimeter grid. Then $(601)^3 \approx 2.17 \cdot 10^8$ possible solutions must be checked with the ambiguity function (9.46).

Least squares ambiguity search technique

The method described here is investigated in further details in Hatch (1990, 1991). The least squares ambiguity technique requires an approximate solution for the position (due to the linearization of the observation equation) which may be obtained from a code range solution. The search area may be established by surrounding the approximate position by a 3σ region. One of the basic principles of the approach is the separation of the satellites into a primary and a secondary group. The primary group consists of four satellites. Based on these four satellites, which should have a good PDOP, the possible ambiguity sets are determined. The remaining secondary satellites are used to eliminate candidates of the possible ambiguity sets.

The set of potential solutions may be found in the following way. Assume the simplified double-difference model (9.15). If the ambiguities are moved to the left side as if they were known, the model reads $\lambda \Phi - N = \varrho$ where all indices have been omitted. For four satellites, three equations of this type may be set up. The three unknown station coordinates contained in the right side of the equation may be solved by linearizing ϱ and inverting the 3×3 design matrix. Specifying and varying the three ambiguities on the left side gives new position solutions where the inverted design matrix remains unchanged. Depending on the variation of the three ambiguities, the set of potential solutions is obtained. Note that Hatch (1990) does not use double-differences but undifferenced phases to avoid any biasing.

From the set of potential solutions, incorrect solutions are removed by taking into account the information of the secondary group of satellites. Sequential least squares adjustment would be appropriately used for this task. Finally, the sum of the squared residuals may be taken as criterion for the quality indicator of the solution. Ideally, only the true set of ambiguities should remain. If this is not the case, then, as described previously, the solution with the smallest sum of squared residuals should be chosen (after comparing it with the second smallest sum).

Fast ambiguity resolution approach

The development of the fast ambiguity resolution approach (FARA) is given in Frei (1991) and summarized in Frei and Schubernigg (1992). Following the latter publication, the main characteristics are (1) to use statistical information from the initial adjustment to select the search range, (2) to use information of the variance-covariance matrix to reject ambiguity sets that are not acceptable from the statistical point of view, and (3) to apply statistical hypothesis testing to select the correct set of integer ambiguities.

Following Erickson (1992a), the FARA algorithm may be partitioned into four steps: (1) computing the float carrier phase solution, (2) choosing ambiguity sets to be tested, (3) computing a fixed solution for each ambiguity set, and (4) statistically testing the fixed solution with the smallest variance.

In the first step, real values for double-difference ambiguities are estimated based on carrier phase measurements and calculated by an adjustment procedure which also computes the cofactor matrix of the unknown parameters and the a posteriori variance of unit weight (a posteriori variance factor). Based on these results, the variance-covariance matrix of the unknown parameters and the standard deviations of the ambiguities may also be computed.

In the second step, the criteria for the ambiguity ranges to be investigated are based on confidence intervals of the real values of the ambiguities. Therefore, the quality of the initial solution of the first step affects the possible ambiguity ranges. In more detail, if σ_N represents the standard deviation of the ambiguity N, then $\pm k\,\sigma_N$ is the search range for this ambiguity where k is derived statistically from Student's t-distribution. This is the first criterion for selecting possible ambiguity sets.

A second criterion is the use of the correlation of the ambiguities. Assuming the double-difference ambiguities N_i and N_j and the difference

$$N_{ij} = N_j - N_i\,, \tag{9.47}$$

the standard deviation follows from the error propagation law as

$$\sigma_{N_{ij}} = \sqrt{\sigma_{N_i}^2 - 2\sigma_{N_iN_j} + \sigma_{N_j}^2} \tag{9.48}$$

where $\sigma_{N_i}^2$, $\sigma_{N_iN_j}$, and $\sigma_{N_j}^2$ are contained in the variance-covariance matrix of the parameters. The search range for the ambiguity difference N_{ij} is $k_{ij}\,\sigma_{N_{ij}}$ where k_{ij} is analogous to the search range for individual double-difference ambiguities. This criterion significantly reduces the number of possible integer sets. An even more impressive reduction is achieved if dual frequency phase measurements are available. Very illustrative figures demonstrating this reduction are given in Frei and Schubernigg (1992).

In the third step, least squares adjustments with fixed ambiguities are performed for each statistically accepted ambiguity set yielding adjusted baseline components and a posteriori variance factors.

In the fourth and final step, the solution with the smallest a posteriori variance is further investigated. The baseline components of this solution are compared with the float solution. If the solution is compatible, it is accepted. As shown in Erickson (1992a), the compatibility may be checked by a χ^2-distribution which tests the compatibility of the a posteriori variance with the a priori variance. Furthermore, another test may be applied to ensure that the second smallest variance is sufficiently less likely than the smallest variance. Note, however, that these two variances are not independent (Teunissen 1996: Sect. 8.2.3).

As seen from the algorithm, FARA only needs data for double-difference phases; thus, in principle, neither code data nor dual frequency data are required; however, these data will increase the number of possible ambiguity sets dramatically (see the second step of the algorithm).

Euler et al. (1990) present a very efficient and rapid search technique, similar to FARA based on the a posteriori variance (resulting from the sum of the squared residual errors). First, an integer set of ambiguities is introduced in the adjustment computation as constraints leading to an initial solution and the corresponding a posteriori variance. The influence of other ambiguity sets on the initial solution and the a posteriori variance is then determined without recomputing the whole adjustment. This influence may be calculated by some simple matrix and vector operations where only a reduced matrix with the dimension of the constraint ambiguities must be inverted. Following Landau and Euler (1992), the computation time for the matrix inversion may be optimized when the Cholesky factorization method is applied which decomposes a symmetric matrix into a product of a lower and an upper triangle matrix. The impact of a changed ambiguity set on the sum of the squared residuals may be reduced by the Cholesky factorization to the computation of an inner product of two vectors. Furthermore, not even the full inner product must be computed in all cases. Based on a threshold, the computation of the inner product for some integer ambiguity sets may be interrupted and the corresponding ambiguity set rejected.

The performance of this method is demonstrated in Landau and Euler (1992) by imposing examples. Assuming six satellites and therefore five double-difference ambiguities with a ± 10 cycle uncertainty each, the total number of possible combinations is 3.2 millions. Using a 486 PC, the computation by the Cholesky factorization took 49.1 seconds. Optimizing the Cholesky factorization by introducing the above mentioned threshold for the inner product, the computation time reduces to 0.2 seconds. For a larger

search window of ± 50 cycles, the corresponding computations amount to 1.5 days for the Cholesky factorization and 3 seconds for the optimized method. The method may be extended to dual frequency data. The appropriate formulas are given in Landau and Euler (1992).

The search techniques described so far performed the search in the ambiguity domain. An alternate technique substitutes the position as known and solves for the ambiguities as unknowns. This could be performed in the following way. Eliminate the ambiguities by forming triple-differences and obtain a first estimate for the position and its standard deviation σ by an adjustment. Now center the approximate position within a cube of dimension $\pm 3\sigma$ in each coordinate direction and partition the cube into a regular grid. The cube, thus, contains a matrix of points where the center point is the triple-difference solution (Fig. 9.3). Each of these grid points is considered a candidate for the correct solution. Consequently, one by one, each candidate position is substituted into the observation equation. Then the adjustment (holding the trial position fixed) is performed and the ambiguities are computed. When all points within the cube have been considered, select the solution where the estimated real values of the ambiguities appear as close as possible to integer values. Now, fix the ambiguities to these integer values and compute (holding the ambiguities fixed) the final position which will, in general, be slightly different from the corresponding grid point of the cube.

Fast ambiguity search filter

Following Chen (1994) and Chen and Lachapelle (1994), the fast ambiguity search filtering algorithm (FASF) is comprised of basically three components: (1) a Kalman filter is applied to predict a state vector which is treated as observable, (2) the search of the ambiguities is performed at every epoch until they are fixed, and (3) the search ranges for the ambiguities are computed recursively and are related to each other.

By applying the Kalman filter, information from the initial epoch to the current epoch is taken into account. The state vector of the Kalman filter also contains the ambiguities which are estimated as real numbers if they cannot be fixed. After fixing the ambiguities, the state vector is modified accordingly. The state vector of the Kalman filter is considered an observable and establishes, along with the regular observables (i.e., double-difference phase equations), the design matrix.

The recursively determined search ranges are based on the a priori geometric information and the effect of other (preliminarily) fixed ambiguities. As an example, take the case of four double-difference ambiguities. The first ambiguity is computed without fixing any other ambiguity. The search range

for the second ambiguity is computed where the first ambiguity is introduced as a known integer quantity (although it may even be the wrong integer number), the search range for the third ambiguity is computed where the first and the second ambiguity are introduced as known integer quantities, and the procedure is continued for the fourth ambiguity. According to Chen and Lachapelle (1994), this concept is denoted as recursive computation of the search range. This recursive computation is similar to nested loops used in computer programs. Referring to the example of the four ambiguities, four nested loops are required where the loop of the first ambiguity is the outermost and the loop of the fourth ambiguity is the innermost loop. It is important to note that the ranges of the loops (apart from the outermost loop) are computed based on the values of the corresponding outer loop indexes. Thus, e.g., the search range of the second loop is determined by using the ambiguity value corresponding to the first (and outermost) loop index.

To avoid very large search ranges, a computational threshold is used. Ambiguities which cross this threshold are not fixed but computed as real numbers. Thus, an attempt to fix the ambiguities is only made if the number of potential ambiguity sets is below this threshold. Under normal circumstances, the number of potential ambiguity sets should decrease with accumulating observations. Ideally, there should finally remain a single potential ambiguity set. In practice, however, this will usually not be the case so that, conventionally, a ratio test of the sum of the squared residuals between the minimum and the second best minimum is calculated. If this ratio fulfills a specified criterion number, the minimum solution is considered to yield the true set of ambiguities.

Once the ambiguities are fixed properly, they are removed from the state vector of the Kalman filter, i.e., from the estimation. Accordingly, the corresponding observation equation is rearranged.

The ranges of the loops for the ambiguities, i.e., the uncertainties, are calculated by using a least squares approach with parameter elimination. First, the parameters representing the station coordinates are eliminated from the normal equations so that the ambiguities are the only remaining parameters of the model based on double-differences. Furthermore, according to the previous discussion on loops associated with the ambiguities, the ambiguities of the outer loops are constrained as integers (even if they may be wrong values). Returning to the example of the four ambiguities, if the range of the third ambiguity is to be determined, the first and the second ambiguity are assumed to be known and introduced as constraints (which is equivalent to removing them from the estimation vector). In fact, this may be done very efficiently as shown in Chen and Lachapelle (1994) where only single rows and columns of the adjustment matrices must be taken into account. As

result of this parameter elimination, a float estimation of the corresponding ambiguity and its variance are finally obtained. Multiplying the variance by a scale factor and subtracting and adding this result with respect to the float solution yields the search range for this specific ambiguity.

Note that if the uncertainty ranges are not calculated correctly, the true ambiguity set will not be found.

Least squares ambiguity decorrelation adjustment method
Teunissen (1993) proposed the idea and further developed the least squares ambiguity decorrelation adjustment (LAMBDA) method. A fairly detailed description of Teunissen's method is (slightly modified) given here. At present, this method is both theoretically and practically at the top level among the ambiguity determination methods.

The conventional formulation of the adjustment by parameters is

$$\underline{n}^T \underline{P}\, \underline{n} = \text{minimum} \tag{9.49}$$

where $\underline{n}$ is the vector of residuals and $\underline{P}$ is the weight matrix. This formulation implies that the weighted sum of squared residuals is minimized. As shown in Sect. 9.3.1, the weight matrix equals the inverse of the cofactor matrix $\underline{Q}$ of observations. Consequently,

$$\underline{n}^T \underline{Q}^{-1}\, \underline{n} = \text{minimum} \tag{9.50}$$

is an equivalent relation. One remark is appropriate here. Imprecisely, the cofactor matrix is frequently also denoted as covariance matrix. Theoretically, the difference between the two matrices is the a priori variance of unit weight acting as a scale factor which may be arbitrarily chosen. Thus, if this scale factor is set equal to one, the cofactor matrix equals the covariance matrix.

Applying least squares adjustment for, e.g., relative positioning based on double-difference phase observations, the unknowns being determined are coordinate increments for the unknown station and double-difference ambiguities. The values obtained from the adjustment procedures are in the sense of this minimum principle the most likely ones. However, the double-difference ambiguities are obtained as real values (often denoted as float ambiguities) but should be integer values. The main objective is, thus, to obtain integer ambiguities which are the most likely ones. Denoting the vector of adjusted float ambiguities by $\underline{\hat{N}}$ and the vector of the corresponding integer ambiguities by $\underline{N}$, the difference between the two vectors may be regarded as residuals of ambiguities. Consequently, it makes sense to minimize these residuals again by the same principle, i.e., the weighted sum of

squared residuals. Explicitly,

$$(\underline{\hat{N}} - \underline{N})^T \underline{Q}_{\hat{N}}^{-1} (\underline{\hat{N}} - \underline{N}) = \text{minimum} \tag{9.51}$$

is obtained where $\underline{Q}_{\hat{N}}$ is the cofactor or covariance matrix (see discussion above!) of the adjusted float ambiguities. Teunissen et al. (1995) introduce the short-hand notation

$$\chi^2(\underline{N}) = (\underline{\hat{N}} - \underline{N})^T \underline{Q}_{\hat{N}}^{-1} (\underline{\hat{N}} - \underline{N}) = \text{minimum} \tag{9.52}$$

and denote the solution of this problem the integer least squares estimate of the ambiguities. Certainly, an approach different from the usual least squares adjustment calculation must be chosen to account for the integer nature of the still unknown ambiguities $\underline{N}$.

The following simple example demonstrates the solution principle. Considering two ambiguities and assuming $\underline{Q}_{\hat{N}}$ as diagonal matrix

$$\underline{Q}_{\hat{N}} = \begin{bmatrix} q_{\hat{N}_1\hat{N}_1} & 0 \\ 0 & q_{\hat{N}_2\hat{N}_2} \end{bmatrix}, \tag{9.53}$$

equation (9.52) yields the result

$$\chi^2(\underline{N}) = \frac{(\hat{N}_1 - N_1)^2}{q_{\hat{N}_1\hat{N}_1}} + \frac{(\hat{N}_2 - N_2)^2}{q_{\hat{N}_2\hat{N}_2}}. \tag{9.54}$$

The minimum is achieved if the N_i are chosen as those integer values being nearest to the real values. In other words, rounding the real value ambiguities to their nearest integer values yields the desired minimum for $\chi^2(\underline{N})$.

Since $\underline{Q}_{\hat{N}}$ was assumed as diagonal matrix, the resulting N_1 and N_2 are still fully decorrelated which is also evident from Eq. (9.54). Geometrically, if two coordinate axes are associated with N_1 and N_2, this equation represents an ellipse centered around the ambiguities $\underline{\hat{N}}$ and with the semiaxes

$$\begin{aligned} a &= \chi(\underline{N})\sqrt{q_{\hat{N}_1\hat{N}_1}} \\ b &= \chi(\underline{N})\sqrt{q_{\hat{N}_2\hat{N}_2}} \end{aligned} \tag{9.55}$$

where $\chi(\underline{N})$ acts as a scale factor. The axes of the ellipse are parallel to the direction of the coordinate axes. This ellipse is regarded as an ambiguity search space. Mathematically, the two integer ambiguities are contained in the two-dimensional integer space.

In reality, $\underline{Q}_{\hat{N}}$ will be a fully occupied symmetric matrix. The result is still an ellipse, but its axes are rotated with respect to the coordinate system associated with N_1 and N_2, which implies a correlation of the two ambiguities so that it is more complicated to find the minimum for $\bar{\chi}^2(\underline{N})$. In other

words, the rounding to the nearest integer principle no longer works. To return to this convenient feature, the idea is to apply a transformation that decorrelates the ambiguities, which means that the transformed covariance matrix of the ambiguities becomes a diagonal matrix.

Finding a transformation that produces a diagonal matrix for $\underline{Q}_{\hat{N}}$ seems to be trivial since an eigenvalue decomposition yields a diagonal matrix as output. Explicitly, each symmetric matrix

$$\underline{Q} = \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix} \tag{9.56}$$

can be transformed into the diagonal matrix

$$\underline{Q}' = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}. \tag{9.57}$$

The eigenvalues are defined by

$$\begin{aligned} \lambda_1 &= \tfrac{1}{2}(q_{11} + q_{22} + w) \\ \lambda_2 &= \tfrac{1}{2}(q_{11} + q_{22} - w) \end{aligned} \tag{9.58}$$

with the auxiliary quantity

$$w = \sqrt{(q_{11} - q_{22})^2 + 4q_{12}^2}\,. \tag{9.59}$$

The two eigenvectors are orthogonal to each other and are defined by the rotation angle φ, which can be calculated by

$$\tan 2\varphi = \frac{2q_{12}}{q_{11} - q_{22}}\,. \tag{9.60}$$

The only problem is that the integer ambiguities $\underline{N}$ must also be transformed and must preserve their integer nature. Thus, an ordinary eigenvalue decomposition will not work.

Generally, the task may be formulated in the following way. The ambiguities $\underline{N}$ and $\hat{\underline{N}}$ are reparameterized by matrix $\underline{Z}$. Note that Teunissen uses the transposed matrix $\underline{Z}^T$, but the principle remains the same. Hence,

$$\begin{aligned} \underline{N}' &= \underline{Z}\,\underline{N} \\ \hat{\underline{N}}' &= \underline{Z}\,\hat{\underline{N}} \\ \underline{Q}_{\hat{N}'} &= \underline{Z}\,\underline{Q}_{\hat{N}}\,\underline{Z}^T \end{aligned} \tag{9.61}$$

where the transformation of the cofactor matrix is obtained by applying the error propagation law. The ambiguities $\underline{N}'$ obtained after transformation

must remain integer values. That restricts the matrix $\underline{Z}$ to a specific class of transformations where three conditions must be fulfilled (Teunissen 1994, 1995). These conditions are: (1) the elements of the transformation matrix $\underline{Z}$ must be integer values, (2) the transformation must be volume preserving, and (3) the transformation must reduce the product of all ambiguity variances.

Note that the inverse of the transformation matrix $\underline{Z}$ must also consist of integer values only, because upon a retransformation of the (determined) integer ambiguities $\underline{N}'$, the integer nature of the ambiguities must be kept.

For the two-dimensional example shown, volume preserving reduces to area preserving of the ellipse represented by the two-dimensional cofactor (covariance) matrix.

If the three conditions are fulfilled, the transformed integer ambiguities are again integer values and the cofactor (covariance) matrix of the transformed ambiguities is more diagonal than the cofactor (covariance) matrix of the original ambiguities (Teunissen 1994).

The Gauss transformation is one of the possible candidates and may either be expressed by

$$\underline{Z}_1 = \begin{bmatrix} 1 & 0 \\ \alpha_1 & 1 \end{bmatrix} \qquad \alpha_1 = -\mathrm{INT}[q_{\hat{N}_1\hat{N}_2}/q_{\hat{N}_1\hat{N}_1}]\,. \tag{9.62}$$

or by the other form

$$\underline{Z}_2 = \begin{bmatrix} 1 & \alpha_2 \\ 0 & 1 \end{bmatrix} \qquad \alpha_2 = -\mathrm{INT}[q_{\hat{N}_1\hat{N}_2}/q_{\hat{N}_2\hat{N}_2}] \tag{9.63}$$

since the role of the two ambiguities may be interchanged. In the transformation (9.62), the ambiguity $\hat{N}_1$ remains unchanged and $\hat{N}_2$ is transformed. Analogously, $\hat{N}_2$ may be kept unchanged and $\hat{N}_1$ will be transformed as achieved by (9.63). For a better distinction of the two transformations, the subscripts 1 and 2 were introduced. The operator INT performs (like an intrinsic function in a computer language) the rounding to the nearest integer. This procedure is required because the elements of the transformation matrix must be integer quantities (otherwise the integer nature of the ambiguities would be destroyed). The theoretical background of the transformation procedure comprises the conditional least squares estimate (Teunissen 1994).

The transformed ambiguities are obtained from

$$\begin{bmatrix} \hat{N}_1' \\ \hat{N}_2' \end{bmatrix} = \begin{bmatrix} 1 & -\mathrm{INT}[q_{\hat{N}_1\hat{N}_2}/q_{\hat{N}_2\hat{N}_2}] \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{N}_1 \\ \hat{N}_2 \end{bmatrix}. \tag{9.64}$$

For a numerical example, Teunissen (1996), Sect. 8.5.2, assumed that after a least squares adjustment the ambiguities

$$\underline{\hat{N}} = \begin{bmatrix} \hat{N}_1 \\ \hat{N}_2 \end{bmatrix} = \begin{bmatrix} 1.05 \\ 1.30 \end{bmatrix}$$

and

$$\underline{Q}_{\hat{N}} = \begin{bmatrix} q_{\hat{N}_1\hat{N}_1} & q_{\hat{N}_1\hat{N}_2} \\ q_{\hat{N}_1\hat{N}_2} & q_{\hat{N}_2\hat{N}_2} \end{bmatrix} = \begin{bmatrix} 53.4 & 38.4 \\ 38.4 & 28.0 \end{bmatrix}$$

were calculated. Now the transformation is applied to $\underline{Q}_{\hat{N}}$. Translating the matrix elements to variances, the ambiguity $\hat{N}_1$ has a larger variance than $\hat{N}_2$. Hence it is preferable first to change $\hat{N}_1$ and keep $\hat{N}_2$ unchanged, i.e., to apply a transformation based on $\underline{Z}_2$. From (9.63),

$$\alpha_2 = -\mathrm{INT}[q_{\hat{N}_1\hat{N}_2}/q_{\hat{N}_2\hat{N}_2}] = -\mathrm{INT}[38.4/28.0] = -1$$

and

$$\underline{Z}_2 = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$$

are obtained. The transformation according to (9.61) reads

$$\underline{Q}_{\hat{N}'} = \underline{Z}_2\,\underline{Q}_{\hat{N}}\,\underline{Z}_2^T = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 53.4 & 38.4 \\ 38.4 & 28.0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$$

and gives

$$\underline{Q}_{\hat{N}'} = \begin{bmatrix} 4.6 & 10.4 \\ 10.4 & 28.0 \end{bmatrix}.$$

The effect of this transformation can be seen best if the ambiguity search space, represented by the standard ellipse (which is centered around the corresponding ambiguities), is considered. The parameters of the standard ellipse follow from (9.56) through (9.60) if $\underline{Q}$ is replaced by $\underline{Q}_{\hat{N}}$ and $\underline{Q}_{\hat{N}'}$, respectively. The eigenvalues of the matrices equal the squared semiaxes of the ellipse and φ defines the direction of the semimajor axis. Explicitly, the data

$$\underline{Q}_{\hat{N}} : \quad a = 9.0, \quad b = 0.5, \quad \varphi = 35°$$
$$\underline{Q}_{\hat{N}'} : \quad a = 5.7, \quad b = 0.8, \quad \varphi = 69°$$

are obtained. Graphically, the standard ellipses are shown in Fig. 9.4. The standard ellipse for $\underline{Q}_{\hat{N}}$ is centered around the ambiguities $\underline{\hat{N}}$, i.e., the origin

is at $\hat{N}_1 = 1.05$ and $\hat{N}_2 = 1.30$. The standard ellipse for $\underline{Q}_{\hat{N}'}$ is centered around the ambiguities $\underline{\hat{N}}'$, i.e., the origin follows from $\underline{\hat{N}}' = \underline{Z}_2 \underline{\hat{N}}$ and amounts to $\hat{N}_1' = -0.25$ and $\hat{N}_2' = 1.30$.

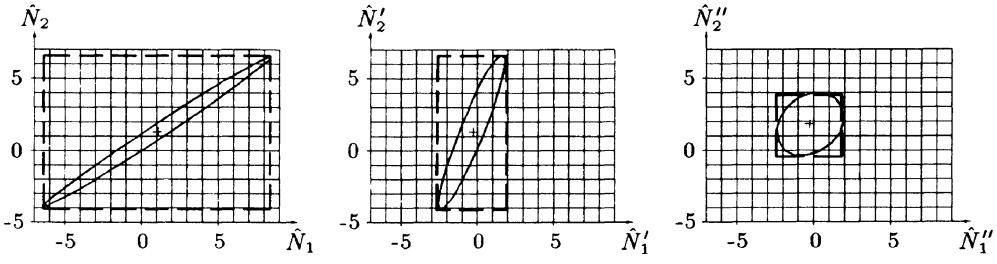


Fig. 9.4. Ambiguity search space for $\underline{Q}_{\hat{N}}$ (left) and transformed ambiguity search spaces for $\underline{Q}_{\hat{N}'}$ (middle) and $\underline{Q}_{\hat{N}''}$ (right)

In Fig. 9.4, search windows are also indicated with sides parallel to the two axes of the two-dimensional integer search space, i.e., two horizontal and two vertical tangents of the ellipse. The "volumes" of the two ellipses are the same because the transformation is volume preserving, but the shape and the orientation of the ellipse has changed. The distance between the two horizontal tangents has not changed because these two tangents bound the search range for the N_2 ambiguity which remained unaltered by the $\underline{Z}_2$ transformation, whereas the distance of the two vertical tangents has changed.

Each grid point represents one pair of ambiguities. Under the assumption that each grid point of the search window must be regarded as a possible candidate to be investigated for a reasonable solution, the advantage of the transformed search space becomes obvious.

From comparing the off-diagonal elements of $\underline{Q}_{\hat{N}}$ and of the transformed $\underline{Q}_{\hat{N}'}$, the decrease of correlation is evident.

Another transformation may now be applied to $\underline{Q}_{\hat{N}'}$. Since ambiguity $\hat{N}_2'$ has a larger variance than $\hat{N}_1'$ it is preferable to change $\hat{N}_2'$ and keep $\hat{N}_1'$ unchanged, i.e., to apply a transformation based on $\underline{Z}_1$. First, from (9.62)

$$\alpha_1 = -\mathrm{INT}[q_{\hat{N}_1\hat{N}_2}/q_{\hat{N}_1\hat{N}_1}] = -\mathrm{INT}[10.4/4.6] = -2$$

is determined giving

$$\underline{Z}_1 = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix}$$

and

$$\underline{Q}_{\hat{N}''} = \underline{Z}_1 \underline{Q}_{\hat{N}'} \underline{Z}_1^T = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 4.6 & 10.4 \\ 10.4 & 28.0 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}$$

where the double prime expresses that the transformation is applied on the once transformed matrix. The result is

$$\underline{Q}_{\hat{N}''} = \begin{bmatrix} 4.6 & 1.2 \\ 1.2 & 4.8 \end{bmatrix}.$$

The standard ellipse for $\underline{Q}_{\hat{N}''}$ is given by $a = 2.4$, $b = 1.9$, $\varphi = 47°$ and is shown in Fig. 9.4. The standard ellipse for $\underline{Q}''_{\hat{N}}$ is centered around the ambiguities $\underline{\hat{N}}''$, i.e., the origin follows from $\underline{\hat{N}}'' = \underline{Z}_1 \underline{\hat{N}}'$ and amounts to $\hat{N}_1'' = -0.25$ and $\hat{N}_2'' = 1.80$. As far as the search window is concerned, the effect may easily be seen from the much smaller search area (represented by the window) of $\underline{Q}_{\hat{N}''}$.

Accordingly, the distance between the two vertical tangents has not changed because these two tangents bound the search range for the N_1 ambiguity which remained unaltered by the $\underline{Z}_1$ transformation, whereas the distance of the two horizontal tangents has changed.

From comparing the off-diagonal elements of $\underline{Q}_{\hat{N}'}$ and of the transformed $\underline{Q}_{\hat{N}''}$, the decrease of correlation is evident. However, the ambiguities are still not fully decorrelated.

The two transformations may also be combined to a single transformation. Using $\underline{Q}_{\hat{N}''} = \underline{Z}_1 \underline{Q}_{\hat{N}'} \underline{Z}_1^T$ and substituting $\underline{Q}_{\hat{N}'} = \underline{Z}_2 \underline{Q}_{\hat{N}} \underline{Z}_2^T$ leads to

$$\underline{Q}_{\hat{N}''} = \underbrace{\underline{Z}_1 \underline{Z}_2}_{\underline{Z}} \underline{Q}_{\hat{N}} \underbrace{\underline{Z}_2^T \underline{Z}_1^T}_{\underline{Z}^T}$$

where

$$\underline{Z} = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -2 & 3 \end{bmatrix}$$

so that now the single transformation matrix $\underline{Z}$ represents the composition of the $\underline{Z}_2$ and the $\underline{Z}_1$ transformation.

The extension of the reparameterization of the ambiguity search space to higher dimensions is possible. Teunissen (1996), Sect. 8.5.3, gives the decorrelating ambiguity transformation $\underline{Z}$ for the three-dimensional case which would apply if double-differences of four satellites are used, and a twelve-dimensional transformation for seven satellites and dual frequency data. Rizos and Han (1995) propose an iterative procedure to generate the decorrelating ambiguity transformation $\underline{Z}$. Note that the ambiguity search space

becomes an ellipsoid for the three-dimensional example and an n-dimensional hyperellipsoid for $n > 3$.

After the decorrelation of the ambiguities by the $\underline{Z}$ transformation, the task of actually solving ambiguity estimates remains. The search can be carried out very efficiently by using the sequential conditional adjustment, a standard technique in adjustment theory. Related to the ambiguity estimation, an overview is given in Jonge and Tiberius (1995) and some details are covered in Teunissen (1996), Sect. 8.3.2. The sequential conditional adjustment determines the ambiguities step by step (i.e., sequential) one after the other. For the i-th ambiguity to be estimated, the previously determined $i - 1$ ambiguities are fixed (i.e., conditional). The sequential conditional least squares adjustment ambiguities are not correlated. This means that the effect of the $\underline{Z}$ transformation will not be destroyed.

Some details on the actual discrete search strategy are given in Teunissen (1994), Teunissen et al. (1994), Teunissen (1996), Sects. 8.3.2, 8.5.3.

In summary, Teunissen's LAMBDA method may be separated into the following steps:

1. A conventional least squares adjustment is carried out to yield the baseline components and float ambiguities.

2. Using the $\underline{Z}$ transformation, the ambiguity search space is reparameterized to decorrelate the float ambiguities.

3. Using the sequential conditional least squares adjustment together with a discrete search strategy, the integer ambiguities are estimated. By the inverse transformation $\underline{Z}^{-1}$, the ambiguities are retransformed to the original ambiguity space where the baseline components are given. Since $\underline{Z}^{-1}$ consists only of integer elements, the integer nature of the ambiguities is kept.

4. The integer ambiguities are fixed as known quantities and another conventional least squares adjustment to determine the final baseline components is performed.

Ambiguity determination with special constraints

Several multiple receiver methods for kinematic applications exist. One common procedure of this technique is to place two or more receivers at fixed locations (usually short distances apart) of the moving object. Since the locations of the antennas are fixed, constraints (e.g., the distance between two antennas) may be formulated which can be used to increase the efficiency of the ambiguity resolution. In principle, the gain by using constraints results

in a reduction of the potential ambiguity sets. This is illustrated briefly by two examples; namely, attitude determination and aircraft-to-aircraft positioning.

The example of attitude determination in a marine environment is taken from Lu and Cannon (1994) and employs the distances between the antennas on a ship as constraints for the ambiguity resolution. Attitude determination is explained in Sect. 12.2. Here, only the principle of the ambiguity resolution with the constraint of the known distance for a single baseline is described. Referring to the double-difference model (9.15), four satellites yielding three double-differences are considered. Analogously to the procedure for the least squares ambiguity search technique described earlier, the equations are reformulated as $\lambda \Phi - N = \varrho$ where all indices have been omitted. The specification of N, the double-difference ambiguities, is accomplished by searching a possible set of ambiguity solutions. The most conservative method to define the search space would be to take the length of the baseline and divide it by the wavelength of the carrier used, i.e., 19 cm in the case of L1. The resulting number of cycles taken once positively and once negatively and considered for each ambiguity defines the search area. Various methods exist to reduce the burden of the search. Jurgens et al. (1991) propose for example a third antenna aligned with the other two antennas (of the baseline to be determined) and located within less than one carrier wavelength from one of the two antennas. In general, the double-difference ambiguities of the two close-by antennas will vanish so that an approximate azimuth and pitch information can be determined and used to reduce the search space for the remaining longer baseline.

Lu and Cannon (1994) and Lu (1995) reduce the search space by introducing the known distance of the baseline. Referring to the system $\lambda \Phi - N = \varrho$, three double-differences are considered and the linearization of ϱ is performed with respect to the reference station of the baseline. Thus, the linearized system may be written as $\underline{w} = \underline{A}\,\underline{x}$ where $\underline{A}$ is a 3×3 design matrix resulting from the linearization, $\underline{x}$ contains the unknown baseline components (since the linearization was carried out with respect to the known station), and the left side of the equation contains the residual vector $\underline{w}$ which also comprises the ambiguities. Since $\underline{x}$ represents the baseline components, the constraint of the length of the baseline, denoted by b, may be introduced by first forming $\underline{A}^{-1}\underline{w} = \underline{x}$ and then $b^2 = \underline{x}^T\underline{x} = \underline{w}^T(\underline{A}\,\underline{A}^T)^{-1}\underline{w}$. This system may be further simplified by applying to $\underline{A}\,\underline{A}^T$ a Cholesky decomposition which reduces $\underline{A}$ to a lower triangle matrix. The advantage obtained from this decomposition is that the third ambiguity may be expressed by a quadratic equation containing the other two ambiguities. Thus, introducing search trials for these two other ambiguities yields two solutions for the third

ambiguity. Therefore, the constraint significantly reduces the search space. Redundant satellites may be used to further reduce the size of the search space.

The performance of this method can best be seen by means of a simple example. Assuming a ± 15 cycle uncertainty for the three unknowns would yield (together with the one ambiguity set obtained by rounding the calculated unknowns to their nearest integer values) $31 \times 31 \times 31 = 29\,791$ possible ambiguity sets, whereas taking into account the constraint as described above reduces the set of possible ambiguities to $31 \times 31 \times 2 = 1\,922$.

The second example presented here refers to the introduction of constraints for an aircraft-to-aircraft positioning as proposed in Lachapelle et al. (1994). The situation is shown in Fig. 9.5. Each of the two aircraft is equipped with two receivers. The corresponding distances of the antennas between i and j on one aircraft and k and ℓ on the other aircraft are known and may be introduced as constraints to determine the double-difference ambiguities for each airplane separately, i.e, the double-differences N_{ij} and $N_{k\ell}$ for the available satellites (which are not indicated here by appropriate superscripts). These resolved ambiguities N_{ij} and $N_{k\ell}$ may now be used to interrelate the two aircraft by constraints. As proposed in Lachapelle et al. (1994), three sets of double-difference ambiguity relations are constrained by using, e.g., $N_{ij} = N_{jk} - N_{ik}$, $N_{ij} = N_{j\ell} - N_{i\ell}$, and $N_{k\ell} = N_{\ell i} - N_{ki}$. Thus, for five satellites there are 4×3 double-difference equations of this type which are used to reduce the number of potential ambiguity solutions. Note that these relations are not independent from each other but may still contribute to average out several error sources like carrier phase noise and multipath.

Examples given in Lachapelle et al. (1994) demonstrate that for two aircraft within 1 km typically 4 to 6 minutes of measurements (with a data rate of 1 Hz) are sufficient to obtain a unique solution. The correctness of the ambiguities may roughly be checked by the double-difference phase residuals which must not show a significant drift over time. A drift would be an indication of wrong ambiguities. The rms of the double-difference phase residuals was in the amount of 0.8 cm.

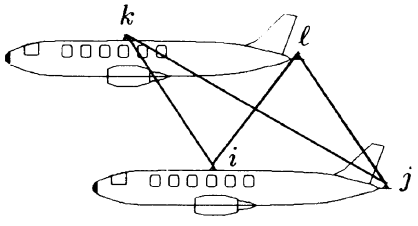


Fig. 9.5. Aircraft-to-aircraft GPS positioning with four receivers

Based on the given data set, several trials were performed by shifting the initial epoch from one trial to the next by 90 seconds. Of these trials, some 50% yielded the same ambiguities. This indicates the correctness of like ambiguity sets; however, this also demonstrates that the reliability of a single solution is not sufficient.

9.2.4 Ambiguity validation

After the determination of the integer ambiguities, it is of interest to validate the quality of the obtained quantities (Wang 1999). Therefore, the uncertainty of the estimated integer ambiguities is to be determined. As pointed out in Joosten and Tiberius (2000), the distribution of the estimated integer ambiguities will be a probability mass function. For a probabilistic measure, the ambiguity success rate is defined which quantifies the probability that the integer ambiguities are correctly estimated. The ambiguity success rate equals the integral of the probability density function of the float ambiguities. The integral extends over the so-called pull-in region in which all float solutions are pulled by the integer least squares criterion to the correct integer ambiguity solution. Due to its definition as probability measure, the success rate is only a single number between 0 and 1 (which may also be expressed as percentage between 0% and 100%).

As mentioned earlier, the ambiguity success rate depends on the functional model, the stochastic model, and the chosen method of integer estimation. Similar to the DOP computations, the success rate may be calculated without actual measurements if the functional and the stochastic model are known.

With respect to the integer estimation method, Teunissen (1999a, b) has proven that the LAMBDA method delivers the optimum success rate of all admissible integer estimators. A proper choice of the weight matrix is also important for the ambiguity resolution. A too optimistic as well as pessimistic precision description will both result in a less than optimal ambiguity success rate. Jonkman (1998) and Teunissen et al. (1998) demonstrate examples of an increased ambiguity success rate by improving the stochastic modeling.

Several methods exist for the computation of the success rate. Joosten and Tiberius (2000) describe a simulation procedure based on a random number generator and somewhere between 100 000 and 1 million samples to achieve a success rate of 99.9 percent. Another method mentioned in Joosten and Tiberius (2000) is the computation of a "sharp lower bound" of the probability of correct integer least squares estimation using conditional standard deviations of the ambiguities which follow directly from the

triangular decomposition of the float ambiguity variance-covariance matrix. Applying the LAMBDA method, this decomposition is available without additional computational effort.

Joosten et al. (1999) stress that the success rate should be considered as the measure for judging the success of ambiguity resolution. When using the standard deviations of the ambiguities, this may yield very misleading results for two reasons: (1) the correlations are neglected when only using the standard deviations, (2) ambiguity transformations change the standard deviations. In contrast to this, the success rate as defined previously is invariant for any ambiguity transformation.

9.3 Adjustment, filtering, and smoothing

9.3.1 Least squares adjustment

Standard adjustment

There are numerous adjustment techniques that can be used, but least squares adjustment with parameters is the only one discussed here. It is based on equations where the observations are expressed as a function of unknown parameters. A Taylor series expansion is usually performed in the case of nonlinear functions. This requires approximate values for the parameters. The Taylor series expansion must be truncated after the second term to obtain a linear function with respect to the unknowns. The resulting linear observation model can be represented in a matrix-vector notation as

$$\underline{\ell} = \underline{A}\,\underline{x} \tag{9.65}$$

where

$\underline{\ell}$ ... vector of observations
$\underline{A}$ ... design matrix
$\underline{x}$ ... vector of unknowns.

By introducing in addition the definitions

σ_0^2 ... a priori variance
$\underline{\Sigma}$ ... covariance matrix,

the cofactor matrix of observations is

$$\underline{Q}_\ell = \frac{1}{\sigma_0^2}\underline{\Sigma}, \tag{9.66}$$

and

$$\underline{P} = \underline{Q}_\ell^{-1} \tag{9.67}$$

is the weight matrix. Assuming n observations and u unknown parameters leads to a design matrix $\underline{A}$ comprising n rows and u columns. For $n > u$, the system (9.65) is redundant (overdetermined) and, in general, nonconsistent because of observational errors or noise. To assure consistency, the noise vector $\underline{n}$ is added to the vector of observations and Eq. (9.65) converts to

$$\underline{\ell} + \underline{n} = \underline{A}\,\underline{x}\,. \tag{9.68}$$

The solution of this system becomes unique by the least squares principle $\underline{n}^T \underline{P}\,\underline{n}$ = minimum. The application of this minimum principle on the observation equations (9.68) leads to the normal equations

$$\underline{A}^T \underline{P}\,\underline{A}\,\underline{x} = \underline{A}^T \underline{P}\,\underline{\ell} \tag{9.69}$$

with the solution

$$\underline{x} = (\underline{A}^T \underline{P}\,\underline{A})^{-1} \underline{A}^T \underline{P}\,\underline{\ell} \tag{9.70}$$

which can be simplified to

$$\underline{x} = \underline{G}^{-1} \underline{g} \tag{9.71}$$

where $\underline{G} = \underline{A}^T \underline{P}\,\underline{A}$ and $\underline{g} = \underline{A}^T \underline{P}\,\underline{\ell}$.

The cofactor matrix $\underline{Q}_x$ follows from $\underline{x} = \underline{G}^{-1} \underline{A}^T \underline{P}\,\underline{\ell}$ by the covariance propagation law as

$$\underline{Q}_x = (\underline{G}^{-1} \underline{A}^T \underline{P})\, \underline{Q}_\ell\, (\underline{G}^{-1} \underline{A}^T \underline{P})^T \tag{9.72}$$

and reduces to

$$\underline{Q}_x = \underline{G}^{-1} = (\underline{A}^T \underline{P}\,\underline{A})^{-1} \tag{9.73}$$

by substituting $\underline{Q}_\ell = \underline{P}^{-1}$.

Sequential adjustment

Assume a partitioning of the observation model (9.68) into two subsets:

$$\underline{\ell} = \begin{bmatrix} \underline{\ell}_1 \\ \underline{\ell}_2 \end{bmatrix} \qquad \underline{n} = \begin{bmatrix} \underline{n}_1 \\ \underline{n}_2 \end{bmatrix} \qquad \underline{A} = \begin{bmatrix} \underline{A}_1 \\ \underline{A}_2 \end{bmatrix}. \tag{9.74}$$

Using the first set only, a preliminary solution $\underline{x}_{(0)}$ can be calculated according to (9.70) and (9.73) by

$$\begin{aligned} \underline{x}_{(0)} &= (\underline{A}_1^T \underline{P}_1\, \underline{A}_1)^{-1} \underline{A}_1^T \underline{P}_1\, \underline{\ell}_1 = \underline{G}_1^{-1} \underline{g}_1 \\ \underline{Q}_{x_{(0)}} &= (\underline{A}_1^T \underline{P}_1\, \underline{A}_1)^{-1} = \underline{G}_1^{-1}\,. \end{aligned} \tag{9.75}$$

Provided that there is no correlation between the two subsets of observations, the weight matrix

$$\underline{P} = \begin{bmatrix} \underline{P}_1 & \underline{0} \\ \underline{0} & \underline{P}_2 \end{bmatrix} \tag{9.76}$$

is a block-diagonal matrix. The matrix $\underline{G}$ and the vector $\underline{g}$ for the adjustment of the full set of observations result from adding the corresponding matrices and vectors for the two subsets:

$$\begin{aligned} \underline{G} &= \underline{A}^T \underline{P}\, \underline{A} = (\underline{A}_1^T \underline{P}_1\, \underline{A}_1 + \underline{A}_2^T \underline{P}_2\, \underline{A}_2) = \underline{G}_1 + \underline{G}_2 \\ \underline{g} &= \underline{A}^T \underline{P}\, \underline{\ell} = (\underline{A}_1^T \underline{P}_1\, \underline{\ell}_1 + \underline{A}_2^T \underline{P}_2\, \underline{\ell}_2) = \underline{g}_1 + \underline{g}_2 \,. \end{aligned} \tag{9.77}$$

If the change of the preliminary solution $\underline{x}_{(0)}$ due to the additional observation set $\underline{\ell}_2$ is denoted as $\Delta\underline{x}$, then

$$(\underline{G}_1 + \underline{G}_2)\,(\underline{x}_{(0)} + \Delta\underline{x}) = \underline{g}_1 + \underline{g}_2 \tag{9.78}$$

is the appropriate formulation of the adjustment. This equation can be slightly rearranged to

$$(\underline{G}_1 + \underline{G}_2)\,\Delta\underline{x} = \underline{g}_1 + \underline{g}_2 - (\underline{G}_1 + \underline{G}_2)\,\underline{x}_{(0)} \tag{9.79}$$

where the right side, cf. Eq. (9.75), can be simplified because of the relation $\underline{g}_1 - \underline{G}_1\,\underline{x}_{(0)} = \underline{0}$ so that

$$(\underline{G}_1 + \underline{G}_2)\,\Delta\underline{x} = \underline{g}_2 - \underline{G}_2\,\underline{x}_{(0)} \tag{9.80}$$

results. Resubstituting $\underline{g}_2$ and $\underline{G}_2$ from (9.77) yields

$$(\underline{G}_1 + \underline{G}_2)\,\Delta\underline{x} = \underline{A}_2^T \underline{P}_2\,\underline{\ell}_2 - \underline{A}_2^T \underline{P}_2\,\underline{A}_2\,\underline{x}_{(0)} \tag{9.81}$$

or

$$(\underline{G}_1 + \underline{G}_2)\,\Delta\underline{x} = \underline{A}_2^T \underline{P}_2\,(\underline{\ell}_2 - \underline{A}_2\,\underline{x}_{(0)}) \tag{9.82}$$

and

$$\Delta\underline{x} = (\underline{G}_1 + \underline{G}_2)^{-1} \underline{A}_2^T \underline{P}_2\,(\underline{\ell}_2 - \underline{A}_2\,\underline{x}_{(0)}) \tag{9.83}$$

or, finally,

$$\Delta\underline{x} = \underline{K}\,(\underline{\ell}_2 - \underline{A}_2\,\underline{x}_{(0)}) \tag{9.84}$$

where

$$\underline{K} = (\underline{G}_1 + \underline{G}_2)^{-1} \underline{A}_2^T \underline{P}_2 \,. \tag{9.85}$$

Note that the term $\underline{A}_2\,\underline{x}_{(0)}$ in (9.84) can formally be considered as prediction for the observations $\underline{\ell}_2$.

The change $\Delta\underline{Q}$ with respect to the preliminary cofactor matrix $\underline{Q}_{x_{(0)}}$ is obtained from the relation

$$\underline{G}\,\underline{Q}_x = (\underline{G}_1 + \underline{G}_2)\,(\underline{Q}_{x_{(0)}} + \Delta\underline{Q}) = \underline{I} \tag{9.86}$$

where $\underline{I}$ denotes the unit matrix. This equation is reformulated as

$$(\underline{G}_1 + \underline{G}_2)\,\Delta\underline{Q} = \underline{I} - (\underline{G}_1 + \underline{G}_2)\,\underline{Q}_{x_{(0)}} \tag{9.87}$$

and, since $\underline{G}_1\,\underline{Q}_{x_{(0)}} = \underline{I}$, this reduces to

$$(\underline{G}_1 + \underline{G}_2)\,\Delta\underline{Q} = -\underline{G}_2\,\underline{Q}_{x_{(0)}} \tag{9.88}$$

or

$$\Delta\underline{Q} = -(\underline{G}_1 + \underline{G}_2)^{-1}\underline{G}_2\,\underline{Q}_{x_{(0)}} \tag{9.89}$$

and, by resubstituting $\underline{G}_2$ from (9.77), the relation

$$\Delta\underline{Q} = -(\underline{G}_1 + \underline{G}_2)^{-1}\underline{A}_2^T\underline{P}_2\,\underline{A}_2\,\underline{Q}_{x_{(0)}} \tag{9.90}$$

follows. Comparing this equation with (9.85), $\underline{K}$ may be substituted and

$$\Delta\underline{Q} = -\underline{K}\,\underline{A}_2\,\underline{Q}_{x_{(0)}} \tag{9.91}$$

results. Matrix $\underline{K}$, which is denoted as gain matrix, satisfies the remarkable relation

$$\underline{K} = (\underline{G}_1 + \underline{G}_2)^{-1}\underline{A}_2^T\underline{P}_2 = \underline{G}_1^{-1}\underline{A}_2^T(\underline{P}_2^{-1} + \underline{A}_2\,\underline{G}_1^{-1}\underline{A}_2^T)^{-1}\,. \tag{9.92}$$

This relation is based on a formula found by Bennet (1965). For additional information see Moritz (1980), Eqs. (19-12) and (19-13). The point of this equation is its application to the inversion of modified matrices of the type $(\underline{C} + \underline{D})$ where $\underline{C}^{-1}$ is known a priori. The identity of (9.92) may be proved by multiplying both sides from left by $(\underline{G}_1 + \underline{G}_2)$ and from right by $(\underline{P}_2^{-1} + \underline{A}_2\,\underline{G}_1^{-1}\underline{A}_2^T)$.

It is essential to learn from Eq. (9.92) that the first form for $\underline{K}$ implies the inversion of a $u \times u$ matrix if u is the number of unknown parameters; whereas, for the second form, an inversion of an $n_2 \times n_2$ matrix is necessary when n_2 denotes the number of observations for the second subset. Therefore, the second form is advantageous as long as $n_2 < u$.

A final remark should conclude the section on the sequential adjustment. Formally, if the substitution $\underline{A}_1 = \underline{I}$ and $\underline{\ell}_1 = \underline{x}_{(0)}$ is performed, then the model for the sequential adjustment is formulated as

$$\begin{aligned} \underline{x}_{(0)} + \underline{n}_1 &= \quad \underline{x} \\ \underline{\ell}_2 \quad + \underline{n}_2 &= \underline{A}_2\, \underline{x}\,. \end{aligned} \tag{9.93}$$

Model (9.93) reflects that the preliminary estimates $\underline{x}_{(0)}$ for the unknown parameters are introduced into the sequential adjustment as observations. This approach is often used in the context with Kalman filtering.

9.3.2 Kalman filtering

Introduction
Consider a dynamic system such as a moving vehicle. The unknown parameters, e.g., the coordinates and the velocity, form the elements of the state vector. This time dependent vector may be predicted for any instant t by means of system equations. The predicted values can be improved or updated by observations containing information on some components of the state vector.

The whole procedure is known as Kalman filtering. It corresponds to the sequential adjustment in the static case. Consequently, optimal estimates of the unknowns on the basis of all observations up to the epoch t are obtained. Note, however, there is no need to store these data for the subsequent epochs.

Prediction
The time dependent state vector $\underline{x}(t)$ comprising the unknown parameters of the dynamic system may be modeled by a system of differential equations of the first order as

$$\dot{\underline{x}}(t) = \underline{F}(t)\, \underline{x}(t) + \underline{w}(t) \tag{9.94}$$

where

$\dot{\underline{x}}(t)$ ... time derivative of the state vector
$\underline{F}(t)$ ... dynamics matrix
$\underline{w}(t)$ ... driving noise.

For the following, at the initial epoch t_0, the state vector $\underline{x}(t_0)$ and its cofactor matrix $\underline{Q}_{x_0}$ are assumed to be known. A general solution for the

system equation (9.94) only exists if the matrix $\underline{F}(t)$ contains periodic or constant coefficients. For the latter case, this solution can be written as

$$\begin{aligned}\underline{x}(t) &= \underline{T}(t, t_0)\,\underline{x}(t_0) + \int_{t_0}^{t} \underline{T}(t, \tau)\,\underline{w}(\tau)\,d\tau \\ &= \underline{T}(t, t_0)\,\underline{x}(t_0) + \underline{e}(t)\,.\end{aligned} \tag{9.95}$$

To express the transition matrix $\underline{T}$ in function of the dynamics matrix $\underline{F}$, the state vector at epoch t is expressed by a Taylor series expansion. Thus,

$$\underline{x}(t) = \underline{x}(t_0) + \dot{\underline{x}}(t_0)\,(t - t_0) + \tfrac{1}{2}\,\ddot{\underline{x}}(t_0)\,(t - t_0)^2 + \ldots \tag{9.96}$$

is obtained. Substituting (9.94), assuming the dynamics matrix to be constant, and neglecting the driving noise $\underline{w}(t)$ gives

$$\underline{x}(t) = \underline{x}(t_0) + \underline{F}(t_0)\,\underline{x}(t_0)\,(t - t_0) + \tfrac{1}{2}\,\underline{F}(t_0)^2\,\underline{x}(t_0)\,(t - t_0)^2 + \ldots \tag{9.97}$$

which, by comparing with (9.95), may be written in the form

$$\underline{x}(t) = \underline{T}(t, t_0)\,\underline{x}(t_0)\,. \tag{9.98}$$

Consequently, by introducing the substitution $\Delta t = t - t_0$, the transition matrix is obtained as an infinite series with respect to $\underline{F}$ by

$$\begin{aligned}\underline{T}(t, t_0) &= \underline{I} + \underline{F}(t_0)\,\Delta t + \tfrac{1}{2}\,\underline{F}(t_0)^2\,\Delta t^2 + \ldots \\ &= \sum_{n=0}^{\infty} \frac{1}{n!}\,\underline{F}(t_0)^n\,\Delta t^n\,.\end{aligned} \tag{9.99}$$

The cofactor matrix $\underline{Q}_x$ of the state vector $\underline{x}(t)$ can be calculated via (9.95) by using the law of covariance propagation yielding

$$\underline{Q}_x = \underline{T}(t, t_0)\,\underline{Q}_{x_0}\,\underline{T}^T(t, t_0) + \underline{Q}_e\,. \tag{9.100}$$

Studying Eqs. (9.95) and (9.100), one may conclude that the essential problems of Kalman filtering are the definition of the transition matrix $\underline{T}$ and of the cofactor matrix $\underline{Q}_e$.

Update

Starting at the initial epoch t_0, the state vector $\underline{x}(t)$ can be predicted for any arbitrary future epoch t by the system equations (9.95). Taking into account the system noise $\underline{e}(t)$, it is assumed that observations $\underline{\ell}(t)$ and the corresponding cofactors $\underline{Q}_\ell$ are available at epoch t. These data may be

related to the updated state vector $\underline{\hat{x}}(t)$ – possibly after a necessary linearization – by the equation

$$\underline{\ell}(t) = \underline{A}\,\underline{\hat{x}}(t)\,. \tag{9.101}$$

Considering both the predicted state vector $\underline{x}(t)$ and the observations $\underline{\ell}(t)$ as stochastic quantities leads to the sequential adjustment problem

$$\begin{aligned} \underline{x}(t) + \underline{n}_x &= \quad \underline{\hat{x}}(t) \\ \underline{\ell}(t) + \underline{n}_\ell &= \underline{A}\,\underline{\hat{x}}(t)\,. \end{aligned} \tag{9.102}$$

This system is equivalent to (9.93). Therefore, the solution can be taken immediately from (9.84) and (9.91) by matching the notations to the present situation, and it follows

$$\begin{aligned} \underline{\hat{x}}(t) &= \underline{x}(t) + \Delta\underline{x}(t) = \underline{x}(t) + \underline{K}\,[\underline{\ell}(t) - \underline{A}\,\underline{x}(t)] \\ \underline{\hat{Q}}_x &= \underline{Q}_x + \Delta\underline{Q}_x \quad = (\underline{I} - \underline{K}\,\underline{A})\,\underline{Q}_x\,. \end{aligned} \tag{9.103}$$

The gain matrix $\underline{K}$ is now given as

$$\underline{K} = \underline{Q}_x\,\underline{A}^T(\underline{Q}_\ell + \underline{A}\,\underline{Q}_x\,\underline{A}^T)^{-1}\,, \tag{9.104}$$

cf. also Eq. (9.92).

Example
Consider a vehicle moving on a straight line with constant velocity v where the motion is affected by the random acceleration a. Also, assume that the (one-dimensional) position $p(t_0)$ and the velocity $v(t_0)$ as well as the corresponding variances σ_p^2, σ_v^2 and that of the noise, σ_a^2, are known at the initial epoch t_0. Furthermore, it is assumed that the position of the vehicle is observed at an epoch $t = t_0 + \Delta t$ and that the observation has the variance σ_ℓ^2 at this epoch.

The state vector consists of the position and the velocity of the vehicle. Thus, for the initial epoch,

$$\underline{x}(t_0) = \begin{bmatrix} p(t_0) \\ v(t_0) \end{bmatrix} \quad \text{and} \quad \dot{\underline{x}}(t_0) = \begin{bmatrix} \dot{p}(t_0) \\ \dot{v}(t_0) \end{bmatrix} = \begin{bmatrix} v(t_0) \\ 0 \end{bmatrix} \tag{9.105}$$

are obtained. The substitution of these vectors and of the random acceleration a into (9.94) yields the dynamics matrix and the driving noise vector for epoch t_0:

$$\underline{F}(t_0) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad \underline{w}(t_0) = a \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{9.106}$$

The transition matrix according to Eq. (9.99) is

$$\underline{T}(t, t_0) = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix} \tag{9.107}$$

where the infinite series has been truncated after the linear term. Assuming a constant acceleration during the integration interval Δt, the noise vector is obtained from (9.95) as

$$\underline{e}(t) = a \begin{bmatrix} \frac{\Delta t^2}{2} \\ \Delta t \end{bmatrix} . \tag{9.108}$$

Note that under the present assumptions the elements of the predicted state vector $\underline{x}(t)$ would also result from the formulas of accelerated motion.

The cofactor matrix $\underline{Q}_x$ of the predicted state vector $\underline{x}(t)$ follows from Eq. (9.100) as

$$\begin{aligned} \underline{Q}_x &= \begin{bmatrix} \sigma_p^2 + \Delta t^2\,\sigma_v^2 + \frac{1}{4}\Delta t^4\,\sigma_a^2 & \Delta t\,\sigma_v^2 + \frac{1}{2}\Delta t^3\,\sigma_a^2 \\ \Delta t\,\sigma_v^2 + \frac{1}{2}\Delta t^3\,\sigma_a^2 & \sigma_v^2 + \Delta t^2\,\sigma_a^2 \end{bmatrix} \\ &= \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12} & Q_{22} \end{bmatrix} . \end{aligned} \tag{9.109}$$

Since the observation equation is $\ell(t) + n(t) = \hat{p}(t)$, the matrix $\underline{A}$ in (9.102) shrinks to the row vector

$$\underline{A} = \begin{bmatrix} 1 & 0 \end{bmatrix} , \tag{9.110}$$

and the gain matrix reduces to the column vector

$$\underline{K} = \frac{1}{Q_{11} + \sigma_\ell^2} \begin{bmatrix} Q_{11} \\ Q_{12} \end{bmatrix} . \tag{9.111}$$

Now, the updated state vector $\hat{\underline{x}}(t)$ and the corresponding cofactor matrix $\hat{\underline{Q}}_x$ can be calculated by Eqs. (9.103).

9.3.3 Smoothing

The process of improving previous estimates for the state vector by a new measurement is called smoothing. Since smoothing is performed backwards in time, it is (contrary to the real-time Kalman filtering) a postmission process. One smoothing technique is presented here for the sake of completeness.

Using the notations $\underline{x}(t)$ for the predicted state vector, $\underline{\hat{x}}(t)$ for the updated state vector, and $\underline{\overset{\circ}{x}}(t)$ for the smoothed state vector, then a formula for optimal smoothing is

$$\underline{\overset{\circ}{x}}(t_i) = \underline{\hat{x}}(t_i) + \underline{D}(t_{i+1}, t_i)\left[\underline{\overset{\circ}{x}}(t_{i+1}) - \underline{x}(t_{i+1})\right] \tag{9.112}$$

where the gain matrix is

$$\underline{D}(t_{i+1}, t_i) = \underline{\hat{Q}}_{x_i}\, \underline{T}(t_{i+1}, t_i)\, \underline{Q}^{-1}_{x_{i+1}} \tag{9.113}$$

At the epoch of the last update measurement, the updated state vector is set identical to the smoothed one, and the backwards algorithm can be started. From Eq. (9.112), one may conclude that the process requires the predicted and updated vectors and the cofactor matrices at the update epochs as well as the transition matrices between the updates. This implies, in general, a large amount of data. This is probably the reason why optimal smoothing is very often replaced by empirical methods.

9.4 Adjustment of mathematical GPS models

9.4.1 Linearization

When the models of Chap. 8 are considered, the only term comprising unknowns in nonlinear form is ϱ. This section explains in detail how ϱ is linearized. The basic formula from Eq. (8.2)

$$\begin{aligned} \varrho_i^j(t) &= \sqrt{(X^j(t) - X_i)^2 + (Y^j(t) - Y_i)^2 + (Z^j(t) - Z_i)^2} \\ &\equiv f(X_i, Y_i, Z_i) \end{aligned} \tag{9.114}$$

shows the unknown point coordinates X_i, Y_i, Z_i in nonlinear form. Assuming approximate values X_{i0}, Y_{i0}, Z_{i0} for the unknowns, an approximate distance $\varrho_{i0}^j(t)$ can be calculated by

$$\begin{aligned} \varrho_{i0}^j(t) &= \sqrt{(X^j(t) - X_{i0})^2 + (Y^j(t) - Y_{i0})^2 + (Z^j(t) - Z_{i0})^2} \\ &\equiv f(X_{i0}, Y_{i0}, Z_{i0})\,. \end{aligned} \tag{9.115}$$

Using approximate values, the unknowns X_i, Y_i, Z_i can be decomposed by

$$\begin{aligned} X_i &= X_{i0} + \Delta X_i \\ Y_i &= Y_{i0} + \Delta Y_i \\ Z_i &= Z_{i0} + \Delta Z_i \end{aligned} \tag{9.116}$$

where now ΔX_i, ΔY_i, ΔZ_i are new unknowns. This means that the original unknowns have been split into a known part (represented by the approximate values X_{i0}, Y_{i0}, Z_{i0}) and an unknown part (represented by ΔX_i, ΔY_i, ΔZ_i). The advantage of this splitting-up is that the function $f(X_i, Y_i, Z_i)$ is replaced by an equivalent function $f(X_{i0} + \Delta X_i, Y_{i0} + \Delta Y_i, Z_{i0} + \Delta Z_i)$ which can now be expanded into a Taylor series with respect to the approximate point. This leads to

$$\begin{aligned} f(X_i, Y_i, Z_i) &\equiv f(X_{i0} + \Delta X_i, Y_{i0} + \Delta Y_i, Z_{i0} + \Delta Z_i) \\ &= f(X_{i0}, Y_{i0}, Z_{i0}) + \frac{\partial f(X_{i0}, Y_{i0}, Z_{i0})}{\partial X_{i0}} \Delta X_i \\ &+ \frac{\partial f(X_{i0}, Y_{i0}, Z_{i0})}{\partial Y_{i0}} \Delta Y_i + \frac{\partial f(X_{i0}, Y_{i0}, Z_{i0})}{\partial Z_{i0}} \Delta Z_i + \ldots \end{aligned} \tag{9.117}$$

where the expansion is truncated after the linear term; otherwise, the unknowns ΔX_i, ΔY_i, ΔZ_i would appear in nonlinear form. The partial derivatives are obtained from (9.115) by

$$\begin{aligned} \frac{\partial f(X_{i0}, Y_{i0}, Z_{i0})}{\partial X_{i0}} &= -\frac{X^j(t) - X_{i0}}{\varrho^j_{i0}(t)} \\ \frac{\partial f(X_{i0}, Y_{i0}, Z_{i0})}{\partial Y_{i0}} &= -\frac{Y^j(t) - Y_{i0}}{\varrho^j_{i0}(t)} \\ \frac{\partial f(X_{i0}, Y_{i0}, Z_{i0})}{\partial Z_{i0}} &= -\frac{Z^j(t) - Z_{i0}}{\varrho^j_{i0}(t)} \end{aligned} \tag{9.118}$$

and are the components of the unit vector pointing from the satellite towards the approximate site. The substitution of Eqs. (9.115) and (9.118) into Eq. (9.117) gives

$$\begin{aligned} \varrho^j_i(t) = \varrho^j_{i0}(t) &- \frac{X^j(t) - X_{i0}}{\varrho^j_{i0}(t)} \Delta X_i - \frac{Y^j(t) - Y_{i0}}{\varrho^j_{i0}(t)} \Delta Y_i \\ &- \frac{Z^j(t) - Z_{i0}}{\varrho^j_{i0}(t)} \Delta Z_i \end{aligned} \tag{9.119}$$

where the equivalence of $f(X_i, Y_i, Z_i)$ with $\varrho^j_i(t)$ has been used. This equation is now linear with respect to the unknowns ΔX_i, ΔY_i, ΔZ_i.

9.4.2 Linear model for point positioning with code ranges

The model is given only in its elementary form and, thus, apart from the geometry, only the clocks are modeled. The ionosphere, troposphere, and

other minor effects are neglected. According to Eq. (8.5), the model for point positioning with code ranges is given by

$$R_i^j(t) = \varrho_i^j(t) + c\,\delta_i(t) - c\,\delta^j(t) \tag{9.120}$$

which can be linearized by substituting (9.119):

$$\begin{aligned} R_i^j(t) = \varrho_{i0}^j(t) &- \frac{X^j(t) - X_{i0}}{\varrho_{i0}^j(t)}\,\Delta X_i - \frac{Y^j(t) - Y_{i0}}{\varrho_{i0}^j(t)}\,\Delta Y_i \\ &- \frac{Z^j(t) - Z_{i0}}{\varrho_{i0}^j(t)}\,\Delta Z_i + c\,\delta_i(t) - c\,\delta^j(t)\,. \end{aligned} \tag{9.121}$$

Leaving the terms containing unknowns on the right side, the equation above is rewritten as

$$\begin{aligned} R_i^j(t) - \varrho_{i0}^j(t) + c\,\delta^j(t) &= -\frac{X^j(t) - X_{i0}}{\varrho_{i0}^j(t)}\,\Delta X_i \\ -\frac{Y^j(t) - Y_{i0}}{\varrho_{i0}^j(t)}\,\Delta Y_i &- \frac{Z^j(t) - Z_{i0}}{\varrho_{i0}^j(t)}\,\Delta Z_i + c\,\delta_i(t) \end{aligned} \tag{9.122}$$

where the satellite clock bias is assumed to be known. This assumption makes sense because satellite clock correctors can be received from the navigation message. Model (9.122) comprises (for the epoch t) four unknowns, namely ΔX_i, ΔY_i, ΔZ_i, $\delta_i(t)$. Consequently, four satellites are needed to solve the problem. The shorthand notations

$$\begin{aligned} \ell^j &= R_i^j(t) - \varrho_{i0}^j(t) + c\,\delta^j(t) \\ a_{X_i}^j &= -\frac{X^j(t) - X_{i0}}{\varrho_{i0}^j(t)} \\ a_{Y_i}^j &= -\frac{Y^j(t) - Y_{i0}}{\varrho_{i0}^j(t)} \\ a_{Z_i}^j &= -\frac{Z^j(t) - Z_{i0}}{\varrho_{i0}^j(t)} \end{aligned} \tag{9.123}$$

help to simplify the representation of the system of equations. Assuming now four satellites numbered from 1 to 4, then

$$\begin{aligned} \ell^1 &= a_{X_i}^1\,\Delta X_i + a_{Y_i}^1\,\Delta Y_i + a_{Z_i}^1\,\Delta Z_i + c\,\delta_i(t) \\ \ell^2 &= a_{X_i}^2\,\Delta X_i + a_{Y_i}^2\,\Delta Y_i + a_{Z_i}^2\,\Delta Z_i + c\,\delta_i(t) \\ \ell^3 &= a_{X_i}^3\,\Delta X_i + a_{Y_i}^3\,\Delta Y_i + a_{Z_i}^3\,\Delta Z_i + c\,\delta_i(t) \\ \ell^4 &= a_{X_i}^4\,\Delta X_i + a_{Y_i}^4\,\Delta Y_i + a_{Z_i}^4\,\Delta Z_i + c\,\delta_i(t) \end{aligned} \tag{9.124}$$

is the appropriate system of equations. Note that the superscripts are the satellite numbers and not exponents! Introducing

$$\underline{A} = \begin{bmatrix} a^1_{X_i} & a^1_{Y_i} & a^1_{Z_i} & c \\ a^2_{X_i} & a^2_{Y_i} & a^2_{Z_i} & c \\ a^3_{X_i} & a^3_{Y_i} & a^3_{Z_i} & c \\ a^4_{X_i} & a^4_{Y_i} & a^4_{Z_i} & c \end{bmatrix} \quad \underline{x} = \begin{bmatrix} \Delta X_i \\ \Delta Y_i \\ \Delta Z_i \\ \delta_i(t) \end{bmatrix} \quad \underline{\ell} = \begin{bmatrix} \ell^1 \\ \ell^2 \\ \ell^3 \\ \ell^4 \end{bmatrix}, \tag{9.125}$$

the set of linear equations can be written in the matrix–vector form

$$\underline{\ell} = \underline{A}\,\underline{x}\,. \tag{9.126}$$

For this first example of a linearized GPS model, the resubstitution of the vector $\underline{\ell}$ and the matrix $\underline{A}$ using (9.123) is given explicitly for one epoch t:

$$\underline{\ell} = \begin{bmatrix} R^1_i(t) - \varrho^1_{i0}(t) + c\,\delta^1(t) \\ R^2_i(t) - \varrho^2_{i0}(t) + c\,\delta^2(t) \\ R^3_i(t) - \varrho^3_{i0}(t) + c\,\delta^3(t) \\ R^4_i(t) - \varrho^4_{i0}(t) + c\,\delta^4(t) \end{bmatrix}$$

$$\underline{A} = \begin{bmatrix} -\dfrac{X^1(t) - X_{i0}}{\varrho^1_{i0}(t)} & -\dfrac{Y^1(t) - Y_{i0}}{\varrho^1_{i0}(t)} & -\dfrac{Z^1(t) - Z_{i0}}{\varrho^1_{i0}(t)} & c \\ -\dfrac{X^2(t) - X_{i0}}{\varrho^2_{i0}(t)} & -\dfrac{Y^2(t) - Y_{i0}}{\varrho^2_{i0}(t)} & -\dfrac{Z^2(t) - Z_{i0}}{\varrho^2_{i0}(t)} & c \\ -\dfrac{X^3(t) - X_{i0}}{\varrho^3_{i0}(t)} & -\dfrac{Y^3(t) - Y_{i0}}{\varrho^3_{i0}(t)} & -\dfrac{Z^3(t) - Z_{i0}}{\varrho^3_{i0}(t)} & c \\ -\dfrac{X^4(t) - X_{i0}}{\varrho^4_{i0}(t)} & -\dfrac{Y^4(t) - Y_{i0}}{\varrho^4_{i0}(t)} & -\dfrac{Z^4(t) - Z_{i0}}{\varrho^4_{i0}(t)} & c \end{bmatrix}. \tag{9.127}$$

From the linear system (9.126), the coordinate differences ΔX_i, ΔY_i, ΔZ_i and the receiver clock error $\delta_i(t)$ for epoch t result. The desired point coordinates are finally obtained by (9.116).

Recall that the selection of the approximate values for the coordinates was completely arbitrary; they could even be set equal to zero (but this might require an iteration).

Point positioning with code ranges is applicable for each epoch separately. Therefore, this model may also be used in kinematic applications.

9.4.3 Linear model for point positioning with carrier phases

The procedure is the same as in the previous section. In Eq. (8.10), the linearization is performed for $\varrho_i^j(t)$, and known terms are shifted to the left side. Multiplying the equation by λ and using $c = \lambda f$, the result is

$$\begin{aligned}\lambda \Phi_i^j(t) - \varrho_{i0}^j(t) + c\,\delta^j(t) = &-\frac{X^j(t) - X_{i0}}{\varrho_{i0}^j(t)}\,\Delta X_i \\ -\frac{Y^j(t) - Y_{i0}}{\varrho_{i0}^j(t)}\,\Delta Y_i &- \frac{Z^j(t) - Z_{i0}}{\varrho_{i0}^j(t)}\,\Delta Z_i + \lambda N_i^j + c\,\delta_i(t)\end{aligned} \tag{9.128}$$

where, compared to point positioning with code ranges, the number of unknowns is now increased by the ambiguities. Considering again four satellites, the system is given in matrix–vector form $\underline{\ell} = \underline{A}\,\underline{x}$ where

$$\underline{\ell} = \begin{bmatrix} \lambda \Phi_i^1(t) - \varrho_{i0}^1(t) + c\,\delta^1(t) \\ \lambda \Phi_i^2(t) - \varrho_{i0}^2(t) + c\,\delta^2(t) \\ \lambda \Phi_i^3(t) - \varrho_{i0}^3(t) + c\,\delta^3(t) \\ \lambda \Phi_i^4(t) - \varrho_{i0}^4(t) + c\,\delta^4(t) \end{bmatrix}$$

$$\underline{A} = \begin{bmatrix} a_{X_i}^1(t) & a_{Y_i}^1(t) & a_{Z_i}^1(t) & \lambda & 0 & 0 & 0 & c \\ a_{X_i}^2(t) & a_{Y_i}^2(t) & a_{Z_i}^2(t) & 0 & \lambda & 0 & 0 & c \\ a_{X_i}^3(t) & a_{Y_i}^3(t) & a_{Z_i}^3(t) & 0 & 0 & \lambda & 0 & c \\ a_{X_i}^4(t) & a_{Y_i}^4(t) & a_{Z_i}^4(t) & 0 & 0 & 0 & \lambda & c \end{bmatrix} \tag{9.129}$$

$$\underline{x}^T = \begin{bmatrix} \Delta X_i & \Delta Y_i & \Delta Z_i & N_i^1 & N_i^2 & N_i^3 & N_i^4 & \delta_i(t) \end{bmatrix}$$

and where the coefficients of the coordinate increments, cf. (9.123), are supplemented with the time parameter t. Obviously, the four equations do not solve for the eight unknowns. This reflects the fact that point positioning with phases in this form cannot be solved epoch by epoch. Each additional epoch increases the number of unknowns by a new clock term. Thus, for two epochs there are eight equations and nine unknowns (still an underdetermined problem). For three epochs there are 12 equations and 10 unknowns, thus, a slightly overdetermined problem. The 10 unknowns in the latter example are the coordinate increments ΔX_i, ΔY_i, ΔZ_i for the unknown point, the integer ambiguities N_i^1, N_i^2, N_i^3, N_i^4 for the four satellites, and the receiver clock biases $\delta_i(t_1)$, $\delta_i(t_2)$, $\delta_i(t_3)$ for the three epochs.

The matrix–vector scheme for the last example is given by

$$\underline{\ell} = \begin{bmatrix} \lambda\,\Phi_i^1(t_1) - \varrho_{i0}^1(t_1) + c\,\delta^1(t_1) \\ \lambda\,\Phi_i^2(t_1) - \varrho_{i0}^2(t_1) + c\,\delta^2(t_1) \\ \lambda\,\Phi_i^3(t_1) - \varrho_{i0}^3(t_1) + c\,\delta^3(t_1) \\ \lambda\,\Phi_i^4(t_1) - \varrho_{i0}^4(t_1) + c\,\delta^4(t_1) \\ \lambda\,\Phi_i^1(t_2) - \varrho_{i0}^1(t_2) + c\,\delta^1(t_2) \\ \lambda\,\Phi_i^2(t_2) - \varrho_{i0}^2(t_2) + c\,\delta^2(t_2) \\ \lambda\,\Phi_i^3(t_2) - \varrho_{i0}^3(t_2) + c\,\delta^3(t_2) \\ \lambda\,\Phi_i^4(t_2) - \varrho_{i0}^4(t_2) + c\,\delta^4(t_2) \\ \lambda\,\Phi_i^1(t_3) - \varrho_{i0}^1(t_3) + c\,\delta^1(t_3) \\ \lambda\,\Phi_i^2(t_3) - \varrho_{i0}^2(t_3) + c\,\delta^2(t_3) \\ \lambda\,\Phi_i^3(t_3) - \varrho_{i0}^3(t_3) + c\,\delta^3(t_3) \\ \lambda\,\Phi_i^4(t_3) - \varrho_{i0}^4(t_3) + c\,\delta^4(t_3) \end{bmatrix} \qquad \underline{x} = \begin{bmatrix} \Delta X_i \\ \Delta Y_i \\ \Delta Z_i \\ N_i^1 \\ N_i^2 \\ N_i^3 \\ N_i^4 \\ \delta_i(t_1) \\ \delta_i(t_2) \\ \delta_i(t_3) \end{bmatrix}$$

$$\underline{A} = \begin{bmatrix} a_{X_i}^1(t_1) & a_{Y_i}^1(t_1) & a_{Z_i}^1(t_1) & \lambda & 0 & 0 & 0 & c & 0 & 0 \\ a_{X_i}^2(t_1) & a_{Y_i}^2(t_1) & a_{Z_i}^2(t_1) & 0 & \lambda & 0 & 0 & c & 0 & 0 \\ a_{X_i}^3(t_1) & a_{Y_i}^3(t_1) & a_{Z_i}^3(t_1) & 0 & 0 & \lambda & 0 & c & 0 & 0 \\ a_{X_i}^4(t_1) & a_{Y_i}^4(t_1) & a_{Z_i}^4(t_1) & 0 & 0 & 0 & \lambda & c & 0 & 0 \\ a_{X_i}^1(t_2) & a_{Y_i}^1(t_2) & a_{Z_i}^1(t_2) & \lambda & 0 & 0 & 0 & 0 & c & 0 \\ a_{X_i}^2(t_2) & a_{Y_i}^2(t_2) & a_{Z_i}^2(t_2) & 0 & \lambda & 0 & 0 & 0 & c & 0 \\ a_{X_i}^3(t_2) & a_{Y_i}^3(t_2) & a_{Z_i}^3(t_2) & 0 & 0 & \lambda & 0 & 0 & c & 0 \\ a_{X_i}^4(t_2) & a_{Y_i}^4(t_2) & a_{Z_i}^4(t_2) & 0 & 0 & 0 & \lambda & 0 & c & 0 \\ a_{X_i}^1(t_3) & a_{Y_i}^1(t_3) & a_{Z_i}^1(t_3) & \lambda & 0 & 0 & 0 & 0 & 0 & c \\ a_{X_i}^2(t_3) & a_{Y_i}^2(t_3) & a_{Z_i}^2(t_3) & 0 & \lambda & 0 & 0 & 0 & 0 & c \\ a_{X_i}^3(t_3) & a_{Y_i}^3(t_3) & a_{Z_i}^3(t_3) & 0 & 0 & \lambda & 0 & 0 & 0 & c \\ a_{X_i}^4(t_3) & a_{Y_i}^4(t_3) & a_{Z_i}^4(t_3) & 0 & 0 & 0 & \lambda & 0 & 0 & c \end{bmatrix} \tag{9.130}$$

where the different time epochs have been inserted accordingly. The solution of this redundant system is performed by least squares adjustment.

9.4.4 Linear model for relative positioning

The previous sections have shown linear models for both code ranges and carrier phases. For the case of relative positioning, the investigation is restricted to carrier phases, since it should be obvious how to change from the more expanded model of phases to a code model. Furthermore, the linearization and setup of the linear equation system remains, in principle, the same for phases and phase combinations and could be performed analogously for each model. Therefore, the double-difference is selected for treatment in detail. The model for the double-difference of Eq. (8.37), multiplied by λ, is

$$\lambda \Phi_{AB}^{jk}(t) = \varrho_{AB}^{jk}(t) + \lambda N_{AB}^{jk} \tag{9.131}$$

where the term ϱ_{AB}^{jk}, containing the geometry, is composed as

$$\varrho_{AB}^{jk}(t) = \varrho_B^k(t) - \varrho_B^j(t) - \varrho_A^k(t) + \varrho_A^j(t) \tag{9.132}$$

which reflects the fact of four measurement quantities for a double-difference. Each of the four terms must be linearized according to (9.119) yielding

$$\begin{aligned}
\varrho_{AB}^{jk}(t) \;=\; & \varrho_{B0}^k(t) - \frac{X^k(t) - X_{B0}}{\varrho_{B0}^k(t)} \Delta X_B - \frac{Y^k(t) - Y_{B0}}{\varrho_{B0}^k(t)} \Delta Y_B \\
& - \frac{Z^k(t) - Z_{B0}}{\varrho_{B0}^k(t)} \Delta Z_B \\
& - \varrho_{B0}^j(t) + \frac{X^j(t) - X_{B0}}{\varrho_{B0}^j(t)} \Delta X_B + \frac{Y^j(t) - Y_{B0}}{\varrho_{B0}^j(t)} \Delta Y_B \\
& + \frac{Z^j(t) - Z_{B0}}{\varrho_{B0}^j(t)} \Delta Z_B \\
& - \varrho_{A0}^k(t) + \frac{X^k(t) - X_{A0}}{\varrho_{A0}^k(t)} \Delta X_A + \frac{Y^k(t) - Y_{A0}}{\varrho_{A0}^k(t)} \Delta Y_A \\
& + \frac{Z^k(t) - Z_{A0}}{\varrho_{A0}^k(t)} \Delta Z_A \\
& + \varrho_{A0}^j(t) - \frac{X^j(t) - X_{A0}}{\varrho_{A0}^j(t)} \Delta X_A - \frac{Y^j(t) - Y_{A0}}{\varrho_{A0}^j(t)} \Delta Y_A \\
& - \frac{Z^j(t) - Z_{A0}}{\varrho_{A0}^j(t)} \Delta Z_A \,.
\end{aligned} \tag{9.133}$$

Substituting (9.133) into (9.131) and rearranging leads to the linear observation equation

$$\begin{aligned}\ell_{AB}^{jk}(t) = \; & a_{X_A}^{jk}(t)\,\Delta X_A + a_{Y_A}^{jk}(t)\,\Delta Y_A + a_{Z_A}^{jk}(t)\,\Delta Z_A \\ & + a_{X_B}^{jk}(t)\,\Delta X_B + a_{Y_B}^{jk}(t)\,\Delta Y_B + a_{Z_B}^{jk}(t)\,\Delta Z_B + \lambda\, N_{AB}^{jk}\end{aligned} \tag{9.134}$$

where the left side

$$\ell_{AB}^{jk}(t) = \lambda\,\Phi_{AB}^{jk}(t) - \varrho_{B0}^{k}(t) + \varrho_{B0}^{j}(t) + \varrho_{A0}^{k}(t) - \varrho_{A0}^{j}(t) \tag{9.135}$$

comprises both the measurement quantities and all terms computed from the approximate values. On the right side of (9.134), the abbreviations

$$\begin{aligned}
a_{X_A}^{jk}(t) &= +\frac{X^k(t) - X_{A0}}{\varrho_{A0}^{k}(t)} - \frac{X^j(t) - X_{A0}}{\varrho_{A0}^{j}(t)} \\
a_{Y_A}^{jk}(t) &= +\frac{Y^k(t) - Y_{A0}}{\varrho_{A0}^{k}(t)} - \frac{Y^j(t) - Y_{A0}}{\varrho_{A0}^{j}(t)} \\
a_{Z_A}^{jk}(t) &= +\frac{Z^k(t) - Z_{A0}}{\varrho_{A0}^{k}(t)} - \frac{Z^j(t) - Z_{A0}}{\varrho_{A0}^{j}(t)} \\
a_{X_B}^{jk}(t) &= -\frac{X^k(t) - X_{B0}}{\varrho_{B0}^{k}(t)} + \frac{X^j(t) - X_{B0}}{\varrho_{B0}^{j}(t)} \\
a_{Y_B}^{jk}(t) &= -\frac{Y^k(t) - Y_{B0}}{\varrho_{B0}^{k}(t)} + \frac{Y^j(t) - Y_{B0}}{\varrho_{B0}^{j}(t)} \\
a_{Z_B}^{jk}(t) &= -\frac{Z^k(t) - Z_{B0}}{\varrho_{B0}^{k}(t)} + \frac{Z^j(t) - Z_{B0}}{\varrho_{B0}^{j}(t)}
\end{aligned} \tag{9.136}$$

have been used. Note that the coordinates of one point (e.g., A) must be known for relative positioning. More specifically, the known point A reduces the number of unknowns by three because of

$$\Delta X_A = \Delta Y_A = \Delta Z_A = 0 \tag{9.137}$$

and leads to a slight change in the left side term

$$\ell_{AB}^{jk}(t) = \lambda\,\Phi_{AB}^{jk}(t) - \varrho_{B0}^{k}(t) + \varrho_{B0}^{j}(t) + \varrho_{A}^{k}(t) - \varrho_{A}^{j}(t)\,. \tag{9.138}$$

Assuming now four satellites j, k, l, m and two epochs t_1, t_2, the matrix–vector system

$$\underline{\ell} = \begin{bmatrix} \ell_{AB}^{jk}(t_1) \\ \ell_{AB}^{jl}(t_1) \\ \ell_{AB}^{jm}(t_1) \\ \ell_{AB}^{jk}(t_2) \\ \ell_{AB}^{jl}(t_2) \\ \ell_{AB}^{jm}(t_2) \end{bmatrix} \qquad \underline{x} = \begin{bmatrix} \Delta X_B \\ \Delta Y_B \\ \Delta Z_B \\ N_{AB}^{jk} \\ N_{AB}^{jl} \\ N_{AB}^{jm} \end{bmatrix}$$

$$\underline{A} = \begin{bmatrix} a_{X_B}^{jk}(t_1) & a_{Y_B}^{jk}(t_1) & a_{Z_B}^{jk}(t_1) & \lambda & 0 & 0 \\ a_{X_B}^{jl}(t_1) & a_{Y_B}^{jl}(t_1) & a_{Z_B}^{jl}(t_1) & 0 & \lambda & 0 \\ a_{X_B}^{jm}(t_1) & a_{Y_B}^{jm}(t_1) & a_{Z_B}^{jm}(t_1) & 0 & 0 & \lambda \\ a_{X_B}^{jk}(t_2) & a_{Y_B}^{jk}(t_2) & a_{Z_B}^{jk}(t_2) & \lambda & 0 & 0 \\ a_{X_B}^{jl}(t_2) & a_{Y_B}^{jl}(t_2) & a_{Z_B}^{jl}(t_2) & 0 & \lambda & 0 \\ a_{X_B}^{jm}(t_2) & a_{Y_B}^{jm}(t_2) & a_{Z_B}^{jm}(t_2) & 0 & 0 & \lambda \end{bmatrix} \tag{9.139}$$

is obtained which represents a determined and, thus, solvable system. Note that for one epoch the system has more unknowns than observation equations.

9.5 Network adjustment

9.5.1 Single baseline solution

The previous sections described the linearization of the observation equations. The adjustment itself (i.e., the solution of the system of linear equations) is a purely mathematical task to be solved by the computer. It is not the objective of this GPS text to investigate different solution strategies. Extensive details on this subject can be found in Leick (1995). Only some general remarks are given here.

The adjustment principle $\underline{n}^T \underline{P}\, \underline{n}$ = minimum requires the implementation of the weight matrix $\underline{P}$ for the solution. As shown in Sect. 8.3.2, phases and single-differences are uncorrelated, whereas double-differences

and triple-differences are correlated. The implementation of the double-difference correlation can be easily accomplished. Alternatively, the double-differences can be decorrelated by using a Gram–Schmidt orthogonalization (Remondi 1984). The implementation of the correlation of the triple-differences is more difficult. Furthermore, it is questionable if it is worthwhile to expend the effort to compute a correct triple-difference correlation, since the noise of the triple-differences will always prevent to obtain a refined solution.

In the case of an observed network, the use of the single baseline method usually implies a baseline by baseline computation for all possible combinations. If n_i denotes the number of observing sites, then $n_i\,(n_i-1)/2$ baselines can be calculated. Note that only $n_i - 1$ of them are theoretically independent. The redundant baselines can either be used for misclosure checks or for an additional adjustment of the baseline vectors.

There are other approaches. Among them, one method restricts the selection to $n_i - 1$ vectors. This means that the optimal baseline configuration must be chosen. Another method is a variation of the previously described one and encompasses more than one session. All possible baselines are computed for each session. Finally, the resulting vectors of all sessions are subject to a common adjustment.

The disadvantage of the simple single baseline solution from the theoretical point of view is that it is not correct. This is due to the correlation of the simultaneously observed baselines. By solving baseline by baseline, this correlation is ignored.

9.5.2 Multipoint solution

In contrast to the baseline by baseline solution, the multipoint solution considers at once all points in the network. The key difference compared to the single baseline solution is that in the multipoint approach the correlations between the baselines are taken into account.

The principal correlations have been shown in Sect. 8.3.2. The same theoretical aspects also apply to the extended case of a network. The examples are kept as simple as possible to avoid the burden of lengthy formulas.

Single-difference example for a network
When three points A, B, C, a single satellite j, and a single epoch t are considered, two independent baselines can be defined. Taking A as reference

site, for the two baselines A–B and A–C the two single-differences

$$\begin{aligned} \Phi^j_{AB}(t) &= \Phi^j_B(t) - \Phi^j_A(t) \\ \Phi^j_{AC}(t) &= \Phi^j_C(t) - \Phi^j_A(t) \end{aligned} \tag{9.140}$$

can be set up for the one satellite at epoch t. By introducing the vector $\underline{SD}$ for the single-differences, the vector $\underline{\Phi}$ for the phases, and a matrix $\underline{C}$ as

$$\underline{SD} = \begin{bmatrix} \Phi^j_{AB}(t) \\ \Phi^j_{AC}(t) \end{bmatrix} \quad \underline{C} = \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \quad \underline{\Phi} = \begin{bmatrix} \Phi^j_A(t) \\ \Phi^j_B(t) \\ \Phi^j_C(t) \end{bmatrix}, \tag{9.141}$$

the relation $\underline{SD} = \underline{C}\,\underline{\Phi}$ can be formed. To find the correlation, the covariance propagation law is applied by $\mathrm{cov}(\underline{SD}) = \underline{C}\,\mathrm{cov}(\underline{\Phi})\,\underline{C}^T$ leading to

$$\mathrm{cov}(\underline{SD}) = \sigma^2\,\underline{C}\,\underline{C}^T \tag{9.142}$$

because of $\mathrm{cov}(\underline{\Phi}) = \sigma^2\,\underline{I}$, cf. Eq. (8.47). Substituting matrix $\underline{C}$ from (9.141) and evaluating the matrix operation yields

$$\mathrm{cov}(\underline{SD}) = \sigma^2 \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{9.143}$$

which shows, as is to be expected, a correlation of the single-differences of the two baselines with a common point. Recall that single-differences of a single baseline are uncorrelated as discovered in Sect. 8.3.2.

Double-difference example for a network

Since double-differences are already correlated for a single baseline, a correlation must be expected for the network too. Nevertheless, the subsequent slightly larger example will demonstrate the increasing complexity. Assume again three points A, B, C with A as reference site for the two baselines A–B and A–C. Consider a single epoch t for four satellites j, k, ℓ, m where j is taken as the reference satellite for the double-differences.

There are $(n_i - 1)(n_j - 1)$ independent double-differences for n_i points and n_j satellites. For the given example, $n_i = 3$ and $n_j = 4$ and, thus, 6

double-differences can be written. Based on Eq. (8.25), these are

$$\begin{aligned}
\Phi_{AB}^{jk}(t) &= \Phi_B^k(t) - \Phi_B^j(t) - \Phi_A^k(t) + \Phi_A^j(t) \\
\Phi_{AB}^{j\ell}(t) &= \Phi_B^\ell(t) - \Phi_B^j(t) - \Phi_A^\ell(t) + \Phi_A^j(t) \\
\Phi_{AB}^{jm}(t) &= \Phi_B^m(t) - \Phi_B^j(t) - \Phi_A^m(t) + \Phi_A^j(t) \\
\Phi_{AC}^{jk}(t) &= \Phi_C^k(t) - \Phi_C^j(t) - \Phi_A^k(t) + \Phi_A^j(t) \\
\Phi_{AC}^{j\ell}(t) &= \Phi_C^\ell(t) - \Phi_C^j(t) - \Phi_A^\ell(t) + \Phi_A^j(t) \\
\Phi_{AC}^{jm}(t) &= \Phi_C^m(t) - \Phi_C^j(t) - \Phi_A^m(t) + \Phi_A^j(t)
\end{aligned} \tag{9.144}$$

for the assumptions made. As in the previous example, a matrix–vector relation is desired. By introducing

$$\underline{C} = \begin{bmatrix}
1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\
1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\
1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1
\end{bmatrix}$$

$$\underline{DD} = \begin{bmatrix}
\Phi_{AB}^{jk}(t) \\
\Phi_{AB}^{j\ell}(t) \\
\Phi_{AB}^{jm}(t) \\
\Phi_{AC}^{jk}(t) \\
\Phi_{AC}^{j\ell}(t) \\
\Phi_{AC}^{jm}(t)
\end{bmatrix} \qquad \underline{\Phi} = \begin{bmatrix}
\Phi_A^j(t) \\
\Phi_A^k(t) \\
\Phi_A^\ell(t) \\
\Phi_A^m(t) \\
\Phi_B^j(t) \\
\Phi_B^k(t) \\
\Phi_B^\ell(t) \\
\Phi_B^m(t) \\
\Phi_C^j(t) \\
\Phi_C^k(t) \\
\Phi_C^\ell(t) \\
\Phi_C^m(t)
\end{bmatrix}, \tag{9.145}$$

the relation

$$\underline{DD} = \underline{C}\,\underline{\Phi} \tag{9.146}$$

is valid. The covariance follows by

$$\mathrm{cov}(\underline{DD}) = \underline{C}\,\mathrm{cov}(\underline{\Phi})\,\underline{C}^T \tag{9.147}$$

which reduces to

$$\mathrm{cov}(\underline{DD}) = \sigma^2\,\underline{C}\,\underline{C}^T \tag{9.148}$$

because of the uncorrelated phases. Explicitly, the matrix product

$$\underline{C}\,\underline{C}^T = \begin{bmatrix} 4 & 2 & 2 & 2 & 1 & 1 \\ 2 & 4 & 2 & 1 & 2 & 1 \\ 2 & 2 & 4 & 1 & 1 & 2 \\ 2 & 1 & 1 & 4 & 2 & 2 \\ 1 & 2 & 1 & 2 & 4 & 2 \\ 1 & 1 & 2 & 2 & 2 & 4 \end{bmatrix} \tag{9.149}$$

is a full matrix as expected. Finally, the weight matrix

$$\underline{P} = \frac{1}{12}\begin{bmatrix} 6 & -2 & -2 & -3 & 1 & 1 \\ -2 & 6 & -2 & 1 & -3 & 1 \\ -2 & -2 & 6 & 1 & 1 & -3 \\ -3 & 1 & 1 & 6 & -2 & -2 \\ 1 & -3 & 1 & -2 & 6 & -2 \\ 1 & 1 & -3 & -2 & -2 & 6 \end{bmatrix} \tag{9.150}$$

is the inverse of (9.149). Beutler et al. (1986) give detailed instructions for a computer implementation and also show some results of network campaigns where the correlations have either been totally neglected, introduced in a single baseline mode, or calculated correctly (Beutler et al. 1987). For small networks with baselines not exceeding 10 km, the differences of the three methods are in the range of a few millimeters. Clearly, the solution without any correlation deviates from the theoretically correct values by a greater amount. It is estimated that the single baseline method deviates from the multibaseline (correlated) solution by a maximum of 2σ.

9.5.3 Single baseline versus multipoint solution

The previous paragraphs have shown the difference between the single baseline and the multipoint solution from the theoretical point of view and mentioned a few results from GPS campaigns. This section presents some arguments for using one or the other method.

- The correlation is not modeled correctly with the single baseline solution because correlations between baselines are neglected.
- The computer program is, without doubt, much simpler for the single baseline approach.
- With modern software and hardware, the computational time is not a real problem.
- Cycle slips are more easily detected and repaired in the multipoint mode.
- It takes less effort in the single baseline mode to isolate bad measurements and possibly to eliminate them.
- The economic implementation of the full correlation for a multipoint solution only works properly for networks with the same observation pattern at each receiver site. In the event of numerous data outages it is better to recalculate the covariance matrix.
- Even in the case of the multipoint approach, it becomes questionable whether the correlations can be modeled properly. An illustrative example is given in Beutler et al. (1990) where single and dual frequency receivers are combined in a network. For the dual frequency receivers, the ionosphere-free combination Lc is formed from L1 and L2 and processed together with the L1 data of the single frequency receivers. Thus, a correlation is introduced because of the L1 data. A proper modeling of the correlation biases the ionosphere-free Lc baseline by the ionosphere of the L1 baseline, an effect which is definitely undesirable.

The factors listed illustrate the difficulty in deciding which method (i.e., single baseline versus multipoint solution) to use.

9.5.4 Least squares adjustment of baselines

Dealing with networks, usually the number of measured baselines will exceed the minimum amount. In this case, redundant information is available and the determination of the coordinates of the network points may be carried out by a least squares adjustment.

According to the brief description of the least squares adjustment in Sect. 9.3.1, observations are required. The baseline vectors $\underline{X}_{ij}$ between the

unknown network points $\underline{X}_j$ and $\underline{X}_i$ may be introduced as observables (in the sense of adjustment theory). Therefore, the expression

$$\underline{X}_{ij} = \underline{X}_j - \underline{X}_i \tag{9.151}$$

referring the observables to the unknowns may be formulated. Since the unknowns appear in linear form, there is no linearization required. In case of redundance, noise (or residuals) $\underline{n}_{ij}$ must be added to the observables to assure consistency, cf. (9.68). Therefore,

$$\underline{X}_{ij} + \underline{n}_{ij} = \underline{X}_j - \underline{X}_i \tag{9.152}$$

is obtained. Making the residuals explicit, the vector equation

$$\underline{n}_{ij} = \underline{X}_j - \underline{X}_i - \underline{X}_{ij} \tag{9.153}$$

is obtained which is composed of the three scalar equations

$$\begin{aligned} n_{X_{ij}} &= X_j - X_i - X_{ij} \\ n_{Y_{ij}} &= Y_j - Y_i - Y_{ij} \\ n_{Z_{ij}} &= Z_j - Z_i - Z_{ij} \,. \end{aligned} \tag{9.154}$$

The components X_{ij}, Y_{ij}, Z_{ij} of the baseline vector are regarded as observables. The elements of the design matrix are the coefficients of the unknowns $\underline{X}_i$ and $\underline{X}_j$ and amount to 0, +1 or −1. The relation (9.153) is solved by applying the least squares principle $\underline{n}^T\underline{n}$ = minimum if equal weights for the baselines are assumed. Otherwise, a weight matrix $\underline{P}$ must be taken into account according to (9.67).

Introducing the coordinate differences X_{ij}, Y_{ij}, Z_{ij} as the only observables, absolute coordinates cannot be derived because the matrix of normal equations becomes singular. Consider the most general case; the rank deficiency of a three-dimensional network amounts to seven corresponding to the seven degrees of freedom of a three-dimensional network or the seven parameters of a similarity transformation in space (three translations, three rotations, and one scale factor).

For relative positioning, orientation and scale of the network of baseline vectors (and also its shape) are determined due to the definition of the satellite orbits, e.g., in the WGS-84. This means that four of seven parameters are determined: three rotations and the scale. The three translations of the whole network, i.e., a shift vector, are still undetermined so that the rank deficiency of the normal equations equals three. Selecting and fixing a single point of the network solves the problem of the shift vector and leads to the minimally constrained solution. Fixing means that the coordinates of the

selected point are considered known which reduces the number of the network unknowns by three and eliminates the rank deficiency. Note, however, that the fixing of coordinates should be restricted to one point. Otherwise, constraints would be induced into the GPS network which could obviate the strong GPS geometry and result in network distortions.

Previously, the problem of correlations was discussed. For single baseline solutions, only correlations between vector components are computed. For multipoint baseline solutions, correlations between the various baselines are computed. In the first case, the correlations may be disregarded, whereas in the latter case, the correlations should be taken into account. Sophisticated software properly employs correlations.

Despite the linearity of the unknowns in (9.152), sometimes the unknowns $\underline{X}_j$ and $\underline{X}_i$ are split into known approximate values $\underline{X}_{j0}$ and $\underline{X}_{i0}$ and unknown increments $\Delta\underline{X}_j$ and $\Delta\underline{X}_i$ so that (9.152) becomes

$$\underline{X}_{ij} + \underline{n}_{ij} = \underline{X}_{j0} + \Delta\underline{X}_j - (\underline{X}_{i0} + \Delta\underline{X}_i) \tag{9.155}$$

which may be rearranged to

$$\underline{n}_{ij} = \Delta\underline{X}_j - \Delta\underline{X}_i + \underline{X}_{ij0} - \underline{X}_{ij} \tag{9.156}$$

where

$$\underline{X}_{ij0} = \underline{X}_{j0} - \underline{X}_{i0} \tag{9.157}$$

has been introduced. Equation (9.156) corresponds to (9.153). For numerical values of the approximate coordinates as needed in (9.157), a representative navigation solution provided by the receivers may be used.

The adjustment described for three-dimensional Cartesian coordinates may also be performed in ellipsoidal coordinates. The principle of formulating the observation equations is precisely the same; however, the expression which relates the observables to the unknowns is more complicated.

9.6 Dilution of precision

The geometry of the visible satellites is an important factor in achieving high quality results especially for point positioning and kinematic surveying. The geometry changes with time due to the relative motion of the satellites. A measure of the geometry is the Dilution of Precision (DOP) factor.

First, the specific case of four satellites is considered. The linearized observation equations for the point positioning model with code ranges are

given by Eq. (9.126), and the solution for the (four) unknowns follows from the inverse relation $\underline{x} = \underline{A}^{-1}\,\underline{\ell}$. The design matrix $\underline{A}$ is given by Eq. (9.127)

$$\underline{A} = \begin{bmatrix} -\dfrac{X^1(t) - X_{i0}}{\varrho^1_{i0}(t)} & -\dfrac{Y^1(t) - Y_{i0}}{\varrho^1_{i0}(t)} & -\dfrac{Z^1(t) - Z_{i0}}{\varrho^1_{i0}(t)} & c \\ -\dfrac{X^2(t) - X_{i0}}{\varrho^2_{i0}(t)} & -\dfrac{Y^2(t) - Y_{i0}}{\varrho^2_{i0}(t)} & -\dfrac{Z^2(t) - Z_{i0}}{\varrho^2_{i0}(t)} & c \\ -\dfrac{X^3(t) - X_{i0}}{\varrho^3_{i0}(t)} & -\dfrac{Y^3(t) - Y_{i0}}{\varrho^3_{i0}(t)} & -\dfrac{Z^3(t) - Z_{i0}}{\varrho^3_{i0}(t)} & c \\ -\dfrac{X^4(t) - X_{i0}}{\varrho^4_{i0}(t)} & -\dfrac{Y^4(t) - Y_{i0}}{\varrho^4_{i0}(t)} & -\dfrac{Z^4(t) - Z_{i0}}{\varrho^4_{i0}(t)} & c \end{bmatrix} \tag{9.158}$$

where the first three elements in each row are the components of the unit vectors $\underline{\varrho}^j_i$, $j = 1, 2, 3, 4$ pointing from the four satellites to the observing site i. The solution fails if the design matrix is singular or, equivalently, if its determinant becomes zero. The determinant is proportional to the scalar triple product

$$\left((\underline{\varrho}^4_i - \underline{\varrho}^1_i), (\underline{\varrho}^3_i - \underline{\varrho}^1_i), (\underline{\varrho}^2_i - \underline{\varrho}^1_i)\right)$$

which can geometrically be interpreted by the volume of a body. This body is formed by the intersection points of the site–satellite vectors with the unit sphere centered at the observing site. The larger the volume of this body, the better the satellite geometry. Since good geometry should mirror a low DOP value, the reciprocal value of the volume of the geometric body is directly proportional to DOP. The critical configuration is given when the body degenerates to a plane. This is the case when the unit vectors $\underline{\varrho}^j_i$ form a cone with the observing site as apex (Wunderlich 1992).

More generally, DOP can be calculated from the inverse of the normal equation matrix of the solution. The cofactor matrix $\underline{Q}_X$ follows from

$$\underline{Q}_X = (\underline{A}^T \underline{A})^{-1}\,. \tag{9.159}$$

In this case, the weight matrix has been assumed to be a unit matrix. The cofactor matrix $\underline{Q}_X$ is a 4×4 matrix where three components are contributed by the site position X, Y, Z and one component by the receiver clock. Denoting the elements of the cofactor matrix as

$$\underline{Q}_X = \begin{bmatrix} q_{XX} & q_{XY} & q_{XZ} & q_{Xt} \\ q_{XY} & q_{YY} & q_{YZ} & q_{Yt} \\ q_{XZ} & q_{YZ} & q_{ZZ} & q_{Zt} \\ q_{Xt} & q_{Yt} & q_{Zt} & q_{tt} \end{bmatrix}, \tag{9.160}$$

the diagonal elements are used for the following DOP definitions:

$$\begin{aligned} \text{GDOP} &= \sqrt{q_{XX} + q_{YY} + q_{ZZ} + q_{tt}} && \ldots \text{ geometric dilution of precision} \\ \text{PDOP} &= \sqrt{q_{XX} + q_{YY} + q_{ZZ}} && \ldots \text{ position dilution of precision} \\ \text{TDOP} &= \sqrt{q_{tt}} && \ldots \text{ time dilution of precision.} \end{aligned} \tag{9.161}$$

It should be noted that the previous DOP explanation using the geometric body refers to GDOP.

These definitions deserve brief explanations in order to avoid confusion. Quite often the elements under the square root are displayed as quadratic terms. This depends on the designation of the elements of the cofactor matrix. Here, the diagonal elements of the cofactor matrix have been denoted as q_{XX}, q_{YY}, q_{ZZ}, and q_{tt}, thus no superscripts appear in the DOP definitions. If the diagonal elements are denoted as, e.g., q_X^2, q_Y^2, q_Z^2, q_t^2, then of course the superscripts also appear in the DOP definitions. The general rule used here is: when computing DOP values, the elements of the cofactor matrix are not squared (e.g., for GDOP the square root of the trace must be calculated).

The DOPs in (9.161) are expressed in the equatorial system. When the topocentric local coordinate system with its axes along the local north, east, and up (i.e., vertical) is used, the global cofactor matrix $\underline{Q}_X$ must be transformed into the local cofactor matrix $\underline{Q}_x$ by the law of covariance propagation. Denoting now as $\underline{Q}_X$ that part of the cofactor matrix that contains the geometrical components (disregarding the time-correlated components), the transformation is

$$\underline{Q}_x = \underline{R}\,\underline{Q}_X\,\underline{R}^T = \begin{bmatrix} q_{xx} & q_{xy} & q_{xh} \\ q_{xy} & q_{yy} & q_{yh} \\ q_{xh} & q_{yh} & q_{hh} \end{bmatrix} \tag{9.162}$$

where the rotation matrix $\underline{R} = [\underline{n}, \underline{e}, \underline{u}]^T$ contains the axes of the local coordinate system as given in Eq. (7.3).

Because of the invariance of the trace of a matrix with respect to rotation, the PDOP value in the local system is identical to the value in the global system. In addition to the PDOP, two further DOP definitions are

given. HDOP, the dilution of precision in the horizontal position, and VDOP, denoting the corresponding value for the vertical component, the height:

$$\begin{aligned} \text{HDOP} &= \sqrt{q_{xx} + q_{yy}} \\ \text{VDOP} &= \sqrt{q_{hh}} \end{aligned} \tag{9.163}$$

The discussion, so far, involves only single epoch point positioning. When designing a survey, it is helpful to know DOP for the full (observation) session. The procedure is to compute the DOP values on an epoch by epoch basis for the desired period. The time increment between epochs can be matched to the specific purposes of the planned survey. In Fig. 9.6, PDOP is shown for an entire day and was computed for 15-minute increments. No measured data are necessary to calculate DOP values! The satellite positions can be calculated either from almanac data or from an appropriate orbit file. Note that DOP computations are not restricted to point positioning but can also be applied to relative positioning. Starting with the design matrix for a baseline vector determination, the cofactor matrix can be computed. These DOP values can be considered as relative DOP values.

The purposes of DOP are twofold. First, it is useful in planning a survey and, second, it may be helpful in interpreting processed baseline vectors. For example, data with poor DOP could possibly be omitted.

Finally, the correlation of DOP with the positioning accuracy is considered. Denoting the measurement accuracy by σ (i.e., the standard deviation), the positioning accuracy follows from the product of DOP and

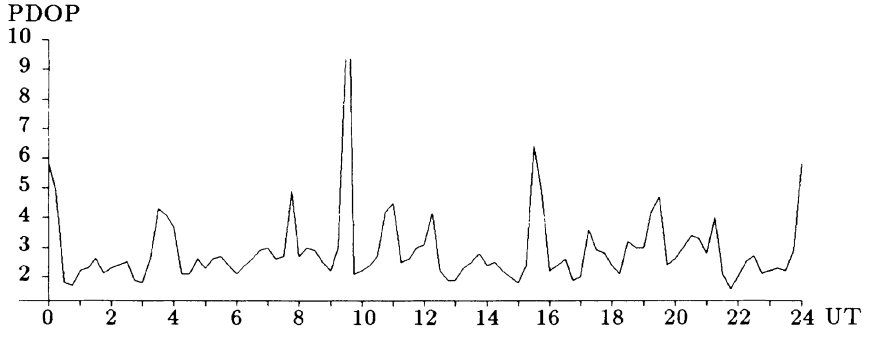


Fig. 9.6. PDOP for Graz, Austria, on September 2, 1995 (cutoff elevation: 15°)

measurement accuracy. Applied to the specific DOP definitions,

GDOP σ	...	geometric accuracy in position and time
PDOP σ	...	accuracy in position
TDOP σ	...	accuracy in time
HDOP σ	...	accuracy in the horizontal position
VDOP σ	...	accuracy in vertical direction

is obtained (Wells et al. 1987). The list of DOP definitions is not restricted to those given here. The meaning of other DOP definitions can be derived from the associated acronyms.

9.7 Accuracy measures

9.7.1 Introduction

Several accuracy measures are in use which are subsequently explained. First, some statistical terms are defined (Kreyszig 1968: Sect. 18.7). Let x be a random (or stochastic) variable and $f(x)$ the density function of its distribution, then the integral

$$P(a < x < b) = \int_a^b f(x)\,dx \tag{9.164}$$

defines the probability that the variable x assumes any value in the interval $a < x < b$. The mean (or expectation) μ of x is given by

$$\mu = \int_{-\infty}^{+\infty} x\, f(x)\,dx \tag{9.165}$$

and the variance σ^2 of x follows from

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2\, f(x)\,dx\,. \tag{9.166}$$

The square root of the variance is denoted as standard deviation σ.

A stochastic variable x with mean μ and variance σ^2 may be converted (or standardized) into the variable $(x-\mu)/\sigma$. This standardized variable has zero mean and unit variance.

The density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2} \tag{9.167}$$

of the most frequently used Gaussian or normal distribution reduces to

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \tag{9.168}$$

for a standardized variable z.

9.7.2 Chi-square distribution

Let x be the sum of the squares of n independent and standardized stochastic variables with Gaussian distribution. The density of the associated Chi-square or briefly χ^2-distribution (Miller 1963: p. 325) is defined by

$$f_n(x) = \frac{1}{2^{n/2}\,\Gamma(n/2)} x^{(n/2-1)}\, e^{-x/2} \tag{9.169}$$

where the number n is denoted as degree of freedom. The Gamma-function $\Gamma(n/2)$ for $n > 0$ is given by

$$\Gamma(n/2) = \int_0^\infty \mu^{(n/2-1)}\, e^{-\mu}\, d\mu \tag{9.170}$$

and has the numerical values $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(2/2) = 1$, $\Gamma(3/2) = \sqrt{\pi}/2$ for the numbers $n = 1, 2, 3$ (Bronstein and Semendjajew 1996: p. 600).

9.7.3 Specifications

One-dimensional accuracy measures
The one-dimensional case is specified by $n = 1$ and $x = z^2$. The χ^2-distribution converts to the Gaussian distribution with zero mean and unit variance which may be proved by transforming the probability $P = \int f_1(x)\, dx$ by substituting $x = z^2$ and $dx = 2z\, dz$. The associated probability follows as

$$P_1(-\alpha < z < \alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\alpha}^{+\alpha} e^{-z^2/2}\, dz\,. \tag{9.171}$$

For nonstandardized variables, α must be multiplied by the standard deviation σ.

Table 9.6. One-dimensional accuracy measures

α	Probability	Notation
0.67	50.0%	Linear Error Probable (LEP)
1.00	68.3%	1σ level
1.96	95.0%	95% confidence level
2.00	95.4%	2σ level
3.00	99.7%	3σ level

Numerical values for the probability P_1 of the normal distribution are extensively tabulated in most textbooks on statistics. Some examples are given in Table 9.6 where the term confidence level is introduced as a synonym for probability.

Two-dimensional accuracy measures
The two-dimensional case is specified by $n = 2$ and $x = z_1^2 + z_2^2$. The density function of the χ^2-distribution becomes $f_2(x) = \frac{1}{2}e^{-x/2}$. The associated probability is given by

$$P_2(0 < x < \alpha) = 1 - e^{-\alpha/2} \tag{9.172}$$

from which the inverse relation $\alpha = \alpha(P_2)$ can be derived.

Numerical values for the probability P_2 are readily obtained from (9.172). Some examples are given in Table 9.7 where the entry $\sqrt{\alpha}$ is chosen based on geometrical reasons.

If z_1 and z_2 are identified with the semiaxes a, b of the standard ellipse (i.e., ellipse of constant probability scaled with 1σ), then x corresponds to the squared mean position error σ_P. This quantity is invariant with respect

Table 9.7. Two-dimensional accuracy measures

$\sqrt{\alpha}$	Probability	Notation
1.00	39.4%	1σ or standard ellipse
1.18	50.0%	Circular Error Probable (CEP)
$\sqrt{2}$	63.2%	Distance RMS (DRMS)
2.00	86.5%	2σ ellipse
2.45	95.0%	95% confidence level
$2\sqrt{2}$	98.2%	2DRMS
3.00	98.9%	3σ ellipse

to rotation and, thus,

$$\sigma_P = \sqrt{a^2 + b^2} = \sqrt{\sigma_x^2 + \sigma_y^2} \tag{9.173}$$

is obtained. The standard ellipse converts to a circle if both semimajor axes have the same length. The radius of the circle scaled with 1.18σ defines the Circular Error Probable (CEP). The radius of the circle scaled with $\sqrt{2}\sigma$ is denoted Distance Root Mean Square (DRMS) or, simply, mean radial error.

Following Mikhail (1976), Sect. 2.5, the Circular Error Probable may be approximated by CEP $\approx 0.59(\sigma_x + \sigma_y)$.

Three-dimensional accuracy measures

The three-dimensional case is specified by $n = 3$ and $x = z_1^2 + z_2^2 + z_3^2$. The density function of the χ^2-distribution becomes more complicated and the associated probability is given by

$$P_3(0 < x < \alpha) = \frac{1}{\sqrt{2\pi}} \int_0^\alpha \sqrt{x}\, e^{-x^2/2}\, dx\,. \tag{9.174}$$

Numerical values for the probability P_3 can be taken from statistical textbooks although tables for the χ^2-distribution with $n = 3$ are rather rare. However, the χ^2-distribution for arbitrary degrees of freedom is standard in most mathematical software packages. Some examples are given in Table 9.8 where again the entry $\sqrt{\alpha}$ is used.

In general, the standard ellipsoid has similar properties as the standard ellipse. The ellipsoid converts to a sphere if the semiaxes of the ellipsoid have same lengths. The radius of the sphere scaled with 1.53σ defines the Spherical Error Probable (SEP). The radius of the sphere scaled with $\sqrt{3}\sigma$ is denoted as Mean Radial Spherical Error (MRSE) which is a single measure for the three-dimensional case. Following Mikhail (1976), Sect. 2.5, the Spherical Error Probable may be approximated by SEP $\approx 0.51(\sigma_x + \sigma_y + \sigma_z)$.

Table 9.8. Three-dimensional accuracy measures

$\sqrt{\alpha}$	Probability	Notation
1.00	19.9%	1σ or standard ellipsoid
1.53	50.0%	Spherical Error Probable (SEP)
$\sqrt{3}$	61.0%	Mean Radial Spherical Error (MRSE)
2.00	73.8%	2σ ellipsoid
2.80	95.0%	95% confidence level
3.00	97.1%	3σ ellipsoid

10 Transformation of GPS results

10.1 Introduction

The (official) reference frame of GPS is the World Geodetic System 1984 (WGS-84). When using GPS, the coordinates of terrestrial sites are obtained in the same reference frame. The surveyor is not, usually, interested in coordinates of the terrestrial points referring to a global frame; rather, the results are preferred in a local coordinate frame either as geodetic (i.e., ellipsoidal) coordinates, as plane coordinates, or as vectors combined with other terrestrial data. Since the WGS-84 is a geocentric system and the local system usually is not, certain transformations are required. The subsequent sections deal with the transformations most frequently used.

10.2 Coordinate transformations

10.2.1 Cartesian coordinates and ellipsoidal coordinates

Denoting the (global) Cartesian coordinates of a point in space by X, Y, Z and assuming an ellipsoid of revolution with the same origin as the Cartesian coordinate system, the point can also be expressed by the ellipsoidal coordinates φ, λ, h (Fig. 10.1). According to Heiskanen and Moritz (1967), p. 182, the relation between the Cartesian coordinates and the ellipsoidal coordinates is given by

$$\begin{aligned} X &= (N+h)\cos\varphi\cos\lambda \\ Y &= (N+h)\cos\varphi\sin\lambda \\ Z &= \left(\frac{b^2}{a^2}N+h\right)\sin\varphi \end{aligned} \tag{10.1}$$

where N is the radius of curvature in prime vertical which is obtained by

$$N = \frac{a^2}{\sqrt{a^2\cos^2\varphi + b^2\sin^2\varphi}}, \tag{10.2}$$

and a, b are the semiaxes of the ellipsoid. Recall that the Cartesian coordinates related to the WGS-84 are also denoted ECEF coordinates and that the origins of the ECEF coordinate system and of the WGS-84 ellipsoid are identical (i.e., geocentric).

B. Hofmann-Wellenhof et al., *Global Positioning System*

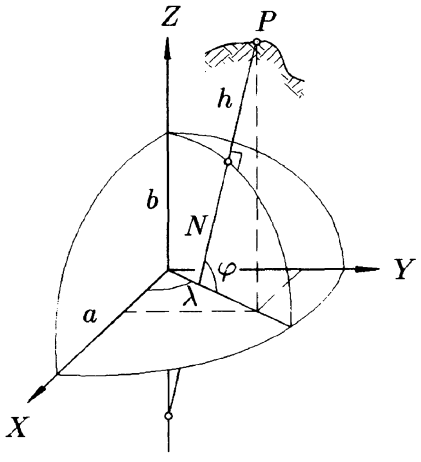


Fig. 10.1. Cartesian coordinates X, Y, Z and ellipsoidal coordinates φ, λ, h

The formulas (10.1) transform ellipsoidal coordinates φ, λ, h into Cartesian coordinates X, Y, Z. For GPS applications, the inverse transformation is more important since the Cartesian coordinates are given and the ellipsoidal coordinates sought. Thus, the task is now to compute the ellipsoidal coordinates φ, λ, h from the Cartesian coordinates X, Y, Z. Usually, this problem is solved iteratively although a solution in closed form is possible. From X and Y, the radius of a parallel,

$$p = \sqrt{X^2 + Y^2} = (N + h)\cos\varphi\,, \tag{10.3}$$

can be computed. This equation is rearranged as

$$h = \frac{p}{\cos\varphi} - N \tag{10.4}$$

so that the ellipsoidal height appears explicitly. Introducing by

$$e^2 = \frac{a^2 - b^2}{a^2} \tag{10.5}$$

the first numerical eccentricity, the relation $b^2/a^2 = 1 - e^2$ follows which can be substituted into the equation for Z in (10.1). The result

$$Z = (N + h - e^2 N)\sin\varphi \tag{10.6}$$

can be written as

$$Z = (N + h)\left(1 - e^2 \frac{N}{N + h}\right)\sin\varphi \tag{10.7}$$

equivalently. Dividing this expression by Eq. (10.3) gives

$$\frac{Z}{p} = \left(1 - e^2 \frac{N}{N+h}\right) \tan\varphi \tag{10.8}$$

which yields

$$\tan\varphi = \frac{Z}{p}\left(1 - e^2 \frac{N}{N+h}\right)^{-1} . \tag{10.9}$$

For the longitude λ, the equation

$$\tan\lambda = \frac{Y}{X} \tag{10.10}$$

is obtained from Eq. (10.1) by dividing the first and the second equation.

The longitude λ can be directly computed from Eq. (10.10). The height h and the latitude φ are determined by Eqs. (10.4) and (10.9). The problem is that both equations depend on the latitude and the height. A solution can be found iteratively by the following steps:

1. Compute $p = \sqrt{X^2 + Y^2}$.

2. Compute an approximate value $\varphi_{(0)}$ from
   $\tan\varphi_{(0)} = \frac{Z}{p}(1 - e^2)^{-1}$.

3. Compute an approximate value $N_{(0)}$ from
   $N_{(0)} = \frac{a^2}{\sqrt{a^2\cos^2\varphi_{(0)} + b^2\sin^2\varphi_{(0)}}}$.

4. Compute the ellipsoidal height by
   $h = \frac{p}{\cos\varphi_{(0)}} - N_{(0)}$.

5. Compute an improved value for the latitude by
   $\tan\varphi = \frac{Z}{p}\left(1 - e^2 \frac{N_{(0)}}{N_{(0)} + h}\right)^{-1}$.

6. Check for another iteration step: if $\varphi = \varphi_{(0)}$ then the iteration is completed otherwise set $\varphi_{(0)} = \varphi$ and continue with step 3.

Many other computation methods have been devised. One example for the transformation of X, Y, Z into φ, λ, h without iteration but with an inherent approximation is

$$\begin{aligned} \varphi &= \arctan \frac{Z + e'^2\, b \sin^3\theta}{p - e^2\, a \cos^3\theta} \\ \lambda &= \arctan \frac{Y}{X} \\ h &= \frac{p}{\cos\varphi} - N \end{aligned} \tag{10.11}$$

where

$$\theta = \arctan \frac{Z\, a}{p\, b} \tag{10.12}$$

is an auxiliary quantity and

$$e'^2 = \frac{a^2 - b^2}{b^2} \tag{10.13}$$

is the second numerical eccentricity. Actually, there is no reason why these formulas are less popular than the iterative procedure since there is no significant difference between the two methods. A computation method with neither iteration nor approximation is, e.g., given by Zhu (1993).

For a numerical example, consider a point with $\varphi = 47°$, $\lambda = 15°$, and $h = 2\,000$ m referring to the WGS-84 ellipsoid. Using the WGS-84 parameters (Sect. 3.2.1) the desired ECEF coordinates are computed from (10.1) as $X = 4\,210\,520.621$ m, $Y = 1\,128\,205.600$ m, $Z = 4\,643\,227.495$ m. It is recommended that the inverse transformation be performed to check the achievable accuracy.

10.2.2 Global coordinates and local level coordinates

The global coordinates are identical with the Cartesian coordinates of the previous section; however, instead of using the components X, Y, Z, the vector notation $\underline{X}$ is preferred. Thus, the vectors $\underline{X}_i$ and $\underline{X}_j$ represent two terrestrial points P_i and P_j. Defining the vector between these two points in the global coordinate system by $\underline{X}_{ij} = \underline{X}_j - \underline{X}_i$, this vector may also be defined in the local level system referenced to the tangent plane at P_i and introducing the notation $\underline{x}_{ij}$.

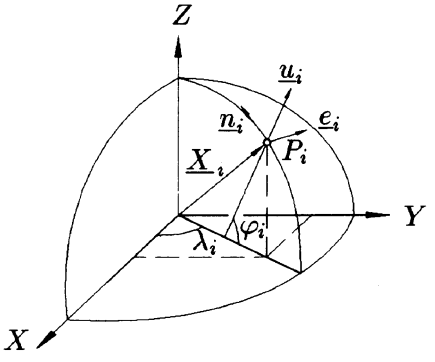


Fig. 10.2. Global and local level coordinates

The axes $\underline{n}_i$, $\underline{e}_i$, $\underline{u}_i$ of the local (tangent plane) coordinate system at P_i corresponding to the north, east, and up direction, are represented in the global system by

$$\underline{n}_i = \begin{bmatrix} -\sin\varphi_i \cos\lambda_i \\ -\sin\varphi_i \sin\lambda_i \\ \cos\varphi_i \end{bmatrix}, \quad \underline{e}_i = \begin{bmatrix} -\sin\lambda_i \\ \cos\lambda_i \\ 0 \end{bmatrix}, \quad \underline{u}_i = \begin{bmatrix} \cos\varphi_i \cos\lambda_i \\ \cos\varphi_i \sin\lambda_i \\ \sin\varphi_i \end{bmatrix} \tag{10.14}$$

where the vectors $\underline{n}_i$ and $\underline{e}_i$ span the tangent plane at P_i (Fig. 10.2). The third coordinate axis of the local level system, i.e., the vector $\underline{u}_i$, is orthogonal to the tangent plane and coincides with the ellipsoidal normal.

The local level system refers to the (natural) plumb line at P_i (more precisely: to the tangent at P_i of the slightly curved plumb line) if the ellipsoidal coordinates φ_i, λ_i in (10.14) are replaced by the astronomical latitude and longitude.

Now the components n_{ij}, e_{ij}, u_{ij} of the vector $\underline{x}_{ij}$ in the local level system are introduced. These coordinates are sometimes denoted as ENU (east, north, up) coordinates. Considering Fig. 10.3, these components are obtained by a projection of vector $\underline{X}_{ij}$ onto the local level axes $\underline{n}_i$, $\underline{e}_i$, $\underline{u}_i$. Analytically, this is achieved by inner products (also denoted as scalar or dot products). Therefore,

$$\underline{x}_{ij} = \begin{bmatrix} n_{ij} \\ e_{ij} \\ u_{ij} \end{bmatrix} = \begin{bmatrix} \underline{n}_i \cdot \underline{X}_{ij} \\ \underline{e}_i \cdot \underline{X}_{ij} \\ \underline{u}_i \cdot \underline{X}_{ij} \end{bmatrix} \tag{10.15}$$

is obtained. Assembling the vectors $\underline{n}_i$, $\underline{e}_i$, $\underline{u}_i$ of the local level system as

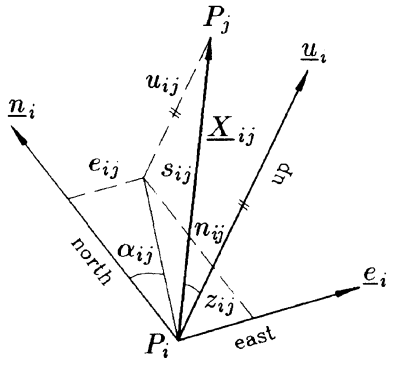


Fig. 10.3. Measurement quantities in the local level system

columns in a matrix $\underline{D}_i$, i.e.,

$$\underline{D}_i = \begin{bmatrix} -\sin\varphi_i \cos\lambda_i & -\sin\lambda_i & \cos\varphi_i \cos\lambda_i \\ -\sin\varphi_i \sin\lambda_i & \cos\lambda_i & \cos\varphi_i \sin\lambda_i \\ \cos\varphi_i & 0 & \sin\varphi_i \end{bmatrix}, \tag{10.16}$$

relation (10.15) may be simplified to

$$\underline{x}_{ij} = \underline{D}_i^T \underline{X}_{ij} . \tag{10.17}$$

The components of $\underline{x}_{ij}$ may also be expressed by the spatial distance s_{ij}, the azimuth α_{ij}, and the zenith angle z_{ij} which is assumed to be corrected for refraction. The appropriate relation is

$$\underline{x}_{ij} = \begin{bmatrix} n_{ij} \\ e_{ij} \\ u_{ij} \end{bmatrix} = \begin{bmatrix} s_{ij} \sin z_{ij} \cos\alpha_{ij} \\ s_{ij} \sin z_{ij} \sin\alpha_{ij} \\ s_{ij} \cos z_{ij} \end{bmatrix} \tag{10.18}$$

where the terrestrial measurement quantities s_{ij}, α_{ij}, z_{ij} refer to P_i, i.e., the measurements were taken at P_i. Inverting (10.18) gives the measurement quantities explicitly:

$$\begin{aligned} s_{ij} &= \sqrt{n_{ij}^2 + e_{ij}^2 + u_{ij}^2} \\ \tan\alpha_{ij} &= \frac{e_{ij}}{n_{ij}} \\ \cos z_{ij} &= \frac{u_{ij}}{\sqrt{n_{ij}^2 + e_{ij}^2 + u_{ij}^2}} . \end{aligned} \tag{10.19}$$

Substituting (10.15) for n_{ij}, e_{ij} and u_{ij}, the measurement quantities may be expressed by the components of the vector $\underline{X}_{ij}$ in the global system.

10.2.3 Ellipsoidal coordinates and plane coordinates

In contrast to the previous section, only points on the ellipsoid are considered. Thus, ellipsoidal latitude φ and longitude λ are of interest here. The objective is to map a point φ, λ on the ellipsoid into a point x, y in a plane. Map projections may be expressed by

$$\begin{aligned} x &= x(\varphi, \lambda; a, b) \\ y &= y(\varphi, \lambda; a, b) \end{aligned} \tag{10.20}$$

in general form. Geodetic applications require conformal mapping. Conformality means that an angle on the ellipsoid is preserved after mapping it into the plane. More precisely, the angle included by two geodesics on the ellipsoid is preserved if the two geodesics are conformally mapped into the plane.

Some of the most important conformal mappings of the ellipsoid onto a plane are subsequently interpreted geometrically, although they are defined analytically. Detailed formulas for many conformal mappings can be found, e.g., in Richardus and Adler (1972) or in Hofmann-Wellenhof et al. (1994).

- Conical projection. Consider a cone which is tangent to the ellipsoid at a selected (standard) parallel. After development of the conical surface, the meridians are straight lines converging at a point called the apex. This point is also the center of circles which represent the projected parallels. The standard parallel is mapped without distortion. One example of this projection is the conformal Lambert projection.

- Cylindrical projection. This is a special case of the conical projection if the apex is moved to infinity so that the cone becomes a cylinder which is tangent at the equator. In the transverse position, the cylinder is tangent at a standard meridian. After development of the cylindrical surface, the standard meridian is mapped without distortion. Two examples of this projetion are the Transverse Mercator projection and the Universal Transverse Mercator (UTM) projection. Because both methods are widely used, more details are given below.

- Azimuthal projection. This is also a special case of the conical projection if the apex is moved to the pole so that the cone becomes a plane which is tangent at the pole. The pole is the center of circles representing the parallels and of straight lines representing the meridians. More generally, the projection plane can be defined as plane tangent at any point on the ellipsoid. One example of this projection is the stereographic projection.

Transverse Mercator projection

This method is also referred to as Gauss–Krüger projection. The ellipsoid is partitioned into 120 zones of 3° longitude each where the central meridian is in the center of each zone. The central meridian is mapped onto the plane without scale distortion and represents the y-axis (north direction). The x-axis is the mapping of the equator. The central meridian of the zone and the equator are special cases since all other meridians and parallels are mapped as curved lines. Because of conformality, the mapped images of the meridians and parallels are orthogonal to each other.

The numbering of the zones is related either to Greenwich or to Ferro. By definition, the latter is situated 17°40′ west of Greenwich.

For the Transverse Mercator projection, the ellipsoidal point φ, λ being mapped into a point y, x of the plane is given by the series expansions

$$\begin{aligned} y &= B(\varphi) + \frac{t}{2} N \cos^2\varphi \, \ell^2 + \frac{t}{24} N \cos^4\varphi \, (5 - t^2 + 9\eta^2 + 4\eta^4) \, \ell^4 \\ &+ \frac{t}{720} N \cos^6\varphi \, (61 - 58\, t^2 + t^4 + 270\, \eta^2 - 330\, t^2\eta^2) \, \ell^6 \\ &+ \frac{t}{40\,320} N \cos^8\varphi \, (1385 - 3111\, t^2 + 543\, t^4 - t^6) \, \ell^8 + \ldots \end{aligned} \tag{10.21}$$

$$\begin{aligned} x &= N \cos\varphi \, \ell + \frac{1}{6} N \cos^3\varphi \, (1 - t^2 + \eta^2) \, \ell^3 \\ &+ \frac{1}{120} N \cos^5\varphi \, (5 - 18\, t^2 + t^4 + 14\, \eta^2 - 58\, t^2\eta^2) \, \ell^5 \\ &+ \frac{1}{5040} N \cos^7\varphi \, (61 - 479\, t^2 + 179\, t^4 - t^6) \, \ell^7 + \ldots \end{aligned}$$

where

$B(\varphi)$	...	arc length of meridian
$N = \dfrac{a^2}{b\sqrt{1+\eta^2}}$	...	radius of curvature in prime vertical
$\eta^2 = e'^2 \cos^2\varphi$	...	auxiliary quantity
$e'^2 = (a^2 - b^2)/b^2$	...	second numerical eccentricity
$t = \tan\varphi$	...	auxiliary quantity
$\ell = \lambda - \lambda_0$	...	longitude difference
λ_0	...	longitude of the central meridian

are used. By convention, y is given first followed by x because the pair of coordinates (y, x) corresponds to (φ, λ). The arc length of meridian $B(\varphi)$ is

the ellipsoidal distance from the equator to the point to be mapped and can be computed by the series expansion

$$B(\varphi) = \alpha\,[\varphi + \beta \sin 2\varphi + \gamma \sin 4\varphi + \delta \sin 6\varphi + \varepsilon \sin 8\varphi + \ldots] \tag{10.22}$$

where

$$\begin{aligned}
\alpha &= \frac{a+b}{2}\left(1 + \frac{1}{4}n^2 + \frac{1}{64}n^4 + \ldots\right) \\
\beta &= -\frac{3}{2}n + \frac{9}{16}n^3 - \frac{3}{32}n^5 + \ldots \\
\gamma &= \frac{15}{16}n^2 - \frac{15}{32}n^4 + \ldots \\
\delta &= -\frac{35}{48}n^3 + \frac{105}{256}n^5 - \ldots \\
\varepsilon &= \frac{315}{512}n^4 + \ldots
\end{aligned} \tag{10.23}$$

and

$$n = \frac{a-b}{a+b}. \tag{10.24}$$

As examples,

	Bessel ellipsoid	WGS-84 ellipsoid
α	6 366 742.5203 m	6 367 449.1458 m
β	$-2.511\,274\,56 \cdot 10^{-3}$	$-2.518\,827\,92 \cdot 10^{-3}$
γ	$2.627\,71 \cdot 10^{-6}$	$2.643\,54 \cdot 10^{-6}$
δ	$-3.42 \cdot 10^{-9}$	$-3.45 \cdot 10^{-9}$
ε	$5 \cdot 10^{-12}$	$5 \cdot 10^{-12}$

(10.25)

are the parameters for the Bessel ellipsoid and the WGS-84 ellipsoid.

Consider the following numerical example where the arc length of the meridian on the Bessel ellipsoid is calculated. For $\varphi = 47°$, the result $B(\varphi) = 5\,206\,717.123$ m is obtained.

The inverse Gauss–Krüger projection involves the mapping of a point y, x in the plane to a point φ, λ on the ellipsoid. The formulas are given by the series expansions

$$\begin{aligned}\varphi &= \varphi_f + \frac{t_f}{2N_f^2}(-1-\eta_f^2)\,x^2 \\ &+ \frac{t_f}{24\,N_f^4}(5+3t_f^2+6\eta_f^2-6t_f^2\,\eta_f^2-3\eta_f^4-9t_f^2\,\eta_f^4)\,x^4 \\ &+ \frac{t_f}{720\,N_f^6}(-61-90\,t_f^2-45\,t_f^4-107\,\eta_f^2+162\,t_f^2\,\eta_f^2 \\ &\qquad +45\,t_f^4\,\eta_f^2)\,x^6 \\ &+ \frac{t_f}{40\,320\,N_f^8}(1385+3633\,t_f^2+4095\,t_f^4+1575\,t_f^6)\,x^8+\ldots \end{aligned} \tag{10.26}$$

$$\begin{aligned}\lambda &= \lambda_0 + \frac{1}{N_f\cos\varphi_f}\,x + \frac{1}{6N_f^3\cos\varphi_f}(-1-2t_f^2-\eta_f^2)\,x^3 \\ &+ \frac{1}{120\,N_f^5\cos\varphi_f}(5+28\,t_f^2+24\,t_f^4+6\eta_f^2+8t_f^2\,\eta_f^2)\,x^5 \\ &+ \frac{1}{5040\,N_f^7\cos\varphi_f}(-61-662\,t_f^2-1320\,t_f^4-720\,t_f^6)\,x^7+\ldots \end{aligned}$$

where the terms with the subscript f must be calculated based on the footpoint latitude φ_f. For the footpoint latitude, the series expansion is given by

$$\varphi_f = \bar{y} + \bar{\beta}\sin 2\bar{y} + \bar{\gamma}\sin 4\bar{y} + \bar{\delta}\sin 6\bar{y} + \bar{\varepsilon}\sin 8\bar{y} + \ldots \tag{10.27}$$

where

$$\begin{aligned}\bar{\alpha} &= \frac{a+b}{2}\left(1+\frac{1}{4}\,n^2+\frac{1}{64}\,n^4+\ldots\right) \\ \bar{\beta} &= \frac{3}{2}\,n-\frac{27}{32}\,n^3+\frac{269}{512}\,n^5+\ldots \\ \bar{\gamma} &= \frac{21}{16}\,n^2-\frac{55}{32}\,n^4+\ldots \\ \bar{\delta} &= \frac{151}{96}\,n^3-\frac{417}{128}\,n^5+\ldots \\ \bar{\varepsilon} &= \frac{1097}{512}\,n^4+\ldots \end{aligned} \tag{10.28}$$

and the relation

$$\bar{y} = \frac{y}{\bar{\alpha}} \tag{10.29}$$

are used. Note that the coefficient $\bar{\alpha}$ is identical to α in (10.23). As examples,

	Bessel ellipsoid	WGS-84 ellipsoid
$\bar{\alpha}$	6 366 742.5203 m	6 367 449.1458 m
$\bar{\beta}$	$2.511\,273\,24 \cdot 10^{-3}$	$2.518\,826\,58 \cdot 10^{-3}$
$\bar{\gamma}$	$3.678\,79 \cdot 10^{-6}$	$3.700\,95 \cdot 10^{-6}$
$\bar{\delta}$	$7.38 \cdot 10^{-9}$	$7.45 \cdot 10^{-9}$
$\bar{\varepsilon}$	$17 \cdot 10^{-12}$	$17 \cdot 10^{-12}$

(10.30)

are the parameters for the Bessel ellipsoid and the WGS-84 ellipsoid.

For a numerical example, consider the coordinate $y = 5\,206\,717.123$ m referred to the Bessel ellipsoid and compute the associated footpoint latitude. The result is $\varphi_f = 47°$.

Universal Transverse Mercator (UTM) system

The UTM system is a modification of the Transverse Mercator system. First, the ellipsoid is partitioned into 60 zones with a width of 6° longitude each. Second, a scale factor of 0.9996 is applied to the conformal coordinates in the plane. The reason for this factor is to avoid fairly large distortions in the outer areas of a zone.

The zone numbering starts with M1 for the central meridian $\lambda_0 = 177°$ W and continues with M2 for the central meridian $\lambda_0 = 171°$ W. An adequate formula to calculate the zone number is given by

$$\text{INT}\left(\frac{180 \pm \lambda}{6}\right) + 1 \tag{10.31}$$

where the plus sign must be used for eastern longitudes and the minus sign for western longitudes.

For a numerical example, consider a point on the WGS-84 ellipsoid with $\varphi = 47°$ N and $\lambda = 16°$ E and compute the UTM coordinates. The zone is M33 according to Eq. (10.31) and the central meridian becomes $\lambda_0 = 15°$ E. The final result $y = 5\,205\,649.348$ m, $x = 76\,025.312$ m is obtained by the formulas (10.21) after multiplication by the scale factor. It is recommended that the inverse mapping be performed to demonstrate that millimeter accuracy is achieved.

10.2.4 Height transformation

In the previous section, a point on the ellipsoid was mapped onto a plane and vice versa. The ellipsoidal height could be completely ignored. In this section, the primary interest is the height. The formula

$$h = H + N \tag{10.32}$$

where

$$\begin{array}{lll} h & \ldots & \text{ellipsoidal height} \\ H & \ldots & \text{orthometric height} \\ N & \ldots & \text{geoidal height (undulation)} \end{array} \tag{10.33}$$

is the relationship between the ellipsoid and the geoid. As shown in Fig. 10.4, this formula is an approximation but is sufficiently accurate for all practical purposes. The angle ε expresses the deflection of the vertical between the slightly curved plumb line and the ellipsoidal normal. This angle does not exceed 30 arcseconds in most areas.

Positioning with GPS results in X, Y, Z coordinates. After applying the transformation (10.11), ellipsoidal heights are obtained. If, additionally, one of the two remaining terms in (10.32) is given, the other one can be calculated. Thus, if the geoid is known, orthometric heights can be derived, or, if orthometric heights are known, geoidal heights can be derived.

As long as geodesists were performing horizontal surveys (e.g., triangulation) separate from vertical surveys (e.g., leveling), the difference between the ellipsoid and the geoid was not significant. This changed with the advent of satellite geodesy which yields both horizontal and vertical information. Now, the geometrically defined (i.e., ellipsoidal) heights obtained from satellite geodesy must be combined with physically defined (i.e., orthometric) heights. Thus, a short explanation is given below for readers not familiar with the ellipsoid and the geoid.

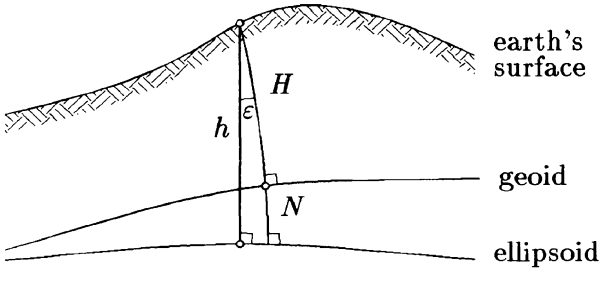


Fig. 10.4. Definition of heights

Ellipsoid

An ellipsoid of revolution is an approximation of the surface of the earth. This surface is formed by choosing a proper sized ellipse and rotating it about its minor axis. The ellipsoid is a convenient mathematical surface which has been subdivided by latitude and longitude to form a coordinate system. Distances and azimuths can be computed on this ellipsoid to millimeter accuracy and hundredths of an arcsecond.

It should be noted that there is a distinction between locally best-fitting (nongeocentric) ellipsoids and a global (geocentric) ellipsoid. Different local ellipsoids are primarily used because they fit a particular portion of the earth better than a global ellipsoid. In fact, the ellipsoid chosen for an area – for example the Clarke ellipsoid for the former North American Datum 1927 (NAD-27) – was picked to minimize the difference between that ellipsoid and the geoid in that area. The center of a local ellipsoid does not coincide with the center of the true earth, but is displaced from the true center by up to some 100 meters.

Fortunately, only one global ellipsoid is used today. When it became possible to connect the datums of the various continents, geodesists decided to choose a single global reference ellipsoid. The present Geodetic Reference System 1980 (GRS-80) was chosen after much discussion. The WGS-84 used for GPS is virtually identical with the GRS-80. The center of the GRS-80 ellipsoid coincides with the true center of the earth, and its surface provides an average fit of the geoid. The fit of this new surface to the geoid is not an improvement for all places on the earth. For example, the old Clarke ellipsoid surface only differs by small amounts from the geoid surface for the U.S.; whereas, the GRS-80 ellipsoid is approximately 30 m lower than the geoid in the eastern U.S. Surveyors using the old NAD-27 datum were accustomed to reduce their distance measurements to sea level when actually they were correcting lengths to map them onto the ellipsoid. Now on the east coast of the U.S., they must subtract approximately 30 m from the elevation to obtain the correct height to use for the reduction to the ellipsoid.

Geoid

The geoid is defined physically and is the surface that is used to represent the actual shape of the earth. Drawings of the geoid show it as a bumpy surface with hills and valleys similar to a topographic model.

The center of the geoid coincides with the true center of the earth and its surface is an equipotential surface. The geoid can be visualized by imagining that the earth were completely covered by water. This water surface would (in theory) be an equipotential surface since the water would flow to

compensate for any height difference that would occur. In actuality, the sea level differs slightly from a true equipotential surface due to bumps formed by different ocean currents, salinities, temperatures, etc.

The geoid is the surface chosen for leveling datums. In many countries, the geoid that most nearly coincides with the average sea level value (measured by tide gauges) is chosen as the zero elevation. Because of this choice, the datum is often referred to as the sea level datum. This does not mean, however, that the zero elevation will coincide with the average level of sea level along a coastline. In fact, differences of about one meter exist between the zero elevation and the average sea level at any given point. The average sea level at a given point along the coast is affected by many factors so that average sea levels along the coast do not form an equipotential surface.

The geoidal surface is very irregular and it is virtually impossible to realize an exact mathematical model for the geoid. A good approximation of the geoid has been achieved through the use of spherical harmonics. "New geoids in the world" including a high resolution geoid model for the U.S. are published in IAG (1995). Current research results are being published by a Working Group of the IAG (International Geoid Service 2000). However, geoid determination will undoubtedly occupy geodesists for the next several years (maybe decades).

10.3 Datum transformations

The coordinate transformations in the previous section dealt with the transition from one type of coordinates to another type of coordinates for the same point. Global coordinates X, Y, Z have been transformed into ellipsoidal coordinates φ, λ, h and into local level coordinates n, e, u, and two-dimensional ellipsoidal surface coordinates φ, λ have been transformed into plane coordinates y, x. Finally, the ellipsoidal height has been transformed either to the orthometric height or to the geoidal height.

A (geodetic) datum defines the relationship between a global and a local three-dimensional Cartesian coordinate system; therefore, a datum transformation transforms one coordinate system of a certain type to another coordinate system of the same type. This is one of the primary tasks when combining GPS data with terrestrial data, i.e., the transformation of geocentric WGS-84 coordinates to local terrestrial coordinates. As mentioned earlier, the terrestrial system uses a locally best-fitting ellipsoid, e.g., the Clarke ellipsoid or the GRS-80 ellipsoid in the U.S. and the Bessel ellipsoid in many parts of Europe. The local ellipsoid is linked to a nongeocentric Cartesian coordinate system where the origin coincides with the center of the ellipsoid. Plane coordinates such as Gauss–Krüger coordinates are ob-

tained by mapping the local ellipsoid into the plane.

In the context of datum transformations, the GPS fiducial point concept must be mentioned although it was primarily used for orbit improvement in the past (e.g., Ashkenazi et al. 1990). Fiducial points are sites whose positions are accurately known from a (GPS independent) method such as VLBI or SLR. The concept of fiducial points is quite simple: during a GPS campaign, at least three fiducial points in the area of the campaign are also equipped with receivers (apart from the points to be determined). This enables the transformation of GPS coordinates into the frame of the fiducial points by a three-dimensional similarity transformation. Note that the geometry of the fiducial points with respect to the remaining points has a strong effect on the GPS accuracy in this region.

10.3.1 Three-dimensional transformation

Consider two sets of three-dimensional Cartesian coordinates forming the vectors $\underline{X}$ and $\underline{X}_T$ (Fig. 10.5). The similarity transformation, also denoted as Helmert transformation, between the two sets can be formulated by the relation

$$\underline{X}_T = \underline{c} + \mu \underline{R} \underline{X} \tag{10.34}$$

where $\underline{c}$ is the translation (or shift) vector, μ is a scale factor, and $\underline{R}$ is a rotation matrix.

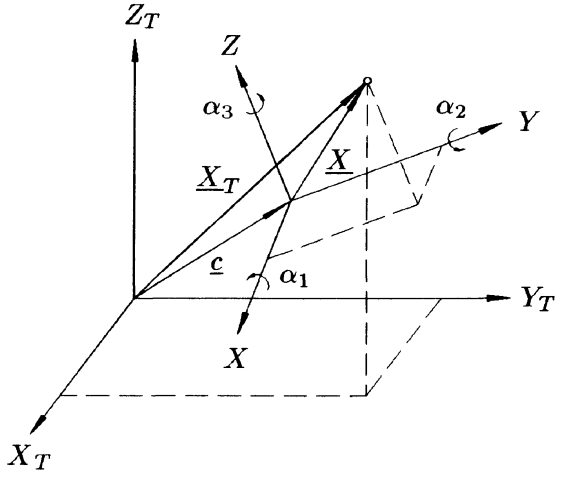


Fig. 10.5. Three-dimensional transformation

The components of the shift vector

$$\underline{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} \tag{10.35}$$

account for the coordinates of the origin of the $\underline{X}$ system in the $\underline{X}_T$ system. Note that a single scale factor is considered. More generally (but with GPS not necessary), three scale factors, one for each axis, could be used. The rotation matrix is an orthogonal matrix which is composed of three successive rotations

$$\underline{R} = \underline{R}_3\{\alpha_3\}\,\underline{R}_2\{\alpha_2\}\,\underline{R}_1\{\alpha_1\}\,. \tag{10.36}$$

Explicitly,

$$\underline{R} = \begin{bmatrix} \cos\alpha_2 \cos\alpha_3 & \cos\alpha_1 \sin\alpha_3 & \sin\alpha_1 \sin\alpha_3 \\ & + \sin\alpha_1 \sin\alpha_2 \cos\alpha_3 & - \cos\alpha_1 \sin\alpha_2 \cos\alpha_3 \\ - \cos\alpha_2 \sin\alpha_3 & \cos\alpha_1 \cos\alpha_3 & \sin\alpha_1 \cos\alpha_3 \\ & - \sin\alpha_1 \sin\alpha_2 \sin\alpha_3 & + \cos\alpha_1 \sin\alpha_2 \sin\alpha_3 \\ \sin\alpha_2 & - \sin\alpha_1 \cos\alpha_2 & \cos\alpha_1 \cos\alpha_2 \end{bmatrix} \tag{10.37}$$

is obtained.

In the case of known transformation parameters $\underline{c}$, μ, $\underline{R}$, a point from the $\underline{X}$ system can be transformed into the $\underline{X}_T$ system by (10.34).

If the transformation parameters are unknown, they can be determined with the aid of common (identical) points. This means that the coordinates of the same point are given in both systems. Since each common point (given by $\underline{X}_T$ and $\underline{X}$) yields three equations, two common points and one additional common component (e.g., height) are sufficient to solve for the seven unknown parameters. In practice, redundant common point information is used and the unknown parameters are calculated by least squares adjustment.

Since the parameters are mixed nonlinearly in Eq. (10.34), a linearization must be performed where approximate values μ_0, $\underline{R}_0$, $\underline{c}_0$ are required. In the case of a datum transformation between WGS-84 and a local system, the approximation $\mu_0 = 1$ is appropriate and the relation

$$\mu = \mu_0 + \Delta\mu = 1 + \Delta\mu \tag{10.38}$$

is obtained. Furthermore, the rotation angles α_i in (10.37) are small and may be considered as differential quantities $\Delta\alpha_i$. Introducing these quantities

into (10.37), setting $\cos\Delta\alpha_i = 1$ and $\sin\Delta\alpha_i = \Delta\alpha_i$, and considering only first-order terms gives

$$\underline{R} = \begin{bmatrix} 1 & \Delta\alpha_3 & -\Delta\alpha_2 \\ -\Delta\alpha_3 & 1 & \Delta\alpha_1 \\ \Delta\alpha_2 & -\Delta\alpha_1 & 1 \end{bmatrix} = \underline{I} + \Delta\underline{R} \tag{10.39}$$

where $\underline{I}$ is the unit matrix and $\Delta\underline{R}$ is a (skewsymmetric) differential rotation matrix. Thus, the approximation $\underline{R}_0 = \underline{I}$ is appropriate. Finally, the shift vector is split up in the form

$$\underline{c} = \underline{c}_0 + \Delta\underline{c} \tag{10.40}$$

where the approximate shift vector

$$\underline{c}_0 = \underline{X}_T - \underline{X} \tag{10.41}$$

follows by substituting the approximations for the scale factor and the rotation matrix into Eq. (10.34).

Introducing the Eqs. (10.38), (10.39), (10.40) into (10.34) and skipping details which can be found for example in Hofmann-Wellenhof et al. (1994) gives the linearized model for a single point i. This model can be written in the form

$$\underline{X}_{T_i} - \underline{X}_i - \underline{c}_0 = \underline{A}_i\,\Delta\underline{p} \tag{10.42}$$

where the left side of the equation is known and may formally be considered as an observation. The design matrix $\underline{A}_i$ and the vector $\Delta\underline{p}$, containing the unknown parameters, are given by

$$\underline{A}_i = \begin{bmatrix} 1 & 0 & 0 & X_i & 0 & -Z_i & Y_i \\ 0 & 1 & 0 & Y_i & Z_i & 0 & -X_i \\ 0 & 0 & 1 & Z_i & -Y_i & X_i & 0 \end{bmatrix} \tag{10.43}$$

$$\Delta\underline{p} = [\ \Delta c_1 \quad \Delta c_2 \quad \Delta c_3 \quad \Delta\mu \quad \Delta\alpha_1 \quad \Delta\alpha_2 \quad \Delta\alpha_3\ ]^T .$$

Recall that Eq. (10.42) is now a system of linear equations for point i. For n common points, the design matrix A is

$$\underline{A} = \begin{bmatrix} \underline{A}_1 \\ \underline{A}_2 \\ \vdots \\ \underline{A}_n \end{bmatrix} . \tag{10.44}$$

In detail, for three common points the design matrix is

$$\underline{A} = \begin{bmatrix} 1 & 0 & 0 & X_1 & 0 & -Z_1 & Y_1 \\ 0 & 1 & 0 & Y_1 & Z_1 & 0 & -X_1 \\ 0 & 0 & 1 & Z_1 & -Y_1 & X_1 & 0 \\ 1 & 0 & 0 & X_2 & 0 & -Z_2 & Y_2 \\ 0 & 1 & 0 & Y_2 & Z_2 & 0 & -X_2 \\ 0 & 0 & 1 & Z_2 & -Y_2 & X_2 & 0 \\ 1 & 0 & 0 & X_3 & 0 & -Z_3 & Y_3 \\ 0 & 1 & 0 & Y_3 & Z_3 & 0 & -X_3 \\ 0 & 0 & 1 & Z_3 & -Y_3 & X_3 & 0 \end{bmatrix} \tag{10.45}$$

which leads to a slightly redundant system. Least squares adjustment yields the parameter vector $\Delta\underline{p}$ and the adjusted values by (10.38), (10.39), (10.40). Once the seven parameters of the similarity transformation are determined, formula (10.34) can be used to transform other than the common points.

For a specific example, consider the task of transforming GPS coordinates of a network to (three-dimensional) coordinates of a (nongeocentric) local system. The GPS coordinates are denoted by $(X, Y, Z)_{\mathrm{GPS}}$ and the local system coordinates are the plane coordinates $(y, x)_{\mathrm{LS}}$ and the ellipsoidal height h_{LS}. To obtain the transformation parameters, it is assumed that the coordinates of the common points in both systems are available. The solution of the task is obtained by the following algorithm:

1. Transform the plane coordinates $(y, x)_{\mathrm{LS}}$ of the common points into the ellipsoidal surface coordinates $(\varphi, \lambda)_{\mathrm{LS}}$ by using the appropriate mapping formulas.

2. Transform the ellipsoidal coordinates $(\varphi, \lambda, h)_{\mathrm{LS}}$ of the common points into the Cartesian coordinates $(X, Y, Z)_{\mathrm{LS}}$ by (10.1).

3. Determine the seven parameters of a Helmert transformation by using the coordinates $(X, Y, Z)_{\mathrm{GPS}}$ and $(X, Y, Z)_{\mathrm{LS}}$ of the common points.

4. For network points other than the common points, transform the coordinates $(X, Y, Z)_{\mathrm{GPS}}$ into $(X, Y, Z)_{\mathrm{LS}}$ via Eq. (10.34) using the transformation parameters determined in the previous step.

5. Transform the Cartesian coordinates $(X, Y, Z)_{\mathrm{LS}}$ computed in the previous step into ellipsoidal coordinates $(\varphi, \lambda, h)_{\mathrm{LS}}$ by (10.11) or by the iterative procedure.

6. Map the ellipsoidal surface coordinates $(\varphi, \lambda)_{\mathrm{LS}}$ computed in the previous step into plane coordinates $(y, x)_{\mathrm{LS}}$ by the appropriate mapping formulas.

The advantage of the three-dimensional approach is that no a priori information is required for the seven parameters of the similarity transformation. The disadvantage of the method is that for the common points ellipsoidal heights (and, thus, geoidal heights) are required. However, as reported by Schmitt et al. (1991), the effect of incorrect heights of the common points often has a negligible affect on the plane coordinates (y, x). For example, incorrect heights may cause a tilt of a 20 km × 20 km network by an amount of 5 m in space; however, the effect on the plane coordinates is only approximately 1 mm.

For large areas, the height problem can be solved by adopting approximate ellipsoidal heights for the common points and performing a three-dimensional affine transformation instead of the similarity transformation.

10.3.2 Two-dimensional transformation

Two-dimensional coordinates are now considered. The two different sets of plane coordinates are represented by $\underline{x}$ and $\underline{x}_T$ (Fig. 10.6). The two-dimensional similarity transformation is defined by

$$\underline{x}_T = \underline{c} + \mu \underline{R}\, \underline{x} \tag{10.46}$$

with a scale factor μ, the shift vector

$$\underline{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \tag{10.47}$$

and the rotation matrix

$$\underline{R} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \tag{10.48}$$

which contains a single rotation angle. Hence, Eq. (10.46) comprises four transformation parameters: the two translation components c_1, c_2, the scale factor μ, and the rotation angle α. Substituting (10.47) and (10.48) into (10.46) gives the transformed components explicitly:

$$\begin{aligned} x_T &= c_1 + \mu\, x\, \cos\alpha - \mu\, y\, \sin\alpha \\ y_T &= c_2 + \mu\, x\, \sin\alpha + \mu\, y\, \cos\alpha\,. \end{aligned} \tag{10.49}$$

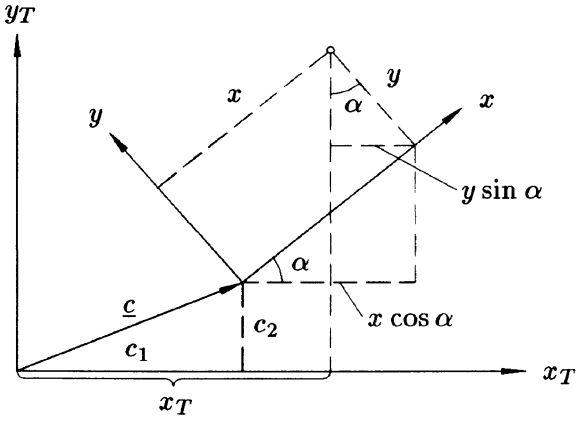


Fig. 10.6. Two-dimensional transformation

These formulas can be verified by the geometry given in Fig. 10.6 where the contributing terms for x_T (without scale factor) are indicated.

In the case of known transformation parameters $\underline{c}$, μ, $\underline{R}$, a point in the $\underline{x}$ system can be transformed into the $\underline{x}_T$ system by (10.46).

If the transformation parameters are unknown, they can – analogously to the three-dimensional case – be determined using common points. Two common points, each yielding two equations, are sufficient to solve for the four unknown parameters. In practice, redundant common point information is used and the unknown parameters are calculated by least squares adjustment.

As seen from (10.49), the unknowns appear again in nonlinear form. However, using the auxiliary unknowns

$$\begin{aligned} p &= \mu \cos \alpha \\ q &= \mu \sin \alpha \,, \end{aligned} \tag{10.50}$$

linear equations

$$\begin{aligned} x_T &= c_1 + p\,x - q\,y \\ y_T &= c_2 + q\,x + p\,y \end{aligned} \tag{10.51}$$

with respect to the unknowns are obtained. In the case of redundant common points, the solution is obtained by the conventional least squares adjustment, where the coordinates x_T and y_T are formally considered as observations. The scale factor and the rotation angle are determined from the auxiliary unknowns by

$$\begin{aligned} \mu &= \sqrt{p^2 + q^2} \\ \tan \alpha &= q/p \,. \end{aligned} \tag{10.52}$$

For a specific example, consider the task of transforming GPS coordinates of a network to (two-dimensional) coordinates of a (nongeocentric) local system. The GPS coordinates are denoted by $(X, Y, Z)_{\mathrm{GPS}}$ and the plane coordinates in the local system are denoted by $(y, x)_{\mathrm{LS}}$. To get the transformation parameters, it is assumed that for some points (common points) the coordinates in both systems are known. Note that no height information in the local system is available. The solution of the task is obtained by the following algorithm:

1. Transform the Cartesian coordinates $(X, Y, Z)_{\mathrm{GPS}}$ of all network points into ellipsoidal coordinates $(\varphi, \lambda, h)_{\mathrm{GPS}}$ by (10.11) or by the iterative procedure. Note that the ellipsoid of the local datum should be used.

2. Disregard for all network points the ellipsoidal heights h_{GPS} and map the ellipsoidal surface coordinates $(\varphi, \lambda)_{\mathrm{GPS}}$ computed in the previous step into plane coordinates $(y, x)_{\mathrm{GPS}}$ by using the appropriate mapping formulas and the ellipsoid of the local datum.

3. Determine the four parameters of a two-dimensional similarity transformation by using the coordinates $(y, x)_{\mathrm{GPS}}$ and $(y, x)_{\mathrm{LS}}$ of the common points.

4. For network points other than the common points, transform the coordinates $(y, x)_{\mathrm{GPS}}$ into $(y, x)_{\mathrm{LS}}$ via Eq. (10.46) using the transformation parameters determined in the previous step.

This data transformation works well (even without using height information) for small sized networks, low elevations, and for areas where geoidal heights do not change. The coordinates $(y, x)_{\mathrm{GPS}}$, however, depend on the dimension and on the displacement of the local ellipsoid. These dependencies may lead to a distortion of the point clusters which, in principle, is not matched by a similarity transformation (Lichtenegger 1991). A numerical example is given in Schmitt et al. (1991) and shows that the use of the WGS-84 ellipsoid and the two-dimensional similarity transformation leads to unacceptable discrepancies in the range of 8 mm to 15 mm for a network covering an area of 200 km $\times$ 200 km.

One way to overcome the problem is to apply other transformations such as the affine transformation between the two coordinate sets in the plane. Another way is to transform the $(y, x)_{\mathrm{LS}}$ coordinates of common points approximately to $(X, Y, Z)_{\mathrm{LS}}$ by using approximate ellipsoidal heights. Now, the parameters of a three-dimensional similarity transformation can be derived which enable the approximate transformation of all GPS coordinates into the local datum. This transformation could also be achieved by only

applying an approximate shift vector to the GPS coordinates. The approximate coordinates are then mapped into the plane, and the final similarity transformation is performed in two-dimensional space.

10.3.3 One-dimensional transformation

One of the distinct features of GPS is that three-dimensional (3D) coordinates are obtained in a common frame. There is no separation between the horizontal coordinates and the height of a point because all three components are calculated together by the same procedure. In classical geodesy, horizontal coordinates and heights were obtained independently.

Now, the question arises why in the previous section a two-dimensional (2D) transformation and now a one-dimensional (1D) transformation are discussed. The answer lies in historical data. Many countries have excellent horizontal control networks available but often fairly poor ellipsoidal heights because of lack of geoidal heights. Thus, it is appropriate to use the 2D transformation if height information is not available. Similarly, the 1D transformation can be used to transform heights without detailed knowledge of the geoid.

Symbolically, the 1D transformation is obtained by 3D $\ominus$ 2D. In more detail, the parameters for the 1D transformation are obtained by "subtracting" the parameters of the 2D transformation from the parameters of the 3D transformation. This looks like

$$\begin{array}{lccccccc} \text{3D} & c_1 & c_2 & c_3 & \mu & \alpha_1 & \alpha_2 & \alpha_3 \\ \text{2D} & c_1 & c_2 & & \mu & & & \alpha_3 \\ \hline \text{1D} & & & c_3 & & \alpha_1 & \alpha_2 & \end{array} \Big\} \ominus \tag{10.53}$$

where for the 2D transformation the rotation angle has been given the corresponding subscript. In other words, the 3D transformation is composed of a 2D transformation for the horizontal coordinates and a 1D transformation for the heights.

Concentrating now on the 1D transformation, from the parameters in (10.53) it can be seen that the transformation consists of a shift along the vertical axis, a tilt (rotation) about the north–south axis, and a tilt about the east–west axis. These three unknowns are determined by using the height information of at least three common points.

Transformation using heights

Assume that for some (common) points in a GPS network the orthometric heights or elevations H_i and the ellipsoidal heights h_i are known. The

mathematical model for the one-dimensional transformation is given by

$$H_i = h_i + \Delta h - y_i \, \Delta\alpha_1 + x_i \, \Delta\alpha_2 \tag{10.54}$$

where Δh is a vertical shift, and $\Delta\alpha_1$ and $\Delta\alpha_2$ are rotation angles about the x-axis and the y-axis. Formally, Eq. (10.54) corresponds to the third component of the three-dimensional similarity transformation without taking into account a scale factor. Additionally, the equation is now expressed in the local coordinate frame with the position coordinates x_i and y_i. These coordinates are required with only low accuracy and can be taken, e.g., from a map. Geometrically, the model equation (10.54) may be interpreted as the equation of a plane which enables the interpolation of geoidal heights $N = h - H$ in other than the common points. The interpolation could be extended to a higher-order surface to take into account more irregular geoid structures.

When a geoid model is available, the ellipsoidal heights h_i could be transformed into approximate elevations H_{i0}. Normally, there are discrepancies between the heights H_i and H_{i0} due to the combined effects of the GPS systematic errors and the errors in the geoid modeling.

The FGCS specifications for GPS surveys require that surveys to be included into the national network be tied to a minimum of four benchmarks well-distributed geometrically throughout the project area (e.g., corners of project area). The additional benchmarks (ties) enable a least squares adjustment of the model equations (10.54) and provide the necessary check on the computation of the rotation of the ellipsoid to the geoid. A good practice is to perform the rotation using three of the elevations and to check the rotated elevation of the fourth point against the true elevation. The two values should agree within a few centimeters under normal conditions. An additional check on the correctness of the transformation is provided by inspecting the magnitude of the two rotation angles computed by the least squares adjustment. Normally, these angles should be less than a few arcseconds.

Usually, the elevations of points in a network of small size, let us say 10 km × 10 km, can be determined with an accuracy of about 3 cm. In cases where the generalized model adequately describes the variation of the geoidal heights, much larger areas can be surveyed using this method with comparable accuracies being achieved.

Transformation using height differences

In the preceding paragraph, the importance of geoidal heights has been stressed. The ellipsoidal height determined by GPS can be transformed to the orthometric height if the geoidal height is known.

There are times, however, when it is only required to measure changes in elevation. For example, when it is desired to measure the subsidence rate of a point (e.g., oil platform). In such cases, the importance of a well-known geoid diminishes because relative heights are considered. For two points,

$$\begin{aligned} H_1 &= h_1 - N_1 \\ H_2 &= h_2 - N_2 \end{aligned} \tag{10.55}$$

are the height relations and

$$H_2 - H_1 = h_2 - h_1 - (N_2 - N_1) \tag{10.56}$$

is the height difference or height change between the points 1 and 2. Here, only the difference of the geoidal heights affects the result. Thus, if the geoidal heights are constant in a local area, meaning that the separation between the geoid and the ellipsoid is constant, they can be ignored. Similarly, if the geoid has a constant slope with the ellipsoid, the heights can be computed accurately by rotating the GPS heights into the geoidal surface as previously described.

10.4 Combining GPS and terrestrial data

10.4.1 Common coordinate system

So far GPS and terrestrial networks have been considered separately with respect to the adjustment. The combination, for example by a datum transformation, was supposed to be performed after individual adjustments. Now the common adjustment of GPS observations and terrestrial data is investigated. The problem encountered here is that GPS data refer to the three-dimensional geocentric Cartesian system WGS-84, whereas terrestrial data refer to the individual local level (tangent plane) systems at each measurement point referenced to plumb lines. Furthermore, terrestrial data are traditionally separated into position and height, where the position refers to an ellipsoid and the (orthometric) height to the geoid.

For a common adjustment, a common coordinate system is required which all observations are transformed to. In principle, any arbitrary system may be introduced as common reference. One possibility is to use two-dimensional (plane) coordinates in the local system as proposed by Daxinger and Stirling (1995). Here, a three-dimensional coordinate system is chosen. The origin of the coordinate system is the center of the ellipsoid adopted for the local system, the Z-axis coincides with the semiminor axis of the ellipsoid, the X-axis is obtained by the intersection of the ellipsoidal Greenwich

meridian plane and the ellipsoidal equatorial plane, and the Y-axis completes the right-handed system. Position vectors referred to this system are denoted by $\underline{X}_{\mathrm{LS}}$ where LS indicates the reference to the local system.

After the decision on the common coordinate system, the terrestrial measurements referring to the individual local level systems at the observing sites must be represented in this common coordinate system. Similarly, GPS baseline vectors regarded as measurement quantities are to be transformed to this system.

10.4.2 Representation of measurement quantities

Distances

The spatial distance s_{ij} as function of the local level coordinates is given in (10.19). If n_{ij}, e_{ij}, u_{ij}, the components of $\underline{x}_{ij}$, are substituted by (10.15), the relation

$$\begin{aligned} s_{ij} &= \sqrt{n_{ij}^2 + e_{ij}^2 + u_{ij}^2} \\ &= \sqrt{(X_j - X_i)^2 + (Y_j - Y_i)^2 + (Z_j - Z_i)^2} \end{aligned} \tag{10.57}$$

is obtained where (10.14) has also been taken into account namely the fact that $\underline{n}_i$, $\underline{e}_i$, $\underline{u}_i$ are unit vectors. Obviously, the second expression arises immediately from the Pythagorean theorem. Differentiation of (10.57) yields

$$ds_{ij} = \frac{X_{ij}}{s_{ij}}(dX_j - dX_i) + \frac{Y_{ij}}{s_{ij}}(dY_j - dY_i) + \frac{Z_{ij}}{s_{ij}}(dZ_j - dZ_i) \tag{10.58}$$

where

$$\begin{aligned} X_{ij} &= X_j - X_i \\ Y_{ij} &= Y_j - Y_i \\ Z_{ij} &= Z_j - Z_i\,, \end{aligned} \tag{10.59}$$

have been introduced accordingly. The relation (10.58) may also be expressed as

$$\Delta s_{ij} = \frac{X_{ij}}{s_{ij}}(\Delta X_j - \Delta X_i) + \frac{Y_{ij}}{s_{ij}}(\Delta Y_j - \Delta Y_i) + \frac{Z_{ij}}{s_{ij}}(\Delta Z_j - \Delta Z_i) \tag{10.60}$$

if the differentials are replaced by differences.

Azimuths

Again the same principle applies: the measured azimuth α_{ij} as function of the local level coordinates is given in (10.19). If n_{ij}, e_{ij}, u_{ij}, the components of $\underline{x}_{ij}$, are substituted by (10.15), the relation

$$\begin{aligned}\tan \alpha_{ij} &= e_{ij}/n_{ij} \\ &= \frac{-X_{ij} \sin \lambda_i + Y_{ij} \cos \lambda_i}{-X_{ij} \sin \varphi_i \cos \lambda_i - Y_{ij} \sin \varphi_i \sin \lambda_i + Z_{ij} \cos \varphi_i}\end{aligned} \tag{10.61}$$

is obtained. After a lengthy derivation, the relation

$$\begin{aligned}\Delta\alpha_{ij} = &\frac{\sin \varphi_i \cos \lambda_i \sin \alpha_{ij} - \sin \lambda_i \cos \alpha_{ij}}{s_{ij} \sin z_{ij}} (\Delta X_j - \Delta X_i) \\ &+ \frac{\sin \varphi_i \sin \lambda_i \sin \alpha_{ij} + \cos \lambda_i \cos \alpha_{ij}}{s_{ij} \sin z_{ij}} (\Delta Y_j - \Delta Y_i) \\ &- \frac{\cos \varphi_i \sin \alpha_{ij}}{s_{ij} \sin z_{ij}} (\Delta Z_j - \Delta Z_i) \\ &+ \cot z_{ij} \sin \alpha_{ij} \, \Delta\varphi_i \\ &+ (\sin \varphi_i - \cos \alpha_{ij} \cos \varphi_i \cot z_{ij}) \, \Delta\lambda_i\end{aligned} \tag{10.62}$$

is obtained.

Directions

Measured directions R_{ij} are related to azimuths α_{ij} by the orientation unknown o_i. The relation reads

$$R_{ij} = \alpha_{ij} - o_i \,, \tag{10.63}$$

and the expression

$$\Delta R_{ij} = \Delta\alpha_{ij} - \Delta o_i \,. \tag{10.64}$$

is immediately obtained.

Zenith angles

The zenith angle z_{ij} as function of the local level coordinates is given in (10.19). If n_{ij}, e_{ij}, u_{ij}, the components of $\underline{x}_{ij}$, are substituted by (10.15), the relation

$$\begin{aligned}\cos z_{ij} &= u_{ij}/s_{ij} \\ &= \frac{X_{ij} \cos \varphi_i \cos \lambda_i + Y_{ij} \cos \varphi_i \sin \lambda_i + Z_{ij} \sin \varphi_i}{\sqrt{X_{ij}^2 + Y_{ij}^2 + Z_{ij}^2}}\end{aligned} \tag{10.65}$$

is obtained where (10.57) and (10.59) have been used. After a lengthy derivation, the relation

$$\begin{aligned}\Delta z_{ij} = &\frac{X_{ij}\cos z_{ij} - s_{ij}\cos\varphi_i\cos\lambda_i}{s_{ij}^2\sin z_{ij}}(\Delta X_j - \Delta X_i)\\ &+\frac{Y_{ij}\cos z_{ij} - s_{ij}\cos\varphi_i\sin\lambda_i}{s_{ij}^2\sin z_{ij}}(\Delta Y_j - \Delta Y_i)\\ &+\frac{Z_{ij}\cos z_{ij} - s_{ij}\sin\varphi_i}{s_{ij}^2\sin z_{ij}}(\Delta Z_j - \Delta Z_i)\\ &-\cos\alpha_{ij}\,\Delta\varphi_i - \cos\varphi_i\sin\alpha_{ij}\,\Delta\lambda_i\end{aligned} \tag{10.66}$$

is obtained.

It is presupposed that the zenith angles are reduced to the chord of the light path. This reduction may be modeled by

$$z_{ij} = z_{ij_{\text{meas}}} + \frac{s_{ij}}{2R_E}k\,, \tag{10.67}$$

where $z_{ij_{\text{meas}}}$ is the measured zenith angle, R_E is the mean radius of the earth, and k is the coefficient of refraction. For k either a standard value may be substituted or the coefficient of refraction is estimated as additional unknown. In the case of estimation, there are several choices, e.g., one value for k for all zenith angles, or one value for a group of zenith angles, or one value per day.

Ellipsoidal height differences

The "measured" ellipsoidal height difference is represented by

$$h_{ij} = h_j - h_i\,. \tag{10.68}$$

The heights involved are obtained by transforming the Cartesian coordinates into ellipsoidal coordinates according to (10.11) or by using the iterative procedure given in Sect. 10.2.1. The height difference is approximately (neglecting the curvature of the earth) given by the third component of $\underline{x}_{ij}$ in the local level system. Hence,

$$h_{ij} = \underline{u}_i \cdot \underline{X}_{ij} \tag{10.69}$$

or, by substituting $\underline{u}_i$ according to (10.14), the relation

$$h_{ij} = \cos\varphi_i\cos\lambda_i\,X_{ij} + \cos\varphi_i\sin\lambda_i\,Y_{ij} + \sin\varphi_i\,Z_{ij} \tag{10.70}$$

is obtained. This equation may be differentiated with respect to the Cartesian coordinates. If the differentials are replaced by the corresponding differences,

$$\begin{aligned}\Delta h_{ij} = &\cos\varphi_j\cos\lambda_j\,\Delta X_j + \cos\varphi_j\sin\lambda_j\,\Delta Y_j + \sin\varphi_j\,\Delta Z_j\\ &-\cos\varphi_i\cos\lambda_i\,\Delta X_i - \cos\varphi_i\sin\lambda_i\,\Delta Y_i - \sin\varphi_i\,\Delta Z_i\end{aligned} \tag{10.71}$$

is obtained where the coordinate differences were decomposed into their individual coordinates.

Baselines
From GPS measurements, baselines $\underline{X}_{ij_{(GPS)}} = \underline{X}_{j_{(GPS)}} - \underline{X}_{i_{(GPS)}}$ in the WGS-84 are obtained. The position vectors $\underline{X}_{i_{(GPS)}}$ and $\underline{X}_{j_{(GPS)}}$ may be transformed by a three-dimensional (7-parameter) similarity transformation to a local system indicated by LS. According to Eq. (10.34), the transformation formula reads

$$\underline{X}_{LS} = \underline{c} + \mu \underline{R} \underline{X}_{GPS} \tag{10.72}$$

where the meaning of the individual quantities is the following:

$\underline{X}_{LS}$	...	Position vector in the local system
$\underline{X}_{GPS}$	...	position vector in the WGS-84
$\underline{c}$	...	shift vector
$\underline{R}$	...	rotation matrix
μ	...	scale factor.

Forming the difference of two position vectors, i.e., the baseline $\underline{X}_{ij}$, the shift vector $\underline{c}$ is eliminated. Using (10.72), there results

$$\underline{X}_{ij_{(LS)}} = \mu \underline{R} \underline{X}_{ij_{(GPS)}} \tag{10.73}$$

for the baseline. Similar to (10.42), the linearized form is

$$\underline{X}_{ij_{(LS)}} = \underline{X}_{ij_{(GPS)}} + \underline{A}_{ij} \Delta \underline{p} \tag{10.74}$$

where now the vector $\Delta \underline{p}$ and the design matrix $\underline{A}_{ij}$ are given by

$$\begin{aligned} \Delta \underline{p} &= [\Delta\mu \quad \Delta\alpha_1 \quad \Delta\alpha_2 \quad \Delta\alpha_3]^T \\ \underline{A}_{ij} &= \begin{bmatrix} X_{ij} & 0 & -Z_{ij} & Y_{ij} \\ Y_{ij} & Z_{ij} & 0 & -X_{ij} \\ Z_{ij} & -Y_{ij} & X_{ij} & 0 \end{bmatrix}_{(GPS)} . \end{aligned} \tag{10.75}$$

Note that the rotations $\Delta\alpha_i$ refer to the axes of the system adherent to GPS. If they should refer to the local system, then the signs of the rotations must be changed, i.e., the signs of the elements of the last three columns of matrix $\underline{A}_{ij}$ must be reversed.

The vector $\underline{X}_{ij_{(LS)}}$ on the left side of (10.74) contains the points $\underline{X}_{i_{(LS)}}$ and $\underline{X}_{j_{(LS)}}$ in the local system. If these points are unknown, then they are replaced by known approximate values and unknown increments

$$\begin{aligned} \underline{X}_{i_{(LS)}} &= \underline{X}_{i0_{(LS)}} + \Delta \underline{X}_{i_{(LS)}} \\ \underline{X}_{j_{(LS)}} &= \underline{X}_{j0_{(LS)}} + \Delta \underline{X}_{j_{(LS)}} \end{aligned} \tag{10.76}$$

where the coefficients of these unknown increments (+1 or −1) together with matrix $\underline{A}_{ij}$ form the design matrix.

The vector $\underline{X}_{ij_{(\mathrm{GPS})}}$ in (10.74) is regarded as measurement quantity. Thus, finally,

$$\underline{X}_{ij_{(\mathrm{GPS})}} = \Delta\underline{X}_{j_{(\mathrm{LS})}} - \Delta\underline{X}_{i_{(\mathrm{LS})}} - \underline{A}_{ij}\,\Delta\underline{p} + \underline{X}_{j0_{(\mathrm{LS})}} - \underline{X}_{i0_{(\mathrm{LS})}} \quad (10.77)$$

is the linearized observation equation.

In principle, any type of geodetic measurement can be employed if the integrated geodesy adjustment model is used. The basic concept is that any geodetic measurement can be expressed as a function of one or more position vectors $\underline{X}$ and of the gravity field W of the earth. The usually nonlinear function must be linearized where the gravity field W is split into the normal potential U of an ellipsoid and the disturbing potential T, thus, $W = U + T$. Applying a minimum principle leads to the collocation formulas (Moritz 1980: Chap. 11).

Many examples integrating GPS and other data can be found in technical publications. For example, attempts to detect earth deformations from GPS and terrestrial data. Another approach is the combination of GPS and gravity data (Hein 1990a). Further contributions are cited in Delikaraoglou and Lahaye (1990) and numerical examples are given, e.g., in Hofmann-Wellenhof et al. (1994), Sect. 3.6.4, Daxinger and Stirling (1995).

11 Software modules

11.1 Introduction

The policy used throughout this book of not endorsing commercial products is also adhered to in this chapter. There are no references given, although publications of authors describing various software packages have been used. The main goal of this chapter is to provide potential GPS users with information concerning the various features of different software modules. The following sections cannot mention all imaginable features, but they will be helpful in determining which features are worthwhile. The main consideration is that the software (and the corresponding receiver) should match user's requirements. Most applications do not require all given features.

The following paragraphs describing various terms are kept as brief as possible. More extensive descriptions are found in other chapters of this book. A summary table is provided in each section. The goal of these tables is twofold. First, they can be used as a checklist when comparing different software and, second, they can assist in the development of new software.

Table 11.1 with keywords as concise and compact as possible gives an overview of the modules described in the following sections.

Table 11.1. Software modules

• Planning
• Data transfer
• Data processing
• Quality control
• Network computations
• Data base management
• Utilities
• Flexibility

11.2 Planning

Satellite visibility
In the planning stage of a survey, it is often necessary to have a visibility table listing the locations of the satellites. Also, rise and set times of the satellites are useful. In addition to the visibility table, a (polar) sky plot of the individual satellites is useful. A helpful option is the ability to enter site

B. Hofmann-Wellenhof et al., *Global Positioning System*

obstructions at both ends of a baseline to determine the "combined" visibility. Especially in populated areas with many buildings, a sophisticated satellite visibility software is of critical importance.

Satellite geometry
The DOP values are highly correlated with satellite geometry and, thus, the number of visible satellites. A table or a plot of DOP helps to select observation windows with good geometry. This is of importance especially for kinematic surveys.

Simulations
Simulations can be used to plan a survey. The distribution of the sites, the observation window, and the session length all contribute to the design matrix from which the cofactor matrix can be computed. The ambiguity success rate expressing the probability that the integer ambiguities are correctly estimated can be computed based on the functional and stochastic model without measurements. In some cases maximum accuracy is desired, while in other cases minimum cost is the objective. Software might also support the planning of vehicle routes to save time and resources.

Table 11.2. Module on planning

Planning
• Satellite visibility
– elevation and azimuth lists
– rise and set times
– sky plot
– site obstruction
• Satellite geometry
– list of DOP
– plot of DOP
• Simulations
– cofactor matrix
– ambiguity success rate
– optimization

11.3 Data transfer

Downloading data
Upon completion of a survey, the data stored in each receiver must usually be downloaded to a computer. There are several possibilities for downloading

these data depending on the storage device. Primarily, the files are directly copied from the receiver to a computer disk. The transferred data should be verified for consistency and integrity. This could be performed either after downloading or before decoding.

Decoding data
The downloaded data are often decoded from a special compact receiver format to a general binary format. Software that yields the navigation message (e.g., broadcast ephemerides), the measured quantities (i.e., code ranges, carrier phases, and Doppler data), and other information (e.g., field log) is used to provide files required by the processing software. An important software feature is the ability to convert the data into the RINEX format. Most data exchanges are based on this format so that data from two different types of receivers can be used for processing.

Table 11.3. Module on data transfer

Data transfer
• Downloading data
– storage device
– copying technique
– transfer checks
• Decoding data
– binary format
– RINEX format

11.4 Data processing

Baseline definition
For projects where more than a single baseline vector has been measured there are different combinations of baselines that can be computed. A sketch of the survey project should be displayed. The processing software should offer the choice of calculating single baseline vectors (with all possible combinations) or of calculating multipoint vectors (for n sites, there are $n-1$ independent baselines). User interaction for a reformulation of baselines must be possible. The ability to combine various sessions is also useful.

Ephemerides
The broadcast ephemerides must be converted to an orbit file where the coordinates of the satellites are available for selected epochs. Satellite coordinates at other epochs are normally interpolated from a few selected epochs.

In addition to the broadcast ephemerides, an option for using the precise orbit is essential for high-accuracy surveys.

Code data processing
Before a baseline vector can be computed, the coordinates of each endpoint of the line are normally first determined using the code pseudoranges. In addition to the three-dimensional positions, the receiver clock offset is obtained. Stations with known coordinates allow for time transfer. Useful modules have been designed to differentially postprocess code data from two or more sites. Software should also offer real-time DGPS option.

Phase data processing
There are numerous options available for processing phase data. The raw phases can be used or they can be combined to form single-, double-, or triple-differences, for either one or two frequencies. There is also a variety of data combinations. Dual frequency phases may be combined to form the geometry-free, the ionosphere-free, wide lane, or narrow lane signals. In addition, phase and code range or phase and Doppler combinations are possible. For high-precision real-time kinematic surveys, software should include RTK modules.

Covariance matrices
Correct modeling of the covariances can be fairly complex for various data combinations. Software should cover the most common cases such as covariances for double-differences in both the single baseline and the multipoint mode.

Parameter estimation
Optimal software should provide flexibility in selecting the parameters to be adjusted and the way that the adjustment will be performed. The following options are some that sophisticated software should contain:

- Computation of baseline vectors. This includes (rapid) static, kinematic, and pseudokinematic surveys. A choice of using the single baseline or multipoint mode should be available. As previously mentioned, the ability to combine various sessions is also useful.

- Ambiguities. In the first adjustment, the ambiguities are usually calculated as real values (i.e., float solution). Subsequently, the ambiguities are fixed as integer values using different techniques. In succeeding adjustments, the ambiguities are substituted as known values (i.e., fixed solution). The ability to bias integers one at a time can prove useful.

For RTK applications, OTF techniques should be included. A reliable ambiguity validation is important.

- Atmosphere. There should be various choices for modeling the atmosphere. In addition to a standard atmosphere model, various models for the ionosphere and the troposphere allow optimizing data processing, especially for long baselines. Sophisticated software is capable of estimating ionospheric y-biases and tropospheric zenith delays.

Table 11.4. Module on data processing

Data processing
• Baseline definition
– project sketch
– single baseline
– multibaselines
– combination of sessions
• Ephemerides
– broadcast ephemerides
– precise ephemerides
• Code data processing
– single point solution
– DGPS
• Phase data processing
– raw phase
– phase differences
– data combinations
– RTK
• Covariance matrices
• Parameter estimation
– computation of baselines
– ambiguities
– atmospheric parameters

11.5 Quality control

Data analysis
Reliable software should provide sophisticated data analysis. Gross errors should automatically be detected and eliminated. Automatic cycle slip detection and repair is essential. Plot options are an excellent tool to visualize

errors such as cycle slips. A listing of the repaired cycle slips can also be helpful for understanding the quality of the data. A color quality indicator for the baselines may account for the ambiguity resolution and the standard deviation.

Statistical data
In addition to the parameter estimation, a posteriori site variance-covariance matrices and interpoint variance-covariance matrices should be provided. Observation residuals resulting from the baseline adjustment can also be used for checking the homogeneity of the data. A plot of these residuals quickly shows outliers, and critical measurements can be isolated. Error ellipses may be displayed for a graphic representation of the baseline standard deviations.

Loop closures
Computing loop misclosures is the most direct method of performing quality control checks. The best loop software allows the user to choose arbitrary loops by specifying various routes by node identifiers.

Table 11.5. Module on quality control

Quality control
• Data analysis
– gross error detection
– cycle slip detection and repair
– listings and plots
• Statistical data
– a posteriori site covariance matrices
– interpoint covariance matrices
– residuals
– plots
• Loop closures

11.6 Network computations

Adjustment of GPS networks
The final results of a baseline vector computation are the differences in ECEF coordinates. The coordinates of points are then determined by combining these coordinate differences in a network adjustment where one (or more) point is held fixed. Recall that the residuals as well as loop misclosures are suitable quality control factors.

Datum and coordinate transformation
A seven parameter similarity transformation (i.e., Helmert transformation) is generally performed in order to shift, scale, and rotate the entire network to fixed control. Appropriate software is an essential part of GPS surveying. Datum transformations should include frequently used coordinate frames (e.g., NAD-83, NAD-27, state plane). The software should also transform ECEF coordinates to latitude, longitude, and height and map geodetic coordinates into the plane (e.g., UTM, Lambert projection, Gauss–Krüger projection).

Hybrid data combination
More flexible network adjustment programs that adjust both GPS vectors and conventionally measured angles, distances, and elevation differences are useful when GPS and conventional measurements are combined. The combination of GPS results with any other kind of data is optimally performed by the least squares collocation method.

Table 11.6. Module on network computations

Network computations
• Adjustment of GPS networks
• Datum and coordinate transformation
• Hybrid data combination

11.7 Data base management

Archiving data and survey results
The data of GPS surveys should be organized in a data base. This permits retrieval of the data for additional computation. For example, recomputation using precise orbits instead of broadcast orbits should be possible. Also, the results of the processing should be archived in the data base. This ensures an easy access to the data for any purpose. File compaction programs can be used to good advantage in archiving data.

Interface to other data bases
To interface with national control points, a plot of both the national network reference points and the project points is useful in determining which points should be selected as tie points. To interface with Geographic Information Systems (GIS), geodetic data are included as overlay. This permits the

viewing of point locations and provides users with a clear understanding of actual site locations.

Table 11.7. Module on data base management

Data base management
• Archiving data and survey results
• Interface to other data bases
– national control points
– GIS

11.8 Utilities

File editing
Occasionally, editing of a data file is necessary. For example, noisy data might be eliminated, data from fragmented files might be combined, or large files might be split into smaller files.

Time conversions
In addition to GPS time, there are other time systems desirable such as the Gregorian date, the modified Julian date, and others. Conversions from one to the other system is often desired.

Documentation of results
The final project documentation should contain a complete list of the results. Various plots are helpful to understand the survey results.

Change of default units
Sometimes it may be desirable to change the default units (e.g., meters, degrees) to user-specific units.

Processing settings
This may include a standard processing setting and a user-defined setting for optimizing results with respect to parameter estimation, processing mode, data set extraction.

Table 11.8. Module on utilities

Utilities
• File editing
• Time conversions
• Documentation of results
• Change of default units
• Processing settings

11.9 Flexibility

This section partly repeats items previously mentioned and stresses some additional aspects which may help evaluate the versatility and flexibility of a particular software package.

- Computer compatibility. There are some minimum computer requirements; however, software should run on a wide variety of computers. As far as possible, the software should be computer independent.
- Software installation. There may be a distinction between a typical (and fully automated) installation and an expert and user defined installation (to omit or include certain specific options).
- Processing time. Computation time is of course a crucial question and is (apart from computer performance) dependent on the complexity of the software. Using a single baseline as a test criterion is a good method of timing various programs.
- Real-time applications. Software and hardware combinations that are capable of operating in real time are becoming more important. These applications include real-time differential navigation and stake-out surveying. Communication links are critical to this type of GPS use.
- Memory amount.
- Limitations. What are the limits for the number of sites, satellites, and sessions?
- Data. It should be possible to add one or more data sets at a point to create a larger single data set. For example, combining two sessions at a point. An option for reformatting the observation data into the RINEX format is essential. It should also be possible to process both single and dual frequency data using the same software. Combining data from different receivers (also for a single baseline) and possibly of different satellite types is highly desirable. Software should be capable of processing several sessions in either batch or manual mode.
- Orbit. It is critical that the software is capable of using both the broadcast and precise ephemerides.
- Parameter estimation. Is it possible to select the number of parameters for the vector computation or the adjustments or is it restricted to fixed parameters?
- Transformations and map projections. Are coordinate transformations widely offered? Transformations from the global system to a local datum are necessary. Map projections should also be available.
- User interactivity. There should be an option to change the amount of (necessary) user interactivity. For normal processing, the user does not need or want much interactivity.

- User friendliness. How much time is necessary to learn how to use the software? Does it need an expert with lots of computer experience? Does it offer help routines?

Table 11.9. Module on flexibility

Flexibility
• Computer compatibility
• Software installation
• Processing time
• Real-time applications
– point positioning
– differential positioning
– relative positioning
• Memory amount
• Limitations on the number of
– sites
– satellites
– sessions
• Data
– additional data
– conversion to RINEX format
– combination of single and dual frequency data
– combination of different receiver types
– combination of different satellite types
– combination of various sessions
• Orbit
– broadcast ephemerides
– precise ephemerides
• Parameter estimation
– number of parameters
– selectable parameters
• Transformations and map projections
• User interactivity option
• User friendliness

12 Applications of GPS

12.1 General uses of GPS

The general overall uses of GPS are numerous and a single textbook cannot address all the GPS applications, especially when users create new ones almost every day. In this chapter, an arbitrary grouping of global, regional, and local uses has been selected. A further additional arbitrary classification of navigation and survey uses is made to help organize the discussion of GPS applications.

Stand-alone users of GPS achieve the limited accuracy provided by the SPS. As mentioned earlier, the positional accuracy is about 10–20 m at the 95% probability level. The official values for position, height, and time have not yet been redefined after SA has been switched off.

An accuracy at the 10 m level suffices for many applications, particularly in navigation. For higher-accuracy requirements, DGPS is adequate for local, regional and even global applications. These applications mainly cover high-precision navigation on land, at sea, and in the air. Today's navigation of civil aircraft is based on a variety of systems such as radio altimeters and inertial systems which form the flight management system. DGPS supports these systems and can also be used to supplement the Instrument Landing System (ILS) which is presently the international standard for approach and landing. An overview of the present DGPS performance is given in Table 12.1, where the individual values are chosen for mnemonic reasons. The accuracy of the heights is worse by a factor of 1.5 to 2.

Relative (static) positioning can be performed on baselines of arbitrary length up to some thousands of kilometers. In many cases, however, baselines do not exceed 20 km. For such baselines, the present achievable accuracies in horizontal position (1σ level) are listed in Table 12.2. The accuracy in height is worse by a factor of 1.5 to 2. The table is partitioned into static and

Table 12.1. Accuracies for DGPS

Observable	Station separation	Horizontal accuracy
Code range	1 000 km	<10 m
Smoothed code ranges	100 km	<1 m
Carrier phases	some 10 km	<0.1 m

B. Hofmann-Wellenhof et al., *Global Positioning System*

Table 12.2. Accuracies for relative positioning

Mode	Horizontal accuracy
Static	$\pm(5\,\text{mm} + 0.5\,\text{ppm})$
Kinematic	$\pm(5\,\text{cm} + 5\,\text{ppm})$

kinematic surveys. The accuracies are based on single frequency receivers tracking 5 satellites with reasonable geometry and under normal ionospheric conditions. Furthermore, it is assumed that the ambiguities have been resolved. The accuracies in Table 12.2 are rather conservative and the values were again chosen for mnemonic reasons. Note that reliable kinematic results are only obtained if the (once resolved) ambiguities are regularly checked using OTF techniques (Bačić et al. 1995, Wylde and Featherstone 1995).

In static mode, the session lengths can be substantially reduced with dual frequency receivers (Table 7.3); however, the positional accuracy is not improved for distances up to 20 km. For longer baselines, dual frequency receivers are required for the mitigation of ionospheric biases. The relative accuracy for baselines of 100 km is in the 0.1 ppm range. For 1 000 km baselines, 0.01 ppm accuracies are achievable.

In kinematic mode, the positional accuracy can be improved by the factor $\sqrt{2n}$ when dual frequency receivers are used and data are collected during n epochs in the stop-and-go mode. Compared to the static mode, the accuracy in the kinematic mode is worse mainly due to multipath and DOP variations. These effects are more or less averaged in the static mode.

The following numerical example is based on a 10 km baseline. In static mode, the positional accuracy of ± 1 cm, corresponding to ± 1 ppm, will be achieved with both single and dual frequency receivers. In (real) kinematic mode, the accuracy for single frequency users decreases to ± 10 cm or ± 10 ppm. However, an accuracy of about ± 2 cm is achievable if dual frequency receivers are used in the stop-and-go mode where a dozen epochs of data might be collected on each stopped position.

12.1.1 Global uses

Navigation

This was the planned primary use of GPS. Both military and civilian uses of the system in this mode are similar in that users wish to know their spatial locations as precisely as possible. For example, all types of aircraft and vessels may use GPS for en route navigation. But as aircraft are ready for landing or as vessels enter restricted waters, accuracy becomes more critical (Table 12.3). Another factor critical to aircraft navigation is the high level

Table 12.3. Current air user requirements in meters

Phase	Category	Position	Height
En route	–	≥100	≥100
Approach	I	17.1	4.1
and	II	5.2	1.7
landing	III	4.0	0.6

of integrity demanded, for example, by the Federal Aviation Administration (FAA). This agency has specified a status warning time of 30 seconds for en route navigation and 2 seconds during approach and landing. This level of integrity is achieved by either Receiver Autonomous Integrity Monitoring (RAIM) which is an internal quality control of the actual GPS solution, or by GPS Integrity Channels (GIC) which use signals from a source external to GPS. In the spring of 1994, FAA officially accepted GPS as part of the U.S. airspace system.

Surveying
Global application of GPS provides a powerful geodetic tool. This science is involved in monitoring global changes over time which is the key for understanding long-term geodynamical phenomena. Applications include measuring crustal deformations, postglacial rebound, volcanic uplift, plate tectonics, and earth rotation. In the past, Very Long Baseline Interferometry (VLBI) and Satellite Laser Ranging (SLR) techniques have been used for this purpose. GPS is not presently capable of replacing these techniques but it will be used to augment them and provide more cost-effective solutions to geodetic problems (Beutler 1996b).

Timing and communications
Another global use of GPS is in global determination of accurate time. High accuracy timing has many scientific applications such as coordinating seismic monitoring and other global geophysical measurements. Inexpensive GPS receivers operating on known stations provide a timing accuracy of about 40 nanoseconds (95% probability level) with only one satellite in view. With more sophisticated techniques, one can globally synchronize clocks even more precisely and an achievable accuracy of 1 nanosecond is considered possible.

Global communication depends upon precise coordination to pack as many data bits into a given period of time as possible. GPS should make it possible to increase the overall efficiency of communication, allowing many more users per unit of time than is presently possible.

12.1.2 Regional uses

Navigation
The regional applications of GPS navigation are enormous and include: exploration, transportation management, structural monitoring, and various types of automation.

Using code ranges, accuracies down to the meter level can be achieved by differential navigation, i.e., DGPS. This involves placing fixed GPS receivers at known locations in a project area and broadcasting range corrections to receivers within the area. The U.S. Coast Guard (USCG) has established a national coastal differential navigation network to assist vessels approaching the coast. Apart from coastal waters, the network also serves navigable inland waters such as lakes and rivers. The maritime DGPS service will be expanded to a nationwide DGPS service to cover all surface areas. Additionally, private firms have set up similar networks in heavily trafficked areas and sell a precise navigation service.

There are numerous uses of GPS in managing regional resource and controlling activities within large areas. One example is the Geographic Information System (GIS) which is a computer based geographically oriented data base that helps manage resources on both regional and local levels. The GIS also provides a means for making rapid informed decisions for planning, development, and tracking of infrastructure. Normally, a GIS is initialized by performing base mapping to create a digital data file that contains all map information. The digital file usually contains roads, buildings, vegetation types, property lines, soil types, and other critical data which can be displayed in various overlays on a computer screen. GPS often plays a key role in providing ground control for the base mapping. The coordinates of photo-identifiable points are determined by GPS, and these points are used to control the scale and orientation of the map. GPS can also be used to update GIS information and enter new information in the file. This is accomplished by placing the GPS receiver over the point one wishes to include in the GIS to determine its location and then keying in an object code for the item so that it is tagged with time and coordinates.

The GIS is an excellent example of the use of GPS in its various modes: (1) point positioning, (2) differential positioning, (3) precise surveying, and (4) combinations of the methods.

The point positioning mode would be used to help locate existing survey marks (or photo-targets) whose coordinates had been previously determined, or whose coordinates had been scaled from a large-scale map. Inexpensive hand-held receivers are being used to locate control points that have defied recovery because land features have greatly changed. Also, this method

facilitates locating property corners whose coordinates have been computed from surveyed bearings and distances from a known point.

Next, the precise survey mode is used to measure the vectors between the known control points (both horizontal and vertical) and selected photo-targets or photo-identifiable points. These points are used in the photogrammetric map compilation to scale and orient the photograph to the chosen datum. The static, pseudokinematic, and kinematic modes could all be used during this phase of the work. The kinematic method has also been used to determine the coordinates of the photocenter by placing one GPS receiver in the aircraft and a second on a ground datum point. For this application, dual frequency receivers being able to resolve cycle slips or interruptions to the received signal during flight are useful. This is accomplished by using smoothed code pseudoranges to compute the vector between fixed and airborne receiver using the differential navigation technique. The integer bias of the wide lane phase and, finally, the integer ambiguity of the base (19 cm) carrier is resolved to reinitialize the airborne receiver on-the-fly. This technique significantly reduces the amount of ground control needed to produce a metric map.

After the map has been produced and the GIS data base is initialized, changes and corrections often must be made to update the data base. It would be inordinately expensive to refly and remap the area, so that GPS can further be used to determine the coordinates of additional points to be placed in the data base. For example, instrumented vans are being used to collect information as they drive along the highway. Videocameras and other devices collect data while GPS keeps track of the position of the van, and, thus, the various features are collected. Several state highway departments are using this technique to develop a data base of all important road features so they can better manage their resources.

Surveying

Geodesists have long desired to measure crustal movements for various scientific purposes. One use would be to predict earthquakes by measuring certain precursor ground movement. GPS is an ideal tool for such studies in that the equipment is relatively inexpensive, portable, and highly accurate. For example, the NGS has measured an array of widely spaced points in the eastern part of the U.S. (Eastern Strain Network) using dual frequency GPS receivers. This array or network will be periodically resurveyed so that crustal movement can be measured. Typically, this information can be used to provide nuclear plant siting information as well as providing a general purpose high accuracy network. The NGS network is tied into or referenced to GPS receivers located at VLBI tracking sites, so the Eastern

Strain Network also provides a multipurpose high precision network. Various other groups have measured similar networks to determine both continental and island crustal motion throughout the world. For example, the Japanese have established an array of permanent GPS tracking stations to provide near real-time crustal motion monitoring. Another example is the geodynamic goal of the Austrian GPS Reference network presented by Stangl et al. (1991).

GPS is the first measuring device that can be used to accurately measure height differences in real time. Studies by the NGS have shown that repeat measurements between stable and subsiding points provide accurate measures of the subsidence. Industry has used GPS to measure the subsidence among groups of offshore oil platforms by making repeat surveys. One example is the subsidence measurement of several North Sea oil platforms with respect to other platforms that were thought to be in a more stable area. Monthly GPS surveys showed measurable height changes, and trends in the results positively proved which platforms were subsiding (relative to other platforms). In this particular case, GPS was probably the only method to make the determination.

As GPS equipment decreases in price, it is expected that receivers will be permanently mounted on structures and at selected sites and used to measure both subsidence and crustal deformation.

12.1.3 Local uses

The difference between regional and local uses of GPS is quite arbitrary because a small state GIS could either be considered regional or local in scope. Also, the difference between navigation and surveying will disappear since navigation has become more accurate. Nevertheless, the grouping is kept for convenience.

Navigation
As stated earlier, the air user accuracy requirements are fairly high during approach and landing (Table 12.3). In order to achieve these accuracies in the navigation mode, fixed receivers must broadcast range corrections (and carrier phases) to incoming aircraft so that they can compute more accurate positions (by DGPS) as they approach the runway. GPS-aided approach and landing will be an economical answer to airport control.

The use of GPS in its point positioning mode is becoming popular for emergency vehicle management. For example, if emergency vehicles are equipped with a C/A-code pseudoranging receiver and a radio transceiver, the position of each vehicle can be determined by GPS as it travels to a des-

ignated area, and the transceiver sends this position to a central dispatcher. The location of all vehicles is then displayed on a screen so that at any given instant the dispatcher can see where his resources are positioned. If, for example, one of the roads is blocked, the dispatcher can reroute the vehicle or he can dispatch a closer vehicle. There is little doubt that such GPS-based Automatic Vehicle Location (AVL) systems (Krakiwsky 1991) will be installed in all major cities in the world in the near future.

Also for farming and forestry, positional information is required. GPS point positioning or DGPS facilitate seed planning, fertilization, plant protection, and automated tractor guidance (Petersen 1991, Auernhammer et al. 1994, Stafford and LeBars 1996).

Surveying

The GIS discussed in the previous section is also being developed for cities and towns so this use will not be covered further. A typical local use of GPS is local property and site survey. Some surveyors are using GPS to place all projects surveyed on a single datum (e.g., NAD-83). An economical way to do this is to place an antenna at a fixed central location (e.g., office) and to determine the precise coordinates of this point by measuring vectors from the nearby fixed control points. After the coordinates of the fixed central site have been determined, one receiver is left there and is run continuously during those times where field crews are engaged in survey activities. Each survey crew takes one or two additional receivers and places them on points in the local scheme being surveyed. For example, a topographic site survey and boundary determination requires the establishment of many temporary and permanent points all connected by conventional survey measurements. Two GPS receivers placed on two widely spaced points in this local scheme could collect data during the time the survey crew made conventional measurements with little additional manpower cost. Upon completion of the conventional survey and simultaneous GPS measurements, the coordinates of all points in the scheme would be determined and also all bearings in the scheme would be precisely related to true north.

An additional local application of GPS is to perform topographic surveys by using the kinematic mode. In areas of relatively few obstructions, a roving GPS receiver can either be carried or placed on a vehicle (e.g., all terrain vehicle) and the terrain traversed by a series of cross sections. The horizontal position and the height of points can typically be determined every second, so a high density of point determinations will result even when the receiver is vehicle mounted and travels at high speed. The processing and plotting of the kinematic cross sections are automated so that the field to finish time can be minimized.

Several organizations are using GPS kinematic surveying to determine the coordinates of the photocenter during aerial mapping flights. There is also consideration being given to using three or more antennas connected to the GPS receiver to determine the orientation angles of the aircraft. Both aspects are treated in more detail in the following two sections.

12.2 Attitude determination

12.2.1 Theoretical considerations

Attitude is defined as the orientation of a specific body frame attached to a land vehicle, ship or aircraft with respect to a reference frame which is usually a north, east, and up local level frame.

The parameters used to define three-dimensional attitude are r, p, y, the angles for roll, pitch, and yaw (or heading). In the case of an aircraft, the roll angle measures the rotation of the aircraft about the fuselage axis, the pitch angle measures the rotation about the wing axis, and the yaw angle measures the rotation about the vertical axis. If all three angles are zero, the aircraft is flying exactly straight and level, and heading on course (Graas and Braasch 1992). Similar reference frames can be developed for other vehicles.

Traditionally, attitude parameters are derived from inertial navigation systems or other electronic devices. With the advent of low-cost high-performance GPS sensors, multiantenna systems which integrate three or more antennas in a proper configuration provide an alternative and cost-effective means to obtain reliable and accurate platform attitude information.

The plane of a rigid platform in space is defined by three antennas A_i in a proper configuration. Two independent baselines provide six equations to determine the three attitude parameters. Thus, three conditions exist: the length of the baselines and the spatial angle between them are invariant against attitude variations.

For the mathematical formulation of three-dimensional attitude determination, the position vector of an antenna A_i in the local level frame is denoted by $\underline{x}_i$ and the corresponding position vector in the body frame by $\underline{x}_i^b$. The vector $\underline{x}_i$ results from relative positioning with GPS carrier phase measurements where the accurate geocentric coordinates of the antennas with respect to each other are transformed to the local level system. The body frame may be realized by the position vectors $\underline{x}_i^b$ of three antennas A_i as shown in Fig. 12.1.

The three attitude parameters correspond to the three Euler angles rotating the local level frame into the body frame. Thus, attitude determination

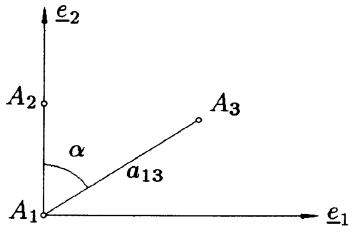


Fig. 12.1. Definition of the body frame

is defined by

$$\underline{x}_{ij}^b = \underline{R}_2\{r\}\,\underline{R}_1\{p\}\,\underline{R}_3\{y\}\,\underline{x}_{ij} \tag{12.1}$$

where relative vectors (i.e., baseline vectors) $\underline{x}_{ij}$ and $\underline{x}_{ij}^b$ have been introduced instead of absolute vectors $\underline{x}_i$ and $\underline{x}_i^b$ to eliminate the shift vector. The three consecutive rotations can be expressed by a single rotation matrix $\underline{R}\{r,p,y\}$. Explicitly

$$\underline{R}\{r,p,y\} = \begin{bmatrix} \cos r\cos y & \cos r\sin y & -\sin r\cos p \\ -\sin r\sin p\sin y & +\sin r\sin p\cos y & \\ -\cos p\sin y & \cos p\cos y & \sin p \\ \sin r\cos y & \sin r\sin y & \cos r\cos p \\ +\cos r\sin p\sin y & -\cos r\sin p\cos y & \end{bmatrix} \tag{12.2}$$

is obtained (Lachapelle et al. 1994).

Direct computation of attitude

The attitude parameters can directly be computed from the local level coordinates without knowledge of the body frame coordinates (Lu et al. 1993). The yaw y and pitch p (corresponding to azimuth and elevation angle) are determined from one baseline only. Selecting for example the baseline between A_1 and A_2 (Fig. 12.1) the relations

$$\begin{aligned} \tan y &= e_{12}/n_{12} \\ \tan p &= \frac{u_{12}}{\sqrt{e_{12}^2 + n_{12}^2}} \end{aligned} \tag{12.3}$$

are obtained where the baseline components are introduced in the north, east, and up local level frame.

In order to obtain roll r, the baseline vector $\underline{x}_{13}$ is first rotated by the yaw and pitch resulting in the vector $\underline{x}^r_{13}$. A third rotation (i.e., roll) rotates this vector to the body frame yielding

$$\begin{bmatrix} a_{13} \sin\alpha \\ a_{13} \cos\alpha \\ 0 \end{bmatrix} = \begin{bmatrix} \cos r & 0 & -\sin r \\ 0 & 1 & 0 \\ \sin r & 0 & \cos r \end{bmatrix} \begin{bmatrix} e^r_{13} \\ n^r_{13} \\ u^r_{13} \end{bmatrix} . \tag{12.4}$$

The roll can be computed from the third line of the above equation by

$$\tan r = -u^r_{13}/e^r_{13} . \tag{12.5}$$

The direct computation method is a function of three antennas only and (as mentioned) does not require a priori knowledge of the body frame. The disadvantage of the method is that redundant antennas are not used.

Least squares estimation of attitude

Least squares estimation of attitude is based on the knowledge of the body frame coordinates of the antennas. These coordinates may be obtained through a survey or initialization process. The GPS-derived local level coordinates can then be treated as observations.

In a slightly different form, Eq. (12.1) may be written as

$$\underline{x}^b_{ij} = \underline{R}\{r, p, y\}\, \underline{x}_{ij} \tag{12.6}$$

where the three rotation matrices are combined into one matrix. Each baseline with known body frame coordinates and GPS-derived local level coordinates gives rise to three equations.

In the most general case, the matrix $\underline{R}$ could be replaced by a matrix where all nine elements are unknown. This corresponds to an affine transformation. At least three baselines (or four antennas) are needed to solve for the nine unknowns.

In the case of rigid platforms, similarity transformations are sufficient. The minimum of three antennas forming two (non-collinear) baseline vectors provide six equations for the three unknown attitude parameters. Thus, the problem is redundant and can be solved by least squares estimation. Common least squares estimation requires linearization of Eq. (12.6) with respect to the attitude parameters. Another approach is the generalized inversion of Eq. (12.6). Denoting n as the number of independent baselines, Eq. (12.6) may be written in the form

$$\underline{A}^b = \underline{R}\{r, p, y\}\, \underline{A} \tag{12.7}$$

where $\underline{A}^b$ is a $3 \times n$ matrix with the baseline vectors in the body frame as column vectors. Analogously, matrix $\underline{A}$ contains the baseline vectors in the local level frame. Solving for the rotation matrix involves multiplying Eq. (12.7) by the (generalized) inverse of the matrix $\underline{A}$. The result is given by (Graas and Braasch 1991)

$$\underline{R} = \underline{A}^b \underline{A}^T (\underline{A}\,\underline{A}^T)^{-1} \tag{12.8}$$

where the superscript T indicates transposition. Denoting the elements of the matrix $\underline{R}$ by R_{ij}, the attitude parameters are obtained by

$$\begin{aligned} \tan r &= -\frac{R_{13}}{R_{33}} \\ \tan p &= \frac{R_{23}}{\sqrt{R_{21}^2 + R_{22}^2}} \\ \tan y &= -\frac{R_{21}}{R_{22}} \end{aligned} \tag{12.9}$$

which may be verified by inspecting the matrix $\underline{R}$, cf. Eq. (12.2).

The advantage of attitude estimation by the least squares technique is that the computation of the attitude is more rigorous since it computes the best estimate for the attitude parameters.

Modeling of wing flexure

If four antennas are mounted, for instance one on each wing tip of an aircraft and the remaining two atop the fuselage, then the roll angle could be determined independently from heading and the pitch angle. In practice, however, the movement of the aircraft's wings during flight may completely prevent the precise determination of the roll angle in this way.

Another technique is reported by Cannon et al. (1994). The nonrigidity of the body frame is taken into account by modeling the wing flexure. Constraining the wing flexure in the vertical component of the body frame leads to the relation

$$\underline{x}_i^b = \underline{x}_{0i}^b - f\,\underline{e}_3 \tag{12.10}$$

where f is a scalar which is estimated in the least squares adjustment. The position vectors $\underline{x}_{0i}^b$ refer to the situation without wing flexure (aircraft at rest) and can be measured directly using a theodolite or can be determined by static GPS observations prior to takeoff.

12.2.2 Practical considerations

An accuracy of ± 1 mm in the relative position of the antennas corresponds to one milliradian or 0.057 degrees in attitude accuracy for an antenna separation of 1 m. For larger antenna separations, the angular accuracy increases. However, for longer baselines, the search for the phase ambiguities becomes more difficult.

Multipath is a significant limiting error source in attitude determination. This effect can be reduced by installing a common groundplane for all antennas or by the use of choke ring antennas.

The key requirement to attitude determination is ambiguity resolution on-the-fly. This is particularly valid for airborne or marine applications. Ambiguity resolution can be speeded up by special antenna configurations as proposed for example by El-Mowafy (1994). Another approach is to incorporate geometric constraints due to the known antenna array geometry (Landau and Ordóñez 1992, Lu 1995). More details on these techniques are given in Sect. 9.2.

An example of a manufactured multiantenna array dedicated to attitude determination is discussed by Quin et al. (1992). This platform consists of four microstrip antennas connected to a 24-channel single frequency receiver. Six channels are assigned to each antenna so that up to six satellites may be tracked. One antenna is used to determine position and velocity of the platform. The baselines between the primary antenna and the other three antennas are used for attitude determination. The redundant fourth antenna allows internal checks. In the case of such a dedicated or self-contained system, single-differencing is possible since the receiver clock offset is common to all antenna measurements. This adds a degree of freedom to each baseline solution and means that only 3 satellites are required to solve the interantenna vector.

The key point for the nondedicated system (i.e., 4 independent receivers) is that off-the-shelf receivers can be used which provides flexibility since these receivers can then be used for other applications. In this case, double-differencing must be applied to account for the receiver clock offsets.

12.3 Airborne GPS for photo-control

Now that the constellation provides 24-hour coverage, GPS is widely used in aircraft (and other vehicles) to provide positions for photographs. Several government organizations as well as private firms now routinely use GPS for precise positioning. The development of accurate (10 cm) C/A-code receivers and dual-band full-wavelength receivers has accelerated the use of GPS for

decimeter accuracy positioning.

The most simple use of the system is to perform DGPS using the C/A-code pseudorange. This technique is extremely robust in that temporary loss of lock does not affect the accuracy of subsequent positions. Furthermore, initialization on the ground prior to flight is unnecessary. Submeter accuracies are obtainable (with proper receivers and software) and provide adequate control for small-scale mapping and control of orthophoto maps.

The more recent and precise positioning method for photo-aircraft is to use dual frequency full wavelength receivers and software that provides on-the-fly resolution of the integer ambiguities. Because the full wavelength carrier phase is available on both frequencies, the 86 cm wide lane observable can be formed. A few epochs of precise code range data combined with Kalman filtering permit the solution of the initial vector between the fixed base station and the roving airborne receiver. This knowledge is then used to define a search box which contains the solution of the wide lane ambiguities. The smaller the search box, the faster the ambiguity is resolved. Once the wide lane ambiguity has been resolved, the L1 ambiguity can be quickly found. Factors which affect the speed and reliability of this method are:

- pseudorange bias and noise,
- carrier phase bias and noise,
- multipath,
- baseline length,
- satellite geometry.

Although the on-the-fly software can be initialized while the aircraft is in motion, the initial integer resolution can be accomplished more rapidly with the aircraft parked on the ground. Normally, only a minute or two are required for this initialization. If the airborne receiver then tracks the satellites during takeoff and climbing, there is a greater chance of success of the mission.

The GPS receiver must be capable of recording the precise time of each aerial exposure. Specially equipped aerial cameras provide a time pulse for the precise instant of exposure of the photograph. This time is then recorded by the GPS receiver so that the coordinates at that instant can be interpolated for the adjacent regular GPS receiver epochs. Normally, receivers are set to record data every 0.5 to 2 seconds. The shorter interval provides more data so that the software can recover more rapidly from temporary loss of lock, and also provides more accurate photo positions since the aircraft trajectory is more correctly defined. A short data rate can fill up the memory during a photographic mission day so that for longer days or briefer epochs an external recording device such as a laptop computer is maybe needed.

The best position for the airborne GPS antenna is directly over the camera. This configuration reduces the error that can occur due to an incorrectly applied offset. A number of research studies investigates the use of multiple antennas and receivers to provide both redundant observations and aircraft attitude information (Blankenberg 1994, Cooper et al. 1994). At present, the best method of determining the attitude of the aircraft is to compute the roll, pitch, and yaw from the photogrammetric block adjustment.

A number of photogrammetric analytical block adjustments have been adapted to accept the airborne GPS data. In addition to providing aircraft attitude information, these adjustments can be used to verify the GPS coordinates. For example, blocks with fixed (ground) control at the corner points will flag any photo-position that is in error. Typically, agreement between the block solution and the GPS positions is at the 5 to 10 centimeter level.

The solution of airborne GPS heights is performed using the ellipsoidal height of the base (ground) station so that all airborne heights are in the ellipsoidal system. Elevations relative to the geoid are normally desired so that the geoidal heights must be applied to the airborne heights (along with the distance from the antenna to the camera). The most accurate method of determining the local geoidal height is to make a ground survey where vertical control points are occupied. The locally derived geoidal heights can then be applied to the appropriate photo-heights to obtain the proper elevation.

GPS controlled photography is also suitable for large-scale mapping. In this instance, greater care should be exercised to ensure that the derived positions meet the desired accuracy requirements. In particular, the ground base station from which the airborne positions are derived should be located as close to the center of the project area as possible. This location will reduce the distance between the aircraft and the base station during the actual photography resulting in the greatest precision (due to reduction of unknown ionospheric and tropospheric errors). Furthermore, this base station site will provide the most accurate determination of the geoidal height. As a practical matter, it is best to have two ground stations to provide redundancy of data. One good scenario is to locate one receiver on the project site and the second receiver at an airport near the site. Furthermore, ground sites should be free from obstructions and in low-multipath environments.

Processing of airborne data is similar to ground kinematic data. With on-the-fly software it is possible to process the data both forward and backward so that the optimum solution of the data can be made. The two solutions can be either averaged or combined in an optimum way to give the best solution. Also, when two ground stations are used, these solutions may be combined to give optimal results.

The GPS airborne control method is extremely cost effective in that a great portion of local control needed for photo-mapping can be eliminated. In many areas where access to the site is limited, the airborne control method will make difficult mapping tasks routine.

12.4 Interoperability of GPS

12.4.1 GPS and Inertial Navigation Systems

One of the main disadvantages of using GPS for surveying, land navigation, and feature coordinate tagging is the temporary outage as the receiver passes under bridges or other obstructions. For example, a highway data collection van would experience loss of positioning when the van passed under trees overhanging the highway. Various devices have been used to compensate for this disadvantage. One simple dead reckoning system to use is an odometer and a (flux gate) magnetic compass to assist in interpolating the position of vehicles during satellite outages (Byman and Koskelo 1991). More sophisticated approaches involve combining GPS with Inertial Navigation Systems (INS).

Basically, a (strapdown) INS consists of two components: (1) gyros to monitor the angular motion of the body frame axes with respect to the local level frame and (2) accelerometers placed on the body frame axes to measure accelerations (i.e., velocity rates). Starting from a known position and integrating (corrected) accelerations twice over time yields differences in position which determine the trajectory of the vehicle. This simple principle becomes fairly complicated in practice, and the reader is referred to the voluminous literature (Schwarz and Lachapelle 1990).

The error of an (unaided) INS generally increases with the square of time due to the double integration over time. What are the advantages of INS compared to GPS? First, inertial systems are autonomous and independent of external sources. Also, there is no visibility problem. Second, INS provides accuracy similar to GPS when used over short time intervals with data rates of generally 64 Hz. Hence, INS can serve as an interpolator for GPS gaps.

When GPS is combined with a precise INS, it is possible to reinitialize the GPS integer cycle count after a temporary loss of lock. For example, it is critical to maintain continuous count of integer cycles when performing a kinematic survey with C/A-code receivers. Accurate inertial systems could overcome the loss of cycle count problem by being able to predict an accurate change in carrier phase and, thus, cycle count during signal outages or massive cycle slips (Sect. 9.1.2).

12.4.2 GPS and GLONASS

Comparison of GPS with GLONASS
The GLONASS is the Russian equivalent of GPS. The systems differ in terms of the control segment, the space segment, and the signal structure, cf. Table 12.4 after Lichtenegger (1999).

GLONASS is operated by the Russian Military Space Forces (MSF). So, both systems are under military control. They apply different reference frames as well as different time references. The GLONASS coordinates are based on the Russian Parameter of the Earth 1990 (PE-90) System and the reference time is UTC(SU) as maintained at the Russian Institute of Metrology for Time and Space (Glavkosmos 1991). Unlike GPS, the reference time of GLONASS is not continuous since leap seconds are inserted.

The space segments are very similar. Nominally, GLONASS is composed of 24 satellites equally spaced in three orbital planes, with 64.8 degree nominal inclination. The nearly circular orbits with altitudes of 19 100 km and periods of 11.25 hours are similar to GPS. The higher inclination of

Table 12.4. GPS versus GLONASS

	GPS	GLONASS
Control Segment		
operator	DoD	MSF
tracking network	global	regional
reference frame	WGS-84	PE-90
time reference	UTC(USNO)	UTC(SU)
FOC	July 17th, 1995	January 18th, 1996
Space Segment		
number of satellites	24	24
SV identification	PRN number	carrier number
mean altitude	20 200 km	19 100 km
orbital planes	6	3
SV spacing	uneven	even
inclination	55°	65°
ground track repetition	1 sidereal day	8 sidereral days
Signal Structure		
signal separation	CDMA	FDMA
carrier frequencies	constant	variable
PRN codes	variable	constant
navigation message	orbital parameters	positions, velocities and accelerations

GLONASS orbits increases the accuracy and availability at higher latitudes. Note that the present number of GLONASS satellites is far below its nominal value.

A substantial difference of the two systems concerns the signal structure. GPS uses Code Division Multiple Access (CDMA) where every satellite transmits the same two carriers which are modulated by PRN-codes specifically assigned to each satellite. By contrast, GLONASS uses Frequency Division Multiple Access (FDMA) where to each satellite two individual carrier frequencies are assigned. The PRN-codes, however, are the same for all satellites.

The individual carrier frequencies f_{L1}^j and f_{L2}^j are defined by

$$\begin{aligned} f_{L1}^j &= 1602.0000 \text{ MHz} + j \cdot 0.5625 \text{ MHz} \\ f_{L2}^j &= 1246.0000 \text{ MHz} + j \cdot 0.4375 \text{ MHz} \end{aligned} \tag{12.11}$$

where j is the carrier number assigned to the specific satellite (e.g., Kleusberg 1990b). The remarkable relation

$$\frac{f_{L1}}{f_{L2}} = \frac{9}{7} \tag{12.12}$$

holds for all satellites j. It is worth noting that the same carrier number is used for certain antipodal satellites (CGSIC 1994, Daly and Misra 1996). As for GPS, two codes are modulated onto the carriers. The free accessible C/A-code with a period of 1 ms has a chip rate of 0.511 MHz and is modulated onto L1 only. The P-code with a period of 1 s has a chip rate of 5.113 MHz, is modulated onto both carriers, and is also freely accessible. Unlike GPS, there is no provision to encrypt the P-code. Note that the chipping rates for GLONASS are approximately two times lower than the corresponding values for GPS.

The structure of the navigation message is different from GPS. Every half hour the satellites transmit their geocentric position, velocity and acceleration vectors. The position vector must be interpolated by the user for observations between these epochs (Stewart and Tsakiri 1998).

Information on the system status for civil users is provided by the Coordination Scientific Information Center of the Ministry of Defence of the Russian Federation in the form of biweekly reports. Another source of information is the Web site at the location http://mx.iki.rssi.ru/SFCSIC.

Integration of GPS and GLONASS

It should be noted that GLONASS has some advantages compared to GPS. Currently, GLONASS signals experience no A-S degradation. Therefore,

dual frequency data are available to the public. All GLONASS satellites are equipped with retroreflectors for laser tracking, while at this time only a few GPS satellites have got this feature.

The GLONASS constellation provides a like number of satellites which can be used in combination with the GPS satellites. Today, receivers are produced with dual GPS/GLONASS capability. These multichannel receivers are capable of tracking any combination of GPS and GLONASS satellites. The main benefit to surveyors is the increase in sky coverage, making it possible to measure baselines that would otherwise be obstructed. Also, the increased number of measurements makes the kinematic method much more viable than it is with just the GPS constellation. Similarly, on-the-fly kinematic surveying with an inexpensive C/A-code receiver (rather than the more expensive dual frequency receivers) can be used for most present land surveying tasks such as boundary, topographic, hydrographic, and construction surveys.

The drawbacks of GPS/GLONASS integration originate from the system differences as listed in Table 12.4. The transformation of the two reference frames is performed by a similarity transformation with given parameters. The offset UTC(USNO) – UTC(SU) is normally taken into account as an additional unknown. If GLONASS phase measurements should be included in a combined solution using double-differences, one has to consider that each satellite uses its own carrier frequency. The model equations for carrier phases measured at the two sites A and B to the satellite j are given by

$$\begin{aligned} \Phi_A^j(t) + f^j\,\delta^j(t) &= \frac{1}{\lambda^j}\,\varrho_A^j(t) + N_A^j + f^j\,\delta_A(t) \\ \Phi_B^j(t) + f^j\,\delta^j(t) &= \frac{1}{\lambda^j}\,\varrho_B^j(t) + N_B^j + f^j\,\delta_B(t) \end{aligned} \tag{12.13}$$

and differ from those in Eq. (8.30) by the superscripct j assigned to the satellite-specific wavelength. The measured phases are scaled to ranges by

$$\tilde{\Phi}^j(t) = \lambda^j\,\Phi^j(t) \tag{12.14}$$

where $\tilde{\Phi}^j(t)$ is output by the receiver. The single difference of the two equations (12.13) leads to

$$\tilde{\Phi}_B^j(t) - \tilde{\Phi}_A^j(t) = \varrho_B^j(t) - \varrho_A^j(t) + \lambda^j\,[N_B^j - N_A^j] + c\,[\delta_B(t) - \delta_A(t)] \tag{12.15}$$

with $c = \lambda^j\,f^j$ being the velocity of light. Introducing the shorthand notations of Sect. 8.3.1, i.e., symbolically $*_{AB}^j = *_B^j - *_A^j$, a more compact form is achieved by

$$\tilde{\Phi}_{AB}^j(t) = \varrho_{AB}^j(t) + \lambda^j\,N_{AB}^j + c\,\delta_{AB}(t)\,. \tag{12.16}$$

Assuming two satellites j, k gives rise to two single-differences (12.16). From these the double-difference

$$\tilde{\Phi}^k_{AB}(t) - \tilde{\Phi}^j_{AB}(t) = \varrho^k_{AB}(t) - \varrho^j_{AB}(t) + \lambda^k N^k_{AB} - \lambda^j N^j_{AB} \quad (12.17)$$

is obtained. After introducing again shorthand notations, cf. (8.38), symbolically $*^{jk}_{AB} = *^k_{AB} - *^j_{AB}$, there results

$$\tilde{\Phi}^{jk}_{AB}(t) = \varrho^{jk}_{AB}(t) + \lambda^k N^k_{AB} - \lambda^j N^j_{AB} \quad (12.18)$$

which may be rearranged by "adding zero" in the form of $-\lambda^k N^j_{AB} + \lambda^k N^j_{AB}$ so that

$$\tilde{\Phi}^{jk}_{AB}(t) = \varrho^{jk}_{AB}(t) + \lambda^k N^{jk}_{AB} + N^j_{AB}(\lambda^k - \lambda^j) \quad (12.19)$$

is finally obtained. This equation differs from the double-difference equation (8.37) by the "single-difference bias" $b_{SD} = N^j_{AB}(\lambda^k - \lambda^j)$. The unknown single-difference N^j_{AB} can be estimated from single-point positioning with an accuracy of about 10 m corresponding to 50 cycles. The wavelength difference for two GLONASS carriers separated by one carrier number corresponds to 0.000 351 cycles, cf. Eq. (12.11). Thus, the result $b_{SD} = 0.02$ cycles is obtained. This shows that for small differences in the carrier numbers b_{SD} acts as a nuisance parameter. For larger numbers, iterative processing has been proposed. In the first step, only GPS satellites or only GLONASS satellites with the smallest differences in the carrier numbers are considered. Hence, the double-differenced ambiguities of these satellites can be resolved and an improved position is obtained leading to a more accurate estimation for N^j_{AB}. The procedure is then continued and stepwise extended to all satellites until all ambiguities have been resolved. More details on this subject can be found in Habrich et al. (1999), Han et al. (1999).

International GLONASS Experiment (IGEX-98)
This project was co-sponsored by the International Association of Geodesy (IAG), the International GPS Service (IGS), the U.S. Institute of Navigation (ION), and the International Earth Rotation Servive (IERS).

The goal of the experiment was the continuous collection of GPS/GLONASS data from a global network of stations, the comparison of hardware and software, the determination of precise GLONASS orbits, and the determination of the transformation parameters between PE-90 and WGS-84 and the time difference between UTC(SU) and UTC(USNO).

The first phase of the experiment was performed between October, 1998 and April, 1999. More than 60 stations (with ITRF coordinates) in over 25 countries participated. Single and (mostly) dual frequency GPS/GLONASS

receivers were used. In addition, Laser tracking was performed in 30 SLR stations around the world.

The most impressive result was the determination of precise GLONASS orbits with improvements from several tens of meters to several decimeters (Weber and Fragner 1999). In order to further refine the results, the experiment has been extended and the second phase will run at least until April, 2003. A detailed report on the organization, preliminary results and future plans of IGEX-98 are given in Slater et al. (1999).

The joint use of GPS and GLONASS is the basis for the future Global Navigation Satellite System (GNSS). This system has some major advantages in terms of improvements in navigation performance, i.e., accuracy, availability, continuity, and integrity.

12.4.3 GPS and other sensors

Several manufacturers have combined GPS and conventional survey instruments. Today, the state-of-the-art surveying instruments measure angles and distances electronically so that the addition of GPS capability fits well with the total survey concept. Such a system is particularly useful once a network of active control points is established. In this case, continuous accurate positioning is possible by computing vectors to the surrounding active control points.

Meanwhile, GPS receivers have been placed on low-altitude satellites with various other sensing devices such as laser altimeters and synthetic aperture radar. Inclusion of GPS to determine positions of various events logged by such equipment add to the data's value and in most cases reduce the costs of data collection.

One special use is to include GPS capability in earth sensing satellites. On-board GPS receivers permit satellites to be accurately positioned with much less effort than the former tracking techniques. For example, the utility of imaging satellites such as the French SPOT satellite was greatly improved by an on-board GPS receiver to determine the precise position of each image.

12.4.4 GPS and the Federal Radionavigation Plan

The U.S. Department of Transportation (DoT) operates navigation systems for both civil and military use, whereas the U.S. Department of Defense (DoD) mainly operates systems for national defense.

In 1980, DoD and DoT jointly produced the Federal Radionavigation Plan (FRP) in order to select a mix of common-use navigation systems. These efforts should increase accuracy, coverage and reliability, while costs are reduced. Since 1980, DoT is conducting open meetings for all users

of radionavigation systems provided by the U.S. Government to revise and update the FRP every two years. The version currently in effect is the 1999 FRP which had a one-year delay.

Today, seven different systems are considered in the FRP. These are GPS (including augmentations), the Long Range Navigation System LORAN-C, the Very High Frequency Omnidirectional Range and Distance Measuring Equipment (VOR/DME), its military counterpart named Tactical Air Navigation (TACAN), the Instrument Landing System (ILS), the Microwave Landing System (MLS), and radiobeacons. For details about these systems, the reader is referred to Schrödter (1994), Department of Defense and Department of Transportation (2000).

It is also the responsibility of the FRP to study the feasibility of replacing some of the existing systems by GPS, particularly since Full Operational Capability (FOC) has been declared. For short-range navigation such as airport-approach and landing, DGPS is appropriate.

12.5 Installation of control networks

GPS is, in fact, a geodetic tool in that it provides precise vector measurements over long as well as short distances. Virtually all GPS processing software employs a three-dimensional model, and results are given in geodetic latitude, longitude, and ellipsoidal height in the WGS-84 which for vectors is (virtually) the same as NAD-83. If the GPS datum does not correspond to the national datum, a three-dimensional similarity transformation must be performed. Afterwards, the ellipsoidal values are mapped onto a projection so that land surveyors who are more familiar with plane coordinates can use the survey results.

GPS provides reference datums by two methods: (1) the passive control networks and (2) the active control networks.

12.5.1 Passive control networks

Virtually all civilized areas have some type of geodetic control networks generally surveyed by triangulation, traverse (or a combination of the two methods), and by spirit leveling. In many cases, the horizontal control network is separate from the vertical control network with perhaps 10% of these points in common. These control networks were used to scale and orient maps to provide a unified reference frame for large-scale projects, and to provide a common datum for property surveys. Geodetic control networks basically serve two functions: (1) they provide a "seamless" datum spanning large areas, and (2) they provide a reference framework which appears to be er-

rorless to its users (surveyors). The apparent errorless aspect of a geodetic network is accomplished by employing state-of-the-art instrumentation and using the utmost care in making observations. Historically, geodetic survey instruments were manufactured by specialists, and observers were rigorously trained in their use. Horizontal accuracies of 1:250 000 of the distance and vertical accuracies on a few millimeters per kilometer were typically obtained.

The advent of GPS changed the concept of what constituted a geodetic control network. The first surveys performed in 1983 were carried out with internal accuracies approaching 1 ppm, and disagreed with the existing control networks by 10 to 20 ppm, in many cases. This discrepancy was not surprising to the geodesists responsible for establishing the control networks, but it did concern GPS surveyors because they could not obtain accurate position checks of their work. Eventually, the concept of supernets or sparse arrays of high accuracy control networks established by GPS began to evolve. Many such supernets have been established in the U.S. and in other countries by state governments. These supernets will supersede the traditional lower accuracy control network.

Horizontal control

In addition to being less accurate than required by the users, the existing control networks have another disturbing feature that dooms them to extinction. The points in the two-dimensional network are usually located on the highest point in the area and are often not easily occupiable by GPS. Also, existing control points are located in heavily treed areas or at sites that are difficult to reach. Since most of the future geodetic (horizontal) surveys will be performed by GPS, the existing network points are (in many cases) inadequate.

The solution to this problem was found by establishing an array of new control points spaced between 25 and 100 km at locations suitable for GPS occupation. In most cases, these points are located at sites that are vehicle accessible and where they will not be disturbed. The internal accuracy of the supernets is normally 0.1 ppm so that in effect they appear to be errorless to users.

The global accuracy of supernets is controlled by referencing them to ITRF points. This international framework of points provide the ultimate geodetic framework, inasmuch as the interrelationship of the network sites is known to a higher accuracy than any other global array of points. The global supernet observations using dual frequency receivers are routinely made by the IGS stations (Sect. 4.4.1). Since the positions of the points in a state supernet can be determined to a few centimeters accuracy with respect to

the IGS stations, it is possible to establish networks in adjoining states at separate times. The seamless feature required for a true geodetic network is, therefore, easily achieved because there is relatively little error in the common boundary between the adjacent states. This feature of GPS supernets is particularly important because the surveys can be performed as needed and do not have to proceed in an orderly progression as conventional surveys.

Vertical control
GPS is a three-dimensional survey system in that it measures a vector between points with roughly the same degree of accuracy in all three components. Therefore, GPS is able to determine elevations as well as latitude and longitude. However, there is a problem in using GPS to establish elevations, since the height differences measured with GPS are referenced to an ellipsoid and not to the geoid to which conventionally leveled heights refer. As explained in more detail in Sect. 10.2.4, the separation between the ellipsoid and geoid differs for every point on the earth and can be found from existing global or local geoid models. One way to model the local or regional geoid is to run levels to a point of known ellipsoidal height. This is normally done for points in a supernet so that each point has an accurate ellipsoidal height and elevation. Based on these points, geodesists are able to refine their geoidal height prediction model to provide a more accurate determination. Many geodesists feel that prediction models in the future generally will have accuracies at the centimeter level in areas where supernets have been established.

Vertical accuracies of a few centimeters can also be achieved by connecting or referencing a GPS survey to supernet points. Since the supernet points have well-known elevations, and interpolation is done between these known points, the elevations of the new points can be accurately determined. Additionally, accuracies of such surveys are improved when a good model of geoidal heights is used.

12.5.2 Active control networks

For postprocessing or real-time DGPS applications, the monitor stations can be arranged in permanently operating networks which are called active control networks or Wide-Area Differential GPS (WADGPS) systems. The latter networks improve the integrity and reliability of DGPS. Usually, the operation of active control networks is fully automated. This means that satellite tracking, collection of dual frequency and ground meteorological data, performance of self-diagnostic tests, and the communication with the master control station in full duplex mode are performed automatically. The

master control station coordinates the observing stations, processes the data, and manages the dissemination of the results. Just as the supernets are making the existing control networks obsolete, the development of active control networks will possibly make the supernets obsolete.

The IGS network serving as a passive control network may also be considered as an active control network on a global scale. The measured code and phase range data at each tracking site are made available to all users through the IGS information system (Sect. 2.4.3) to allow for relative positioning in postprocessing mode.

An example of a regional active control network is the Canadian Active Control System (CACS) (Delikaraoglou et al. 1990). This system has been under development since 1985. One of the original goals of the network was to improve the satellite orbits for the Canadian region. Most of the approximately twenty regularly distributed monitor stations (with the master control station in Ottawa) are also part of the IGS network.

Another example for an active control network is the Continuously Operating Reference Station (CORS) Network managed by the U.S. National Geodetic Survey (NGS). The CORS system is currently mostly composed of the about 50 stations of the U.S. Coast Guard (USCG) that have been established for maritime DGPS service. NGS provides the data for postprocessing applications.

The Permanent GPS Geodetic Array (PGGA) in California is an active GPS network to monitor the boundary between the North American and Pacific tectonic plate. The network consists of more than 25 sites. The research is performed in cooperation with NASA's Dynamics of Solid Earth (DOSE) program, the U.S. Geological Survey (USGS), and the National Science Foundation. The data, which are archived daily, are also used for IGS computations (Gurtner 1995).

Another prototype of an active control network is the system developed by the Texas Highway Department. The goal was to continuously track all satellites from nine sites located throughout the state and then use the phase data to compute vectors from these fixed sites to other receivers surveying the various highway projects. The Texas system is a good example of the advantages and disadvantages of this concept. The advantage is that survey crews do not have to search for and occupy control points near their project area. They simply occupy the points they need coordinates for at the given project site. The vectors from the fixed stations surrounding the project site are combined with the (short) interstation vectors to give the best-fit solution for that particular survey. The cost effectiveness of this approach is positive when numerous crews take advantage of the system so that the cost of maintaining the fixed sites is spread over a large number of units. The

disadvantage of the Texas system is that the Highway Department is not prepared to provide its internal service to the public, as it is not funded to perform this service. The active control network of the Highway Department is a single-user private network which may have to be duplicated by various other users. This is not the only example of such duplication. Several other organizations are also setting up various types of active control points to serve their particular needs.

In practice, any moderately sized organization (e.g., company) will be able to establish its own tracking network. For example, a firm could place three receivers at equally spaced points that would cover its operations area. The precise coordinates of these fixed sites could be determined by combining a day's observation with data from neighboring IGS stations. Once the coordinates of the local tracking stations have been determined, they can be used as tracking stations themselves. A local tracking station can be as simple as an antenna mounted on the roof and a GPS receiver placed in a drawer and connected to a modem. At the end of the survey day, users would call each local tracker and download desired files. Recalibration of local networks could be performed periodically by recomputing the vectors from the IGS stations.

Today, most of the active control networks offer DGPS or (for shorter distances from the reference station) RTK services where pseudorange corrections or observed carrier phases are transmitted in real time to the user. Commercial DGPS services with worldwide coverage have been installed (Sluiter 1993). Examples for continental or regional DGPS services are the CACS and the service being installed by the USCG for harbor approach of maritime navigation. An overview of the present status of permanent GPS arrays can be found in Jong (1999).

13 Future of GPS

13.1 New application aspects

The future uses of GPS are limited only by one's imagination. Many of the present uses were described in various articles written as early as 1982 when this new system first demonstrated that high accuracies are achievable. With further reduction of equipment costs and both the modernization and augmentation of GPS, numerous additional applications will develop. This includes terrestrial, marine, and aviation applications.

Some examples for applications on land are vehicle navigation and information systems including Intelligent Vehicle/Highway Systems (IVHS) and Intelligent Transportation Systems (ITS). Another land application which has been mentioned is the use of GPS to automate various types of machinery. For example, it should be possible to automate the grading and paving equipment used for road building. Equipment could be run around the clock without operators with GPS performing all motion operations based upon a digital terrain model stored in the computer of the equipment. There will also be numerous applications in survey and geodesy as well as precise time determination and – one of the fastest growing GPS applications – time transfer. Available resources of telephone companies, power companies, and many others are enhanced by precise timing. Other future GPS applications concern atmosphere sounding. The data will contribute to a better understanding of the structure of the atmosphere leading for example to improved models for weather analysis.

Marine applications will include vessel navigation and information systems, precise harbor entrance systems, and oceanography in general. DGPS will mostly be used for these purposes, and dense networks of monitor stations are to be established along the coasts. One example is the network of real-time DGPS beacon sites along the U.S. coasts operated by the USCG.

For aviation, GPS will be integrated into other navigation systems like INS to fulfill the high reliability and integrity requirements. Applications will include en route navigation and surveillance, approach and landing, collision avoidance, and proximity warning. Aircraft could be operated in an automated mode with takeoffs and landings being performed by integrating GPS and computer units.

The use of GPS will increasingly be extended to space for precise positioning of (e.g., earth remote sensing) satellites, for attitude determination of spacecraft, and for missile navigation.

B. Hofmann-Wellenhof et al., *Global Positioning System*

13.2 GPS modernization

In January 1999, the U.S.A. announced a $400 million initiative to modernize GPS. Key feature of the initiative is the implementation of new signal structures in future satellites.

13.2.1 Future GPS satellites

The first Block IIR satellites are already in orbit, while the Block IIF or follow-on satellites are under construction. They will be launched from 2007 onwards. The Block III satellites carrying GPS into 2030 and beyond are presently being designed.

The next generation of satellites will have many improvements over the present satellites. It is planned to include the capability to transmit data between satellites to make the system more independent. The Autonomous Navigation (Auto-Nav) capability will allow the satellites to essentially position themselves without extensive ground tracking. In summary, the future satellites will have the following advantages:

- Navigation accuracy is maintained for six months without ground support. No survivable control and no user modifications are required.
- Uplink jamming concerns are minimized.
- One upload per spacecraft per month instead of one or even more per day is performed.
- Need for overseas stations to support navigation uploads is reduced.
- Improved navigation accuracy is achieved.

These features mainly benefit the military use of the system, since civilians will still be required to provide their own ephemerides for accurate surveys.

13.2.2 Augmented signal structure

Presently, civil users have unlimited access only to the C/A-code on the carrier L1 and, hence, only to one of the two signals transmitted by the GPS satellites.

The call for additional civil signals has led to a study of the National Research Council (1995). The output of this study contains several recommendations as turning off the Selective Availability as well as an additional (L5) carrier frequency that would presumably be available to civil users. The L5 frequency, originally designated as L4 frequency, would be chosen to give civil users the ability to correct for ionospheric effects and to assist in the resolution of carrier phase ambiguities.

In 1997, the Air Force initiated a review of the GPS capabilities. According to this study, in March 1998 the IGEB announced the implementation of a second civil signal in form of the C/A-code modulated onto the carrier L2. In addition, it was announced that there would be a third carrier L5 modulated with another civil code.

According to the modernization initiative released in 1999, the IGEB concept will be realized with the following specifications. Future GPS signals will be transmitted by three (civil) carriers L1, L2, L5 specified as

$$\begin{aligned} \text{L1} &= 154 f_0 = 1575.42 \text{ MHz} \\ \text{L2} &= 120 f_0 = 1227.60 \text{ MHz} \\ \text{L5} &= 115 f_0 = 1176.45 \text{ MHz} \end{aligned}$$

where $f_0 = 10.23\,\text{MHz}$ denotes the basic GPS frequency. The carrier L5 is placed in an aeronautical radionavigation service protected band and was recently allocated by the World Radio Conference (WRC) (Vorhies 2000).

The carriers L1 and L2 will be modulated with the (centered) C/A-code, while L5 will be modulated with a new civil code similar to the P-code. The existing military Y-code will be replaced by new (split) M-codes. This means that the civil signal on L1 and L2 will be sandwiched between two military codes.

The linear combination of L2 with L5 results in an extra-wide lane signal with a wavelength of about 5.9 m. Hence, this combination facilitates ambiguity resolution. By contrast, the linear combination of L1 with L5 will be used as ionosphere-free combination. The common process of code range and phase data from all three carriers will be performed in the Three-Carrier Ambiguity Resolution (TCAR) approach.

The second carrier modulated with the centered civil and split military signal with increased power will be implemented into the last 12 Block IIR satellites scheduled for launch beginning in 2005. The third carrier with another civil signal will be available in the Block IIF satellites (Department of Defense and Department of Transportation 2000).

13.3 GPS augmentation

The acceptance of GPS as the the primary or even sole means of navigation for aviation or for harbor approach needs improvements in terms of accuracy, availability, continuity, and integrity for safety reasons. This may be achieved by augmenting GPS and/or by developing a Global Navigation Satellite System (GNSS).

13.3.1 Ground-based augmentation

Ground-Based Augmentation Systems (GBASs) will be established in local areas such as the vicinity of airports. They will include ground-based pseudolites to transmit GPS-like signals and reference stations to broadcast differential corrections to the aircraft.

Examples for GBASs are the Local Area Augmentation Systems (LAASs) developed under the auspices of the U.S. Federal Aviation Administration (FAA). After 2003, such systems will be installed at the major U.S. airports. LAASs are considered to enable precision approach and landing in accordance with category II/III requirements (Table 12.3).

13.3.2 Satellite-based augmentation

Satellite-Based Augmentation Systems (SBASs) operate on a global or regional scale with a space segment consisting of geostationary (GEO) satellites. These satellites have ranging capabilities like GPS satellites and are equipped with dedicated packages to transmit WADGPS corrections and integrity messages.

One SBAS example is the Wide Area Augmentation System (WAAS) currently being established by the FAA in the contiguous U.S.A. The system consists of a ground network of reference stations along with the master station (CGSIC (1996). This station collects the data, formats the DGPS corrections, and uploads the messages to the geostationary satellites (Loh et al. 1995). WAAS will provide navigation services for all phases of flight except category II/III precision approach. The declaration of the full operational capability (FOC) of the WAAS was originally scheduled for 2001. Today, FOC is expected to be declared in the time frame of 2005 (Department of Defense and Department of Transportation 2000).

The International Maritime Satellites (Inmarsat), a system of geostationary communication satellites, augment GPS in the WAAS. They are equipped with repeaters which transmit C/A-code signals on L1 (Kinal and Singh 1990, Nagle et al. 1992/93). The primary purpose of these satellites, however, is integrity monitoring. A detailed description of Inmarsat can be found in Ackroyd and Lorimer (1990).

13.4 GNSS

13.4.1 GNSS development

The first steps towards GNSS are the GPS/GLONASS integration and the augmentation of the existing systems. Regional and global concepts for the

realization are being discussed (Lichtenegger 1999).

In the initial phase, GPS and GLONASS are augmented with geostationary satellites. Examples for such realizations are the U.S. WAAS, the European Geostationary Navigation Overlay Service (EGNOS), and the Japanese augmentation system MSAS based on their multifunctional transportation satellites.

The second phase is considered (at least outside the U.S.A.) as the transition from military systems to a system under complete civil control. The European contribution to this step is the Galileo project, which has entered the definition phase in 1999 and is expected to be declared operational in 2008. For more details on Galileo the reader is referred to Hofmann-Wellenhof (1999).

A critical issue for GNSS is the allocation of frequency bands for the individual satellite systems. Such an allocation depends on international agreements and is the responsibility of the International Telecommunication Union (ITU). The most recent frequency allocation was made during the WRC 2000 and the result is summarized in Table 13.1 according to Vorhies (2000).

Table 13.1. Frequency allocation

Designation	Bandwidth (MHz)	Allocation
L5-band	1164–1188	L5 (GPS)
	1188–1214	E5 (Galileo)
L2-band	1215–1239	L2 (GPS)
	1237–1260	G2 (GLONASS)
	1254–1258	E4 (Galileo), narrow band
	1260–1300	E6 (Galileo)
L1-band	1559–1563	E2 (Galileo), narrow band
	1563–1587	L1 (GPS)
	1587–1591	E1 (Galileo), narrow band
	1593–1610	G1 (GLONASS)

13.4.2 GNSS/Loran-C integration

Loran-C (and its Russian counterpart Chayka) is a terrestrial radionavigation system where a number of stations transmit high-power pulses in the low-frequency (i.e., 90–110 kHz) band. The broadcasting stations are

organized into chains with one master station and between two and five secondary stations. The cesium clocks of the reference stations are synchronized within a few tens of nanoseconds. Roving Loran-C receivers are equipped with low-cost clocks to measure the signal runtime from the reference stations. Since two clocks are involved, one at the reference station and one at the roving station, the final observables are pseudoranges. Thus, positioning with Loran-C resembles GPS positioning. For more details on Loran-C, the reader is referred to Department of Defense and Department of Transportation (2000).

The existing infrastructure of Loran-C may be used as a regional augmentation system for GNSS. Such an approach is Eurofix, a project under the leadership of the Delft University of Technology, The Netherlands (Offermans et al. 2000). The Loran-C signals are used as carriers for a multichannel long-range (up to 1 000 km off the reference stations) communication system whereby the normal navigation mode of Loran-C is preserved. So, the GNSS/Loran-C integration offers the advantages of a differential GNSS service through Loran-C over a large coverage area, improved availability and continuity by using calibrated Loran-C, and external integrity information.

13.5 Hardware and software improvements

13.5.1 Hardware

Many of today's hardware developments were predicted by GPS pioneers. For example, the rapid cost reduction in equipment, predicted in Wells et al. (1987), in many ways parallels the development of small computers. The continued production of special computer chips will both decrease the cost and size, and increase the reliability and field worthiness.

The first GPS receivers for commercial use cost in excess of $ 100 000 and possessed much less capability than receivers presently costing $ 10 000. Original Equipment Manufacturer (OEM) units are now (2000) being marketed at a cost of less than $ 1 000 (with demonstrated survey accuracies). If one plots the receiver cost against time, the curve shows a sharp (exponential) decline during the first few years (1983–1987) and then becomes asymptotic with the receiver prices slowly decreasing. Today, the point has been reached where volume production by a manufacturer could reduce the receiver cost to a fraction of the present level. The GPS price curve closely resembles the decrease in the price of survey totalstations or hand-held calculators. The cost analysis of GPS is somewhat complicated by the fact that the processing software is bundled in the sales price. When the cost of software is a significant portion of the receiver cost, and the receiver is mainly

composed of integrated circuit chips, the cost of receivers can be appreciably lowered by quantity production.

The present multichannel dual frequency receivers measure and record all the basic observables. It appears that the resolution will soon be improved to 0.1% of the wavelength (or better). This capability translates into a precision of a few centimeters in the code pseudorange, thus, providing instantaneous carrier phase ambiguity resolution. Additionally, multichannel receivers with up to 36 digital programmable channels (measuring any combination of the L1, L2, and L5 observables) will routinely be manufactured. Multiantenna arrays connected to a single receiver will be improved for highly accurate attitude determination. It is anticipated that such receivers will be modified to determine azimuths to distant azimuth marks; thus, reducing the task of surveying a geodetic control network by 50%.

Improvements of the kinematic GPS technique also affect the design of future GPS receivers. Automatic bandwidth adjustment and more rigorous cycle slip detection will be the norm of the future. The size (and power requirements) of receivers will also be reduced. The (no longer futuristic) receiver with stake out capability is a small antenna mounted on top of a range pole with the electronics placed inside the pole to reduce the bulk carried by the rodman. Such a receiver combined with a receiver incorporated in a conventional totalstation instrument provides the surveyor with the ultimate surveying capability and significantly increases the productivity.

As the cost of receivers decrease, instruments will also be placed on structures such as bridges and dams to provide near real-time motion detection. It has been hypothesized for a number of years that instrumenting dams with GPS receivers tied into a central control network could provide a warning system to foretell dam failure.

The GLONASS satellite system was briefly discussed previously. Some manufacturers are offering combined GPS/GLONASS geodetic receivers. These combined receivers significantly improve receiver performance. For example, doubling the number of satellites in view at a given time permits the receiver to be used in many areas where the use of a standard GPS reveiver would be impossible (e.g., in urban canyons). Also, the added satellites significantly speed up the on-the-fly kinematic surveying.

In addition to speed of operation and increased capabilities, manufacturers are also working on the weight and size of GPS equipment. The use of dual frequency receivers to perform centimeter navigation in (near) real time is now quite common. However, the roving receiver, batteries, radios, and other equipment must be carried in a backpack weighing about 10 kilos. Lower-cost dual frequency receivers weighing a few kilos will make (near) real-time centimeter navigation appreciably more practical so that GPS will

be used by a greater proportion of the surveying and mapping community for projects ranging from precise engineering stake out to GIS data collection.

13.5.2 Software

The first GPS processing software (for Macrometer) required about one hour to process a baseline on a large computer. Today, a baseline is processed in seconds using a PC. Some GPS receivers presently include baseline processing software in the receiver. For example, several manufacturers offer receivers capable of performing real-time kinematic (RTK) surveying using their normal hardware and upgraded receiver software. This trend will still increase in the future and additional processing software will be included with the receiver. This software will also allow GPS and GLONASS (and Galileo) data to be combined resulting in real-time centimeter survey.

In the future, the IGS tracking network and active control networks will be expanded and improved. Software will be developed so that surveyors can download data from one or more active control points in (near) real time using a mobile telephone. They will then be able to compute accurate positions on site to any predetermined degree of accuracy. Improved orbits determined from these networks will provide centimeter accuracy positions over large areas so that the concept of survey accuracy will virtually disappear, since all GPS survey points will exceed the present first-order accuracy.

For WAAS applications, software is being developed which computes pseudorange, orbital, and ionospheric corrections at points in a network of fixed ground stations. These corrections are uploaded to geostationary satellites and are rebroadcast to users. They permit the GPS user to determine worldwide instantaneous submeter positions (if the satellite signal can be received). Improvements in the accuracy of this system could reduce the cost and increase the reliability of collecting GIS data by hand-held GPS receivers.

13.6 Conclusion

The present status of GPS is sufficiently exciting for most surveyors; however, future developments will make the present use of the system appear quite ordinary. The enlarged number of next generation satellites will meet all future requirements for navigation, surveying, timing, and additionally provide many new revolutionary uses. Substantially improved hardware and software components will provide desired results more quickly, in real time, at lower cost. Real-time data transfer systems will be standard, reliable,

and simple to use. Accuracy is expected to increase by at least one order of magnitude even with lower costs. In addition, GPS will achieve a higher level of integrity and its output will be more reliable due to extensive quality control. As a consequence, the differences between navigation and surveying will be less pronounced. Apart from the present tasks, new application will arise which today are undreamed of.

The future uses of GPS are virtually unlimited. Its use into the next decades, however, is difficult to comprehend. The one undeniable fact is that surveyors will be fully challenged by this technology and will receive benefits commensurate with their efforts to understand and adapt to this new system.

References

Abidin HZ (1993): On the construction of the ambiguity searching space for on-the-fly ambiguity resolution. Navigation, 40(3): 321–338.

Abidin HZ, Wells DE, Kleusberg A (1992): Some aspects of "on the fly" ambiguity resolution. In: Proceedings of the Sixth International Geodetic Symposium on Satellite Positioning, Columbus, Ohio, March 17–20, vol 2: 660–669.

Ackroyd N, Lorimer R (1990): Global navigation – a GPS user's guide. Lloyd's of London, London New York Hamburg Hong Kong.

Altmayer C (2000): Cycle slip detection and correction by means of integrated systems. In: Proceedings of the 2000 National Technical Meeting of the Institute of Navigation, Anaheim, California, January 26–28, 134–144, available on CD-ROM.

Arinc Research Corporation (2000): Interface control document, Navstar GPS space segment / navigation user interfaces, ICD-GPS-200, revision C (IRN-200C-004). Available under the Web site http://www.navcen.uscg.mil/gps.

Arnold K (1970): Methoden der Satellitengeodäsie. Akademie, Berlin.

Ashby N (1987): Relativistic effects in the Global Positioning System. In: Relativistic effects in geodesy, Proceedings of the International Association of Geodesy (IAG) Symposia of the XIX General Assembly of the IUGG, Vancouver, Canada, August 10–22, vol 1: 41–50.

Ashjaee J (1993): An analysis of Y-code tracking techniques and associated technologies. Geodetical Info Magazine, 7(7): 26–30.

Ashjaee J, Lorenz R (1992): Precision GPS surveying after Y-code. In: Proceedings of ION GPS-92, Fifth International Technical Meeting of the Satellite Division of the Institute of Navigation, Albuquerque, New Mexico, September 16–18, 657–659.

Ashkenazi V, Moore T, Ffoulkes-Jones G, Whalley S, Aquino M (1990): High precision GPS positioning by fiducial techniques. In: Bock Y, Leppard N (eds): Global Positioning System: an overview. Springer, New York Berlin Heidelberg Tokyo, 195–202 [Mueller II (ed): IAG Symposia Proceedings, vol 102].

Auernhammer H, Muhr T, Demmel M (1994): GPS and DGPS as a challenge for environment friendly agriculture. In: The German Institute of Navigation (ed): Proceedings of the Third International Conference on Land Vehicle Navigation, Dresden, Germany, June 14–16, 81–91.

Avdis V, Billiris H, Hurst K, Kastens K, Paradissis D, Veis G (1990): Selection, monumentation, and documentation of GPS sites in the circum-Aegean region, Greece. Newsletter of the Space Geodetic Measurement Sites Subcommission (SGMS), 1(2): 7–9.

Bačić Ž, Kalafut M, Lichtenegger H, Wagner J (1995): Some investigations on precise kinematic GPS surveys. In: Proceedings of the International Symposium on GPS Technology Applications, Bucharest, Romania, September 26–29, 219–224.

Bartone C, Graas F van (1998): Airport pseudolites for local area augmentation. In: Proceedings of IEEE PLANS, Publication 98CH36153, Palm Springs, California, April 20–23, 479–486.

Bauer M (1994): Vermessung und Ortung mit Satelliten, 3rd edn. Wichmann, Karlsruhe.

Bennet JM (1965): Triangular factors of modified matrices. Numerische Mathematik, 7: 217–221.

Beutler G (1991): Himmelsmechanik I. Mitteilungen der Satelliten-Beobachtungsstation Zimmerwald, Bern, vol 25.

Beutler G (1992): Himmelsmechanik II. Mitteilungen der Satelliten-Beobachtungsstation Zimmerwald, Bern, vol 28.

Beutler G (1996a): GPS satellite orbits. In: Kleusberg A, Teunissen PJG (eds): GPS for geodesy. Springer, Berlin Heidelberg New York Tokyo, 37–101 [Bhattacharji S, Friedman GM, Neugebauer HJ, Seilacher A (eds): Lecture Notes in Earth Sciences, vol 60].

Beutler G (1996b): The GPS as a tool in global geodynamics. In: Kleusberg A, Teunissen PJG (eds): GPS for geodesy. Springer, Berlin Heidelberg New York Tokyo, 379–407 [Bhattacharji S, Friedman GM, Neugebauer HJ, Seilacher A (eds): Lecture Notes in Earth Sciences, vol 60].

Beutler G, Davidson DA, Langley RB, Santerre R, Vanicek P, Wells DE (1984): Some theoretical and practical aspects of geodetic positioning using carrier phase difference observations of GPS satellites. University of New Brunswick, Canada, Technical Report vol 109.

Beutler G, Gurtner W, Bauersima I, Rothacher M (1986): Efficient computation of the inverse of the covariance matrix of simultaneous GPS carrier phase difference observations. Manuscripta geodaetica, 11: 249–255.

Beutler G, Bauersima I, Gurtner W, Rothacher M (1987): Correlations between simultaneous GPS double difference carrier phase observations in the multistation mode: implementation considerations and first experiences. Manuscripta geodaetica, 12: 40–44.

Beutler G, Gurtner W, Rothacher M, Wild U, Frei E (1990): Relative static positioning with the Global Positioning System: basic technical considerations. In: Bock Y, Leppard N (eds): Global Positioning System: an overview. Springer, New York Berlin Heidelberg Tokyo, 1–23 [Mueller II (ed): IAG Symposia Proceedings, vol 102].

Bevis M (1991): GPS Networks: the practical side. EOS Transactions, American Geophysical Union, 72(6): 49.

Bevis M, Businger S, Herring TA, Rocken C, Anthes RA, Ware RH (1992): GPS meteorology: remote sensing of atmospheric water vapor using the Global Positioning System. Journal of Geophysical Research 97(D14): 15787–15801.

Blankenberg LE (1994): GPS-supported aerial triangulation – results from test flight Fredrikstad. In: Proceedings of the XX Congress of FIG (International Federation of Surveyors), Commission 5, Survey Instruments and Methods, Melbourne, Australia, March 5–12, 553.1/1–553.1/12.

Bock Y (1996): Reference systems. In: Kleusberg A, Teunissen PJG (eds): GPS for geodesy. Springer, Berlin Heidelberg New York Tokyo, 3–36 [Bhattacharji S, Friedman GM, Neugebauer HJ, Seilacher A (eds): Lecture Notes in Earth Sciences, vol 60].

Borge TK, Forssell B (1994): A new real-time ambiguity resolution strategy based on polynomial identification. In: Proceedings of the International Symposium on Kinematic Systems in Geodesy, Geomatics and Navigation, Banff, Canada, August 30 through September 2, 233–240.

Boucher C, Altamimi Z, Sillard P (eds) (1999): The 1997 International Terrestrial Reference Frame (ITRF97). Observatoire de Paris, IERS Technical Note 27.

Breuer B, Campbell J, Müller A (1993): GPS-Meß- und Auswerteverfahren unter operationellen GPS-Bedingungen. Journal for Satellite-Based Positioning, Navigation and Communication, 2(3): 82–90.

Bronstein IN, Semendjajew KA (1996): Teubner-Taschenbuch der Mathematik. Teubner, Stuttgart Leipzig.

Brouwer D, Clemence GM (1961): Methods of celestial mechanics. Academic Press, New York.

Brunner FK, Gu M (1991): An improved model for the dual frequency ionospheric correction of GPS observations. Manuscripta geodaetica, 16: 205–214.

Brunner FK, Welsch WM (1993): Effect of the troposphere on GPS measurements. GPS World, 4(1): 42–51.

Byman P, Koskelo I (1991): Mapping Finnish roads with differential GPS and dead reckoning. GPS World, 2(2): 38–42.

Cannon ME, Lachapelle G (1993): GPS – theory and applications. Lecture Notes for a seminar on GPS given at Graz in spring 1993.

Cannon ME, Sun H, Owen T, Meindl M (1994): Assessment of a non-dedicated GPS receiver system for precise airborne attitude determination. In: Proceedings of ION GPS-94, 7th International Technical Meeting of the Satellite Division of the Institute of Navigation, Salt Lake City, Utah, September 20–23, part 1: 645–654.

CGSIC (1994): Summary record of the 24th meeting of the Civil GPS Service Interface Committee (CGSIC), Salt Lake City, Utah, September 19–20.

CGSIC (1995): Summary record of the 26th meeting of the Civil GPS Service Interface Committee (CGSIC), Palm Springs, California, September 11–12.

CGSIC (1996): Summary report of the 27th meeting of the Civil GPS Service Interface Committee (CGSIC), Falls Church, Virginia, March 19–21.

Chen D (1994): Development of a fast ambiguity search filtering (FASF) method for GPS carrier phase ambiguity resolution. Reports of the Department of Geomatics Engineering of the University of Calgary, vol 20071.

Chen D, Lachapelle G (1994): A comparison of the FASF and least-squares search algorithms for ambiguity resolution on the fly. In: Proceedings of the International Symposium on Kinematic Systems in Geodesy, Geomatics and Navigation, Banff, Canada, August 30 through September 2, 241–253.

Chin M (1991): CIGNET report. GPS Bulletin, 4(2): 5–11.

Collins J (1989): Fundamentals of GPS baseline and height determinations. Journal of Surveying Engineering, American Society of Civil Engineers, 115(2): 223–235.

Colombo OL, Bhapkar UV, Evans AG (1999): Inertial-aided cycle-slip detection/correction for precise, long-baseline kinematic GPS. In: Proceedings of ION GPS-99, 12th International Technical Meeting of the Satellite Division of the Institute of Navigation, Nashville, Tennessee, September 14–17, 1915–1921, available on CD-ROM.

Conley R (2000): Life after selective availability. Newsletter of the Institute of Navigation, 10(1): 3–4.

Conley R, Lavrakas JW (1999): The world after selective availability. In: Proceedings of ION GPS-99, 12th International Technical Meeting of the Satellite Division of the Institute of Navigation, Nashville, Tennessee, September 14–17, 1353–1361, available on CD-ROM.

Cooper MAR, Cross PA, Dahl PA, Corbett S, Short T (1994): GPS kinematic positions and orientations of airborne sensors: some preliminary results. In: Proceedings of the XX Congress of FIG (International Federation of Surveyors), Commission 5, Survey Instruments and Methods, Melbourne, Australia, March 5–12, 503.1/1–503.1/10.

Counselman CC, Gourevitch SA (1981): Miniature interferometer terminals for earth surveying: ambiguity and multipath with the Global Positioning System. IEEE Transactions on Geoscience and Remote Sensing, GE–19(4): 244–252.

Daly P, Misra PN (1996): GPS and Global Navigation Satellite System (GLONASS). In: Parkinson BW, Spilker JJ (eds): Global Positioning System: Theory and applications. American Institute of Aeronautics and Astronautics, Washington DC, vol 2: 243–272.

Daxinger W, Stirling R (1995): Kombinierte Ausgleichung von terrestrischen und GPS-Messungen. Österreichische Zeitschrift für Vermessung und Geoinformation, 83(1+2): 48–55.

Deines SD (1992): Missing relativity terms in GPS. Navigation, 39(1): 111–131.

Delikaraoglou D, Lahaye F (1990): Optimization of GPS theory, techniques and operational systems: progress and prospects. In: Bock Y, Leppard N (eds): Global Positioning System: an overview. Springer, New York Berlin Heidelberg Tokyo, 218–239 [Mueller II (ed): IAG Symposia Proceedings, vol 102].

Delikaraoglou D, Dragert H, Kouba J, Lochhead K, Popelar J (1990): The development of a Canadian GPS Active Control System: status of the current array. In: Proceedings of the Second International Symposium on Precise Positioning with the Global Positioning System, Ottawa, Canada, September 3–7, 190–202.

DeLoach SR, Remondi B (1991): Decimeter positioning for dredging and hydrographic surveying. In: Proceedings of the First International Symposium on Real Time Differential Applications of the Global Positioning System. TÜV Rheinland, Köln, vol 1: 258–263.

Department of Commerce (2000): Civilian benefits of discontinuing selective availability. Press release, May 1.

Department of Defense (1995): Global Positioning System Standard Positioning Service – signal specification, 2nd edition. Available from the U.S. Coast Guard's GPS Web site http://www.navcen.uscg.mil/gps.

Department of Defense and Department of Transportation (2000): 1999 Federal Radionavigation Plan. U.S. National Technical Information Service, Springfield, Virginia, DOT-VNTSC-RSPA-98-1/DOD-4650.5.

Dierendonck AJ van (1999): Understanding GPS receiver terminology – a tutorial. In: GPS World's Big Book of GPS, 48–59, Advanstar Communications, Cleveland, Ohio.

Dierendonck AJ van, Braasch MS (1997): Evaluation of GNSS receiver correlation processing techniques for multipath and noise mitigation. In: Proceedings of the 1997 National Technical Meeting of the Institute of Navigation, Santa Monica, California, 207–215.

Eissfeller B (1993): Stand der GPS-Empfänger-Technologie. In: Institute of Geodesy (ed): Proceedings of the Geodetic Seminar on Global Positioning System im praktischen Einsatz der Landes- und Ingenieurvermessung, Munich, May 12–14. Schriftenreihe der Universität der Bundeswehr München, vol 45: 29–55.

El-Mowafy A (1994): Kinematic attitude determination from GPS. Reports of the Department of Geomatics Engineering of the University of Calgary, vol 20074.

El-Sheimy N (2000): An expert knowledge GPS/INS system for mobile mapping and GIS applications. In: Proceedings of the 2000 National Technical Meeting of the Institute of Navigation, Anaheim, California, January 26–28, 816–824, available on CD-ROM.

Erickson C (1992a): Investigations of C/A code and carrier measurements and techniques for rapid static GPS surveys. Reports of the Department of Geomatics Engineering of the University of Calgary, vol 20044.

Erickson C (1992b): An analysis of ambiguity resolution techniques for rapid static GPS surveys using single frequency data. In: Proceedings of ION GPS-92, Fifth International Technical Meeting of the Satellite Division of the Institute of Navigation, Albuquerque, New Mexico, September 16–18, 453–462.

Essen L, Froome KD (1951): The refractive indices and dielectric constants of air and its principal constituents at 24000 Mc/s. Proceedings of Physical Society, vol 64(B): 862–875.

Euler H-J, Goad CC (1991): On optimal filtering of GPS dual frequency observations without using orbit information. Bulletin Géodésique, 65: 130–143.

Euler H-J, Landau H (1992): Fast GPS ambiguity resolution on-the-fly for real-time applications. In: Proceedings of the Sixth International Geodetic Symposium on Satellite Positioning, Columbus, Ohio, March 17–20, vol 2: 650–659.

Euler H-J, Sauermann K, Becker M (1990): Rapid ambiguity fixing in small scale networks. In: Proceedings of the Second International Symposium on Precise Positioning with the Global Positioning System, Ottawa, Canada, September 3–7, 508–523.

Fenton PC, Townsend BR (1994): NovAtel Communications Ltd. – what's new? In: Proceedings of the International Symposium on Kinematic Systems in Geodesy, Geomatics and Navigation, Banff, Canada, August 30 through September 2, 25–29.

Fliegel HF, Feess WA, Layton WC, Rhodus NW (1985): The GPS radiation force model. In: Proceedings of the First International Symposium on Precise Positioning with the Global Positioning System, Rockville, Maryland, April 15–19, vol 1: 113–119.

Frei E (1991): GPS – Fast Ambiguity Resolution Approach "FARA": theory and application. Paper presented at XX General Assembly of the IUGG, IAG-Symposium GM 1/4, Vienna, August 11–24.

Frei E, Schubernigg M (1992): GPS surveying techniques using the "Fast Ambiguity Resolution Approach (FARA)". Paper presented at the 34th Australian Surveyors Congress and the 18th National Surveying Conference at Cairns, Australia, May 23–29.

Gambis D, Ray J (2000): Explanatory supplement to IERS Bulletins A and B. Observatoire de Paris, IERS Information, March.

Garin L, Rousseau J (1997): Enhanced strobe correlator multipath rejection for code and carrier. In: Proceedings of ION GPS-97, 10th International Technical Meeting of the Satellite Division of the Institute of Navigation, Kansas City, Montana, September 16–19, 559–568.

Geiger A (1988): Einfluss und Bestimmung der Variabilität des Phasenzentrums von GPS-Antennen. Eidgenössische Technische Hochschule Zürich, Institute of Geodesy and Photogrammetry, Mitteilungen vol 43.

Gendt G, Reigber C, Dick G (1999): GPS meteorology – IGS contribution and GFZ activities for operational water vapor monitoring. In: Proceedings of the Fifth International Seminar "GPS in Central Europe", Reports on Geodesy, Warsaw University of Technology, 5(46): 53–62.

Gerdan GP (1995): A comparison of four methods of weighting double difference pseudo range measurements. Trans Tasman Surveyor, 1(1): 60–66.

Gianniou M (1996): Genauigkeitssteigerung bei kurzzeit-statischen und kinematischen Satellitenmessungen bis hin zur Echtzeitanwendung. Deutsche Geodätische Kommission bei der Bayerischen Akademie der Wissenschaften, Reihe C, vol 458.

Glavkosmos (1991): Global Satellite Navigation System GLONASS. Interface Control Document (second wording), RTCA Paper No. 518-91/SC159-317. Reprinted 1994 by Navtech Seminars & Navtech Book and Software Store, Inc., Arlington, Virginia.

Graas F van, Braasch MS (1991): GPS interferometric attitude and heading determination: Initial flight test results. Navigation, 38(4): 297–316.

Graas F van, Braasch MS (1992): Real-time attitude and heading determination using GPS. GPS World, 3(3): 32-39.

Grafarend EW, Schwarze V (1991): Relativistic GPS positioning. In: Caputo M, Sansò F (eds): Proceedings of the geodetic day in honor of Antonio Marussi. Accademia Nazionale dei Lincei, Rome. Atti dei Convegni Lincei, vol 91: 53–66.

Graviss LP (1992): GPS development program status. In: Proceedings of ION GPS-92, Fifth International Technical Meeting of the Satellite Division of the Institute of Navigation, Albuquerque, New Mexico, September 16–18, 3–16.

Gurtner W (1995): The role of permanent GPS stations in IGS and other networks. In: Proceedings of the Third International Seminar on GPS in Central Europe, Penc, Hungary, May 9–11, 221–239.

Gurtner W (1998): RINEX: the receiver independent exchange format version 2. Available under http://igscb.jpl.nasa.gov/

Gurtner W, Mader G (1990): Receiver independent exchange format version 2. GPS Bulletin, 3(3): 1–8.

Habrich H, Beutler G, Gurtner W, Rothacher M (1999): Double difference ambiguity resolution for GLONASS/GPS carrier phase. In: Proceedings of ION GPS-99, 12th International Technical Meeting of the Satellite Division of the Institute of Navigation, Nashville, Tennessee, September 14–17, 1609–1618, available on CD-ROM.

Han S, Rizos C (1997): Comparing GPS ambiguity resolution techniques. GPS World, 8(10): 54–61.

Han S, Dai L, Rizos C (1999): A new data processing strategy for combined GPS/GLONASS carrier phase-based positioning. In: Proceedings of ION GPS-99, 12th International Technical Meeting of the Satellite Division of the Institute of Navigation, Nashville, Tennessee, September 14–17, 1619–1627, available on CD-ROM.

Hatch R (1990): Instantaneous ambiguity resolution. In: Schwarz KP, Lachapelle G (eds): Kinematic systems in geodesy, surveying, and remote sensing. Springer, New York Berlin Heidelberg Tokyo, 299–308 [Mueller II (ed): IAG Symposia Proceedings, vol 107].

Hatch R (1991): Ambiguity resolution while moving – experimental results. In: Proceedings of ION GPS-91, Fourth International Technical Meeting of the Satellite Division of the Institute of Navigation, Albuquerque, New Mexico, September 11–13, 707–713.

Hatch R (1994): Comparison of several AROF kinematic techniques. In: Proceedings of ION GPS-94, 7th International Technical Meeting of the Satellite Division of the Institute of Navigation, Salt Lake City, Utah, September 20–23, part 1: 363–370.

Hatch R, Keegan R, Stansell TA (1992): Kinematic receiver technology from Magnavox. In: Proceedings of the Sixth International Geodetic Symposium on Satellite Positioning, Columbus, Ohio, March 17–20, vol 1: 174–183.

Hatch R, Jung J, Enge P, Pervan B (2000): Civilian GPS: the benefits of three frequencies. GPS Solutions, 3(4): 1–9.

Hein GW (1990a): Bestimmung orthometrischer Höhen durch GPS und Schweredaten. Schriftenreihe der Universität der Bundeswehr München, vol 38-1: 291–300.

Hein GW (1990b): Kinematic differential GPS positioning: applications in airborne photogrammetry and gravimetry. In: Crosilla F, Mussio L (eds): Il sistema di posizionamento globale satellitare GPS. International Centre for Mechanical Sciences (CISM), Collana di Geodesia e Cartografia, Udine, Italy, 139–173.

Hein GW (1995): Comparison of different on-the-fly ambiguity resolution techniques. In: Proceedings of ION GPS-95, 8th International Technical Meeting of the Satellite Division of the Institute of Navigation, Palm Springs, California, September 12–15, part 2: 1137–1144.

Heiskanen WA, Moritz H (1967): Physical geodesy. Freeman, San Francisco London.

Henstridge F (1991): Getting started in GPS. Professional Surveyor, 11(4): 4–9.

Herring TA (1992): Modeling atmospheric delays in the analysis of space geodetic data. In: Munck JC de, Spoelstra TAT (eds): Refraction of transatmospheric signals in geodesy. Netherlands Geodetic Commission, Delft, new series, vol 36: 157–164.

Hilla S, Jackson M (2000): The GPS toolbox. GPS Solutions, 3(4): 71–74.
Hofmann-Wellenhof B (1999): GPS or GNSS – that is the question! In: Proceedings of the Fifth International Seminar "GPS in Central Europe", Reports on Geodesy, Warsaw University of Technology, 5(46): 11–22.
Hofmann-Wellenhof B, Lichtenegger H (eds) (1988): GPS – Von der Theorie zur Praxis. Mitteilungen der geodätischen Institute der Technischen Universität Graz, vol 62.
Hofmann-Wellenhof B, Remondi BW (1988): The antenna exchange: one aspect of high-precision GPS kinematic survey. In: Groten E, Strauß R (eds): GPS-techniques applied to geodesy and surveying. Springer, Berlin Heidelberg New York Tokyo, 261–277 [Bhattacharji S, Friedman GM, Neugebauer HJ, Seilacher A (eds): Lecture Notes in Earth Sciences, vol 19].
Hofmann-Wellenhof B, Kienast G, Lichtenegger H (1994): GPS in der Praxis. Springer, Wien New York.
Hopfield HS (1969): Two-quartic tropospheric refractivity profile for correcting satellite data. Journal of Geophysical Research, 74(18): 4487–4499.
IAG (1995): New geoids in the world. International Association of Geodesy, Bulletin d'Information 77, IGES Bulletin 4.
International Geoid Service (2000): Bulletin no 10. Politecnico di Milano, Italy.
Jakowsky N (1996): TEC monitoring by using satellite positioning systems. In: Kohl H, Rüster R, Schlegel K (eds): Modern ionosphere science. ProduServ GmbH Verlagsservice. Berlin, 371–390.
Janes HW, Langley RB, Newby SP (1991): Analysis of tropospheric delay prediction models: comparisons with ray-tracing and implications for GPS relative positioning. Bulletin Géodésique, 65: 151–161.
Jin XX, Jong CD de (1996): Relationship between satellite elevation and precision of GPS code observations. The Journal of Navigation, 49: 253–265.
Jong K de (1999): Permanent GPS arrays. Geoinformatics, 2(3): 22–23.
Jong K de (2000): Selective availability turned off. Geoinformatics, 3(5): 14–15.
Jonge P de, Tiberius C (1995): Integer ambiguity estimation with the Lambda method. In: Beutler G, Hein GW, Melbourne WG, Seeber G (eds): GPS trends in precise terrestrial, airborne, and spaceborne applications. Springer, New York Berlin Heidelberg Tokyo, 280–284 [Mueller II (ed): IAG Symposia Proceedings, vol 115].
Jonkman NF (1998): Integer GPS ambiguity estimation without the receiver-satellite geometry. Delft Geodetic Computing Centre, LGR Series, vol 18.
Joos G (1956): Lehrbuch der Theoretischen Physik, 9th edn. Akademische Verlagsgesellschaft Geest & Portig K-G, Leipzig.
Joosten P, Tiberius C (2000): Fixing the ambiguities – are you sure they're right? GPS World, 11(5): 46–51.
Joosten P, Teunissen PJG, Jonkman N (1999): GNSS three carrier phase ambiguity resolution using the LAMBDA-method. In: Proceedings of GNSS'99, 3rd European Symposium on Global Navigation Satellite Systems, Genova, Italy, October 5–8, part 1: 367–372.
Jurgens RD, Rodgers CE, Fan LC (1991): GPS azimuth determining system (ADS), cycle resolution, system design, and army test results. In: Proceedings of the National Technical Meeting of the Institute of Navigation, Phoenix, Arizona, January 22–24, 45–51.

Kasties G, Harrer S (1991): Differential GPS for real time flight path surveillance. In: Proceedings of the First International Symposium on Real Time Differential Applications of the Global Positioning System. TÜV Rheinland, Köln, vol 1: 205–237.

Kaula WM (1966): Theory of satellite geodesy. Blaisdell, Toronto.

Kee C (1996): Wide area differential GPS. In: Parkinson BW, Spilker JJ (eds): Global Positioning System: theory and applications. American Institute of Aeronautics and Astronautics, Washington DC, vol 2: 81–115.

Keegan R (1990): P-code aided Global Positioning System receiver. U.S. Patent Office, Patent no. 4,972,431.

Kinal GV, Singh JP (1990): An international geostationary overlay for GPS and GLONASS. Navigation, 37(1): 81–93.

King RW, Masters EG, Rizos C, Stolz A, Collins J (1987): Surveying with Global Positioning System. Dümmler, Bonn.

Klees R (1990): Anwendung des NAVSTAR/Global Positioning System in der kanadischen Landes-, Kataster- und Stadtvermessung. Allgemeine Vermessungsnachrichten, 97(3): 117–120, and 97(4): 138–157.

Kleusberg A (1990a): A review of kinematic and static GPS surveying procedures. In: Proceedings of the Second International Symposium on Precise Positioning with the Global Positioning System, Ottawa, Canada, September 3–7, 1102–1113.

Kleusberg A (1990b): Comparing GPS and GLONASS. GPS World, 1(6): 52.

Kleusberg A (1994): Die direkte Lösung des räumlichen Hyperbelschnitts. Zeitschrift für Vermessungswesen, 119(4): 188–192.

Klobuchar J (1986): Design and characteristics of the GPS ionospheric time-delay algorithm for single-frequency users. In: Proceedings of PLANS'86 – Position Location and Navigation Symposium, Las Vegas, Nevada, November 4–7, 280–286.

Knight D (1994): A new method of instantaneous ambiguity resolution. In: Proceedings of ION GPS-94, 7th International Technical Meeting of the Satellite Division of the Institute of Navigation, Salt Lake City, Utah, September 20–23, part 1: 707–716.

Koch K-R (1987): Parameter estimation and hypothesis testing in linear models. Springer, Berlin Heidelberg New York London Paris Tokyo.

Kozai Y (1959): On the effects of the sun and the moon upon the motion of a close earth satellite. Smithsonian Astrophysical Observatory, Special Report vol 22.

Krakiwsky EJ (1991): GPS and vehicle location and navigation. GPS World, 2(5): 50–53.

Kreyszig E (1968): Advanced engineering mathematics, 2nd edn. Wiley, New York London Sydney.

Kunysz W (2000): A novel GPS survey antenna. In: Proceedings of the 2000 National Technical Meeting of the Institute of Navigation, Anaheim, California, January 26–28, 698–705, available on CD-ROM.

Lachapelle G (1990): GPS observables and error sources for kinematic positioning. In: Schwarz KP, Lachapelle G (eds): Kinematic systems in geodesy, surveying, and remote sensing. Springer, New York Berlin Heidelberg Tokyo, 17–26 [Mueller II (ed): IAG Symposia Proceedings, vol 107].

Lachapelle G (1991): Capabilities of GPS for airborne remote sensing. Canadian Journal of Remote Sensing, 17(4): 305–312.

Lachapelle G, Cannon ME, Erickson C, Falkenberg W (1992): High precision C/A code technology for rapid static DGPS surveys. In: Proceedings of the Sixth International Geodetic Symposium on Satellite Positioning, Columbus, Ohio, March 17–20, vol 1: 165–173.

Lachapelle G, Sun H, Cannon ME, Lu G (1994): Precise aircraft-to-aircraft positioning using a multiple receiver configuration. Canadian Aeronautics and Space Journal, 40(2): 74–78.

Landau H (1988): Zur Nutzung des Global Positioning Systems in Geodäsie und Geodynamik: Modellbildung, Software-Entwicklung und Analyse. Schriftenreihe der Universität der Bundeswehr München, vol 36.

Landau H, Euler H-J (1992): On-the-fly ambiguity resolution for precise differential positioning. In: Proceedings of ION GPS-92, Fifth International Technical Meeting of the Satellite Division of the Institute of Navigation, Albuquerque, New Mexico, September 16–18, 607–613.

Landau H, Ordóñez JMF (1992): A new algorithm for attitude determination with GPS. In: Proceedings of the Sixth International Geodetic Symposium on Satellite Positioning, Columbus, Ohio, March 17–20, vol 2: 1036–1038.

Langley RB (1991): The GPS receiver: an introduction. GPS World, 2(1): 50–53.

Langley RB (1993): Communication links for DGPS. GPS World, 4(5): 47–51.

Langley RB (1996): Propagation of the GPS signals. In: Kleusberg A, Teunissen PJG (eds): GPS for geodesy. Springer, Berlin Heidelberg New York Tokyo, 103–140 [Bhattacharji S, Friedman GM, Neugebauer HJ, Seilacher A (eds): Lecture Notes in Earth Sciences, vol 60].

Leick A (1995): GPS satellite surveying, 2nd edn. Wiley, New York Chichester Brisbane Toronto Singapore.

Leitinger R (1996): Tomography. In: Kohl H, Rüster R, Schlegel K (eds): Modern ionosphere science. ProduServ GmbH Verlagsservice, Berlin, 346–370.

Levanon N (1999): Instant active positioning with one LEO satellite. Navigation 46(2): 87–95.

Li Z, Schwarz KP, El-Mowafy A (1993): GPS multipath detection and reduction using spectral technique. Paper presented at the General Meeting of the IAG at Beijing, P.R. China, August 8–13.

Lichten SM, Neilan RE (1990): Global networks for GPS orbit determination. In: Proceedings of the Second International Symposium on Precise Positioning with the Global Positioning System, Ottawa, Canada, September 3–7, 164–178.

Lichtenegger H (1991): Über die Auswirkung von Koordinatenänderungen in der Referenzstation bei relativen Positionierungen mittels GPS. Österreichische Zeitschrift für Vermessungswesen und Photogrammetrie, 79(1): 49–52.

Lichtenegger H (1995): Eine direkte Lösung des räumlichen Bogenschnitts. Österreichische Zeitschrift für Vermessung und Geoinformation, 83(4): 224–226.

Lichtenegger H (1998): DGPS fundamentals. Reports on Geodesy, Warsaw University of Technology, 11(41): 7–19.

Lichtenegger H (1999): Europe on the way to GNSS. In: Proceedings of the International Symposium on Modern Information and GPS Technology, Sofia, Bulgaria, November 11–12, 11–21.

Logsdon T (1992): The NAVSTAR Global Positioning System. Van Nostrand Reinhold, New York.

Loh R, Wullschleger V, Elrod B, Lage M, Haas F (1995): The U.S. Wide-Area Augmentation System (WAAS). Navigation, 42(3): 435–465.

Lu G (1995): Development of a GPS multi-antenna system for attitude determination. Reports of the Department of Geomatics Engineering of the University of Calgary, vol 20073.

Lu G, Cannon ME (1994): Attitude determination using a multi-antenna GPS system for hydrographic applications. Marine Geodesy, 17: 237–250.

Lu G, Cannon ME, Lachapelle G, Kielland P (1993): Attitude determination in a survey launch using multi-antenna GPS technology. In: Proceedings of the National Technical Meeting of the Institute of Navigation, San Francisco, California, January 20–22, 251–259.

McCarthy DD (ed) (1996): IERS Conventions. Observatoire de Paris, IERS Technical Note 21.

Mader GL (1990): Ambiguity function techniques for GPS phase initialization and kinematic solutions. In: Proceedings of the Second International Symposium on Precise Positioning with the Global Positioning System, Ottawa, Canada, September 3–7, 1233–1247.

Malys S, Slater J (1994): Maintenance and enhancement of the World Geodetic System 1984. In: Proceedings of ION GPS-94, 7th International Technical Meeting of the Satellite Division of the Institute of Navigation, Salt Lake City, Utah, September 20–23, part 1: 17–24.

Melchior P (1978): The tides of the planet earth. Pergamon Press, Oxford New York Toronto Sydney Paris Frankfurt.

Merminod B, Grant DB, Rizos C (1990): Planning GPS-surveys using appropriate precision indicators. CISM Journal, 44(3): 233–249.

Mervart L (1995): Ambiguity resolution techniques in geodetic and geodynamic applications of the Global Positioning System. PhD dissertation, University of Berne, Switzerland.

Mervart L (1999): Experience with SINEX format and proposals for its further development. In: Proceedings of the Fifth International Seminar "GPS in Central Europe", Reports on Geodesy, Warsaw University of Technology, 5(46): 103–110.

Mikhail EM (1976): Observations and least squares. IEP, New York.

Miller KS (1963): Engineering mathematics. Dover Publications, New York.

Moelker D (1997): Multiple antennas for advanced GNSS multipath mitigation and multipath direction finding. In: Proceedings of ION GPS-97, Kansas City, Montana, September 16–19, 541–550.

Montenbruck O (1984): Grundlagen der Ephemeridenrechnung. Sterne und Weltraum Vehrenberg, München.

Moritz H (1980): Advanced physical geodesy. Wichmann, Karlsruhe.

Moritz H, Mueller II (1988): Earth rotation – theory and observation. Ungar, New York.

Mueller II (1991): International GPS Geodynamics Service. GPS Bulletin, 4(1): 7–16.

Mueller T (1994): Wide area differential GPS. GPS World, 5(6): 36–44.

Mueller T, Biester M, Loomis P (1994): Performance comparison of candidate U.S. Coast Guard WADGPS network architectures. In: Proceedings of the National Technical Meeting of the Institute of Navigation, San Diego, California, January 24–26.

Nagle JR, Dierendonck AJ van, Hua QD (1992/93): Inmarsat-3 navigation signal C/A-code selection and interference analysis. Navigation, 39(4): 445–461.

National Imagery and Mapping Agency (1997): Department of Defense World Geodetic System 1984 – its definition and relationship with local geodetic systems, 3rd edn. NIMA Technical Report TR 8350.2, Bethesda, Maryland. Available as PDF file from the NIMA's Web site http://www.nima.mil/.

National Research Council (1995): The Global Positioning System – a shared national asset. Recommendations for technical improvements and enhancements. National Academy, Washington DC.

Nayak RA, Cannon ME, Wilson C, Zhang G (2000): Analysis of multiple GPS antennas for multipath mitigation in vehicular navigation. In: Proceedings of the 2000 National Technical Meeting of the Institute of Navigation, Anaheim, California, January 26–28, 284–293, available on CD-ROM.

Nee RDJ van (1992): Multipath effects on GPS code phase measurements. Navigation, 39(2): 177–190.

Neilan RE, Moore A (1999): International GPS Service tutorial. Paper presented at the International Symposium on GPS, Tsukuba, Japan, October 18–22.

Niell AE (1996): Global mapping functions for the atmosphere delay at radio wavelengths. Journal of Geophysical Research, 101(B2): 3227–3246.

Offermans GWA, Helwig AWS, Willigen D van (2000): Eurofix system overview: differential GNSS and integrity service through Loran-C. In: Proceedings of the International Symposium on Integration of Loran-C/Eurofix and EGNOS/Galileo. Bonn, Germany, March 22–23, 281–295.

Olsen DL (1992): FRP update – what to expect in the 1992 U.S. Federal Radionavigation Plan. In: Summary record of the 20th meeting of the Civil GPS Service Interface Committee, Albuquerque, New Mexico, September 14–15.

Petersen C (1991): Precision GPS navigation for improving agricultural productivity. GPS World, 2(1): 38–44.

Philippov V, Sutiagin I, Ashjaee J (1999): Measured characteristics of dual depth dual frequency choke ring for multipath rejection in GPS receivers. In: Proceedings of ION GPS-99, 12th International Technical Meeting of the Satellite Division of the Institute of Navigation, Nashville, Tennessee, September 14–17, 793–796, available on CD-ROM.

Quin X, Gourevitch S, Ferguson K, Kuhl M, Ladd J (1992): Dynamic short baseline calibration and attitude determination using Ashtech 3DF system. In: Proceedings of the Sixth International Geodetic Symposium on Satellite Positioning, Columbus, Ohio, March 17–20, vol 1: 190–199.

Ray JK, Cannon ME, Fenton PC (1999): Mitigation of static carrier-phase multipath effects using multiple closely spaced antennas. Navigation, 46(3): 193–201.

Remondi BW (1984): Using the Global Positioning System (GPS) phase observable for relative geodesy: modeling, processing, and results. University of Texas at Austin, Center for Space Research.

Remondi BW (1990a): Pseudo-kinematic GPS results using the ambiguity function method. National Information Center, Rockville, Maryland, NOAA Technical Memorandum NOS NGS-52.

Remondi BW (1990b): Recent advances in pseudo-kinematic GPS. In: Proceedings of the Second International Symposium on Precise Positioning with the Global Positioning System, Ottawa, Canada, September 3–7, 1114–1137.

Remondi BW (1991a): Kinematic GPS results without static initialization. National Information Center, Rockville, Maryland, NOOA Technical Memorandum NOS NGS-55.

Remondi BW (1991b): NGS second generation ASCII and binary orbit formats and associated interpolation studies. Paper presented at the XX General Assembly of the IUGG at Vienna, Austria, August 11–24.

Richardus P, Adler RK (1972): Map projections for geodesists, cartographers and geographers. North-Holland, Amsterdam London.

Rizos C, Han S (1995): A new method for constructing multi-satellite ambiguity combinations for improved ambiguity resolution. In: Proceedings of ION GPS-95, 8th International Technical Meeting of the Satellite Division of the Institute of Navigation, Palm Springs, California, September 12–15, part 2: 1145–1153.

RTCM (1998): RTCM recommended standards for differential GNSS (Global Navigation Satellite Systems), version 2.2. RTCM paper 11-98/SC-104-STD, Radio Technical Commission for Maritime Services, Special Committee No. 104, Washington DC.

Saastamoinen J (1973): Contribution to the theory of atmospheric refraction. Bulletin Géodésique, 107: 13–34.

Schaer S (1997): How to use CODE's global ionosphere maps. Astronomical Institute, University of Berne, see under http://www.aiub.unibe.ch.

Schaer S (1999): Mapping and predicting the earth's ionosphere using the Global Positioning System. Schweizerische Geodätische Kommission, Geodätisch-geophysikalische Arbeiten in der Schweiz, vol 59.

Schmitt G, Illner M, Jäger R (1991): Transformationsprobleme. Deutscher Verein für Vermessungswesen, special issue: GPS und Integration von GPS in bestehende geodätische Netze, vol 38: 125–142.

Schrödter F (1994): GPS-Satelliten-Navigation: Technik, Systeme, Geräte, Funktionen und praktischer Einsatz. Franzis, Poing.

Schupler BR, Clark TA (1991): How different antennas affect the GPS observable. GPS World, 2(10): 32–36.

Schwarz KP, Lachapelle G (eds) (1990): Kinematic systems in geodesy, surveying, and remote sensing. Springer, New York Berlin Heidelberg Tokyo [Mueller II (ed): IAG Symposia Proceedings, vol 107].

Schwarz KP, Cannon ME, Wong RVC (1987): The use of GPS in exploration geophysics – a comparison of kinematic models. Paper presented at the XIX General Assembly of the IUGG at Vancouver, Canada, August 10–22.

Schwarz KP, El-Sheimy N, Liu Z (1994): Fixing GPS cycle slips by INS/GPS: methods and experience. In: Proceedings of the International Symposium on Kinematic Systems in Geodesy, Geomatics and Navigation, Banff, Canada, August 30 through September 2, 265–275.

Scott C (1993): An overview of the communication options for broadcast DGPS corrections. Australian Journal of Geodesy Photogrammetry and Surveying, 58: 69–84.

Seeber G (1993): Satellite geodesy: foundations, methods, and applications. Walter de Gruyter, Berlin New York.

Seidelmann PK (ed) (1992): Explanatory supplement to the Astronomical Almanac. University Science Books, Mill Valley, California.

Seidelmann PK, Fukushima T (1992): Why new time scales? Astronomy and Astrophysics, 265: 833–838.

Sjöberg LE (1997): On optimality and reliability for GPS base ambiguity resolution by combined phase and code observables. Zeitschrift für Vermessungswesen, 122(6): 270–275.

Sjöberg LE (1998): A new method for GPS phase base ambiguity resolution by combined phase and code observables. Survey Review, 34(268): 363–372.

Sjöberg LE (1999): Triple frequency GPS for precise positioning. In: Krumm F, Schwarze VS (eds): Quo vadis geodesia ...? Festschrift for Erik W. Grafarend on the occasion of his 60th birthday. Schriftenreihe der Institute des Studiengangs Geodäsie und Geoinformatik, Universität Stuttgart, Part 2, Report vol 1999.6-2: 467–471.

Slater JA, Willis P, Beutler G, Gurtner W, Lewandowski W, Noll C, Weber R, Neilan RE, Hein G (1999): The International GLONASS Experiment (IGEX-98): organization, preliminary results and future plans. In: Proceedings of ION GPS-99, 12th International Technical Meeting of the Satellite Division of the Institute of Navigation, Nashville, Tennessee, September 14–17, 2293–2302, available on CD-ROM.

Sluiter PG (1993): The world of DGPS. Geodetical Info Magazin, 7(7): 57–61.

Spilker JJ (1996a): GPS signal structure and theoretical performance. In: Parkinson BW, Spilker JJ (eds): Global Positioning System: theory and applications. American Institute of Aeronautics and Astronautics, Washington DC, vol 1: 57–119.

Spilker JJ (1996b): Tropospheric effects on GPS. In: Parkinson BW, Spilker JJ (eds): Global Positioning System: theory and applications. American Institute of Aeronautics and Astronautics, Washington DC, vol 1: 517–546.

Stafford JV, LeBars JM (1996): A GPS backpack system for mapping soil and crop parameters in agricultural fields. The Journal of Navigation, 49(1): 9–21.

Stangl G, Hofmann-Wellenhof B, Pesec P, Sünkel H (1991): Austrian GPS reference network – concept, realization, and first results. Paper presented at the XX General Assembly of the IUGG at Vienna, Austria, August 11–24.

Stewart M, Tsakiri M (1998): GLONASS broadcast orbit computation. GPS Solutions, 2(2): 16–27.

Stubbe P (1996): The ionosphere as a plasma laboratory. In: Kohl H, Rüster R, Schlegel K (eds): Modern ionosphere science. ProduServ GmbH Verlagsservice, Berlin, 274–321.

Szabo DJ, Tubman AM (1994): Kinematic DGPS positioning strategies for multiple reference station coverage. In: Proceedings of the International Symposium on Kinematic Systems in Geodesy, Geomatics and Navigation, Banff, Canada, August 30 through September 2, 173–183.

Teunissen PJG (1993): Least squares estimation of the integer GPS ambiguities. Paper presented at the General Meeting of the IAG at Beijing, P.R. China, August 8–13.

Teunissen PJG (1994): A new method for fast carrier phase ambiguity estimation. In: Proceedings of PLANS'94 – Position Location and Navigation Symposium, Las Vegas, Nevada, April 11–15, 562–573.

Teunissen PJG (1995): The invertible GPS ambiguity transformations. Manuscripta geodaetica, 20: 489–497.

Teunissen PJG (1996): GPS carrier phase ambiguity fixing concept. In: Kleusberg A, Teunissen PJG (eds): GPS for geodesy. Springer, Berlin Heidelberg New York Tokyo, 263–335 [Bhattacharji S, Friedman GM, Neugebauer HJ, Seilacher A (eds): Lecture Notes in Earth Sciences, vol 60].

Teunissen PJG (1999a): An optimality of the integer least-squares estimator. Journal of Geodesy, 73: 587–593.

Teunissen PJG (1999b): A theorem on maximizing the probability of correct integer estimation. Artificial Satellites – Journal of Planetary Geodesy, 34(1): 3–9.

Teunissen PJG, Jonge PJ de, Tiberius CCJM (1994): On the spectrum of the GPS DD-ambiguities. In: Proceedings of ION GPS-94, 7th International Technical Meeting of the Satellite Division of the Institute of Navigation, Salt Lake City, Utah, September 20–23, part 1: 115–124.

Teunissen PJG, Jonge PJ de, Tiberius CCJM (1995): A new way to fix carrier-phase ambiguities. GPS World, 6(4): 58–61.

Teunissen PJG, Jonkman NF, Tiberius CCJM (1998): Weighting GPS dual frequency observations: bearing the cross of cross-correlation. GPS Solutions, 2(2): 28–37.

Tiberius CCJM (1998): Recursive data processing for kinematic GPS surveying. Netherlands Geodetic Commission, Publications on Geodesy, vol 45.

Townsend BR, Fenton PC, Dierendonck AJ van, Nee DJR van (1995): Performance evaluation of the multipath estimating delay lock loop. Navigation, 42(3): 503–514.

Townsend B, Wiebe J, Jakab A (2000): Results and analysis of using the MEDLL receiver as a multipath meter. In: Proceedings of the 2000 National Technical Meeting of the Institute of Navigation, Anaheim, California, January 26–28, 73–79, available on CD-ROM.

Tranquilla JM, Carr JP (1990/91): GPS multipath field observations at land and water sites. Navigation, 37(4): 393–414.

Unguendoli M (1990): A rational approach to the use of a large number of GPS receivers. Bulletin Géodésique, 64(4): 303-312.

Vollath U, Birnbach S, Landau H, Fraile-Ordoñez JM, Martín-Neira M (1999): Analysis of three-carrier ambiguity resolution technique for precise relative positioning in GNSS-2. Navigation, 46(1): 13–23.

Vorhies J (2000): WRC 200 results – GPS. Available from the IGEB's Web site htpp://www.igeb.gov.

Walsh D (1992): Real time ambiguity resolution while on the move. In: Proceedings of ION GPS-92, Fifth International Technical Meeting of the Satellite Division of the Institute of Navigation, Albuquerque, New Mexico, September 16–18, 473–481.

Wang J (1999): Modelling and quality control for precise GPS and GLONASS satellite positioning. PhD dissertation, Curtin University of Technology, Australia.

Wanninger L (1999): The performance of virtual reference stations in active geodetic GPS-networks under solar maximum conditions. In: Proceedings of ION GPS-99, 12th International Technical Meeting of the Satellite Division of the Institute of Navigation, Nashville, Tennessee, September 14–17, 1419–1427, available on CD-ROM.

Weber L, Tiwari A (1994): An ideal link for DGPS: the Radio Data System. In: Proceedings of the XX Congress of FIG (International Federation of Surveyors), Commission 5, Survey Instruments and Methods, Melbourne, Australia, March 5–12, 504.3/1–504.3/10.

Weber R, Fragner E (1999): Combined GLONASS orbits. In: Slater JA, Noll CE, Gowey KT (eds): IGEX-98 Workshop Proceedings, IGS Central Bureau, Pasadena, California, 233–246.

Wei M, Schwarz KP (1995a): Analysis of GPS-derived acceleration from airborne tests. Paper presented at the IAG Symposium G4, XXI General Assembly of IUGG, Boulder, Colorado, July 2–14.

Wei M, Schwarz KP (1995b): Fast ambiguity resolution using an integer nonlinear programming method. In: Proceedings of ION GPS-95, 8th International Technical Meeting of the Satellite Division of the Institute of Navigation, Palm Springs, California, September 12–15, part 2: 1101–1110.

Wells DE, Beck N, Delikaraoglou D, Kleusberg A, Krakiwsky EJ, Lachapelle G, Langley RB, Nakiboglu M, Schwarz KP, Tranquilla JM, Vanicek P (1987): Guide to GPS positioning. Canadian GPS Associates, Fredericton, New Brunswick, Canada.

Wielen R, Schwan H, Dettbarn C, Lenhardt H, Jahreiß H, Jährling R (1999): Sixth Catalogue of Fundamental Stars (FK6), Part I: Basic Fundamental Stars with direct solutions. Veröffentlichungen Astronomisches Rechen-Institut, Heidelberg, vol 35, Braun, Karlsruhe.

Willis P, Boucher C (1990): High precision kinematic positioning using GPS at the IGN: recent results. In: Bock Y, Leppard N (eds): Global Positioning System: an overview. Springer, New York Berlin Heidelberg Tokyo, 340–350 [Mueller II (ed): IAG Symposia Proceedings, vol 102].

Wu JT, Wu SC, Haj GA, Bertiger WI, Lichten SM (1993): Effects of antenna orientation on GPS carrier phases. Manuscripta geodaetica, 18: 91–98.

Wunderlich T (1992): Die gefährlichen Örter der Pseudostreckenortung. Habilitation Thesis, Technical University Hannover.

Wylde GP, Featherstone WE (1995): An evaluation of some stop-and-go kinematic GPS survey options. Australian Surveyor, 40(3): 205–212.

Zhu J (1993): Exact conversion of earth-centered, earth-fixed coordinates to geodetic coordinates. Journal of Guidance, Control, and Dynamics, 16(2): 389–391.

Zhu SY, Groten E (1988): Relativistic effects in GPS. In: Groten E, Strauß R (eds): GPS-techniques applied to geodesy and surveying. Springer, Berlin Heidelberg New York Tokyo, 41–46 [Bhattacharji S, Friedman GM, Neugebauer HJ, Seilacher A (eds): Lecture Notes in Earth Sciences, vol 19].

Subject index

Springer-Verlag and the Environment

WE AT SPRINGER-VERLAG FIRMLY BELIEVE THAT AN international science publisher has a special obligation to the environment, and our corporate policies consistently reflect this conviction.

WE ALSO EXPECT OUR BUSINESS PARTNERS – PRINTERS, paper mills, packaging manufacturers, etc. – to commit themselves to using environmentally friendly materials and production processes.

THE PAPER IN THIS BOOK IS MADE FROM NO-CHLORINE pulp and is acid free, in conformance with international standards for paper permanency.