

GLOBAL POSITIONING SYSTEM

Signals, Measurements, and Performance

Second Edition

**Pratap Misra
Per Enge**

EXHIBIT

2007

GLOBAL POSITIONING SYSTEM

Signals, Measurements, and Performance

Second Edition

Pratap Misra

Lincoln Laboratory

Massachusetts Institute of Technology

Per Enge

Department of Aeronautics and Astronautics

Stanford University



Ganga-Jamuna Press

*To Prakash,
Hülya, Ray, and Tara*

To Elaine and Nick

ISBN: 0-9709544-1-7

Copyright © 2001, 2004, 2006 by Pratap Misra and Per Enge

Designed by Randall Warners

All rights reserved. No part of this work may be reproduced or stored or transmitted by any means, including photocopying, without the written permission of the copyright holders. Translation in any language is prohibited without the permission of the copyright holders.

Printed in the United States of America

Cover illustration courtesy of The Boeing Company

Ganga-Jamuna Press
P.O. Box 692
Lincoln, Massachusetts 01773
GPStextbook@G-JPress.com

Contents

PART I Fundamentals

Chapter 1

Introduction	3
1.1 A Brief History of Navigation	4
1.1.1 Longitude and Time	
1.1.2 Astronomical Methods	
1.1.3 Twentieth Century Developments: Inertial Navigation and Radio	
1.2 Methods of Radionavigation	12
1.2.1 Trilateration	
1.2.2 Hyperbolic Positioning	
1.2.3 Doppler Positioning	
1.3 Radionavigation Systems.....	16
1.3.1 Terrestrial Radionavigation Systems: Loran and Omega	
1.3.2 Satellite Navigation Systems: Transit, GPS, and GNSS	
1.4 Summary	25
Homework Problems	
References	

Chapter 2

GPS in 2005: An Overview	29
2.1 Objectives, Policies, and Status.....	30
2.2 System Architecture	32
2.2.1 Space Segment	
2.2.2 Control Segment	
GPS Coordinate Frame and Time Reference	
2.2.3 User Segment	
2.3 Signals	37
2.3.1 Signal Structure	
2.3.2 Anti-Spoofing (AS) and (the Late) Selective Availability (SA)	
2.3.3 Signal Power	
2.4 Receivers, Measurements, and Performance.....	43
2.4.1 Evolution of Receiver Technology	
2.4.2 Signal Acquisition and Tracking	
2.4.3 Estimation of Position, Velocity, and Time (PVT)	
2.4.4 Positioning Accuracy	
2.5 Differential GPS (DGPS).....	49

	Local-Area Differential GPS, Wide-Area Differential GPS, DGPS Positioning Accuracy	
2.6	Civil Applications	53
	Timing, Precise Positioning, Aviation and Space Navigation, Land and Maritime Navigation, Consumer Market	
2.7	GPS at a Glance	59
2.8	Summary	61
	Homework Problems	
	References	

Chapter 3

	Future Global Navigation Satellite Systems	67
3.1	Performance Metrics	68
3.2	Frequency Allocations	70
3.3	Spreading Codes and Ranging Signals	71
3.4	GPS Modernization	73
	3.4.1 New Civil Signals and Their Benefits	
	L2C Signal, L5 Signal	
	3.4.2 New Military Signals	
	3.4.3 Control Segment Modernization	
	3.4.4 GPS III	
	3.4.5 Development Timetable	
3.5	GLONASS	79
	3.5.1 System Segments	
	3.5.2 Frequency Plan and Signal Structure	
	3.5.3 Development Timetable	
3.6	Galileo	81
	3.6.1 System Segments	
	3.6.2 Services	
	3.6.3 Frequency Plan and Signal Structure	
	3.6.4 Development Timetable	
3.7	Compatibility and Interoperability of GPS, GLONASS, and Galileo	86
3.8	Performance of GPS+GLONASS+Galileo	87
3.9	Summary	89
	Homework Problems	
	References	

Chapter 4

	GPS Coordinate Frames, Time Reference, and Orbits	91
4.1	Global Coordinate Systems	93
	4.1.1 Terrestrial and Inertial Reference Systems	
	Conventional Terrestrial Reference System (CTRS); Conventional Inertial Reference System (CIRS)	
	4.1.2 Geodetic Coordinates, Geoid, and Datums	
	Ellipsoid and Ellipsoidal Coordinates; Definition of Height*; Regional Datums and Map Projections*	
	4.1.3 World Geodetic System 1984 (WGS 84)	
4.2	Time References and GPS Time	105

	4.2.1 Time Scales: Astronomical and Atomic; Solar and Sidereal Times; Atomic Time; Definition of Time Epoch*	
	4.2.2 Stability Measures of Frequency Sources*	
	4.2.3 Oscillators and Their Stability*	
	4.2.4 GPS Time	
4.3	GPS Orbits and Satellite Position Determination	115
	4.3.1 Kepler's Laws*	
	4.3.2 Ideal Elliptical Orbits: Keplerian Elements	
	4.3.3 Satellite Position and Velocity	
	4.3.4 Perturbed Keplerian Orbits*	
	4.3.5 GPS Orbital Parameters	
	Precise Ephemerides, Almanac	
	4.3.6 GPS Navigation Data Message	
	Conveying Satellite Time to a Receiver: Z-Count	
4.4	GPS Satellite Constellation and Visibility Displays	129
4.5	Summary	133
	Appendix 4.A Coordinate Conversion	
	Homework Problems	
	References	

PART II Estimation of Position, Velocity, and Time

Chapter 5

	GPS Measurements and Error Sources	147
5.1	Measurement Models	148
	5.1.1 Code Phase Measurements	
	Constructing Pseudorange Measurements	
	5.1.2 Carrier Phase Measurements	
	5.1.3 An Instructive Model for the Code and Carrier Measurements	
	5.1.4 Error Sources and Models	
5.2	Control Segment Errors: Satellite Clock and Ephemeris	156
5.3	Signal Propagation Modeling Errors	157
	5.3.1 Signal Refraction, Wave Propagation, and Dispersive Media	
	5.3.2 Ionospheric Delay	
	Phase Advance and Group Delay; Obliquity Factor; Delay Estimation with Dual-Frequency Measurements; Broadcast Model	
	5.3.3 Tropospheric Delay	
	Dry and Wet Delays; Tropospheric Models; Mapping Functions	
5.4	Measurement Errors	174
	5.4.1 Receiver Noise	
	5.4.2 Multipath	
	5.4.3 Measurement Error Models	
5.5	User Range Error (URE)	177
5.6	Measurement Error: Empirical Data	179
5.7	Combining Code and Carrier Measurements	181
	5.7.1 Single-Frequency Measurements	
	5.7.2 Dual-Frequency Measurements	
5.8	Error Mitigation: Differential GPS (DGPS)	185
	5.8.1 Error Mitigation	
	5.8.2 Local-Area DGPS and Relative Positioning	

5.8.3	Wide-Area DGPS	
5.9	Summary	194
	Homework Problems	
	References	
Chapter 6		
	PVT Estimation	199
6.1	Position Estimation with Pseudoranges	200
6.1.1	Linear Model for Position Estimation	
6.1.2	RMS Positioning Error	
	Satellite Geometry; Dilution of Precision (DOP); Distribution of DOPs;	
	What's DOP Good For?	
6.1.3	Price of an Inexpensive Receiver Clock	
6.1.4	Other Performance Measures and Specifications	
6.1.5	Empirical Positioning Results	
	SPS Position Estimates; DGPS Position Estimates	
6.2	Position and Velocity from Pseudorange Rates	218
6.2.1	Velocity Estimation	
6.2.2	Position Estimation	
6.3	Time Transfer	220
6.4	Summary	224
	Appendix 6.A. Parameter Estimation	
	Homework Problems	
	References	
Chapter 7		
	Precise Positioning with Carrier Phase	233
7.1	Carrier Phase and Integer Ambiguity Resolution:	
	A Simple Model	234
7.2	Carrier Phase Measurements and Precise Positioning	238
7.2.1	Carrier Phase Measurements	
7.2.2	Precise Relative Positioning and Navigation	
7.3	Elimination of Nuisance Parameters	241
7.3.1	Single Difference	
	Estimation of Position and Change in Position: The Role of Geometric Diversity	
7.3.2	Double Difference	
7.3.3	Triple Difference	
7.3.4	Integer Ambiguity Resolution and Position Estimation	
7.4	Resolving Ambiguities One at a Time	251
7.4.1	Using Code Measurements to Estimate Integers	
7.4.2	Dual-Frequency Measurements: Wide Lining	
7.4.3	Three-Frequency Measurements: L1, L2, and L5	
7.5	Resolving Ambiguities as a Set	258
7.5.1	Linear Model for Position Estimation	
7.5.2	Float Solution	
7.5.3	Search Techniques	
	Constraints on the Integers*; Local Minima Search (LMS) Algorithm*	
	Correlations among the Double-Difference Measurements*; LAMBDA Method*	
7.6	Precise Point Positioning	272
7.6.1	Measurement Models	
7.6.2	Online Positioning Services	

7.7	Summary	275
	Homework Problems	
	References	

PART III GPS Signals

Chapter 8		
	Signals and Linear Systems	283
8.1	Overview	286
8.1.1	Linear Time-Invariant Systems	
8.1.2	Sinusoids	
8.1.3	Singularity Functions: Unit Step, Unit Pulse, and Impulse	
8.1.4	Signal Power and Energy	
8.2	Convolution	294
8.2.1	Superposition of Impulse Responses	
8.2.2	Example: Moving Averages	
8.3	Transfer Functions and Basis Functions	299
8.3.1	Response to a Single Imaginary Exponential	
8.3.2	Response to a Single Cosine Wave	
8.3.3	Response to a Single Complex Exponential	
8.3.4	Vector Representations of Signals	
8.4	Fourier Series	303
8.4.1	Definition and Discussion	
8.4.2	Example: Square Wave or Sampling Waveform	
8.4.3	Parseval's Theorem	
8.5	Fourier Transforms	307
8.5.1	Derivation from Fourier Series	
8.5.2	Energy Spectrum	
8.5.3	Fourier Transform Properties	
8.5.4	Transforms of Key Functions	
8.5.5	Modulation	
8.5.6	Convolution Revisited	
8.5.7	Ideal Filters	
8.5.8	Moving Averages and Butterworth Filters	
8.5.9	Bandwidth of Signals and Filters	
8.6	Random Signals	321
8.6.1	Moments	
8.6.2	Tone Interference	
8.6.3	Noise in Linear Systems	
8.6.4	White Noise and Noise-Equivalent Bandwidth	
8.7	Laplace Transforms	329
8.7.1	Definition and Discussion	
8.7.2	Key Properties and Transforms	
8.7.3	Example: Moving Averages Revisited	
8.7.4	Solving Linear Differential Equations	
8.7.5	Characteristic Equation	
8.7.6	Connection Between the Laplace and Fourier Transforms	
8.7.7	Initial and Final Value Theorems	
8.8	Summary	339
	Homework Problems	
	References	

Chapter 9	
GPS Signals.....	345
9.1 Civil Signal on L1.....	347
9.1.1 Time Domain Description	
9.1.2 Amplitude Spectrum	
9.2 Auto-Correlation.....	353
9.2.1 Random Sequences	
9.2.2 Ranging Precision	
9.2.3 Signal Acquisition	
9.3 Cross-Correlation and Channel Sharing.....	360
9.4 Maximal Length Linear Shift Register Sequences.....	362
9.5 Gold Codes of Length 31 and 1023.....	364
9.5.1 Construction	
9.5.2 Correlation Functions	
9.5.3 Code Spectrum	
9.6 Power Spectral Density.....	368
9.6.1 Long Codes	
9.6.2 C/A-Codes*	
9.7 Narrowband Radio Frequency Interference.....	373
9.7.1 Spreading the Interference	
9.7.2 Impact of Code and Line Spectra*	
9.8 P(Y) Codes on L1 and L2.....	378
9.9 New Civil Signals for GPS.....	379
9.9.1 New Civil Signal at L2	
9.9.2 New Civil Signal at L5	
9.9.3 Navigation Data and Data-Free Transmission	
9.10 Binary Offset Carrier Signals*.....	389
9.11 Summary.....	389
Homework Problems	
References	

Chapter 10	
Signal-to-Noise Ratio and Ranging Precision.....	393
10.1 Signal Path Loss and Transmit Antenna Gain.....	395
10.2 Received Signal Power and Receiver Antenna Gain.....	398
10.3 Noise.....	402
10.3.1 Noise Temperature and Noise Figure	
10.3.2 Noise in a Cascade of Subsystems	
10.4 Noise Analysis of a GPS Receiver.....	408
10.5 Delay Lock Loops and Ranging Precision.....	411
10.6 Ranging Precision in the Presence of White Noise.....	417
10.7 Ranging Precision in the Presence of Signal Reflections (Multipath).....	420
10.7.1 Long-Delay Multipath	
10.7.2 Short-Delay Multipath	
10.7.3 Multipath-Limiting Antennas	
10.8 Summary.....	425
Homework Problems	
References	

PART IV Receivers

Chapter 11	
Signal Conditioning and Acquisition.....	431
11.1 Signal Conditioning.....	434
11.1.1 Frequency Down Conversion	
11.1.2 Image Frequencies	
11.1.3 Sampling	
11.2 Signal Acquisition.....	442
11.2.1 Inphase and Quadrature Processing and Doppler Removal	
11.2.2 Ambiguity Function	
11.2.3 Ambiguity Function for a Length-31 Gold Code	
11.2.4 Ambiguity Function for Random Codes	
11.2.5 Search Area	
11.3 Statistical Analysis of Signal Acquisition.....	450
11.3.1 Union Bound	
11.3.2 Coherent Analysis	
11.3.3 Noncoherent Analysis	
11.3.4 Discussion	
11.4 Summary.....	457
Appendix 11.A Moments for the Coherent Metrics	
Appendix 11.B Densities and Moments for Noncoherent Metrics	
Homework Problems	
References	
Chapter 12	
Signal Tracking.....	467
12.1 Overview of a Signal Tracker.....	469
12.1.1 Correlators	
12.1.2 Discriminators	
12.1.3 Linear Models	
12.2 Delay Lock Loops.....	475
12.2.1 Coherent Delay Lock Loop	
12.2.2 Early Power Minus Late Power Delay Lock Loop	
12.2.3 Linear Models for the Delay Lock Loop	
12.2.4 Step Response of the Unaided Delay Lock Loop	
12.2.5 Rate-Aided Delay Lock Loop	
12.2.6 Performance in Additive White Noise	
12.3 Phase Lock Loops.....	482
12.3.1 Coherent Phase Lock Loop	
12.3.2 Costas Phase Lock Loop	
12.3.3 Navigation Data Recovery	
12.3.4 Step Response of the Second-Order Loop	
12.3.5 Steady State Error for the Second-Order Loop	
12.3.6 Performance of the Costas Loop in Additive White Noise	
12.3.7 Choosing Loop Bandwidth	
12.4 Summary.....	492
Appendix 12.A Frequency Lock Loop	
Homework Problems	
References	

PART I

Fundamentals

Chapter 13	
Coping with Radio Frequency Interference and Signal Obstructions.....	499
13.1 Overview	502
13.1.1 Nominal Signal-to-Noise Ratios	
13.1.2 Signal Obstructions	
13.1.3 Radio Frequency Interference	
13.2 Terrestrial Radio Propagation	509
13.2.1 Two-Path Model: Exact Expression	
13.2.2 Two-Path Model: Approximation for Long Range	
13.2.3 Three-Path Model	
13.3 Antennas	513
13.3.1 Two-Element Nulling Antenna	
13.3.2 Ring Nulling Antenna	
13.4 Assisted GPS	518
13.4.1 Supplant the GPS Navigation Message	
13.4.2 Support Two-Second Integration Times	
13.5 Inertial Aiding	527
13.5.1 Basics	
13.5.2 Gyros and the Sagnac Effect	
13.5.3 Combining GPS and Inertial Measurements	
13.5.4 Error Growth in One Dimension without Tilt	
13.5.5 Error Growth in One Dimension with Tilt	
13.6 Tone Interference and Adaptive A/D Converters	541
13.7 Summary	544
Homework Problems	
References	
Appendix A GPS Data Sets on the CD.....	551
Dr. Guttorm R. Opshaug and Dr. Keith Alter	
Index.....	561

Navigation is the art and science of charting a course from point A to point B and staying on that course. The subject has a long and fascinating history, what with some of our forebears crossing vast oceans guided only by the stars. But these were brave souls for the stars couldn't be counted upon to be visible. The technology of the twentieth century has now solved this problem *nearly* completely by placing artificial stars in the sky. These stars shine all the time, radiating extraordinarily faint radio signals. What the signals lack in raw power, however, they make up in cleverness of design, and provide far more information than the sailors of old ever got from the stars on the clearest of nights.

The NAVSTAR Global Positioning System (GPS) is the first of this new breed of global navigation satellite systems to become operational. Buy a \$100 GPS receiver and a map, and you wouldn't be lost as long as you have a clear view of the sky. Or, buy a pair of more capable receivers for about \$5000 each and, with a careful analysis of the measurements, you would be able to tell if the earth under your feet moved a few millimeters while you weren't looking.

In Part I we begin our study of the engineering principles of GPS with a discussion of the fundamentals. Chapter 1 is brief survey of the history of navigation from watching the heavens to measuring radio signals. Chapter 2 offers an overview of GPS, describing briefly the system, signals, receivers, performance, and applications. Chapter 3 gives a similar overview of the global navigation satellite systems now under development, including the GPS upgrade. Chapter 4 examines three topics basic to GPS: coordinate frames in which to specify a position, time scales, and satellite orbits.

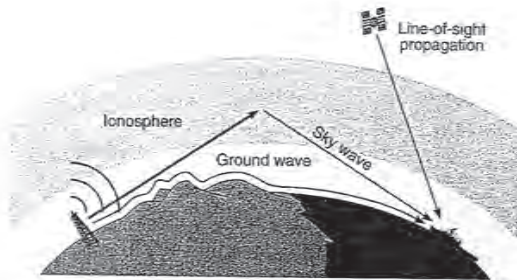


Figure 1.4 Propagation of radio waves.

VLF and LF bands. The signals of ground-based, short-range, line-of-sight radio guidance systems used in aviation (VOR, ILS, and MLS) range from VHF to SHF. The satellite navigation systems (Transit, GPS and GLONASS) have used VHF and UHF bands.

Radionavigation systems have historically been owned and operated by governments for reasons of safety and security, and to promote commerce. Loran, Omega, Transit, and GPS all started out as military systems which were subsequently opened for civil use. An exception was Decca, a commercial hyperbolic navigation system developed in the U.K. during and after World War II which guided maritime traffic in coastal areas in many parts of the world for almost fifty years.

1.2 Methods of Radionavigation

The use of radio waves for obtaining position estimates (also called *fixes*) is nearly as old as that for communication. The first radio aids determined position of a ship or aircraft by measuring bearings to two or more radio beacons with a directional antenna. The development of radio nav aids for aviation accelerated during and after World War II resulting in VHF omnidirectional radio range (VOR), Tactical Air Navigation (Tacan), Instrument Landing System (ILS), and Microwave Landing System (MLS). All are ground-based, short-range, line-of-sight systems [Enge *et al.* (1995)], which are not discussed in this book. We'll narrow our focus to long-range, global radionavigation systems and, then, quickly to GPS.

Navigators on land or over oceans are generally satisfied with the horizontal or two-dimensional (2-D) position estimates. However, the geodesists and surveyors preparing topographic maps require heights. The aviators need altitude in real time. In both cases, the horizontal and vertical positions were determined in the past separately, using different techniques and technologies. It takes a satellite-based navigation system to offer three-dimensional (3-D) position.

1.2.1 Trilateration

The simplest of the principles of radio wave propagation is that these waves travel at a known speed. Therefore, if the transit time of a signal from a transmitting station can be measured, the distance between the transmitter and the observer can be determined. Given distances to three transmitters at known locations, the observer can compute his position unambiguously.

Estimation of a position based on measurement of distances is referred to as *trilateration*. A radionavigation system based on this idea is referred to as a *time-of-arrival* (TOA) system. GPS is a TOA system.

The basic idea of trilateration is illustrated in Figure 1.5 for 2-D positioning. An observer measures his distance (or range) from station S1, whose position coordinates are known. It follows that the observer must be located somewhere on a circle of radius r_1 centered at S1. In other words, we have identified a so-called line of position (LOP), which in this case is a circle. Range measurement from a second station S2 gives another circular LOP, and reduces the uncertainty in the observer's position to the two points (P and P') where the circles intersect. The observer may be able to reject one of the estimates on the basis of prior information. If not, a third measurement would determine the position uniquely.

Mathematical modeling is straightforward. The observer's position is determined by solving a set of quadratic equations

$$\sqrt{(x_k - x)^2 + (y_k - y)^2} = r_k \quad (1.1)$$

where (x_k, y_k) are the known station coordinates; $k = 1, 2, 3$; (x, y) are the observer's coordinates to be determined; and r_k is the measured range.

Extension to 3-D positioning is straightforward, in principle. In order to estimate the vertical position, it is required that the elevation angle of at least one of the stations relative to the observer be 'large' [Section 6.1.2]. The terrestrial radionavigation systems are limited in terms of transmitter heights and, therefore, are limited to 2-D navigation. With space-based signal transmitters, each range measurement would specify a surface of position, a sphere. Intersection of three such spheres would identify a point on or near the earth unambiguously.

In order to measure the signal transit time, it is required that the clocks at the transmitters and the receiver be maintained in synchronization. The radio signals travel at about 3×10^8 m/s, and a synchronization error of $1 \mu\text{s}$ would result in an error of 300 m in distance measurement. A transmitter can be synchronized to another unit at a known distance from which

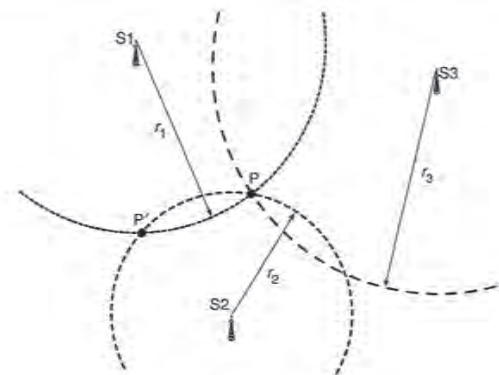


Figure 1.5 Trilateration.

it receives a signal. The demands on a receiver clock, however, must be kept to a minimum to keep the price of a receiver down if the system is planned for wide use. As discussed later in this chapter, GPS minimizes demands on the receiver clock at the expense of an additional radio transmitter and increased computational complexity in the receiver. The latter is easily accommodated in this age of microprocessors, but lack of raw compute power in the early radionavigation receivers put this approach at a disadvantage in comparison to an alternate approach discussed next.

1.2.2 Hyperbolic Positioning

These systems are based on measurement of the *difference* in the times of arrival of signals from two transmitting stations. The transmitters are synchronized; a receiver clock has to measure a time difference, typically a few milliseconds, and does not have to be synchronized to the transmitter clocks. A radionavigation system based on this idea is called a *time-difference-of-arrival* (TDOA) system. Loran is such a system. Omega, the first truly global radionavigation system, now discontinued, was also a TDOA system. Both are described briefly in the next section.

The basic idea is illustrated in Figure 1.6 for 2-D positioning. An observer measures the difference between his ranges to two stations at known locations. The locus of the points having the same range difference from the pair of stations is a hyperbola. That's the first LOP. The measurement of a range difference from a second pair of stations would specify another hyperbola as a second LOP. The observer lies at the intersection of the two hyperbolas and, therefore, this approach to positioning is called *hyperbolic positioning*. At a minimum, we need three stations forming two station pairs. There are position ambiguities due to multiple intersections of the LOPs. These ambiguities can be resolved on the basis of prior knowledge (i.e., a rough idea of where the user is), or additional measurements.

For transmitter pairs (M, X) and (M, Y), Figure 1.6 shows two families of hyperbolic LOPs x_i and y_i corresponding to the different values of the range difference. For example, hyperbola

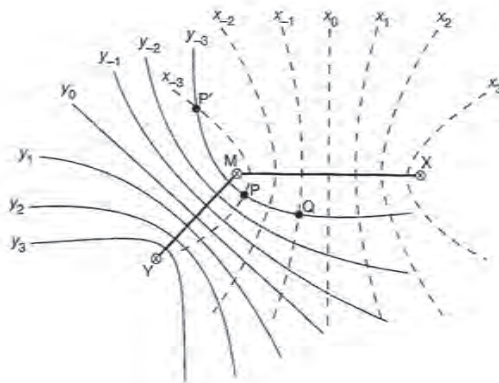


Figure 1.6 Hyperbolic positioning.

marked x_i is the locus of points that are i units closer to transmitter X than to transmitter M. A user at position P may find it difficult to rule out position estimate P', and may require another measurement. A user at Q would have an easier time of dealing with such an ambiguity.

As before, we represent the user coordinates as (x, y) and the known coordinates of the three transmitters M, X, and Y in Figure 1.6 as (x_k, y_k) , $k = 1, 2$, and 3, respectively. Let d_{21} be the measured difference between the observer's distances to stations 2 and 1; d_{31} is similarly defined. The user position is obtained by solving the following equations for (x, y) .

$$\begin{aligned}\sqrt{(x_2 - x)^2 + (y_2 - y)^2} - \sqrt{(x_1 - x)^2 + (y_1 - y)^2} &= d_{21} \\ \sqrt{(x_3 - x)^2 + (y_3 - y)^2} - \sqrt{(x_1 - x)^2 + (y_1 - y)^2} &= d_{31}\end{aligned}\quad (1.2)$$

1.2.3 Doppler Positioning

The *Doppler effect* is another principle of propagation of radio waves (actually waves, in general) exploited for radionavigation. Recall that the Doppler effect is the change in the apparent frequency of a signal received by an observer due to relative motion between a transmitter and the observer. *Doppler shift*, defined as the difference between the frequency of the received signal and the frequency at the source, is a part of our everyday experience, but is less intuitive as a basis for a global navigation system. In fact, the first satellite navigation system, called Transit, was based on the measurement of Doppler shift. This approach to positioning is called *Doppler positioning*.

The basic principle of Doppler positioning is illustrated in Figure 1.7 for one-dimensional (1-D) positioning based on the apparent pitch of the whistle of a locomotive on a straight and level track passing an observer at a steady high speed. For an observer standing in the middle of the track (not recommended), the pitch starts out high and then drops sharply as the locomotive runs over him and recedes. The classical explanation of the apparent change in pitch

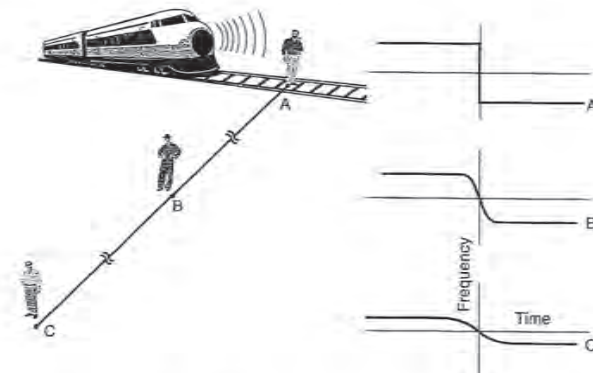


Figure 1.7 Doppler positioning: The Doppler shift profile can be related to the observer's distance from the track.

is as follows. As the locomotive approaches, more cycles are received in a time interval than transmitted due to the shrinking distance between the transmitter and the observer, and the pitch seems higher; the opposite is true as the locomotive recedes.

The transmitted and received frequencies, f_T and f_R , respectively, are related by

$$f_R = f_T \left(1 - \frac{r}{v_s} \right) \quad (1.3)$$

where r is the (changing) line-of-sight distance between the transmitter and the receiver, r is the range rate, and v_s is the speed of propagation of the waves. The range rate can also be thought of as the line-of-sight component of the transmitter-receiver relative velocity vector. The Doppler shift is

$$(f_R - f_T) = -\frac{r}{\lambda} \quad (1.4)$$

where λ is the wavelength of the transmitted signal.

The change in pitch for the observer standing in the middle of the track is a step function (see Figure 1.7). An observer standing away from the track also experiences Doppler shift, but the change between the high and low pitch is smooth. The farther away the observer is from the track, the lower the range rate, and the smaller and slower the change in pitch, as shown in Figure 1.7. The shape of the Doppler shift curve determines the observer's distance from the track. This is 1-D positioning, except for the fact that we have an ambiguity; we can't tell which side of the track the observer is on.

Armed with a watch and the train's schedule, we can turn the above scheme into a 2-D positioning system as follows. Record the time when the pitch of the whistle was midway between the high and the low. Check the train's schedule to determine its position along the track. The along-track position would be in error to the extent the train does not follow the time table, or the observer's watch is biased. Now we have the along-track and cross-track positions of the observer. That's essentially how Transit, the first satellite navigation system, discussed below, worked.

1.3 Radionavigation Systems

1.3.1 Terrestrial Radionavigation Systems: Loran and Omega

Loran

The radionavigation era began in earnest during World War II with the development of *Gee* in Great Britain to guide aircraft and *Loran* (Long-range navigation system) in the United States primarily for ships. Both systems were designed as high-power hyperbolic systems transmitting pulse signals. We discuss Loran, a later version of which remains in service.

Loran-A was developed at the Radiation Laboratory (the legendary RadLab) of Massachusetts Institute of Technology (MIT). It was developed, tested, and deployed with extraordinary speed. By the end of the war, the North Atlantic and parts of the Pacific were covered by Loran-A. Loran-C, the system in use today, was developed in the late 1950s and early 1960s to

expand the range and to improve the accuracy of Loran-A. Loran-C remained a military system until 1974 when it became a Federally provided radionavigation system for civil marine use in the U.S. coastal areas. The system was subsequently expanded to cover all of the coastal and inland navigable waters. In the mid-1980s, the coverage was increased further to fill the mid-continent gap, and the Federal Aviation Administration (FAA) approved the use of Loran-C for navigation in the U.S. airspace.

Loran-C is a hyperbolic system operating in the LF band at 90–110 kHz. The system comprises a set of chains of transmitters. A typical chain consists of a master and two to three secondary transmitting stations separated by about 1000 km, each transmitting at about 1 MW of peak power. The U.S. Coast Guard operates 29 transmitting stations composing 13 chains covering the U.S. coastal waters, conterminous 48 states, the Aleutians, and stretching into the Bering Sea [Loran-C (1992)]. Loran-C never became a global system, though it had the capability. Currently, the system covers much of the northern hemisphere. Loran-C provides 2-D rms positioning accuracy of about 250 m.

The synchronized transmitters of a chain radiate pulses of RF energy. A receiver measures the time difference (TD) between arrival of pulses from the master and secondary stations. Each measured TD defines a hyperbolic LOP for the user. Intersection of two LOPs defines the user position in 2-D. The Loran signal has both ground wave and sky wave components. The ground wave is generally stable and predictable. The sky wave component is not suited for precise positioning because of uncertainties associated with the height of the ionosphere, and the Loran pulse is designed to prevent the delayed sky wave from interfering with the TD measurements from the ground wave [Enge *et al.* (1995)].

In the mid-1990s, the U.S. Government decided to decommission Loran in 2000, but subsequently reversed the decision. The current view is that Loran can serve as a sort of insurance against catastrophic loss of GPS service. The GPS-Loran combination is seen as a robust system, both for navigation in safety-of-life applications such as civil aviation and as a source of precise time. The Loran infrastructure in the United States is now being upgraded, and the system is likely to be around until 2010, and perhaps later. A number of Loran stations continue to be operated in Europe, Asia, and the Middle East. A Russian system similar to Loran, called *Chayka*, continues to be operational.

Omega

Omega was the first worldwide, continuously available radionavigation system. It was developed in the early sixties by the U.S. Navy, the U.S. Coast Guard, and six international partners. The system became operational in the 1970s although the last permanent station (located in Australia) did not come on line until 1982. Omega was decommissioned in 1997.

Omega was realized with eight ground-based transmitters, two of which were located in the United States. Each station operated in continuous-wave mode, transmitting synchronized time-shared signals at five frequencies in the VLF band: 10.2 kHz, 11.05 kHz, 11.333 kHz, 13.6 kHz, and a unique frequency to identify the station. The radiated power of the transmitting stations was about 10 kW. The signals propagated around the globe between the ground and the ionosphere in sky wave mode, trapped in the earth-ionosphere 'duct.' The system was used for transoceanic shipping and saw use in civil aviation after the FAA approved its use as sole means for oceanic navigation and as supplemental for en route navigation [FRP (1994)].

The positioning error of a radionavigation system depends upon the uncertainties in the

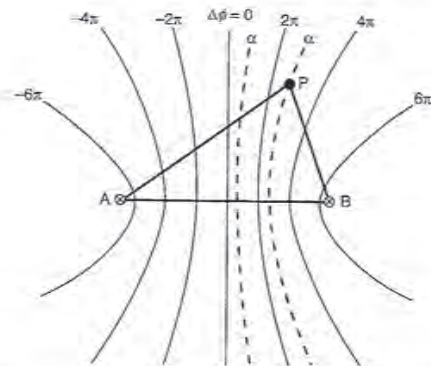


Figure 1.8 Omega lane boundaries and lane ambiguity. A user measuring phase difference α (modulo 360°) has to narrow the position uncertainty enough to identify the correct lane.

signal propagation path. For Omega, empirical propagation models were developed on the basis of worldwide measurements over extended periods. These models were built into the receivers in the form of propagation tables and formulas, and were used to correct the measurements. The system provided 2-D rms positioning accuracy of 2–4 km.

Omega provides us with a simple context in which to introduce two ideas which will come up again when we get to GPS. The first deals with the ambiguity of phase measurements, and the second with differential corrections to improve the quality of the position estimates. We discuss both below [Omega (1983), Enge *et al.* (1995)].

Omega was a hyperbolic navigation system with one subtle difference: Rather than measuring time difference of arrival, an Omega receiver measured the phase difference $\Delta\phi$ between the sinusoidal signals. A line of constant phase difference is a hyperbola, but not a unique one. Lines of constant phase difference recur over the coverage area. Figure 1.8 shows hyperbolic LOPs corresponding to 0° phase difference. In general, given the phase difference $\Delta\phi$, the 'correct' LOP could be any of the hyperbolas corresponding to phase differences $\Delta\phi \pm \text{integer multiples of } 360^\circ$. The region between the lines of zero phase difference is called a *lane*. In order for a phase measurement to provide an unambiguous position estimate, the user has to determine the correct lane first.

For 10.2 kHz signals, the width of a lane along the line joining the two transmitters is about 15 km (one-half of a wavelength). If the *a priori* position uncertainty is less than 7.5 km, the correct lane is identified and phase difference measurements can be used to determine position. If not, we have *lane ambiguity*, and must define wider lanes by using phase differences for a different pair of frequencies. Consider phase measurements at 10.2 and 13.6 kHz. Lines corresponding to constant phase difference at both frequencies repeat after about 45 km (one-half wavelength of the difference frequency of 3.4 kHz). If the initial position uncertainty is greater yet, phase measurements at 10.2 and 11.333 kHz can be used to give a lane width of about 130 km. Given an initial user position, the receiver can keep track of the lanes crossed in the course of a voyage and produce subsequent position estimates without any ambiguity.

The Omega signals propagated over thousands of kilometers under conditions that could change significantly. The propagation models used to correct the measurements basically accounted for the average conditions. The actual propagation conditions could deviate considerably from the average, and this uncertainty was the largest source of error in the Omega measurements. The conditions, however, change slowly, and are similar for users within tens of kilometers of each other.

By monitoring Omega signals at a fixed, known position, the measured carrier phase difference could be compared with that predicted by the models used to build the propagation tables. The discrepancy between the two was the measurement error which is common to all Omega users in the area. The measured error, if broadcast to the users in the area as a *differential correction*, allowed them to mitigate their measurement errors and improve the quality of their position estimates. This system, called *Differential Omega*, was implemented in parts of the Atlantic Ocean and the Mediterranean and Caribbean Seas, where the corrections were broadcast over marine radiobeacons.

The idea of differential corrections is relevant to other radionavigation systems as well. The performance of Loran can be improved significantly in differential mode. The use of differential GPS is already widespread [Section 5.8].

1.3.2 Satellite Navigation Systems: Transit, GPS, and GNSS

The space age began in 1957 with the launch of *Sputnik 1* by the Soviet Union, opening a new frontier for both the Cold War and commerce. Both the United States and the Soviet Union allocated vast resources over the next ten years to the space race. An important benefit to flow from this intense competition was that the potential for space-based communication, navigation, and surveillance systems was recognized and realized quickly.

Transit

The genesis of Transit was a brilliant discovery followed by an equally brilliant deduction. First the discovery: The pattern of Doppler shifts in the signals transmitted by *Sputnik 1* measured from a single ground station at a known position was enough to determine the satellite's orbit. Now the deduction: If the satellite orbit were known, a radio receiver measuring Doppler shifts could determine its position on the earth. It was not immediately clear, however, if the idea could actually be realized.

In 1958, there was much to be learned about designing reliable, long-life spacecraft, launching them into space, keeping the antennas pointed toward the earth, and determining accurate orbits, given the non-uniform nature of the earth's gravity field. The U.S. Navy, however, had a vital need for a navigation system to guide a new class of submarines carrying Polaris missiles, and the first satellite navigation program was born. The system was named *Navy Navigation Satellite System*, but came to be better known as *Transit*.

Transit was a pioneering achievement of the Applied Physics Laboratory of The Johns Hopkins University, which was responsible for the whole thing from the initial concept developed in 1958 to the experimental satellites launched in 1961–1962, and the final system which became operational in 1964. Of course, the credit must be shared with the enterprising leadership of the U.S. Navy, which backed a revolutionary idea and moved swiftly to turn it into an operational system within six years. Our description is based on a lucid account by Stansell (1983).

Transit was realized with four to seven satellites in low-altitude (1100 km), nearly circular, polar orbits. Each satellite broadcast signals at 150 MHz and 400 MHz with a total transmitted power of one watt. Only one satellite was in view at a time and a user waited up to 100 minutes between successive satellite passes to determine position. After a satellite came in view, the receiver recorded continuously the Doppler shift of the received signal and the navigation message giving the satellite position during the satellite pass lasting about 10–20 minutes. Actually, the receivers determined a Doppler count, or integrated Doppler, over an interval instead.

$$\text{Doppler count} = \int_{t_{i-1}}^{t_i} (f_T - f_R) dt \quad (1.5)$$

where f_T and f_R are the transmitted and received frequencies, respectively. In measuring Doppler count, the receiver effectively formed a beat-frequency signal by differencing the received signal from a sinusoidal signal generated by the receiver clock with a nominal frequency f_T . The cycles of the beat signal, counted as so many whole cycles and a partial cycle, gave the Doppler count. From the definition of Doppler shift (1.4),

$$\begin{aligned} \text{Doppler count} &= \frac{1}{\lambda} \int_{t_{i-1}}^{t_i} \dot{r}(t) dt \\ &= \frac{1}{\lambda} [r(t_i) - r(t_{i-1})] \end{aligned} \quad (1.6)$$

The Doppler count is thus a direct measure of change in distance between the receiver and the satellite over a time interval. For each wavelength the satellite moves away, the Doppler count goes up by one, and vice versa.

The measurements from a satellite pass were subsequently processed to compute 2-D position of a stationary or slow-moving user. Dual-frequency measurements allowed for correction for the ionospheric propagation delay [Section 5.3.2]. The user's velocity, if known, could be included in the mathematical model for positioning. Any error in the velocity estimate resulted in penalty in positioning accuracy. (Rule of thumb: 1 km/h velocity error could translate into a 200 m error in the position estimate.) As in the previous example [Section 1.2.3], the shape of the Doppler shift curve provided a measure of the user's distance from the satellite's ground track, and the receiver clock and the satellite almanac determined the position along the ground track. The Doppler shift profiles for two points placed symmetrically about the satellite ground track are different due to the earth's rotation. This automatically resolved the position ambiguity we had observed in relation to the train track in our earlier example.

Transit was used by the U.S. submarine fleet to update a ship's position and reset the inertial navigation system. The 2-D positioning accuracy of Transit was about 25 m (rms) for a stationary user. The system saw limited civil use in the maritime industry and geodesy starting in 1967. A Transit receiver tracking and recording measurements from multiple satellite passes from a fixed location over several days could achieve a 3-D accuracy of 5 m. This mode of usage is called *absolute positioning* or *fixed-site survey*. An alternative is *relative positioning*: estimating the position of a point relative to another point. Accuracy of 1 m was achieved in 3-D relative positioning with Transit at a distance of hundreds of kilometers by tracking multiple

satellite passes concurrently at both points.

There was a Cold War echo from the Soviet Union in the form of two systems virtually identical to Transit: *Parus* for the Soviet Navy and *Tsikada* for merchant ships. Transit was decommissioned in 1996. Parus appears to be active in 2005, but not Tsikada.

The Doppler-based system worked well for ships at sea which required infrequent position updates and could track a signal continuously during the 10–20 minute satellite pass. This technique is not suited for aircraft or mobile users requiring frequent or continuous positioning.

GPS

The success of Transit begat follow-on programs for space-based navigation systems. The U.S. Navy and the Air Force each had a development program in the late 1960s. The two were eventually combined into GPS. A dozen years had passed since the idea of Transit was conceived, and much had been learned about designing spacecraft, launching them in space, and tracking and maintaining them in orbit.

It is instructive to reconstruct the considerations of the GPS designers that led to the choice of system architecture. We do so below, and also introduce some new concepts.

- Active or passive system: A passive system broadcasts signals and a user determines his position by simply 'listening' to them. An active system interacts with each user (e.g., a user may be required to interrogate the system). An active system can handle a fixed, finite number of concurrent users; a passive system can service an unlimited number of users. Besides, a military user wouldn't want to give away his position by radiating signals.
- Positioning Method: Doppler, hyperbolic, or trilateration. In 1970, the clock technology had improved to a point that time-synchronized signals could be transmitted from satellites. Trilateration was the logical choice. GPS is, therefore, described as a *passive ranging system*.
- Pulsed versus continuous wave (CW) signals: Pulsed signals could be time-coordinated so that each satellite would have a separate transmission time. Spread spectrum signaling, however, allows simultaneous signal transmissions on one radio frequency. GPS is the first widespread application of this signaling concept called code division multiple access (CDMA) [Chapter 9].
- Carrier frequency: L-band (1 GHz–2 GHz) was optimum. GPS needed 20 MHz of spectrum, and L-band was relatively uncluttered in the early 1970s (no longer true). At higher frequencies, ranging error due to ionospheric refraction decreases but space loss (i.e., attenuation of signal power due to distance traveled) [Section 10.1] and atmospheric attenuation increase.
- Satellite constellation and orbits: Each user needs four or more satellites in view in order to determine her position (see below). Such worldwide coverage had to be provided economically. The choice of orbits determines satellite design, number of satellites, and costs of launch and maintenance.
 - Low earth orbit (LEO): With altitude < 2000 km, a LEO satellite is only visible for 10–20 minutes at a time, and the receiver would be acquiring new satellites constantly. Doppler rates would be high. Orbital perturbations due

to the atmospheric drag would be high, too. It would take a constellation of 100–200 satellites to provide global coverage for navigation. On the plus side, launch costs would be lower and the satellites could get by with low-power transmitters.

- Medium earth orbit (MEO): With altitude of 5000–20,000 km and 2–4 orbits per day, the satellites are typically visible for several hours in each pass. Launch costs are higher than those for a LEO, but a smaller satellite constellation (24–36 satellites) would be adequate.
- Geostationary orbit (GEO): A satellite in a 36,000-km orbit over the equator appears fixed to an observer on the earth. Global coverage can be provided with a small number of satellites. On the negative side, GEO satellites offer poor coverage at higher latitudes, and have higher launch costs.

These considerations led to a choice of a MEO constellation of 24 satellites for GPS. The Department of Defense (DoD) approved the basic architecture in 1973, and the first satellite was launched in 1978. The system was declared operational in 1995. The cost of development of GPS has been reported to be about \$10 billion; the annual operation and maintenance cost is estimated to be about \$500 million. As an aside, we should note that the full name given to GPS was *NAVSTAR, the Global Positioning System*. NAVSTAR is sometimes represented as an acronym for *Navigation System with Time and Ranging*. Though apt, the DoD did not intend NAVSTAR as an acronym [Parkinson (1996)].

Three engineers/scientists have been honored widely for their roles in the development of GPS: Roger L. Easton [Easton (1980)], Ivan A. Getting [Getting (1993)], and Bradford W. Parkinson [Parkinson (1996)]. Among the many honors Easton has received is the 2004 National Medal of Technology for “his invention of ‘Navigation System using Satellites and Passive Ranging Techniques’ and his subsequent development of Time Navigation and Navigation Technology Satellites that formed the technological basis for the modern GPS.”

Getting (1912–2003), President of the Aerospace Corporation in the 1960s and 1970s, is credited with advocacy and leadership in the development of GPS. The 50th GPS satellite (GPS IIR-11), launched on 20 March 2004, was dedicated to him. Parkinson, the first Director of the GPS Joint Program Office (GPS JPO) in the 1970s, served as the chief architect during its development and implementation, and shepherded the program through the DoD acquisition process with great skill. Among the many honors Getting and Parkinson received is the shared \$500,000 Charles Stark Draper Prize of the National Academy of Engineering in 2003 “for the concept and development of the Global Positioning System.” (It’s interesting that the White House and the NAE saw the responsibility for the development of GPS so differently.)

Four technologies which saw spectacular developments in the second half of the twentieth century and turned the ancient idea of trilateration into GPS are:

- stable space platforms in predictable orbits,
- ultra-stable clocks,
- spread spectrum signaling,
- integrated circuits.

Recall that trilateration required measurement of ranges to three position references (i.e.,

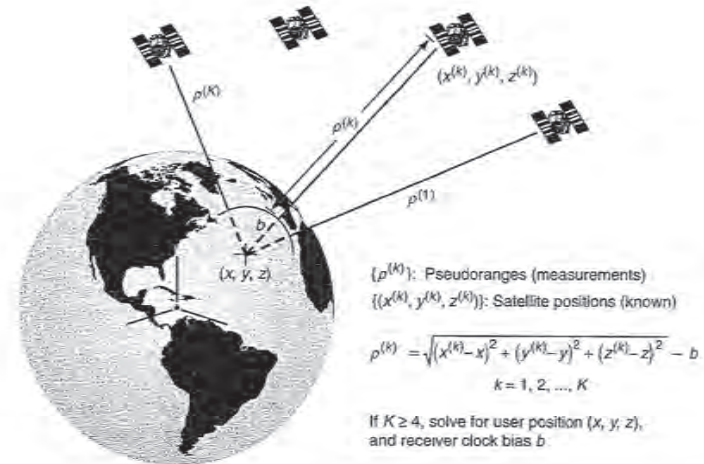


Figure 1.9 The principle of satellite navigation. The user-satellite range measurements based on the times of transmission and receipt of signals are biased by a common amount and are called pseudoranges. Pseudorange measurements are needed from at least four satellites to estimate the user position and receiver clock bias.

objects at known positions). In GPS, the position references are the satellites, which are actually moving in space at a speed of about 4 km/s. The position of a satellite at any instant, however, can be estimated with an error no larger than a few meters based on predictions made 24–48 hours earlier. The distance between the user and a satellite is measured in terms of transit time of the signal from the satellite to the user. The transmission times are imprinted upon the signals in accordance with nearly perfect and nearly perfectly synchronized atomic clocks carried aboard the satellites. The precise estimation of the arrival times is made possible by transmitting spread spectrum signals, which have wide bandwidths but each satellite can transmit its unique signal on the common frequency band. Finally, the spectacular developments in microelectronics have made the receiver light, compact, and an order-of-magnitude less expensive than thought possible in 1980.

In order to measure the true transit time of a signal from a satellite to a receiver, clearly, the clocks in the satellite and the receiver must be maintained in synchronism. Fortunately, this onerous requirement is easily sidestepped, allowing use of inexpensive quartz oscillators in the receivers. The bias in the receiver clock at the instant of the measurements affects the observed transit times for all satellites equally. The corresponding measured ranges are thus all too short, or too long, by a common amount, and are called *pseudoranges*. The receiver clock bias thus becomes the fourth unknown to be estimated, in addition to the three coordinates of position. A user therefore needs a minimum of four satellites in view to estimate his four-dimensional position: three coordinates of spatial position, plus time. An idealized geometrical view of the pseudorange measurements and the resulting equations to be solved for the user position and receiver clock bias are given in Figure 1.9.

Table 1.2 A performance summary of radionavigation systems

System	Coverage		Fix dimensions	Positioning accuracy (rms)
	Worldwide	Continuous		
Loran-C	No	Yes	2-D	250 m
Omega	Yes	Yes	2-D	2–4 km
Transit	Yes	No	2-D	25 m
GPS	Yes	Yes	3-D plus time	Horizontal: 5 m Vertical: 7.5 m

We have now introduced GPS and the radionavigation systems which were its progenitors: Loran-C, Omega, and Transit. The performance characteristics of these systems are summarized in Table 1.2. The positioning accuracy numbers are based on FRP (1994).

GNSS

GPS is the first of a new generation of navigation satellite systems to become operational. Similar systems and satellite-based augmentations are being developed and deployed by governments, international consortia, and commercial interests. The generic name given to these systems is *Global Navigation Satellite Systems* (GNSS). A brief survey appears below. An expanded discussion is found in Chapter 3.

While GPS was under development, the Soviet Union undertook to develop a similar system called *GLONASS*. GLONASS, like GPS, was designed primarily for the military, and a subset of its signals was offered for civil use in the era of *glasnost* and *perestroika* in the late 1980s, apparently as an afterthought. Since the dissolution of the Soviet Union, the Russian Federation has assumed responsibility for GLONASS. The civil user community was excited at the prospect of having access to two autonomous systems as GLONASS built up to a full constellation of 24 prototype satellites in 1996. Unfortunately it didn't last [Langley (1997)]. This once high-prestige system has suffered from lack of resources in the changed political and economic climate, and appears to have been kept going in skeletal form by a dedicated GLONASS cadre. Between 2000 and 2005, the Russians managed only one launch per year, each placing three new satellites in orbit. While a couple of GPS+GLONASS receiver models have been on the market since the mid-1990s, the user community and receiver manufacturers have seemed wary of GLONASS. But there appears to be new optimism in 2005 among GLONASS watchers [Section 3.5].

A more promising GNSS now under development is the European system, Galileo, planned as "an open, global system, fully compatible with GPS, but independent from it." After several years of intense discussions and serious haggling among the member states, the European Union (EU) gave the final green light in 2002 to proceed with the development of the system. The EU has two pressing reasons for this initiative. First, concerns of sovereignty and security require that Europe must control its safety-critical navigation systems, and EU has no control over GPS. Secondly, this step is intended to help the EU countries become players in

the technology and commerce associated with satellite navigation. As was to be expected, the United States, with a vested interest in maintaining the preeminence of GPS, was not enthusiastic about the idea of a rival, but now appears to have made peace with it.

Galileo was conceived as a joint public-private enterprise under civilian control, to be financed and managed by the European Commission (EC), the European Space Agency (ESA), and industry. The estimated cost of developing and deploying the system is €3.2 billion. While the industry was slow to commit funds in the early phases of the program, EC and ESA committed over €1.1 billion in 2002 to proceed with the development phase of the system. The deployment phase will begin in 2006. The system could be operational in 2008, though 2010 appears more realistic [Section 3.6].

Galileo would be a second-generation GNSS, following the trail blazed by GPS and benefiting from the technological advances since the mid-1970s when GPS was designed. As we'll see in Chapter 3, GPS isn't sitting still either. All agree that Galileo and GPS must be compatible and interoperable. The United States, fearing that Galileo could complicate the situation with regard to selective denial of satellite navigation service to an adversary, pressured the Europeans with a demand that the Galileo signals must also be 'jammable' without harm to GPS military signals.

China has developed a regional satellite navigation system called *BeiDou*. They launched two GEO satellites late in 2000 and another in May 2003, apparently completing the system which provides positioning capability over China and neighboring areas. BeiDou (Big Dipper-1) is an active system, which, unlike GPS, GLONASS, and Galileo, interacts with the users. The Mission Control Center determines a position estimate and transmits it to each user. The system accuracy is reported to be "dozens of meters."

The Japanese government in conjunction with an industry consortium is undertaking development of *Quasi-Zenith Satellite System* (QZSS), a regional system to transmit ranging signals over Japan from satellites in highly elliptical orbits, and to transmit differential corrections to GPS and other GNSS satellites. The first satellite launch is planned for 2008, with two more to follow. The cost of development, deployment, and maintenance of the system is expected to be ¥170 billion (\$1.57 billion) over twelve years. Other augmentation systems are described briefly in Chapter 2.

The idea of satellite navigation is not limited to the earth. After two recent high-profile failures in the Mars program, NASA has considered a system of low-orbiting satellites around Mars to provide GPS-like navigation signals and a better communications link to the control center.

GPS is likely to remain the only operational system of its kind at least until 2008, when Galileo is planned to be ready. GPS, however, is evolving into a more capable system. Satellites to be launched in 2005 and thereafter will broadcast additional signals. The completed system with considerably expanded capabilities for both military and civil users is expected to emerge between 2015 and 2020. The plans for GPS modernization are described in Chapter 3.

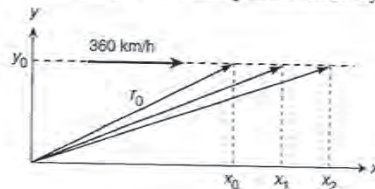
1.4 Summary

In this chapter, we surveyed the history of navigational tools and technology developed over centuries. Especially noteworthy are the efforts to find longitude at sea. There are heroic tales of craftsmen who built clocks with great mechanical ingenuity to keep time at sea, and the

competing astronomers who preferred to measure lunar distance from a star and then do a prodigious amount of calculation by hand. It's all fascinating stuff, and we literally ran through it in a half-dozen pages. The twentieth century brought new technologies of radio and inertial navigation with enormous consequences for war and commerce. To all who would look impatiently at their wristwatches, cell phones, PDAs, or car navigation systems with moving maps for directions to their destinations, we recommend J.E.D. Williams' wonderful monograph for balance and perspective. As for GPS, we'll get plenty of it in the 500-plus pages that follow.

Homework Problems

- 1-1. Given that 1 minute of latitude is approximately equal to 1 nautical mile (1852 meters), how many significant digits after the decimal must be included for a latitude represented in *degrees* to describe a fix that is accurate to 1 cm? How many significant digits are required after the decimal in the *arc-seconds* field if the latitude is represented in degrees, arc-minutes, and arc-seconds to describe a fix that is accurate to 1 cm? Note: 1 degree = 60 arc-minutes, 1 arc-minute = 60 arc-seconds (you may find arc-minutes referred to as 'minutes' and arc-seconds referred to as 'seconds').
- 1-2. Suppose you start from location 45° N , 120° W (lat, long) and fly at an altitude of 10 km with ground speed of 885 km/h for eight hours at a constant heading of 45° from true north. Where would you end up? Assume a spherical earth with radius of 6371 km. Two helpful notes: (i) Ground speed in aviation means the speed of an aircraft relative to the surface of the earth (to be distinguished from air speed, which means the speed of an aircraft relative to its surrounding air mass). (ii) It's safe to say that you'll end up pretty far north. Don't worry too much about precision—you get full credit if your answer is within a kilometer of the exact answer.
- 1-3. An aircraft is carrying a transmitter that is broadcasting a single tone at 100 MHz. The aircraft flies away from you on a straight line at constant altitude with ground speed of 360 km/h. You measure the following Doppler shifts from 100 MHz spaced 0.1 s apart: -33.1679 Hz , -33.1711 Hz , and -33.1743 Hz . Determine the range rates in m/s that correspond to these Doppler measurements. In the figure below, the aircraft altitude is y_0 , and its horizontal distance from observer at the three instants of Doppler measurements is shown as x_0 , x_1 , and x_2 , respectively. Set up two linear equations that relate x_1 and x_2 to x_0 . Can you set up two non-linear equations that relate x_0 and y_0 to the measurements? For extra credit, solve the equations iteratively.



- 1-4. A *pseudolite* (short for pseudo-satellite) consists of a generator of a GPS-like signal and a transmitter. Pseudolites are used to augment the GPS signals. Suppose an observer is constrained to be on the line joining two pseudolites PL1 and PL2, which are separated by 1 km. The pseudolite clocks are perfectly synchronized but the observer's clock may have an unknown bias with respect to the pseudolite clocks. Estimate the observer's position and clock bias given that the pseudoranges to PL1 and PL2 are (a) 550 m and 500 m, respectively, and (b) 400 m and 1400 m, respectively.
- 1-5. You set out from town A and head east to town B 120 km away. Your vehicle has an odometer that is not particularly accurate (it could be off by 1–2 km, or more, after driving 50 km). You carry a decent watch, which is great at keeping time over short intervals, but it has been months since you last reset it. In other words, you can measure time intervals accurately, but do not know exactly what time it is. A short time into your journey, the car breaks down.
- (a) According to your odometer, you have traveled 56 km. Estimate your position.
- (b) As you push your car to the shoulder of the road, a red bus zooms by heading from A to B. You glance at your watch and notice that it is exactly 21 minutes past the hour. You know that the red buses are prompt, and they leave town A every hour on the hour traveling at exactly 3 km/min. Can you estimate your position without using the odometer information?
- (c) Estimate your position and clock bias based on all the information so far. (Hint: Write two equations that relate your position and clock bias to the available information. These equations are sometimes referred to as navigation equations.)
- (d) At 25 minutes past the hour by your watch, you observe a blue bus zoom past at 2.5 km/min, going from B to A. Blue buses leave town B every hour on the hour promptly and drive to town A at a constant speed. Estimate your position and clock bias based on all the information so far. (Hint: Recall that your watch can measure time intervals accurately.)
- (e) How would your solution be affected if your watch were exactly five minutes faster and all the clocks in town A and town B were running five minutes fast?
- (f) Now suppose that your odometer never worked and the only vehicles you see are the identical yellow cabs of carrier L1. These cabs leave town A every minute on the minute and travel at exactly 1 km/min to town B. Can you estimate your position and clock error? Would it help if there were identical green cabs of carrier L2 leaving town B every minute on the minute and traveling at 1 km/min to town A? Explain briefly.

References

- Allan, David W., Neil Ashby, and Clifford C. Hodge (1997). *The Science of Timekeeping*. Application Note 1289, Hewlett-Packard Corporation.
- Bowditch, Nathaniel (1802). *The American Practical Navigator*, issued by the Defense Mapping Agency Hydrographic Office, Washington, DC, 1984.
- Brown, Lloyd A. (1956). The Longitude, in *The World of Mathematics*, J.R. Newman (ed.), Simon and Schuster, pp. 780–819.
- Easton, R.L. (1980). The Navigation Technology Program, in *Global Positioning System, Vol. I*, The Institute of Navigation, pp. 15–20.
- Engel, Per, Eric Swanson, Richard Mullin, Ken Ganther, Anthony Bommarito, and Robert Kelly (1995). Terrestrial Radio Navigation Technologies, *Navigation*, vol. 42, no. 1, pp. 61–108.
- Farrell, Jay A., and Matthew Barth (1999). *The Global Positioning System and Inertial Navigation*, McGraw Hill.
- Federhen, H.M. (1993). *Worldwide Radionavigation Systems*, IDA Paper P-2852, Institute for Defense Analysis, Alexandria, VA.
- Forssell, Börje (1992). *Radionavigation Systems*, Prentice Hall.
- FRP (1994). *Federal Radionavigation Plan*, U.S. Departments of Defense and Transportation.
- Getting, Ivan A. (1993). The Global Positioning System, *IEEE Spectrum*, vol. 30, no. 12, pp. 36–47.
- Greenspan, Richard L. (1995). Inertial Navigation Technology from 1970 to 1995, *Navigation*, vol. 42, no. 1, pp. 165–185.
- Howse, Derek (1980). *Greenwich Time and the Discovery of the Longitude*, Oxford University Press.
- IEEE (1983). *Proceedings of the IEEE*, Special Issue on Global Navigation Systems, vol. 71, no. 10.
- Kayton, Myron (ed.) (1989). *Navigation: Land, Sea, Air, and Space*, IEEE Press.
- Langley, Richard B. (1997). GLONASS: Review and Update, *GPS World*, vol. 8, no. 7, pp. 46–51.
- Loran-C (1992). *Loran-C User Handbook*, U.S. Coast Guard.
- Omega (1983). *Omega Global Radio Navigation, A Guide for Users*, U.S. Coast Guard.
- Parkinson, Bradford W. (1996). Introduction and Heritage of NAVSTAR, the Global Positioning System, in *Global Positioning System: Theory and Applications I*, B.W. Parkinson, J.J. Spilker, P. Axelrad, and P. Enge (eds.), American Institute of Aeronautics and Astronautics, pp. 3–28.
- Snyder, John P. (1987). *Map Projections—A Working Manual*, U.S. Geological Survey Professional Paper 1395, U.S. Government Printing Office.
- Stansell, Thomas A. (1983). *The TRANSIT Navigation Satellite System*, Magnavox Technical Report, R-5933A.
- Williams, J.E.D. (1992). *From Sails to Satellites: The Origin and Development of Navigational Science*. Oxford University Press.

Chapter 2

GPS in 2005: An Overview

- 2.1 Objectives, Policies, and Status
- 2.2 System Architecture
 - 2.2.1 Space Segment
 - 2.2.2 Control Segment
 - GPS Coordinate Frame and Time Reference
 - 2.2.3 User Segment
- 2.3 Signals
 - 2.3.1 Signal Structure
 - 2.3.2 Anti-Spoofing (AS) and (the Late) Selective Availability (SA)
 - 2.3.3 Signal Power
- 2.4 Receivers, Measurements, and Performance
 - 2.4.1 Evolution of Receiver Technology
 - 2.4.2 Signal Acquisition and Tracking
 - 2.4.3 Estimation of Position, Velocity, and Time (PVT)
 - 2.4.4 Positioning Accuracy
- 2.5 Differential GPS (DGPS)
 - Local-Area Differential GPS, Wide-Area Differential GPS, DGPS Positioning Accuracy
- 2.6 Civil Applications
 - Timing, Precise Positioning, Aviation and Space Navigation, Land and Maritime Navigation, Consumer Market
- 2.7 GPS at a Glance
- 2.8 Summary
 - Homework Problems
 - References

Our objective in this chapter is quite ambitious. We aim to provide a compact, technical overview of GPS as it exists in the summer of 2005: the system, signals, measurements, performance, and applications. The subsequent chapters pick up many of the main ideas introduced herein and provide a comprehensive treatment. As such, this chapter is intended to serve both as an introduction to GPS and a road map to this book.

The satellites now scheduled for launch starting in September 2005 will broadcast additional signals for the military and civil users. But it'll be a half-dozen years before there are enough satellites broadcasting the new signals to make them widely useful. We describe these new signals in Chapter 3, where we also introduce the signals and services planned for GLONASS and Galileo.

2.1 Objectives, Policies, and Status

The principal objective of the U.S. Department of Defense (DoD) in developing GPS was to offer the U.S. military accurate estimates of position, velocity, and time. In quantitative terms, this statement was interpreted broadly as providing estimates with position error of 10 meters, velocity error of 0.1 meters/second, and time error of 100 nanoseconds, all in the root-mean-square (rms) sense. These estimates were to be available to an unlimited number of users all over the globe continuously and nearly instantaneously. The planned military use also required the system to be usable on high-dynamics platforms, and the signals to have a measure of resistance to jamming and interference. And, finally, the adversaries of the United States were to be denied the full benefits of the system.

The DoD also planned to provide the civil users of GPS a 'reasonable' accuracy consistent with the national security considerations. These considerations were formalized in a policy to offer two kinds of service:

- *Standard Positioning Service (SPS)* for peaceful civil use,
- *Precise Positioning Service (PPS)* for the DoD-authorized users.

Access to the full capabilities of the system (i.e., PPS) is restricted by cryptographic techniques. The system transmits encrypted signals intended for DoD-authorized users equipped with the appropriate decryption keys. This feature is called *Anti-Spoofing (AS)*. Each satellite transmits one signal for civil use under SPS. It rankled the civil user community that the SPS signal was degraded throughout the 1990s by introducing controlled errors to reduce its precision and limit the civil users to accuracy of 100 meters. Such errors could be removed by the DoD-authorized users. This feature, called *Selective Availability (SA)*, was deactivated by a Presidential Order on 2 May 2000. Both SA and AS are discussed below.

We should note as an aside that there was considerable tension between DoD and the potential civil users in the early years of GPS. The U.S. military at first appeared reluctant—even hostile—to the idea of sharing their marvelous new toy. During 1980–1981, when Charles C. Counselman III (MIT) and Peter MacDoran (Jet Propulsion Laboratory) built receivers and demonstrated that GPS could be used for precise positioning even if the signals were encrypted, they were required to register with the U.S. Government as arms manufacturers. Happily, the situation changed, albeit slowly, and the present relationship between the U.S. military and the civil users can be described as harmonious.

The policy for civil use of GPS was first announced by the DoD in the late 1970s with an assurance that the SPS signals will remain freely available. In 1983, Korean Airline Flight 007 went off course into Soviet airspace, apparently due to navigation problems, and was shot down. This disaster drew attention to the potential benefits of GPS for civil aviation, and the U.S. policy on civil use of GPS was reaffirmed by President Reagan. At the time, the system

was ten years away from being declared operational. Subsequently, the United States made a formal commitment in 1991 to the International Civil Aviation Organization (ICAO), a specialized agency of the United Nations, to make "GPS-SPS available for the foreseeable future on a continuous, worldwide basis, and free of direct user fees."

While free signals and growing markets for GPS products and services created general enthusiasm in the civil sector worldwide, the governments, particularly in Europe, were reluctant to dismantle the existing terrestrial navigation aids (navaids) and become dependent upon a system controlled by the military of a foreign government. Many held the view that technology as basic and vital as GPS required an international institutional framework for development of policies, regulations, and standards, but the U.S. Government wasn't about to share control of GPS. The European Union (EU) therefore decided to develop a satellite navigation system of their own, to be called *Galileo*. As was to be expected, the U.S. Government was not enthusiastic about this idea, but appears now to have made peace with it.

The U.S. policy on GPS is based on balancing the basic requirement of retaining the military advantage of this technology with considerations of commerce and international policy. An interesting account of the issues and considerations is found in the report of a study conducted in 1994–1995 at the direction of the U.S. Congress by the National Academy of Public Administration (NAPA) and the National Research Council (NRC) of the National Academy of Sciences and Engineering [NAPA-NRC (1995)]. This report has been influential in shaping the evolution of both GPS and the U.S. policy on its use. The report concluded that SA was not serving its intended purpose, and recommended that it be discontinued immediately. The report recommended that "immediate steps should be taken to obtain authorization to use an L-band frequency for an additional GPS signal" for civil use. The report also recommended several measures on governance of GPS to achieve the national goals and to promote its international acceptance.

The Presidential Decision Directive (PDD) of 1996 [US GPS Policy (1996)] adopted most of the recommendations of the NAPA-NRC report, and paved the way for discontinuance (in 2000) of SA. According to the PDD, one of the U.S. policy objectives is to promote integration of GPS into peaceful civil, commercial, and scientific applications worldwide, and to advocate acceptance of GPS as an international standard. Accordingly, changes were made in the governance of GPS. The management responsibility was assigned to the Interagency GPS Executive Board (IGEB), co-chaired by representatives of the DoD and the Department of Transportation (DOT). But that didn't mean DoD and DOT had the same clout in managing GPS. The funds to maintain and modernize GPS, and to operate IGEB, all came out of the DoD budget. We are reminded of an old proverb, something about paying the piper and calling the tune.

The Presidential Decision Directive of 1996 was superseded in December 2004 by the U.S. Space-Based Positioning, Navigation, and Timing Policy [US PNT Policy (2004)]. The new complexities are reflected in the fact that while the PDD barely filled three pages, the PNT Policy takes up 11 pages. There is no sharp departure in policy, only more things to address, especially on security. Navigation warfare [Section 3.4], not mentioned in the 1996 document, makes repeated appearances. The new policy includes a restatement of issues of free and open use of civil signals, denial of hostile use, and a clear statement of cooperation with Galileo. The new policy identifies GPS as a component of the "U.S. Critical Infrastructure" to give the system a special status for the purpose of funding under U.S. laws. A National Space-based Positioning, Navigation, and Timing Executive Committee, co-chaired by the DoD and

DOT, will take over the responsibilities of IGEB and may project a higher profile. Its members will include representatives from the various Departments and Agencies of the Federal Government [http://www.pnt.gov/]. FRP (2005) is the official source of the U.S. radionavigation policy and plans.

GPS was declared operational in 1995 and, since then, it has performed virtually flawlessly, offering a service better than that promised in its performance specifications [Section 2.4.2]. The system emerged unscathed from two “rollover” events. First, there was the end-of-week rollover (EOW or W1K) in August 1999 [Section 4.2.4]; the second was the Y2K event. It is to the credit of the U.S. Air Force Space Command, which is responsible for launching the satellites and operating the system, that glitches due to equipment malfunction and operational lapses have been rare. The United States appears committed to providing GPS SPS without interruption or degradation, but may deny the service in selected geographical areas if required by military needs.

GPS has now proven its worth as a military system in three conflicts. The first test came in the Gulf War (1990–1991) in which GPS helped ground troops navigate in the featureless desert while the system only had a partial satellite constellation and was years away from being declared ready for operational use. In Kosovo in 1999, GPS-guided munition (“smart bombs”) appear to have changed the rules of war by proving that a war can be won with minimum risk of casualties for the winning side through an air campaign by targeting bombs with care and delivering them with precision so as to minimize collateral damage on the ground (a euphemism for civilian deaths). And, in Iraq War (2003), GPS was critical to the campaign of “shock and awe,” guiding and tracking aircraft, ships, missiles, armored vehicles, and troops. Based on accounts in the press, the Iraq War also saw the first attempt, albeit clumsy and futile, at jamming GPS signals.

Literature on GPS has grown to be voluminous in recent years, and includes several influential collections of papers and books offering in-depth treatments of the system and its applications from different viewpoints [El-Rabbany (2002), Farrell and Barth (1999), Forssell (1992), Grewal *et al.* (2001), Hoffmann-Wellenhof *et al.* (1997), IEEE (1999), ION (1980), Kaplan (1996), Leick (2004), Parkinson *et al.* (1996), Seeber (2003), Strang and Borre (1997), Teunissen and Kleusberg (1998), Tsui (2004), Van Dierendonck (1997), Wells (1986), and Xu (2003)]. We owe a debt to the authors of these books and papers which have influenced the account that follows.

2.2 System Architecture

GPS consists of three segments: the Space Segment, the Control Segment, and the User Segment. The Space Segment comprises the satellites and the Control Segment deals with the management of the satellite operations. The DoD is responsible for both. The User Segment covers activities related to the development of military and civil GPS user equipment (i.e., receivers). The development of receivers and services in the civil sector is essentially left to market forces. The civil sector, however, clearly benefited from the investments by the DoD in the 1970s and 1980s in development of military receivers. This favor is now being returned as the design of a new generation of GPS military user equipment benefits from the extraordinary developments in the 1990s in civil receiver design and manufacturing.

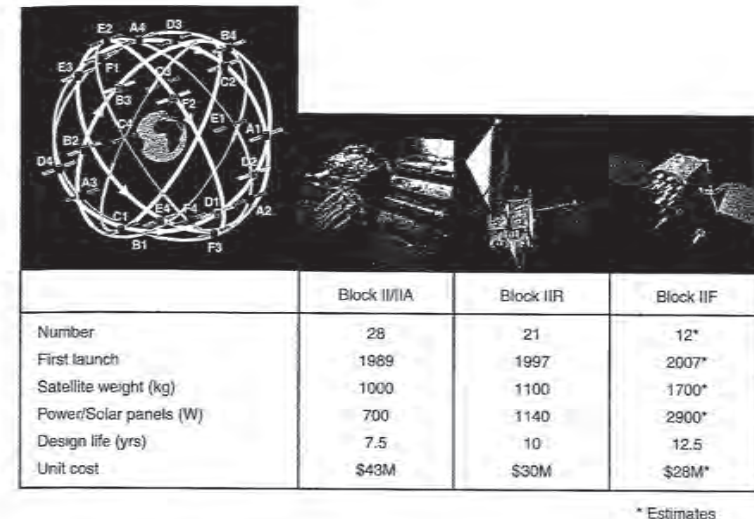


Figure 2.1 The GPS Space Segment consists of a baseline constellation of 24 satellites distributed in six orbital planes. Each new batch of satellites was designed to provide additional capabilities and, it should be noted, came with a lower price tag.

2.2.1 Space Segment

The baseline constellation comprises 24 satellites fielded in nearly circular orbits with a radius of 26,560 km, a period of approximately twelve hours, and stationary ground tracks. The constellation is shown in Figure 2.1. The satellites are arranged in six orbital planes inclined at 55° relative to the equatorial plane, with four primary satellite slots distributed unevenly in each orbit. Each satellite is identified with a two-character code: a letter identifies the orbital plane (A through F); a number identifies the satellite number in the plane (1 through 4 in the baseline constellation). There is a spare satellite slot in each orbital plane, designated by the two-character code as A5, B5, etc. The system can support a constellation of up to thirty satellites on orbit.

With the baseline constellation, almost all users with a clear view of the sky have a minimum of four satellites in view. It's more likely that a user would see six to eight satellites. The satellites broadcast ranging signals and navigation data allowing the users to measure their pseudoranges and to estimate their positions in passive, listen-only mode.

An initial batch of ten prototype or developmental satellites, called Block I satellites, was launched between 1978 and 1985, and used to demonstrate the feasibility of GPS. The prototypes were followed by production models named Block II and IIA, the first launch of which occurred in February 1989. Over the next six years, the system built up to 24 Block II/IIA satellites on orbit, and was declared operational in April 1995. The next generation of satellites,

called Block IIR (the 'R' stands for replenishment), entered service in 1997. The Block IIR satellites now on order will sustain the constellation at least until 2007.

A new generation of GPS satellites, called Block IIF (the 'F' stands for follow-on), is now in production, with delivery of the first six planned for 2007. The DoD had contracted in 1996 with Rockwell International (now a part of The Boeing Company) for six Block IIF satellites with an option to buy up to 27 more. Early in 2000, the DoD announced its decision to change course. It appears now that the DoD plans to acquire only twelve IIFs. The decision was apparently motivated by changes since 1996 in the requirements under the GPS modernization plan. As we'll discuss in Chapter 3, the modernization plan calls for two new civil signals and two new military signals. These changes would have required extensive modifications to the Block IIF design. The DoD will now start with a clean slate the process of acquiring a new generation of GPS satellites to be called GPS III.

The table in Figure 2.1 summarizes the basic features of Block II/IIA, IIR, and IIF satellites. The satellite weight refers to on-orbit weight. The power generated by the solar panels refers to levels early in mission life. Each successive batch of GPS satellites was designed with higher capabilities, longer service life and, it is worth noting, a lower price tag [Fisher and Ghassemi (1999)]. A satellite in orbit represents (in round numbers) a \$100 million asset, counting the cost of building and launching it.

GPS SPS doesn't promise or require a 24-satellite constellation and the United States retains some latitude on maintenance of the satellite constellation. "The intent of the U.S. Government is to manage the constellation such that it never drops below 22 healthy satellites in nominal slots" [SPS (2001)]. Actually, the number of active satellites on orbit since 1995 has exceeded 24, and there have been serious proposals to raise the constellation strength to 27, or higher.

Early in September 2005, the constellation comprised 29 satellites: 17 Block II/IAs and 12 Block IIRs. The number of working satellites at any time can vary due to the requirements of maintenance. Satellites are taken off-line for short periods, announced in advance. Sudden catastrophic failures can reduce the constellation size as well, but the GPS satellites have been consistently reliable, and their life expectancy so far has exceeded design life significantly. Out of ten Block II/IIA satellites (design life: 7.5 years) launched in 1989–1990, all but one were in service at the beginning of 2000, and two made it to 2005.

GPS constellation status and scheduled outage updates are available from the website of the U.S. Coast Guard Navigation Center (NAVCEN) [<http://www.navcen.uscg.gov>].

2.2.2 Control Segment

At the heart of the Control Segment is the Master Control Station (MCS), located at the Schriever (formerly named Falcon) Air Force Base near Colorado Springs, Colorado. The Master Control Station operates the system and provides command and control functions. The specific functions of the Control Segment are:

- to monitor satellite orbits,
- to monitor and maintain satellite health,
- to maintain GPS Time,
- to predict satellite ephemerides and clock parameters,
- to update satellite navigation messages,

- to command small maneuvers of satellites to maintain orbit, and relocations to compensate for failures, as needed.

The satellite signals are tracked from U.S. Air Force monitor stations spread around the globe in longitude: Hawaii, Colorado Springs, Cape Canaveral, Ascension Island, Diego Garcia, and Kwajalein. In 2005, this network was expanded to include six monitor stations operated by the National Geospatial-Intelligence Agency (NGA) of the DoD: Washington, D.C., United Kingdom, Argentina, Ecuador, Bahrain, and Australia. The twelve-station network allows the system operators to watch each satellite from at least two monitor stations at all times. Five more NGA-operated monitor stations will be added to the network in the future.

The monitor stations are unmanned and are operated remotely from the Master Control Station. The Control Segment monitors the PPS signals and removes a satellite from service if anomalous behavior is detected. Equipment at a monitor station consists essentially of GPS receivers with cesium standards, meteorological instruments, and communications gear to transmit the measurements to the Master Control Station via ground and satellite links.

Dedicated ground antennas for communications with the satellites via S-band radio links are co-located with monitor stations at Ascension Island, Cape Canaveral, Diego Garcia, and Kwajalein. These ground antennas are operated remotely from the Master Control Station to receive telemetry from the satellites on the status of its subsystems and functions, to uplink commands, and to upload data to update the navigation messages broadcast by the satellites. [The Air Force Satellite Control Network (AFSCN) Automated Remote Tracking Station (ARTS) at Schriever Air Force Base can also be operated as a GPS ground antenna for command and control.]

The elements of the Control Segment (monitor stations, ground antennas, Master Control Station) and their functions are shown in Figure 2.2. The Control Segment is also referred to as the Operational Control Segment or Operational Control System, both abbreviated as OCS.

The estimation and prediction of satellite orbits and clock biases are based on data from the monitor stations. GPS Time [Section 4.2.4] is defined on the basis of a set of atomic clocks in the satellites and monitor stations. Synchronization of the satellite clocks is accomplished by estimating the time offset, drift, and drift rate of each satellite clock relative to GPS Time, and transmitting the parameters of a model of this bias in the satellite's navigation message. These parameters are a part of the navigation message transmitted by each satellite [Section 4.3.6]. The ephemeris and clock parameters broadcast by the satellites are computed by the Master Control Station and uploaded to the satellites via one of the ground antennas.

The navigation messages broadcast by the satellites are currently uploaded once daily at a minimum. The cross-link communication and ranging capability in Block IIR satellites was designed to allow the satellites to update their navigation messages autonomously and operate over extended periods without contact with the Control Segment. This Autonav function appears not to be exercised and may not be operative.

GPS Coordinate Frame and Time Reference

While on the subject of infrastructure required to support GPS operations, we mention two important ideas briefly and develop them further in Chapter 4. First, in order to represent a position or velocity, we need a coordinate frame. Before the advent of the satellite navigation systems, such coordinate frames were defined on a country or regional basis. A global positioning

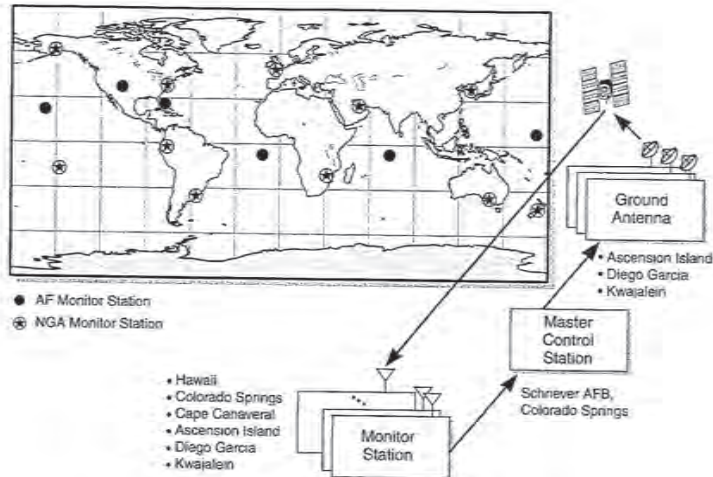


Figure 2.2 The elements of the GPS Control Segment and their functions.

system requires a global coordinate frame in which to express the position of its satellites and the users. GPS provided the impetus for development by the Defense Mapping Agency (DMA) of the DoD of the World Geodetic System 1984 (WGS 84), which has now become a *de facto* international standard, a great accomplishment in itself [Section 4.1]. DMA was subsequently reorganized and renamed the National Imagery and Mapping Agency (NIMA) and, in 2003, the National Geospatial-Intelligence Agency (NGA).

Clocks are at the heart of GPS. A receiver measures ranges to satellites by determining the transit times of signals generated aboard the satellites by synchronized clocks and carrying precise timing marks. Roughly speaking, meter-level positioning accuracy requires measurement of ranges accurate to meter level, which, in turn, requires nanosecond-level synchronization of the satellite clocks and measurement of transit time at nanosecond level. Our second important idea, therefore, deals with the definition and measurement of time. GPS defines a time reference of its own called GPS Time (GPST) and keeps track of the offset between GPST and the international civil standard called UTC [Section 4.2].

2.2.3 User Segment

The success of GPS in large-scale civil use is attributable almost entirely to the revolution in integrated circuits, which has made the receivers compact, light, and an order-of-magnitude less expensive than thought possible twenty years ago. The first receivers designed for precise positioning were introduced in the mid-1980s, and were priced at over \$100,000. Receivers with much higher capabilities are now available for less than \$5000. As late as 1980, it had only been hoped that the receiver manufacturers would be able to produce a basic GPS receiver for the mass market for about \$2000. Price barriers, however, fell quickly. An important

industry milestone was reached in 1992 with the introduction of the first hand-held receiver priced below \$1000. In 1997, the industry breached the \$100 barrier with a pocket-size receiver running on two AA batteries.

Commerce in GPS-related products and services grew rapidly in the 1990s. There are now hundreds of GPS receiver models on the market. It is estimated that more than one million receivers have been produced each year since 1997. According to the U.S. Department of Commerce, the annual worldwide sales of GPS products and services exceeded \$16 billion in 2003. GPS is on its way to becoming a part of our daily lives as an essential element of the commercial and public infrastructure.

While the GPS industry is yet to produce its Bill Gates and Michael Dell, there is money to be made in this business, as shown by Gary Burrell and Min Kao, the co-founders of Garmin International, manufacturer since 1991 of GPS receivers for the consumer and aviation markets. Both made the Forbes magazine's list of "400 Richest People" in 2003 as near-billionaires.

In the last twenty years, several generations of receivers based on newer technologies have been produced, but we are not ready to discuss the receivers yet. First, we have to understand the nature of the signals, which we take up next.

2.3 Signals

Currently, each GPS satellite transmits continuously using two radio frequencies in the L-band referred to as Link 1 (L1) and Link 2 (L2). The L-band covers frequencies between 1 GHz and 2 GHz, and is a subset of the ultra-high frequency (UHF) band. The center frequencies of L1 and L2 are as below:

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

Two signals are transmitted on L1, one for civil users, and the other for DoD-authorized users. The lone signal on L2 is intended for the DoD-authorized users only. The structure of these signals is discussed briefly below. [Actually, the satellites transmit additional RF signals at frequencies referred to as L3 and L4. These signals are associated with classified payloads aboard the satellites: L3 is associated with the Nuclear Detonation Detection System (NDS or Nudet), and L4 is associated with a payload called Reserve Auxiliary Package (RAP).]

Under the GPS modernization program, additional navigation signals will be broadcast by Block IIR-M (the "M" is for modernized) satellites to be launched starting in 2005 and by Block IIF satellites to be launched starting in 2007. These signals are described briefly in Chapter 3.

2.3.1 Signal Structure

The structure of the GPS signals is described exhaustively in Interface Specifications. The Interface Specification for the GPS signals is public [IS-GPS-200D (2004)]. Each GPS signal consists of three components (see Figure 2.3):

- *Carrier*: RF sinusoidal signal with frequency f_{L1} or f_{L2}
- *Ranging code*: Associated with each Service (i.e., GPS and PPS) is a family of

binary codes called *pseudo-random noise* (PRN) sequences or PRN codes. The PRN codes have special mathematical properties which allow all satellites to transmit at the same frequency without interfering with each other [Chapter 9]. These codes also allow precise range measurements, and mitigate the deleterious effects of reflected and interfering signals received by a GPS antenna [Chapter 10]. The SPS codes are called coarse/acquisition codes (C/A-codes) and PPS codes are referred to as precision (encrypted) codes, or P(Y)-codes. Each satellite transmits a unique C/A-code on L1 and unique P(Y)-codes on both L1 and L2.

Each C/A-code is a unique sequence of 1023 bits, called *chips* in this context, which is repeated each millisecond. The duration of each C/A-code chip is about 1 μ s. Equivalently, the *chip width* or wavelength is about 300 m. The rate of the C/A-code chips, called *chipping rate*, is 1.023 MHz [or megachips/second (Mcps)]. A coarse/acquisition code is so called because (i) it's coarse as compared to a precision code, and (ii) it was originally intended to serve as a necessary stepping stone for acquisition of a P(Y)-code. Newer technology now allows direct acquisition of the P(Y)-code.

A P-code is a unique segment of an extremely long ($\approx 10^{14}$ chips) PRN sequence. The chipping rate is 10.23 Mcps, ten times that for a C/A-code, and the chip width is about 30 m. The smaller wavelength results in greater precision in the range measurements than that for the C/A-codes. The P-codes repeat after one week. Actually, since 1994, the satellites have transmitted encrypted versions of the P-codes, called Y-codes, about which we know nothing.

We note here in passing that the length of a PRN code and its chipping rate determine the acquisition and tracking characteristics of the signal in the presence of noise and interference [Sections 3.3, 9.3, 9.4].

- **Navigation data:** a binary-coded message consisting of data on the satellite health status, ephemeris (satellite position and velocity), clock bias parameters, and an almanac giving reduced-precision ephemeris data on all satellites in the constellation. The navigation message is transmitted at a leisurely 50 bits per second (bps), with a bit duration of 20 ms. It takes 12.5 minutes for the entire message to be received. The essential satellite ephemeris and clock parameters are repeated each thirty seconds [Section 4.3.6]. A faster data rate would get the message out quicker, but would require a stronger signal to ensure demodulation with low error rate.

These three components of a signal (namely, carrier, code, and navigation data) are derived coherently from one of the atomic standards aboard the satellite. The frequency of the atomic standards aboard a satellite is 10.23 MHz. The relationship with chipping rates is obvious. Note also the relationship with the L1 and L2 carrier frequencies:

$$\begin{aligned} f_{L1} &= 1575.42 \text{ MHz} = 2 \times 77 \times 10.23 \text{ MHz} \\ f_{L2} &= 1227.60 \text{ MHz} = 2 \times 60 \times 10.23 \text{ MHz} \end{aligned}$$

The code is combined with the binary navigation data using modulo-2 addition: If the code chip and the data bit are the same (both are 0s or both are 1s), the result is 0; and if both are

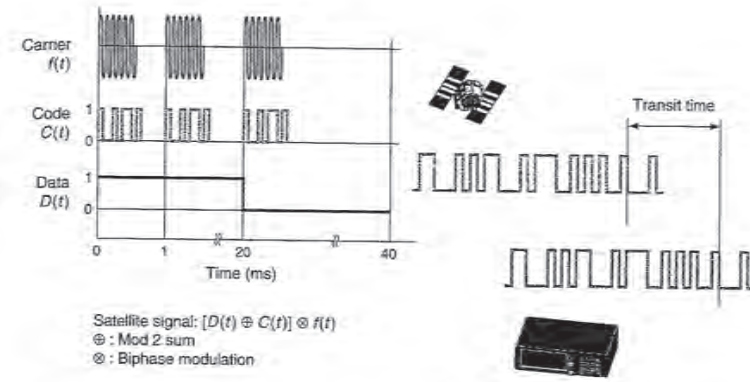


Figure 2.3 The structure of the signal available for civil use and estimation of its transit time from the satellite to a user. Each GPS signal comprises three components: an RF carrier, a unique binary pseudo-random noise (PRN) code, and a binary navigation message. The transit time is estimated by correlating the received signal with its replica generated by the receiver.

different, the result is 1. The composite binary signal is then impressed upon the carrier in a process called modulation. The specific form of modulation used is called *binary phase shift keying* (BPSK): a 0 bit leaves the carrier signal unchanged; and a 1 bit multiplies the carrier by -1 , which is equivalent to shifting the phase of the sinusoidal signal by 180° . At bit transitions from 0 to 1, or from 1 to 0, the phase of the carrier signal is shifted by 180° .

The carrier frequency in Figure 2.3 is 1575.42 MHz (L1). During a 20-ms data bit duration of the navigation message, the code repeats twenty times. In addition to a C/A-code, each satellite also transmits a P(Y)-code on L1. This is accomplished by generating two carrier signals on L1: one as generated by the clock (in-phase component), and the other is obtained by shifting it in phase by 90° (quadrature component). The in-phase component is modulated by a C/A-code, and the quadrature component is modulated by a P(Y)-code. The phase shift makes the two carriers 'orthogonal' in a sense, allowing a receiver to separate their modulating signals. The signal on L2 carries a P(Y)-code only. The three BPSK-modulated signals transmitted by each satellite are shown in Figure 2.4 showing 180° shifts in carrier phase associated with bit transitions.

The L1 and L2 signals leaving the antenna of the k th satellite can be modeled as

$$\begin{aligned} s_{L1}^{(k)}(t) &= \sqrt{2P_C} x^{(k)}(t) D^{(k)}(t) \cos(2\pi f_{L1} t + \theta_{L1}) \\ &\quad + \sqrt{2P_{Y1}} y^{(k)}(t) D^{(k)}(t) \sin(2\pi f_{L1} t + \theta_{L1}) \\ s_{L2}^{(k)}(t) &= \sqrt{2P_{Y2}} y^{(k)}(t) D^{(k)}(t) \sin(2\pi f_{L2} t + \theta_{L2}) \end{aligned} \quad (2.1)$$

where P_C , P_{Y1} , and P_{Y2} are the signal powers for signals carrying C/A-code on L1, and P(Y)-codes on L1 and L2, and, respectively; $x^{(k)}$ and $y^{(k)}$ are the C/A- and P(Y)-code sequences assigned to satellite number k ; $D^{(k)}$ denotes the navigation data bit stream; f_{L1} and f_{L2} are the

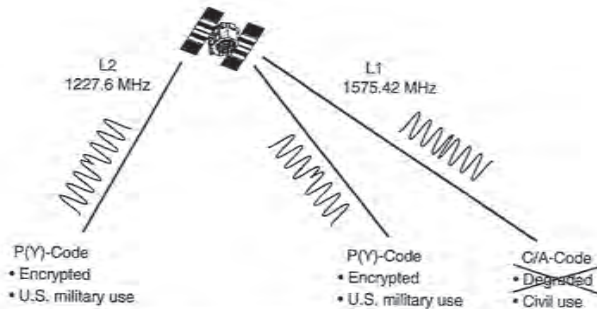


Figure 2.4 In mid-2005, each GPS satellite transmitted three signals, two on L1 and one on L2 frequency. The BPSK-modulated signals are shown. The signal carrying C/A-code on L1 was degraded purposely throughout the 1990s, but this practice has now ended. Access to P(Y)-code is limited to the DoD-authorized users via encryption.

carrier frequencies corresponding to L1 and L2, respectively; and θ_{L1} and θ_{L2} are phase offsets. The signal powers are discussed below. Note, however, that in order to express the BPSK signals as (2.1), we have switched the binary values of the codes and navigation data to ± 1 . From our old notation, a bit 0 maps into +1; and a bit 1 maps into -1. It is left as an exercise to show that (2.1) in fact represents the BPSK signals described earlier.

The modulation of a carrier by a binary code spreads the signal energy, initially concentrated at a single frequency, over a wide frequency band: over 2 MHz for the C/A-code and about 20 MHz for the P(Y)-code, centered at the carrier frequency, as shown in Figure 2.5. While the signal power is unchanged, this step reduces the power spectral density below that for the background RF radiation. Such signals, referred to as spread spectrum signals, have many properties which make them attractive for use in communication and navigation. The signal energy can be 'de-spread' in a receiver, if the code is known. In principle, keeping a code secret limits access to the signal. A comprehensive discussion of the spread spectrum signals is given in Chapter 9.

Let us take a quick look at two special properties of the PRN sequences which take values ± 1 . First, PRN sequences are nearly *orthogonal* to each other: The sum of the term-by-term products of the two sequences, shifted arbitrarily relative to each other, is nearly zero. For satellites k and l , which are assigned unique PRN sequences called C/A-codes $x^{(k)}$ and $x^{(l)}$,

$$\sum_{i=0}^{1022} x^{(k)}(i) \cdot x^{(l)}(i+n) \approx 0, \text{ for all } n, k \neq l \quad (2.2)$$

where $x^{(*)}(1023+m) = x^{(*)}(m)$. The left-hand side of (2.2) defines the *cross-correlation function* of the two sequences for shift n . The PRN sequences are nearly uncorrelated for all shifts. The PRN sequences associated with the P(Y)-codes are also similarly orthogonal to each other. The orthogonality of these codes allows all satellites to broadcast simultaneously at the same frequency without interfering with each other.

The second important property is that a PRN sequence is nearly uncorrelated with itself, except for zero shift. For a C/A-code

$$\sum_{i=0}^{1022} x^{(k)}(i) \cdot x^{(k)}(i+n) \approx 0, \text{ for all } |n| \geq 1 \quad (2.3)$$

The left-hand side of (2.3) defines the *auto-correlation function* of a sequence for shift n . The auto-correlation function of a PRN is nearly zero except for zero shift where it has a sharp peak. These cross- and auto-correlation properties of a PRN are exploited in acquisition and tracking of signals, discussed briefly in Section 2.4.1 and in greater detail in Chapters 9 and 10.

2.3.2 Anti-Spoofing (AS) and (the Late) Selective Availability (SA)

The main mechanism for limiting access to the full capabilities of GPS has been encryption of the P-code broadcast on both L1 and L2. This feature is referred to as Anti-spoofing (AS). An encrypted P-code is referred to as a Y-code. The main purpose of AS is to protect a user from spurious GPS signals with misleading data that an adversary may be tempted to transmit. AS has been active nearly continuously since 1994. Access to Y-coded signals requires the cryptographic key, which the DoD makes available to authorized users.

SPS limits civil users to the C/A-coded signal on L1 but, as we will see in Chapters 5–7, dual-frequency measurements are essential for precise positioning. Receiver manufacturers have, therefore, devised proprietary techniques to gain access to measurements on both L1 and L2. These techniques exploit in different ways the fact that the same P(Y)-code is being transmitted by a satellite on both frequencies. But there is a price: The L2 measurements are more fragile and noisier than they would be if the code were known, and the dual-frequency receivers cost a lot more. These techniques do not compromise AS because the structure of the Y-codes remains unrevealed and no spoofing signals can be generated.

Throughout the 1990s, the signals available for unrestricted use were purposefully degraded under the policy of Selective Availability (SA) by adding controlled errors in the measurements. These errors were significantly larger than the errors inherent to the system. The *net result of SA was about a five-fold increase in positioning error*. The signal degradation was achieved by 'dithering' the satellite clock and, therefore, the timing marks on the ranging signals, affecting the C/A-code, P(Y)-code, and carrier phase measurements equally. Another mechanism designed to degrade performance, though apparently not used often, was to broad-

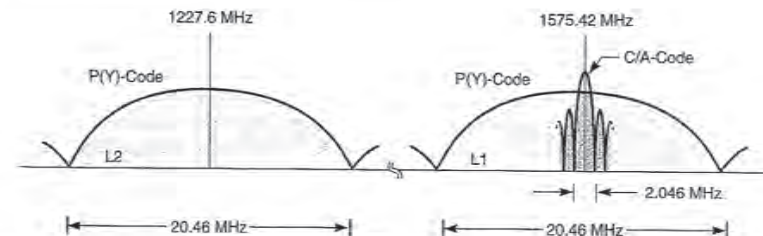


Figure 2.5 Power spectra of signals transmitted by a GPS satellite in mid-2005. The energy of the signal for civil users carrying a C/A-code on L1 is spread mainly over a 2-MHz-wide frequency band. The bandwidths of the signals for military users on L1 and L2 carrying a P(Y)-code are ten times wider.

cast erroneous or imprecise values of the ephemeris parameters. The DoD-authorized users were able to undo SA. Actually, civil users could undo SA as well, but it required additional expense. A user with several hundred dollars to spare for a radio receiver was able to obtain differential corrections in real time, as discussed below. As might have been expected, SA made few friends among civil users.

SA was deactivated on 2 May 2000 in accordance with a Presidential decree. Perhaps the European plans to develop Galileo accelerated the U.S. move to drop SA. Regardless of the political calculation, the departure of SA marked an important milestone for the civil users of GPS. Accuracy is addictive and the return of SA is unthinkable. To give an idea to a newcomer of the bad old days, we present several examples of SA-corrupted data and the corresponding position estimates later in this chapter.

With SA gone, the main advantage of PPS over SPS is robustness: The jamming resistance of PPS is higher, and spoofing it is a challenge. PPS also offers improved positioning performance due to dual-frequency measurements which allow compensation for ionospheric propagation effects [Section 5.3.2]; faster codes which lead to higher precision of range measurements [Section 10.6]; and lower error due to multipath [Section 10.7].

2.3.3 Signal Power

The GPS signals received on the earth are extremely weak. The RF power at the antenna input port of a satellite is about 50 watts, of which about half is allocated to the C/A-code. The satellite antenna is designed to spread the RF signal roughly evenly over the surface of the earth below. The satellites are said to bathe the earth with a gentle, uniform radiation. That, however, may be too sensuous an imagery for engineers.

Before we discuss the signal power levels, let us digress to introduce some terminology. In order to deal simply with a wide range of power levels, electrical engineers express power ratios on a logarithmic scale in units of decibel (dB), defined as

$$\left(\frac{P_1}{P_0}\right)_{dB} = 10 \log_{10} \left(\frac{P_1}{P_0}\right) \quad (2.4)$$

where P_1 and P_0 are the power levels to be compared. Absolute values of power can be expressed similarly in relation to 1 watt or 1 milliwatt in units of dBW or dBm, respectively. Consider a signal with power P_1 of 0.1 watts. This power level can also be represented as -10 dBW or 20 dBm. A second signal, with a power P_2 of 100 watts is 30 dB more powerful than the first signal. A third signal, with 200-watt power (P_3), is 3 dB stronger than the second signal. We can capture these relationships as follows.

$$P_1 = -10 \text{ dBW}, \quad \frac{P_2}{P_1} = +30 \text{ dB}, \quad \frac{P_3}{P_2} = +3 \text{ dB}$$

The GPS specifications on the minimum received power level for the users on the earth are given in Table 2.1. The signal powers are only around 10^{-16} watts! Interestingly, 10^{-16} watts is enough power to navigate with if we were among friends and people of good will. Adversaries and malicious hackers, however, may be tempted to exploit this vulnerability of the system.

The GPS signals are well below the background RF noise level sensed by an antenna

Table 2.1 GPS minimum received signal power specifications

Frequency band	C/A-code (dBW)	P(Y)-code (dBW)
L1	-158.5	-161.5
L2	—	-164.5

[Chapter 10]. It is the knowledge of the signal structure (i.e., PRN code) that allows a receiver to extract the signal buried in noise and make precise measurements. The signal boost so realized is called *processing gain*. If the noise level is raised by RF interference or jamming, the available processing gain may not be enough to extract the signal.

The low signal power is the Achilles' heel of GPS, especially in military use. Even in civil use, there is concern about the vulnerability of the signals as the national commercial and public infrastructure comes to rely more and more on GPS. Frequency diversity and increased signal power are important considerations in the modernization plans to make the GPS service more robust [Chapter 3].

2.4 Receivers, Measurements, and Performance

2.4.1 Evolution of Receiver Technology

Several generations of GPS receivers came to market between 1980 and 2005. The receivers available today bear the same resemblance to the early receivers as the laptop and palm computers do to the minicomputers of the early 1980s. The advent of very large scale integration (VLSI) has led to powerful microprocessors and memory chips, which have changed the look and feel of all electronic equipment, including GPS receivers.

The first set of GPS receivers was built in the 1970s for the DoD to prove that GPS would work (see Figure 2.6). In the early 1980s, the GPS constellation consisted of a few Block I satellites. There was no rush to produce receivers for the potentially huge automobile market because the system was at least ten years away from being ready for operational use and the U.S. policy on civil use was still unclear. Besides, the requisite digital street maps and databases did not yet exist. The geodesy and surveying community, however, saw the potential of GPS to offer extraordinary precision in positioning and was eager to use the system even though four satellites were in view typically for only a few hours per day. The market was limited and the focus was on precise measurements.

In 1982, two GPS receivers designed for precise geodesy came on the market: Macrometer V-1000, designed by Charles C. Counselman III of MIT and built by the Steinbrecher Corporation, and the Texas Instruments TI 4100. These receivers, bulky by today's standards, were revolutionary instruments which demonstrated millimeter-level positioning accuracy with GPS using carrier phase measurements [Chapter 7]. A receiver took up quite a bit of rack space, weighed tens of kilograms, and consumed over 100 watts of power. Each was priced at over \$100 000.

The Macrometer did not require knowledge of the PRN codes and offered carrier phase



Figure 2.6 Early GPS receivers were bulky as shown in this U.S. Air Force photograph (circa 1980).

measurements on L1 from up to six satellites by implementing the so-called squaring channel: by squaring a signal, the binary modulation due to the code and navigation message disappears, and what we are left with is a carrier at twice the frequency. Satellite ephemeris and clock synchronization, available with code measurements, had to be provided externally by other means. The TI 4100 provided C/A- and P-code pseudoranges and carrier phase measurements from four satellites at a time. The receiver actually had a single hardware channel which was switched rapidly among the satellites.

In 1989, the DoD began to launch Block II satellites to populate the constellation fully and announced plans to declare the system operational. The U.S. policy on civil use of GPS had become clear by then. Digital maps and databases were beginning to appear, and receiver manufacturers switched into high gear in anticipation of large-volume production. The focus shifted to digital processing of signals using application-specific integrated circuits (ASIC) which would lead eventually to credit card-size GPS receivers and single-chip designs to be incorporated in pagers and cellular telephones. These receivers typically provide code and carrier phase measurements for the SPS signals from all satellites in view.

2.4.2 Signal Acquisition and Tracking

The basic functions of a GPS receiver are:

- to capture the RF signals transmitted by the satellites spread out in the sky,
- to separate the signals from satellites in view,
- to perform measurements of signal transit time and Doppler shift,
- to decode the navigation message to determine the satellite position, velocity, and clock parameters,
- to estimate the user position, velocity, and time.

The signals are gathered by a hemispherical antenna. Given a recent almanac and a rough idea of the user location, the receiver determines which satellites are in view. Given the satellite ID, the receiver knows the structure of the C/A-code being transmitted by it, and attempts to 'tune' to it to *acquire* the signal, and from then on *track* changes in it continuously.

To acquire a signal, the receiver generates a replica of the known C/A-code, and attempts to align it with the incoming code by sliding the replica in time and computing the correlation. From the auto-correlation property of the signal (2.2), the correlation function exhibits a sharp peak when the code replica is aligned with the code received from the satellite. The uncertainty in matching the replica with the incoming code is limited to only 1023 code chips, and the process of aligning them is generally quick. Direct acquisition of a P(Y)-code is difficult by design due to the length of the code. The signal acquisition is accomplished in two steps and is based on the known timing relationship between the C/A- and P(Y)-codes. First, the receiver acquires the C/A-code and, then, with the aid of the timing information in the navigation message, acquires the P(Y)-code. Direct acquisition of the P(Y)-code would be accomplished in the newer receivers with better clocks and thousands of parallel correlators.

Code tracking is implemented as a feedback control loop, called a *delay lock loop*, which continuously adjusts the replica code to keep it aligned with the code in the incoming signal [Section 12.2]. After the alignment is accomplished, the PRN code is removed from the signal, leaving the carrier modulated by the navigation message. This signal is now tracked with another feedback control loop called a *phase lock loop* [Section 12.3]. Essentially, the receiver generates a sinusoidal signal to match the frequency and phase of the incoming signal, and in the process extracts the navigation message. The Doppler shift is measured in the phase-lock loop. Receivers to be used for precise positioning also keep a continuous track of the Doppler count [Section 1.3.2] or delta pseudorange. Figure 2.7 represents a conceptual view of these tracking loops.

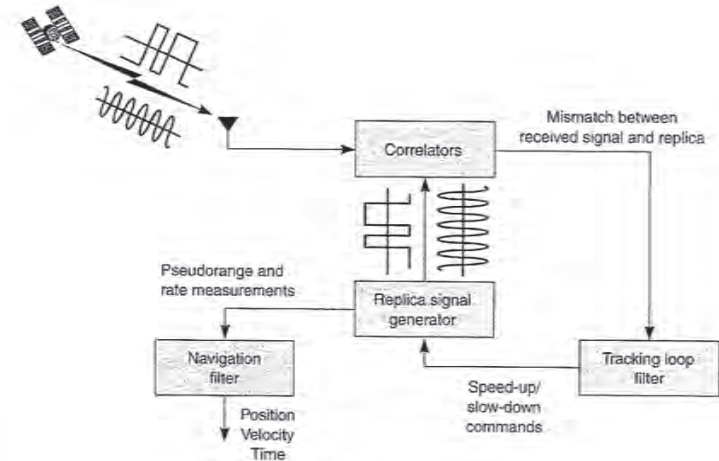


Figure 2.7 A conceptual view of a GPS receiver.

The measurement of transit time for a signal modulated by a C/A-code is conceptually quite simple. The time shift required to align the receiver-generated code replica and the signal received from the satellite is the apparent transit time of the signal modulo 1 ms (see Figure 2.3). The PRN code chips are generated at the satellite at precisely known instants in accordance with the satellite clock and, therefore, the receiver can 'read' the satellite clock time to determine when a chip was generated. The time of reception is determined from the receiver clock. Multiplying the apparent transit time by the speed of light gives pseudorange. Pseudoranges measured from four, or preferably more, satellites are used to compute position [Section 6.1]. The Doppler shift, caused by the relative motion of a satellite and the user, is the projection of the relative velocity on the line of sight, and can be converted into pseudorange rate. Given the pseudorange rates corresponding to four satellites, and the satellite velocity vectors (derived from the navigation message), a user can compute his velocity [Section 6.2]. A GPS receiver does this automatically, continuously, and virtually instantaneously.

2.4.3 Estimation of Position, Velocity, and Time (PVT)

The quality of the PVT estimates obtained by a user from GPS depends basically upon two factors: (i) number of satellites in view and their spatial distribution in the sky, and (ii) quality of the range and range rate measurements. The spatial distribution of the satellites relative to the user is referred to as *satellite geometry*. The satellite geometry changes with time as the satellites rise, move across the sky, and set. Roughly speaking, the geometry is good if the satellites are on all sides of the user, some high in the sky and several low. If a significant part of the sky is somehow blocked, the user may still be able to compute PVT estimates if four or more satellites are in view, but there would generally be an accuracy penalty for poor geometry [Section 6.1.2].

The second factor determining the quality of the PVT estimates is the quality of the pseudorange and Doppler measurements. There are several sources of biases and random errors, which affect the measurements:

- Errors in the navigation message parameters which specify satellite position and signal transmission time introduce errors in the pseudorange measurements. These errors are often referred to as signal-in-space (SIS) errors.
- Propagation delays in the ionosphere and the troposphere, signal distortion due to multipath, and receiver noise also introduce measurement errors. Ionospheric propagation delay, in particular, can be large, ranging from several meters to several tens of meters, depending upon the state of the ionosphere and the elevation of the satellite. This error, however, can be removed substantially by a user equipped with a dual-frequency receiver.

We focus on these errors in Chapter 5, and examine ways to mitigate them. For now, as an academic exercise, we illustrate the nature and size of the error due to Selective Availability (SA), the now-discontinued policy of purposeful degradation of GPS signals referred to earlier. The pseudorange measurement errors are plotted in Figure 2.8 for three satellites, two with SA active and one with SA inactive (or SA off). These measurements were obtained in 1997 from an antenna at a surveyed location. Use of an external cesium atomic standard with the GPS receiver essentially removed the receiver clock bias.

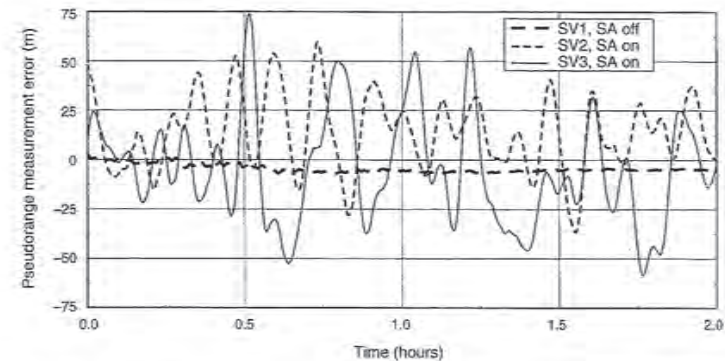


Figure 2.8 Pseudorange errors measured for three satellites, two with SA active and one with SA inactive. The SA-induced pseudorange error remained about 25 m rms throughout the 1990s. SA was terminated on 2 May 2000.

Figure 2.8 shows the reason for the civil community's resentment over SA, which was the dominant source of error. Range measurement error for the satellite with SA off appears essentially to be a 5-m bias holding steady for about ninety minutes. The explanation is that, unlike the SA error, the signal propagation errors change slowly and, if not compensated for, appear as biases. As expected, the positioning accuracy of SPS improved dramatically when SA was 'set to zero' on 2 May 2000, as illustrated in Figure 2.9.

We now revise the simple model for pseudorange measurements used earlier in Figure 1.9 to include the measurement errors discussed above. The pseudorange measurement from the k th satellite is modeled as

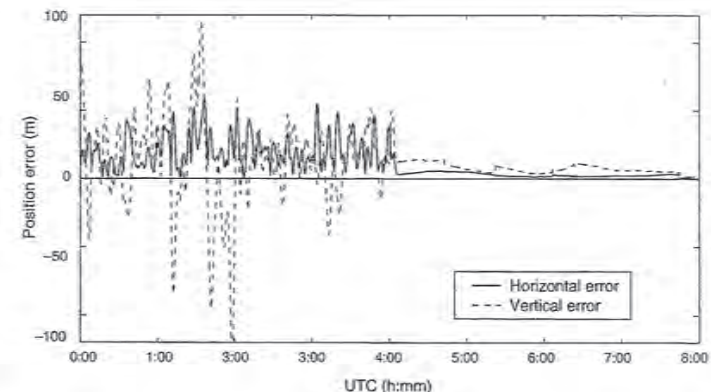


Figure 2.9 Change observed in positioning accuracy of SPS when SA was terminated on 2 May 2000 at about 04:00 hours UTC. (Courtesy of Dr. Sherman Lo, Stanford University)

$$\rho^{(k)} = \sqrt{(x^{(k)} - x)^2 + (y^{(k)} - y)^2 + (z^{(k)} - z)^2} + b + \varepsilon^{(k)} \quad (2.5)$$

$k = 1, 2, \dots, K$, where $(x^{(k)}, y^{(k)}, z^{(k)})$ are the satellite coordinates broadcast in the navigation message; (x, y, z) are the user coordinates to be determined; b is the unknown receiver clock bias; and $\varepsilon^{(k)}$ represents the unknown error in the measurement. We need a minimum of four such equations to solve for the four unknowns. Indeed, if $K = 4$, the best we can do is to pretend that there is no error term present in (2.5). For $K > 4$, we solve the system of equations recognizing that the measurements are 'noisy' and look for a solution (x, y, z, b) which fits the measurements best in some sense. There are several approaches. The most commonly used approach goes all the way back to Gauss and is referred to as *least squares*.

The *residual* is defined as the discrepancy in (2.5) when we substitute the estimates for the unknown parameters. It's the difference between the pseudorange computed from the estimated values of the user position and clock bias, and its measured value. The parameter values which minimize the sum of the squares of the residuals are called least-squares estimates. The least-squares approach turns the estimation problem into a simple minimization problem: Determine the vector (x, y, z, b) which minimizes

$$\sum_{k=1}^K \left(\sqrt{(x^{(k)} - x)^2 + (y^{(k)} - y)^2 + (z^{(k)} - z)^2} + b - \rho^{(k)} \right)^2$$

The problem is solved easily by using the *Newton-Raphson method*, as described in Chapter 6. The components of the minimizing vector, which we represent as $(\hat{x}, \hat{y}, \hat{z}, \hat{b})$, are our "best" estimates of position and clock bias in least-squares sense. We also analyze in Chapter 6 how the measurement errors and satellite geometry determine the quality of these position estimates.

2.4.4 Positioning Accuracy

The performance obtained from GPS is dynamic, changing both with time and place as the satellite geometry and measurement errors change. A global characterization of the performance requires knowledge of the satellite constellation and receiver capabilities, and is given in statistical terms, e.g., as root-mean-square (rms) error or 95th percentile of the error distribution. The latter is often referred to as 95% error. In the 1990s, SA was the dominant source of error, and the performance specifications of 95th percentile for the horizontal and vertical error were 100 m and 156 m, respectively, defined for measurements taken over 24 hours at any point in the world with an unobstructed view of the sky.

In the post-SA world, the performance specifications account only for the satellite orbit and clock errors for which the Control Segment assumes responsibility. These are referred to as signal-in-space (SIS) errors. A user would typically also suffer the consequences of errors in ionospheric and tropospheric modeling, multipath, and receiver noise, all discussed in Chapter 5.

Table 2.2 gives the current SPS performance specifications [SPS (2001)] and the accuracy commonly achieved, the latter in round numbers. The specifications only consider the signal-in-space errors and appear overly conservative. Happily, most users achieve better performance. No performance specifications are given for velocity estimation. For moderate dynamics, velocity accuracy of 0.01 m/s has been reported [Section 6.2.1].

The SPS performance will improve in a few years when the satellites start transmitting

Table 2.2 GPS/SPS performance for global positioning and time dissemination

Error (95%)	Specifications*	Empirical data
Position		
Horizontal	13 m	10 m
Vertical	22 m	15 m
Time	40 ns	30 ns

* Signal-in-space errors only

additional civil signals. And, of course, much better performance is available today to the users of differential GPS, discussed next.

2.5 Differential GPS (DGPS)

The accuracy requirements of the different navigation and positioning applications vary widely. Horizontal positioning accuracy of tens of meters is generally more than adequate for navigation in wide-open spaces: maritime navigation on the open seas; aircraft navigation in en route, terminal, and non-precision approach phases of flight; and recreational use by hikers and backpackers. Many important applications, however, require greater accuracy. Under poor visibility conditions, harbor entry by ships and taxiway guidance on airport surface require meter-level horizontal accuracy, and precision approaches by aircraft typically require meter-level vertical accuracy. Automobile navigation over roads and highways has a similar accuracy requirement. [FRS (2001) defines the various phases of flight and states the navigation accuracy requirements of each.]

In view of the discussion in Section 2.4.2, the quality of the GPS position estimates can be improved by reducing measurement errors and/or improving satellite geometry. We discuss below how this may be done.

- Mitigation of the measurement errors turns out to be simpler: The errors associated with the worst error sources are similar for users located 'not far' from each other, and change 'slowly' in time. In other words, the errors are correlated both spatially and temporally. Clearly, we can estimate the error in a measurement if the receiver location is known. Such error estimates can be used as *differential corrections* if made available to the GPS users in the area, allowing them to mitigate errors in their measurements. That's differential GPS, generally abbreviated as dGPS or DGPS.
- Satellite geometry can be improved by adding satellites to the constellation to provide additional ranging signals. A user can also improve the geometry by deploying pseudo-satellites, called *pseudolites*, which transmit GPS-like signals. The pseudolites can be deployed on the ground, in the air, or on a ship. A GPS receiver has to be modified to receive and process these signals.

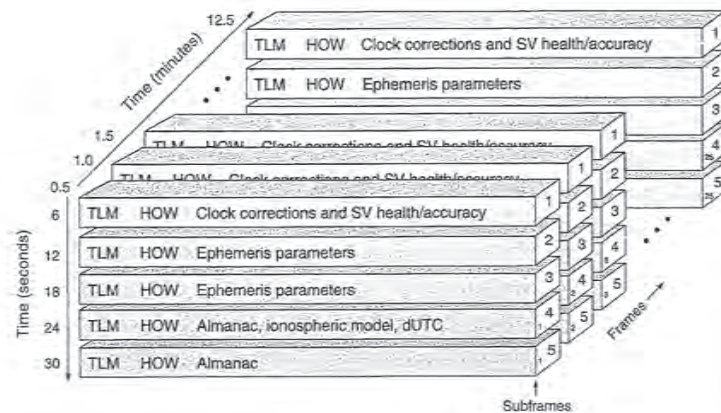


Figure 4.13 GPS navigation message organization: frames and subframes. Subframes 1, 2, and 3 repeat every 0.5 minutes; subframes 4 and 5 repeat every 12.5 minutes. Subframes 1, 2, and 3 are specific to the transmitting satellite; subframes 4 and 5 are common to all satellites. (Courtesy of Dr. Frank van Diggelen, Global Locate)

ten 30-bit words (see Figure 4.13). Subframes 1–3 typically repeat the same information from frame to frame. Subframes 4–5 of the consecutive frames, however, contain different ‘pages’ of the navigation message. It takes 25 frames (12.5 minutes) to transmit the complete navigation message. A unit of 25 frames is called a Master Frame. The information content of the various subframes is summarized below. A detailed description is given in the Interface Specification [IS-GPS-200D].

Subframe 1:	satellite clock corrections, health indicators, age of data
Subframes 2–3:	satellite ephemeris parameters
Subframe 4:	ionosphere model parameters, UTC data, almanac and health status data for satellites numbered 25 and higher
Subframe 5:	almanac and health status data for satellites numbered 1–24

The first two words of each six-second subframe have a special significance: Telemetry word (TLM) and Hand-over word (HOW). The TLM contains a fixed 8-bit synchronization pattern. The HOW word, discussed below, conveys the satellite time to a receiver.

Conveying Satellite Time to a Receiver: Z-Count

Each satellite keeps time in accordance with an atomic standard on board. Parameters used to relate the satellite time to GPST and UTC were discussed in Section 4.2.4. In the next chapter, we’ll be concerned with the transit time of a signal from a satellite to the receiver, and we’ll want to determine the time (per satellite clock) at the start of transmission of a navigation data bit. We have avoided all bit-and-byte-level discussion of the navigation message so far, referring the reader instead to the IS, but we make an exception here and give a brief account of the

parameters which give us the satellite time associated with the beginning of each subframe.

The satellite time is kept as number of 1.5-second epochs since the zero time point (currently, midnight between 21 August and 22 August 1999 [Section 4.2.4]) by a 29-bit number called Z-count. The 10 MSBs of the Z-count specify the week number (modulo 1024). The 19 LSBs specify the Time of Week (TOW) count in 1.5-second epochs. The first 17 bits of HOW, the second word in each subframe, are the 17 MSBs of TOW, specifying the time of week in 6-second epochs or the number of 6-second subframes since the beginning of the week and start of transmission of the next subframe. The Z-count increases by four between two consecutive subframes.

We’ll resume this discussion of how to compute the transit time of a signal in Section 5.1.

4.4 GPS Satellite Constellation and Visibility Displays

A constellation is specified by defining the nominal orbits and positions of the satellite slots within each orbit, and tolerances for insertion and maintenance of the satellites within their assigned slots. The orbital characteristics of the baseline, 24-satellite GPS constellation are listed below [SPS (2001)].

- Semi-major axis: 26,560 km (± 50 km for Block II satellites)
The orbital altitude above the earth corresponds roughly to three earth radii.
- Period: one-half mean sidereal day, or about 11 h 58 m.
The earth spin period in space is about 23h 56m, so a satellite would repeat the same ground track after two revolutions. A stationary user would see the spatial distribution of the satellites above her repeat after 23h 56m. (The ground tracks appear to repeat from day to day, but not from month to month due to the slow westward drift, referred to above.)
- Eccentricity < 0.02 .
- Six orbital planes, named A through F, all with inclination $55^\circ (\pm 3^\circ)$ relative to the equatorial plane. The right ascensions of the ascending node for the six orbital planes are separated by 60° in the equatorial plane.
- Four satellites per plane, distributed unevenly (see Figure 4.14). Two of the satellites are spaced between 30.0° and 32.1° . The remaining three inter-satellite spacings are between 92.38° and 130.98° . Relative spacing of the satellites and their ground tracks are maintained within their nominal values by adjusting their semi-major axes periodically.

The rationale for the uneven spacing of the satellites is to minimize the effect of single satellite failures out of the 24-satellite constellation. Apparently, a symmetric arrangement increases the chances of poor geometry when satellites fail. The baseline constellation geometry was obtained from a constrained optimization, and is claimed to offer the best global coverage defined as “PDOP < 10 for the best four satellites above 5° elevation with a single satellite failure.” [Massatt and Zeitzev (1998)].

Each satellite is identified with a two-character code: A letter identifies the orbital plane (A through F), and a number identifies the satellite slot number in the plane (1 through 4). The assignment of slot numbers in an orbital plane is not progressive (see Figure 4.14 and Table

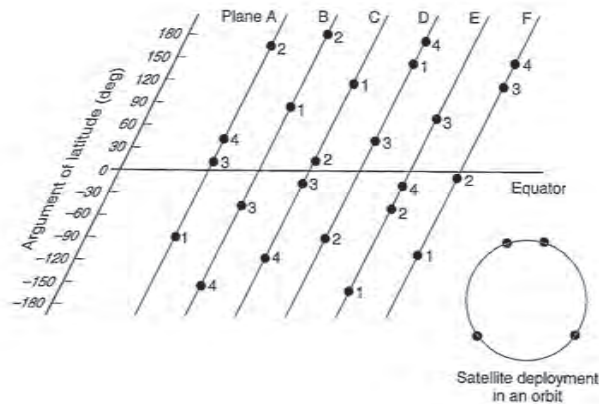


Figure 4.14 The baseline GPS constellation consists of six orbital planes with four satellites in each plane. Each satellite is identified with a two-character code: A letter identifies the orbital plane (A through F), and a number identifies the satellite slot number in the plane (1 through 4). [Adapted from Massatt and Zeitew (1998)].

4.4): A3, A4, A2, A1; B1, B2, B4, B3; etc. There is no logic behind the numbering scheme. It just evolved to be what it is.

Each satellite has associated with it a unique PRN number corresponding with the PRN code transmitted by the satellite. There is also an SVN or NAVSTAR number assigned to each

Table 4.4 Orbital parameters for the baseline 24-satellite GPS constellation

Semi-major axis: 26,559.8 km; eccentricity: 0; inclination: 55°; argument of perigee: 0 [SPS (2001)]

Slot ID	Right ascension (deg)	Mean anomaly (deg)	Slot ID	Right ascension (deg)	Mean anomaly (deg)
A3	272.85	11.68	D1	92.85	135.23
A4	272.85	41.81	D4	92.85	167.36
A2	272.85	161.79	D2	92.85	265.45
A1	272.85	268.13	D3	92.85	35.16
B1	332.85	80.96	E1	152.85	197.05
B2	332.85	173.34	E2	152.85	302.60
B4	332.85	204.38	E4	152.85	333.69
B3	332.85	309.98	E3	152.85	66.07
C1	32.85	111.88	F1	212.85	238.89
C4	32.85	241.56	F2	212.85	345.23
C3	32.85	339.67	F3	212.85	105.21
C2	32.85	11.60	F4	212.85	135.35

satellite. Table 4.4 gives the orbital parameters for the baseline 24-satellite constellation corresponding to epoch 1993 July 1 0h. A graphical representation is given in Figure 4.14.

Several graphical devices and displays are used to help an observer determine which satellites would be visible from her position at a certain time, and how they would be distributed in the sky. Such analysis is commonly based on a recent almanac from a GPS satellite. A commonly used bar chart (see Figure 4.15) shows the rise and set times of the satellites for a location, typically over a twenty-four-hour period. The bars shift to the left by about four minutes from one day to the next.

An observer can determine from Figure 4.15 the number of satellites in view at a particular time but not their distribution in the sky. In order to see the changing 'geometry' of the satellites, we need a sky plot (or polar plot) shown in Figure 4.16. A sky plot shows the track of each satellite in the sky above an observer for the time period of interest. The outer circle corresponds to the observer's horizon, and the center corresponds to the zenith. The three inner circles correspond to different elevation angles. The position of a satellite is thus characterized by azimuth and elevation angle (both terms are defined in Appendix 4.A). We know intuitively that it would be better for position estimation if the satellites were 'well distributed' in the sky rather than all bunched up. We will revisit this notion in Chapter 6.

Satellite visibility plots like Figures 4.15 and 4.16 were mainstays for planning data collection in the early days of GPS while the constellation was sparse. Since 1995, however, the constellation has comprised 25-plus satellites and the users have come to expect that they will see 6–8 satellites unless the view of the sky is obstructed. Figure 4.16 can be customized to represent such obstructions and, with additional annotation (time, direction of satellite motion, etc.), can still be a valuable tool for planning data collections.

A constellation designer would be interested in a global characterization of the number of satellites in view. A plot of the ground tracks of the satellites is a useful graphical device for

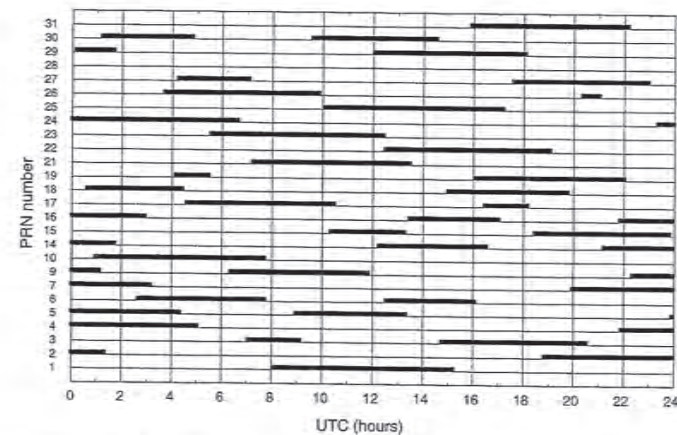


Figure 4.15 Satellite visibility (rise and set times) at Lexington, Massachusetts, plotted for a day in 1998.

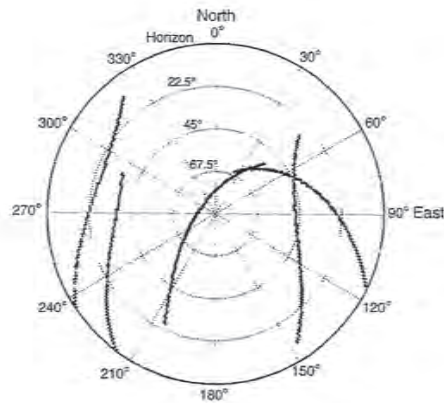
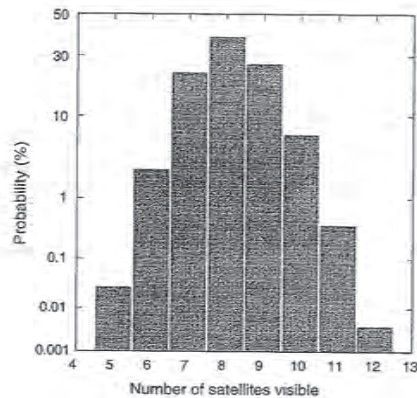


Figure 4.16 Sky view of satellite tracks.

this purpose. The *subsattellite point* is defined as the point at which the line from the satellite to the center of the earth intersects the earth's surface. An observer at a subsattellite point on the equator would see the satellite directly overhead, and nearly so at other latitudes (within 15 arc-minutes at mid-latitudes). A ground track is the locus of the subsattellite point as the satellite revolves in its orbit and the earth rotates. Because of uneven phasing of the satellites in each orbital plane, the longitudes of ascending node are all different, and there are 24 separate ground tracks. It is a messy picture we will skip.

Figure 4.17 A histogram of the number of satellites in view of users worldwide (24-satellite baseline constellation, satellite elevation angle $> 5^\circ$).

We present instead in Figure 4.17 a histogram of the number of satellites visible to users around the world with a clear view of the sky. This display is based on the baseline 24-satellite GPS constellation (Table 4.4), and we count only those satellites which are above 5° in elevation. According to this model, about 99.99% of GPS users around the world see six or more satellites, and 70% see eight or more. Figure 4.17 was generated in a simulation consisting of the following steps: (i) select a user location on the globe at random; (ii) select a time of day at random; (iii) compute satellite positions corresponding to the time selected and determine how many were in the user's view; and (iv) repeat steps (i) through (iii) a large number of times.

4.5 Summary

We have covered a lot of ground in this sprawling chapter in an attempt to fill in the background. The main ideas are:

- definition of WGS 84,
- GPS Time (GPST) reference and its relationship with other time references,
- GPS navigation message parameters and computation of satellite position and velocity at an epoch.

In order to specify the position of a point accurately, we first have to define a coordinate frame precisely. For GPS, the framework for specifying a position is provided by the World Geodetic System 1984 (WGS 84). WGS 84 comprises definitions of (i) an earth-centered, earth-fixed (ECEF) Cartesian coordinate frame in which the positions of the GPS satellites and the users are represented; (ii) an ellipsoid which serves as a geometric model for the shape of the earth, and (iii) the geoid which serves as the zero-level surface for measurement of height globally.

Measurement of time and time intervals is at the heart of GPS. The signals transmitted by GPS satellites are referenced to GPS Time. GPST is a composite time defined on the basis of the times kept by a set of atomic clocks at the monitor stations and in the satellites. Unlike UTC, the current civil standard, GPST makes no provisions for leap seconds to stay in sync with the earth's rotation. GPST, however, is steered to stay within $1 \mu\text{s}$ of UTC (modulo one second). GPST identifies a particular epoch in terms of two parameters: (i) the number of seconds that have elapsed since the previous Saturday/Sunday midnight, and (ii) a week number counted from the Standard Epoch of GPS (modulo 1024).

Each GPS satellite transmits a message at 50 bps. This navigation message provides the values of parameters and coefficients of models. Of particular interest is a sixteen-parameter set which allows a receiver to determine the satellite position and velocity at a time epoch.

Chapter 5

GPS Measurements and Error Sources

5.1 Measurement Models

- 5.1.1 Code Phase Measurements
 - Constructing Pseudorange Measurements
- 5.1.2 Carrier Phase Measurements
- 5.1.3 An Instructive Model for the Code and Carrier Measurements
- 5.1.4 Error Sources and Models

5.2 Control Segment Errors: Satellite Clock and Ephemeris

5.3 Signal Propagation Modeling Errors

- 5.3.1 Signal Refraction, Wave Propagation, and Dispersive Media
- 5.3.2 Ionospheric Delay
 - Phase Advance and Group Delay; Obliquity Factor;
 - Delay Estimation with Dual-Frequency Measurements; Broadcast Model
- 5.3.3 Tropospheric Delay
 - Dry and Wet Delays; Tropospheric Models; Mapping Functions

5.4 Measurement Errors

- 5.4.1 Receiver Noise
- 5.4.2 Multipath
- 5.4.3 Measurement Error Models

5.5 User Range Error (URE)

5.6 Measurement Error: Empirical Data

5.7 Combining Code and Carrier Measurements

- 5.7.1 Single-Frequency Measurements
- 5.7.2 Dual-Frequency Measurements

5.8 Error Mitigation: Differential GPS (DGPS)

- 5.8.1 Error Mitigation
- 5.8.2 Local-Area DGPS and Relative Positioning
- 5.8.3 Wide-Area DGPS

5.9 Summary

- Homework Problems
- References

GPS provides two types of measurements. Code tracking provides estimate of apparent transit time of a signal. Carrier phase tracking provides measurement of the received carrier phase relative to the phase of a sinusoidal signal generated by the receiver clock. Both are biased (and, in case of carrier phase, ambiguous) estimates of the instantaneous user-satellite range. How these measurements are formed in a receiver is discussed in Chapter 12 and in a comprehensive paper by Ward (1996).

In this chapter, we examine the measurements and develop simple mathematical models to relate them to receiver-satellite ranges and range rates. This process requires us to think about the sources of errors in these models. Analysis of the errors and schemes for their alleviation take up the rest of the chapter. These models are used in the next chapter for estimation of position, velocity, and time (PVT).

In this chapter and the next two, we sometimes speak loosely of the user position or receiver position, but what we mean is the position of the receiver antenna. When dealing with centimeter- and millimeter-level position estimates in Chapter 7, we must be even more specific and refer to the electrical phase center of the antenna.

5.1 Measurement Models

5.1.1 Code Phase Measurements

A basic measurement made by a GPS receiver is the apparent transit time of the signal from a satellite to the receiver, defined as the difference between signal reception time, as determined by the receiver clock, and the transmission time at the satellite, as marked on the signal. It is measured as the amount of time shift (modulo 1 ms) required to align the C/A-code replica generated at the receiver with the signal received from the satellite. This measurement is biased due to the fact that the satellite and receiver clocks are not synchronized, and each keeps time independently. Each satellite generates its signals in accordance with a clock on board. The receiver generates a replica of each signal in accordance with its own clock. The corresponding biased range, or pseudorange, is defined as the transit time so measured multiplied by the speed of light in a vacuum.

There are three time scales to deal with: two of these are the times kept by the satellite and receiver clocks; the third is a common time reference, GPS Time (GPST), introduced in Section 4.2.4 as a composite time scale derived from the times kept by clocks at GPS monitor stations and aboard the satellites.

Let τ be the transit time associated with a specific code transition of the signal from a satellite received at time t per GPST. Let $t^s(t - \tau)$ be the corresponding emission time, and $t_u(t)$ be the arrival time measured by the user's receiver clock. The measured apparent range ρ , called pseudorange, is determined from the apparent transit time as

$$\rho(t) = c[t_u(t) - t^s(t - \tau)] \quad (5.1)$$

Both t and τ are unknown, and are to be estimated. The range to a satellite is about 20,000 km when overhead, and about 26,000 km when rising or setting. The signal transit time varies between about 70 ms and 90 ms.

In this section, we deal with measurements from a GPS satellite in a generic way, making no reference to the satellite ID or carrier frequency (L1 or L2). We employ superscript s to

identify a term associated with a satellite, and subscript u to identify a term associated with the user or receiver, where needed for clarity. Additional notation will be introduced in later sections as needed to deal with dual-frequency measurements from multiple satellites.

The time scales of the receiver and the satellite clocks can be related to GPST as

$$t_u(t) = t + \delta t_u(t) \quad (5.2)$$

$$t^s(t - \tau) = (t - \tau) + \delta t^s(t - \tau) \quad (5.3)$$

where δt_u is the receiver clock bias and δt^s is the bias in the satellite clock, both measured relative to GPST, as shown in Figure 5.1. In our notation, both δt_u and δt^s reflect the amounts by which the satellite and receiver clocks are advanced in relation to GPST. The satellite clock bias (δt^s) is estimated by the Control Segment and specified in terms of the coefficients of a quadratic polynomial in time. The values of these coefficients are broadcast in the navigation message [Section 4.3.6].

Accounting for the clock biases, the measured pseudorange (5.1) can be written as

$$\begin{aligned} \rho(t) &= c[t + \delta t_u(t) - (t - \tau + \delta t^s(t - \tau))] + \varepsilon_p(t) \\ &= c\tau + c[\delta t_u(t) - \delta t^s(t - \tau)] + \varepsilon_p(t) \end{aligned} \quad (5.4)$$

Symbol ε is used throughout Part II to denote unmodeled effects, modeling errors, and measurement errors; subscripts, and occasionally overbar and tilde, are used to distinguish among the different scenarios. The transit time multiplied by the speed of light in a vacuum can be modeled as

$$c\tau = r(t, t - \tau) + I_p(t) + T_p(t) \quad (5.5)$$

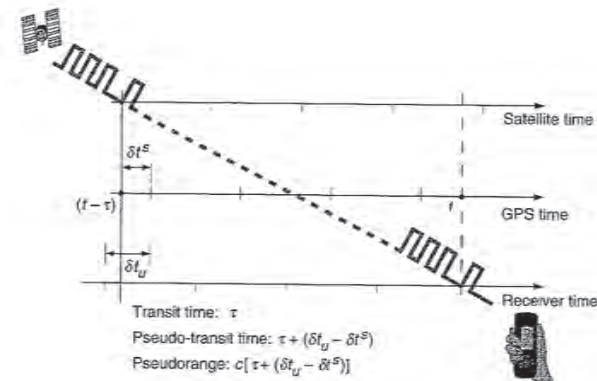


Figure 5.1 A conceptual view of the pseudorange measurements. The receiver and satellite clocks are unsynchronized. The apparent transit time (pseudo-transit time) is the difference between the signal reception time according to the receiver clock and the transmission time imprinted upon the signal in accordance with the satellite clock.

where $r(t, t - \tau)$ is the geometric (or, true) range between the user position at time t and the satellite position at $(t - \tau)$. I_p and T_p reflect the delays associated with the transmission of the signal through the ionosphere and the troposphere, respectively. Both terms are positive, and we will discuss both later in this chapter. For simplicity, we drop explicit reference to the measurement epoch t , and rewrite the model for the measured pseudorange as

$$\rho = r + c[\delta t_u - \delta t^s] + I_p + T_p + \varepsilon_p \quad (5.6)$$

Ideally, we would have liked to measure r , the true range to the satellite. What we have instead is ρ , the pseudorange, a biased and noisy measurement of r . How accurate an estimate of position, velocity, or time we obtain from these measurements would depend upon our ability to compensate for, or eliminate, the biases and errors. Equation (5.6) is basic, and we will return to it again and again.

The receiver clocks are generally basic quartz crystal oscillators, and tend to drift. The receiver manufacturers attempt to limit the deviation of the receiver clock from GPST, and schedule simultaneous measurements from all satellites typically at 1 Hz at epochs that are within ± 1 ms of the GPST seconds. One approach to maintaining the receiver clock within a certain range of GPST is to steer the receiver clock 'continuously.' Such steering is implemented in software. The second approach is to let the clock drift until it reaches a certain threshold (typically, 1 ms), and then reset it with a jump to return the bias to zero. How a receiver keeps track of the bias between its clock and GPST is discussed in Chapter 6. An example of pseudorange measurements with a receiver using the second approach is shown in Figure 5.2.

Figure 5.2 shows pseudorange measurements simultaneously from three satellites which rose about the same time but were in different orbits. One came overhead and stayed in view for almost seven hours. The other two stayed lower in the sky and were seen for shorter periods. The discontinuities common to all three sets of measurements are due to the resetting of the receiver clock. It is left as an exercise to determine if the receiver clock is running fast or slow, and to estimate its frequency offset from the nominal value of 10.23 MHz.

Constructing Pseudorange Measurements

The measurement of pseudoranges is scheduled in accordance with the receiver clock. In (5.1), determining $t_u(t)$ is straightforward, but $t^s(t - \tau)$ is a little tricky. We now resume the discussion we began toward the end of Section 4.3.6. The measurement epoch would typically occur in the middle of a C/A-code chip. The corresponding transmission time per satellite clock is reconstructed as follows in a process akin to determining time from an analog clock—by reading the hour hand, the minute hand, and the second hand.

Recall from Section 4.3.6 that the satellite time is transmitted in the navigation message in the form of Z-count, which increments in units of 1.5 seconds, and is specified at the beginning of each subframe. The Z-counts associated with two successive subframes differ by four (i.e., six seconds). In order to determine the satellite clock time associated with the current measurement epoch, we need the Z-count associated with the *current* subframe, plus the satellite time elapsed since the beginning of the subframe. This elapsed time can be measured in terms of its components as follows: the whole number of navigation message data bits transmitted in the current subframe, the whole number of C/A-code periods since the beginning of the current navigation message bit, the number of whole chips in the current C/A-code cycle, and the fraction of the current C/A-code chip (see Figure 5.3).

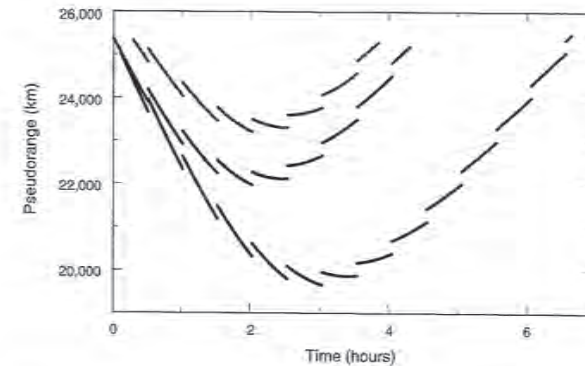


Figure 5.2 Pseudorange measurements from three satellites obtained with a stationary receiver. The variation in the pseudorange is mainly due to change in geometric range resulting from the satellite motion and rotation of the earth. The discontinuities are common to all measurements, and are due to 1-millisecond jumps in the receiver clock. (Some receivers 'steer' the receiver clock 'continuously' to control its offset from GPS Time.)

$$\begin{aligned} t^s(t - \tau) = & \text{Z-count} \times 1.5 \\ & + \text{number of navigation data bits transmitted} \times 20 \times 10^{-3} \\ & + \text{number of C/A-code repeats} \times 10^{-3} \\ & + \text{number of whole C/A-code chips} / (1.023 \times 10^6) \\ & + \text{fraction of a C/A-code chip} / (1.023 \times 10^6) \text{ seconds} \end{aligned}$$

The Z-count is read off the navigation data. The number of navigation data bits in the subframe transmitted and the number of C/A-code periods in the current navigation data bit are counted by the receiver software. As we'll see in Chapter 12, the delay lock loop (DLL) provides the last two terms.

5.1.2 Carrier Phase Measurements

A measurement much more precise than that of code phase is the phase of the carrier received from a satellite. The carrier phase measurement by a GPS receiver, also called carrier beat phase measurement, is the difference between the phases of the receiver-generated carrier signal and the carrier received from a satellite at the instant of the measurement. As we'll see, this measurement is an indirect and ambiguous measurement of the signal transit time. There are some subtleties associated with the carrier phase measurements. We begin with basic definitions and deal first with simple, idealized scenarios.

We measure carrier phase in terms of the number of cycles generated or received since the starting point of an interval. The phase at time t is defined as

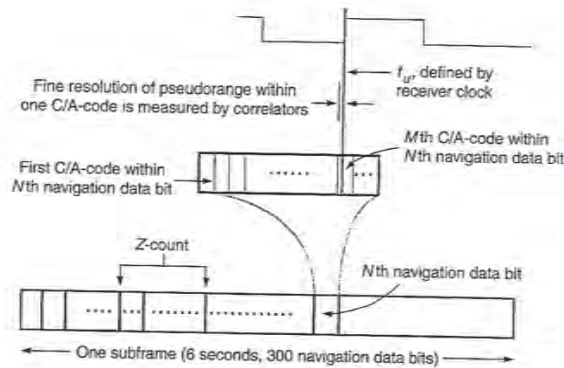


Figure 5.3 Pseudorange determination simplified. The transmission time per satellite clock is reconstructed in a process similar to reading time from an analog clock—by reading the hour hand, the minute hand, and the second hand.

$$\phi(t) = \phi(t_0) + \int_{t_0}^t f(s) ds$$

where $f(s)$ is the time-varying frequency and $\phi(t_0)$ is the initial phase offset. In this model, the timer used to measure the epochs t_0 and t is assumed to be perfect. If the interval $(t - t_0)$ is 'short' and the signal generator is 'highly stable,' we can write

$$\phi(t) = \phi(t_0) + f \cdot (t - t_0)$$

where f is the instantaneous frequency. But what if we were measuring time intervals by counting these very cycles generated by an (imperfect) oscillator with nominal frequency specified as f_0 ? In this case

$$\phi(t) = \phi(t_0) + f_0 \cdot (t - t_0)$$

This is how we'll model the phase of the carrier generated by the receiver clock, which is also our timer.

Consider first an idealized case of error-free measurements with perfect and synchronized satellite and receiver clocks, and no relative motion between the satellite and the user. In this model, the carrier phase measurement (i.e., the difference between the phases of the receiver generated carrier and the received carrier) would remain fixed at a fraction of a cycle, and the distance between a satellite and the receiver would be an unknown number of whole cycles plus the measured fractional cycle. This measurement, however, contains no information regarding the number of whole cycles, referred to as *integer ambiguity*. Now suppose that the carrier phase is tracked while the receiver or the satellite moves so that the distance between them grows by a wavelength. The corresponding carrier phase measurement would now be a full cycle plus the fractional-cycle phase measured before the movement began.

To measure carrier phase in GPS, a receiver acquires phase lock with the satellite signal, measures the initial fractional phase difference between the received and receiver-generated signals, and from then on tracks changes in this measurement, counting full cycles and keeping track of the fractional cycle at each measurement epoch.

An important idea of this section is that the phase of the received signal at any instant can be related to the phase at the satellite at the time of transmission by the transit time of the signal, and it's τ we are after. Staying with our idealized case of perfect and synchronized clocks, the carrier phase measurement in units of cycles can be written as

$$\phi(t) = \phi_r(t) - \phi^s(t - \tau) + N \quad (5.7)$$

where $\phi_r(t)$ is the phase of the receiver-generated signal; $\phi^s(t - \tau)$ is the phase of the signal received from the satellite at time t , or the phase of the signal at the satellite at time $(t - \tau)$; τ is the transit time of the signal; and N is the integer ambiguity. Estimation of N is referred to as *integer ambiguity resolution*.

Simplifying the above expression by writing

$$\phi^s(t - \tau) = \phi^s(t) - f \cdot \tau$$

we obtain

$$\begin{aligned} \phi(t) &= f \cdot \tau + N \\ &= \frac{r(t, t - \tau)}{\lambda} + N \end{aligned} \quad (5.8)$$

where f and λ are the carrier frequency and wavelength, respectively, and $r(t, t - \tau)$, as before, is the geometric range between the user position at time t and the satellite position at $(t - \tau)$. As in the previous section, we are dealing with the measurements at a receiver from a single satellite, and we identify terms with the receiver or the satellite, if needed for clarity.

Now we account for the clock biases, initial phase offsets, atmospheric propagation delays, and measurement errors. Dropping explicit reference to the measurement epoch, (5.8) can be rewritten as

$$\phi = \lambda^{-1} [r + I_\phi + T_\phi] + \frac{c}{\lambda} (\delta t_u - \delta t^s) + N + \epsilon_\phi \quad (5.9)$$

where I_ϕ and T_ϕ are the ionospheric and tropospheric propagation delays in meters, respectively; and c is the speed of light in a vacuum. Note that the carrier phase measurement in (5.9) is in units of cycles. The integer term N could have gone on either side of the equation. We keep it on the right and leave its sign unspecified.

The change in carrier phase measurement over a time interval corresponds to changes in both the user-satellite range and receiver clock bias, and is referred to as *integrated Doppler* or *delta pseudorange*. The rate of change of the carrier phase measurement gives the *pseudorange rate*, which is made up of the actual range rate plus the receiver clock frequency bias.

Equation (5.9) appears similar to that for the pseudorange measurements based on code tracking (5.6). Both the code and carrier phase measurements are corrupted by the same error sources, but there is an important difference. Code tracking provides essentially unambiguous pseudoranges which, as we'll see, are coarse by comparison. The carrier phase measurements are extremely precise, but are encumbered with integer ambiguities. The integers remain con-

stant as long as the carrier tracking loop maintains lock. Any break in tracking, no matter how short, could change the integer values.

In order to take full advantage of the precision of these measurements to obtain accurate position estimates, we have to compensate for the various errors and resolve the integer ambiguities. As will be clear following the discussion of measurement errors in the next section, we cannot expect to determine the value of integer N in (5.9) in real time, and we'd have to look for alternative approaches to take advantage of the precision of these measurements.

One way to get at least a partial benefit of the precise carrier phase measurements without being bogged down by the integer ambiguities is via delta pseudoranges obtained from the change in carrier phase measurement over a time interval. From (5.9), the change in carrier phase measurements between time instants t_0 and t_1 is

$$\phi(t_1) - \phi(t_0) \approx \lambda^{-1} [r(t_1) - r(t_0)] + \frac{c}{\lambda} [\delta t_u(t_1) - \delta t_u(t_0)] + \tilde{\epsilon}_\phi$$

The integer ambiguity term drops out if the carrier is tracked continuously between t_0 and t_1 . The error in the above representation is related to the changes during the observation interval in the satellite clock bias and the ionospheric and tropospheric propagation delays. This idea is discussed further in Section 5.7.

5.1.3 An Instructive Model for the Code and Carrier Measurements

We can think of the code and carrier phase measurements as though obtained with a measuring tape. In this conceptual exercise suggested by van Diggelen (1997), one end of the tape measure is attached to the satellite and you are holding the other end, including the spool, allowing the tape to roll forward or back freely as you and the satellite move. The difference in the two types of measurements can be illustrated by the different markings on the tape measure, as shown in Figure 5.4.

For code measurements, the tape has coarse tick marks with resolution of one meter, and the tick marks are labeled from zero to 26,560,000 m to indicate the distance to a satellite. At a measurement epoch, you simply look at the tape and read off the range. Actually, the code measurements entail little conceptual difficulty and this analogy was hardly necessary. So, let's move on to the carrier phase measurements.

The tape used for the carrier phase measurements is peculiar: It has very fine tick marks in units of cycles with resolution of 0.01 cycle, but no labels (1 cycle $\approx$ 19 cm at L1, and 24 cm at L2). At a measurement epoch, when you look at the tape all you can make out is a partial cycle. You know that the range to the satellite is a certain (large) number of whole cycles plus the observed partial cycle. While this partial cycle can be measured very precisely, you have no idea of the number of whole cycles between you and the satellite. The situation repeats at the next measurement epoch.

How to take advantage of these very fine but ambiguous measurements? There is a way. The trick is not to take your eyes off the tape between measurement epochs. At the first measurement epoch, make a note of the partial cycle. From that point on, count the number of whole cycle marks as they go by. At each successive measurement epoch, record the change in range in terms of the number of whole cycles and a partial cycle. Each of these measurements, when combined with the initial unknown whole number of cycles, represents range to the satellite at that instant. A collection of such measurements over a 'short' period from all

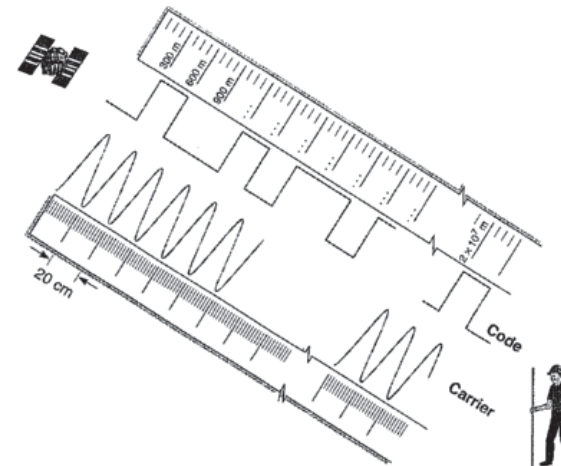


Figure 5.4 A conceptual representation of the code and carrier phase measurements with a tape measure.

satellites in view can be used to estimate the unknown whole cycles, and determine your position very precisely, as we'll discuss in Chapter 6. The problem is that if you took your eyes off the tape, or even blinked, you would miss the count of whole cycles, and will have to start all over again.

5.1.4 Error Sources and Models

Measurement errors are often categorized as *noise* or *bias*. Noise generally refers to a quickly varying error that averages out to zero over a 'short' time interval, where short is defined in relation to the integration time or smoothing time of a receiver [Chapter 11]. A bias tends to persist over a period of time. This distinction is helpful in understanding the effects of the measurement errors, and we'll return to this point in Section 5.5 after we have examined the errors in measurement of pseudoranges in the next section. The errors can be grouped as follows:

- errors in the parameter values broadcast by a satellite in its navigation message for which the Control Segment is responsible,
- uncertainties associated with the propagation medium which affect the travel time of the signal from a satellite to the receiver,
- receiver noise which affects the precision of a measurement, and interference from signals reflected from surfaces in the vicinity of the antenna.

We examine these error sources below. Our objective is to understand the nature and the size of the pseudorange error introduced by each source and analyze the efficacy of different approaches to error mitigation.

5.2. Control Segment Errors: Satellite Clock and Ephemeris

The ephemeris and clock parameter values broadcast by the satellites are computed by the Control Segment on the basis of measurements at GPS monitor stations. The current values of these parameters are obtained from a Kalman filter used to generate the states of the satellites (position and velocity) and their clocks (phase bias, frequency bias, and frequency drift rate). A prediction model is then used to generate the ephemeris and clock parameters to be uploaded to the satellites and broadcast by them in the 50 bps navigation message. There are errors associated with both the estimation of the current values of the parameters, and prediction of their values for the future. The prediction error grows with the age of data (AoD), defined as the time since the last parameter upload. Clearly, the more accurate are the models used to estimate and predict the ephemeris and clock parameters, and the more frequent the data uploads to the satellites, the lower the Control Segment errors.

The rms ranging error attributed to the clock and ephemeris parameter errors was limited to 6 m for the Precise Positioning Service (PPS) by specification. Actually, the frequency standards aboard the satellites have consistently exceeded their specifications, and the ephemeris prediction errors have been kept low by frequent uploads. As a result, Block II/IIA and IIR satellites beat the specification, limiting the Control Segment error in recent years to under 3 m rms. The specification has now been tightened to 3 m for Block IIF.

The ephemeris error is usually decomposed into components along three orthogonal directions defined relative to the satellite orbit: radial, along-track and cross-track (see Figure 5.5). In estimation of an orbit based on range measurements, the radial component of the ephemeris error tends to be the smallest. The along-track and cross-track components can be several times larger. This is fortunate because, as shown in Figure 5.5, the error in a pseudorange measure-

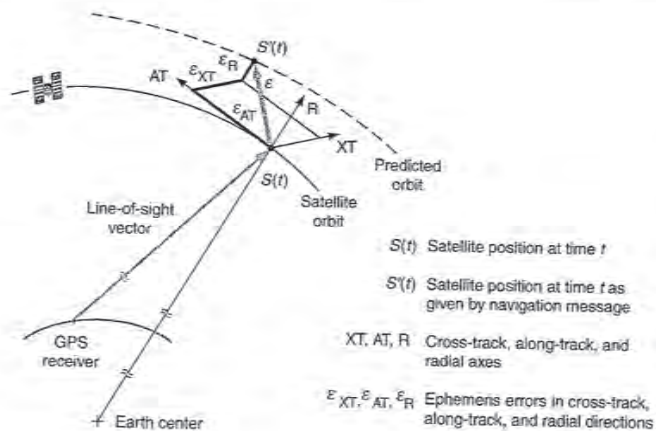


Figure 5.5 Ephemeris error components in the along-track (AT), cross-track (XT), and radial (R) directions. The radial component is the principal source of error in the range measurement. The components of the along-track and cross-track errors along the line of sight are small.

ment is the projection of the satellite position error vector on the satellite-receiver line of sight, which depends mostly upon the radial component of the ephemeris error. The components of the along-track and cross-track errors along the line of sight are small.

The range error due to the errors in the clock and ephemeris parameters is defined as the root-sum-square (rss) value of the clock error and the line-of-sight component of the ephemeris error. The size of this error is estimated and tracked by the Control Segment in real time within 1 m rms. With typical once-a-day data uploads, the current estimates of the rms range errors due to the ephemeris and clock parameters are about 1.5 m each. (The 3-D ephemeris error over a day is typically 3–5 m rms.)

The Control Segment monitors the growth in parameter errors by comparing the broadcast values to the best current estimates available. If the estimated range error for a satellite exceeds a threshold, a 'contingency data upload' is scheduled. (The threshold was revised in 1997 from 8 m to 5 m.) New initiatives would reduce the navigation parameter errors further [Malys *et al.* (1997)].

The Block IIR and IIF satellites are capable of operating in Autonav mode while out of contact with the Control Segment over extended periods, though this function currently appears to be inoperative. In Autonav mode, cross-link ranging among the satellites would be used to correct the ephemerides stored in the on-board computers. The Block IIF satellites are planned to maintain the clock and ephemeris errors below 3 m for up to sixty days out of contact with the Control Segment, not counting the error due to polar motion [Section 4.1] and changes in UT1 [Section 4.2], which cannot be estimated in Autonav mode. In normal mode of operation, the IIFs would limit AoD to three hours by uploading the ephemeris and clock data to all satellites via ultra-high-frequency cross links [Fisher and Ghassemi (1999)].

5.3 Signal Propagation Modeling Errors

5.3.1 Signal Refraction, Wave Propagation, and Dispersive Media

The GPS signals are affected by the medium through which they travel from the satellites to a receiver. The travel distance ranges from about 20,000 km when a satellite is overhead to about 26,000 km when it is rising or setting. All but the final 5% of the signal travel can be regarded as in a vacuum or free space, through which the electromagnetic signals travel with a constant speed $c = 299,792,458$ m/s, the well-known universal constant. Closer to the surface of the earth (Figure 5.6), at a height of about 1000 km, the signals enter an atmosphere of charged particles, called the ionosphere. Later, at a height of about 40 km, the signals encounter an electrically neutral gaseous atmosphere to be referred to as the troposphere.

The atmosphere changes the velocity (speed and direction) of propagation of radio signals. This phenomenon is referred to as *refraction*. The change in speed of propagation changes the signal transit time, which is the basic measurement from GPS. We take a slight detour in this section to review refraction and wave propagation from freshman physics.

The *refractive index* of a medium (n) is defined as the ratio of the speed of propagation of the signal in a vacuum (c) to the speed in the medium (v).

$$n = \frac{c}{v} \quad (5.10)$$

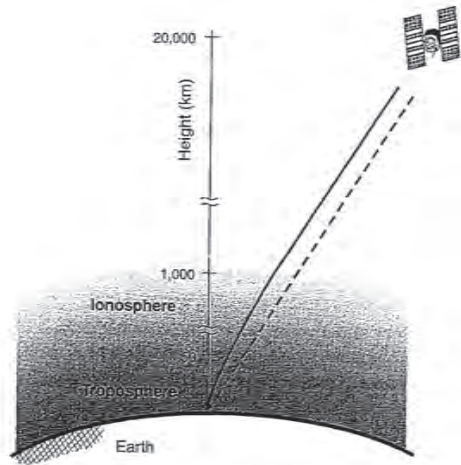


Figure 5.6 Refraction of GPS signals in the earth's atmosphere results in changes to both speed and direction. Increase in path length due to bending of the signal ray, greatly exaggerated above, is generally insignificant. The effect of the change in speed of propagation, however, can result in pseudo-range measurement error of several meters or more.

Actually, the ionosphere and troposphere are not uniform in composition and the refractive index changes all along the path of a signal. Change in signal speed changes the travel time of the signal and, therefore, changes the apparent range to the satellite as computed from (5.1). Changing refractive index in a medium results in bending of the path of a signal ray (Snell's law), making the path longer than the geometrical straight-line path (Figure 5.6). The transit time along this curved path, however, is shorter than that for a straight-line path (Fermat's principle of least time). The effect of signal bending, however, is very small except at very low elevation angles and we will assume the signal paths to be straight lines.

Given the refractive index profile along the propagation path of a signal, the travel time

$$\tau = \frac{1}{c} \int_{\text{Satellite}}^{\text{Receiver}} n(l) dl \quad (5.11)$$

where the integration is along the signal propagation path, and $n(l)$ represents the (changing) refractive index of the medium. In the rest of this section, we will use 'R' for the receiver and 'S' for the satellite in the integration limits for simplicity, or ignore the limits altogether. We can write an expression for the excess delay in signal propagation due to refraction as

$$\Delta\tau = \frac{1}{c} \int_S^R [n(l) - 1] dl \quad (5.12)$$

or, equivalently, effective increase in path length

$$\Delta\rho = \int_S^R [n(l) - 1] dl. \quad (5.13)$$

If the refractive index of a medium depends upon the frequency of the signal, the medium is said to be *dispersive*. We know, for example, that a glass prism is dispersive for visible light, and so are water droplets hanging in the air. For radio signals in the L-band, the ionosphere is dispersive, but the troposphere is not.

In order to determine the excess travel time for the GPS signals as they propagate through the ionosphere and troposphere, we have to determine the refractive index along the propagation path. That is the main subject of Sections 5.3.2 and 5.3.3. But first we review wave propagation and set the stage for discussion of *code-carrier divergence*, an important phenomenon associated with the propagation of GPS signals through the ionosphere.

Let us consider a monochromatic sinusoidal wave moving in the x -direction.

$$s(x, t) = s_0 \cos(\omega t - kx + \phi_0) \quad (5.14)$$

where s_0 is the amplitude of the wave; ω is the *circular frequency*; t is time; k is called the *wave number*, to be addressed presently; and ϕ_0 is the phase offset. Given below are the familiar, basic relationships among the parameters of wave motion: ω , k , frequency f , (temporal) period T , and wavelength (or, spatial period) λ .

$$\omega = 2\pi/T$$

$$k = 2\pi/\lambda$$

$$T = 1/f = 2\pi/\omega$$

Returning to (5.14), as t or x changes, so does the phase $(\omega t - kx + \phi_0)$ of the wave. The phase is constant if $(\omega t - kx)$ is constant, for example, if

$$x = \frac{\omega}{k} t = v_p t$$

where v_p is the *phase velocity* of the wave. (Actually, the correct term is phase speed, velocity being associated with a vector.)

$$v_p = \lambda \cdot f = \frac{\lambda}{T} = \frac{\omega}{k}$$

We can rewrite (5.14) as

$$s(x, t) = s_0 \cos[\omega(t - x/v_p) + \phi_0] \quad (5.15)$$

As the wave moves, the whole sinusoidal pattern moves along the x -direction with speed v_p . The distance between the wave crests is the wavelength (or, spatial period) λ . The frequency f with which the wave crests pass a stationary observer is $f = v_p/\lambda$.

Now let us look at the propagation of a modulated signal through a dispersive medium. Consider a simple amplitude-modulated (AM) signal

$$s(t) = \cos(\omega_m t) \cos(\omega_c t) \quad (5.16)$$

where ω_c is the carrier frequency and ω_m is the modulation frequency, $\omega_m \ll \omega_c$. The signal is plotted in Figure 5.7(a). [In the context of a GPS SPS signal, ω_c and ω_m are akin to the carrier frequency (≈ 1.5 GHz) and the chipping rate (≈ 1 MHz) of the code, respectively.] The reason for this choice of the signal becomes clear when we invoke a simple trigonometric identity to rewrite (5.16) as

$$s(t) = \frac{1}{2} [\cos(\omega_c + \omega_m)t + \cos(\omega_c - \omega_m)t]$$

The amplitude-modulated signal (5.16) is simply the average of two sinusoidal signals of slightly different frequencies. We now consider transmission of this signal through a dispersive medium where each of the sinusoidal signals travels at a slightly different speed. Let the wave number corresponding to frequency ω_c be k , and the wave numbers corresponding to $(\omega_c + \omega_m)$ and $(\omega_c - \omega_m)$ be $(k + \Delta k)$ and $(k - \Delta k)$, respectively.

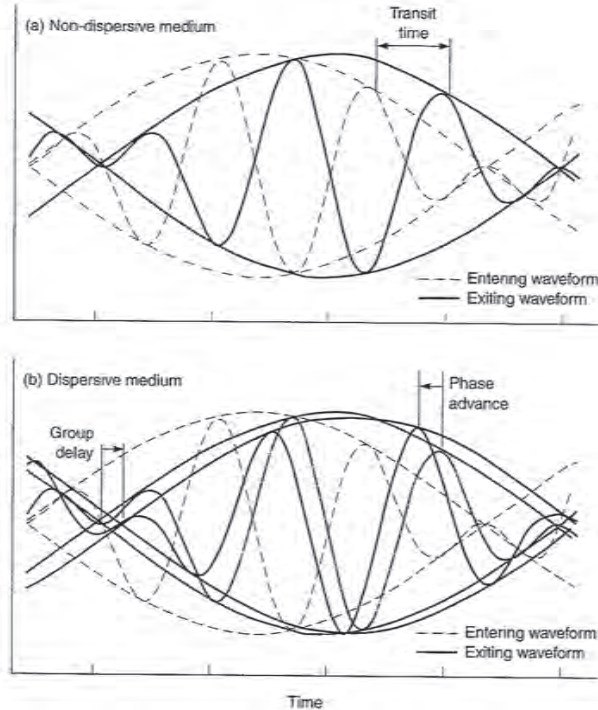


Figure 5.7 Propagation of a modulated signal in (a) free space, and (b) a dispersive medium which advances the carrier phase and delays the modulation function.

After the two frequency components have traveled a distance x in this dispersive medium

$$s(x, t) = \frac{1}{2} [\cos((\omega_c + \omega_m)t - (k + \Delta k)x) + \cos((\omega_c - \omega_m)t - (k - \Delta k)x)]$$

Again, by a simple trigonometric identity

$$\begin{aligned} s(x, t) &= \cos(\omega_m t - \Delta k \cdot x) \cos(\omega_c t - kx) \\ &= \cos \omega_m \left(t - \left(\frac{\Delta k}{\omega_m} \right) x \right) \cos \omega_c \left(t - \frac{k}{\omega_c} x \right) \end{aligned}$$

As ω_m gets smaller, $\Delta k/\omega_m$ approaches, in the limit, $dk/d\omega$, and

$$s(x, t) = \cos \omega_m \left(t - \frac{dk}{d\omega} x \right) \cos \omega_c \left(t - \frac{k}{\omega_c} x \right) \quad (5.17)$$

Equation (5.17) is an important result: In a dispersive medium, the carrier and the modulation travel at different speeds. The carrier (frequency ω_c) travels with phase velocity

$$v_p = \frac{\omega_c}{k},$$

and the modulation function $\cos(\omega_m t)$ travels with group velocity

$$v_g = \frac{d\omega}{dk}$$

where the derivative is evaluated at $\omega = \omega_c$. We can rewrite (5.17) as

$$s(x, t) = \cos \omega_m \left(t - \frac{x}{v_g} \right) \cos \omega_c \left(t - \frac{x}{v_p} \right) \quad (5.18)$$

Consider first the simple case of free space where $v_p = v_g$. We have a superposition of two waves of slightly different frequencies, $(\omega_c + \omega_m)$ and $(\omega_c - \omega_m)$, traveling at the same speed. The net result, shown in Figure 5.7(a) as a function of t for a fixed x , is a beat signal or a wave with a slowly pulsating intensity (or amplitude modulation). In other words, a wave whose frequency is ω_c , the average of the two frequencies, and which oscillates in strength with a frequency ω_m , one-half of the difference of the two frequencies. We see an envelope riding on top of the carrier waves, both propagating at the same speed.

Let's consider next the case of a dispersive medium in which the two waves of slightly different frequencies travel at slightly different speeds. The difference is that we have an amplitude-modulated signal with the modulation traveling at a different speed than the carrier. Figure 5.7(b) is a plot of (5.18), as before, as a function of t for a fixed x . As the waves travel along, the envelope riding on them travels slower ($v_g < v_p$). The important point for our immediate purpose is related to measurement of phases of the carrier and the modulation (i.e., the code) of GPS signals. As shown in Figure 5.7(b), the carrier phase would be measured too short and the code phase would be measured too long, as compared to measurements in Figure 5.7(a). This phenomenon is referred to as *code-carrier divergence*.

Earlier we defined the refractive index of a medium as $n = c/v$. For a dispersive medium, the phase and group velocities are different, and we define the phase and group refractive indices as

$$n_p = \frac{c}{v_p} \quad \text{and} \quad n_g = \frac{c}{v_g}$$

It is left as an exercise to show from the definitions of phase and group refractive indices that they are related as

$$n_g = n_p + f \frac{dn_p}{df} \quad (5.19)$$

We will use this relationship below when we examine transmission of the carrier and modulation of GPS signals through the ionosphere.

5.3.2 Ionospheric Delay

The ionosphere, extending from a height of about 50 km to about 1000 km above the earth, is a region of ionized gases (free electrons and ions). The ionization is caused by the sun's radiation, and the state of the ionosphere is determined primarily by the intensity of the solar activity. The ionosphere is composed of layers (named D, E, F1, and F2) at different heights, each with different rates of production and loss of free electrons. The peak electron density (number of electrons/m³) occurs in height range 250–400 km (F2 layer).

The physical characteristics of the ionosphere change widely between day and night. As the sun rises, its ultraviolet (UV) radiation starts breaking up gas molecules (mainly H₂ and He at higher altitudes, and O₂ and N₂ lower down) into ions and free electrons. The electron density builds up, with a peak around 2 p.m. local time, and then starts declining. At night, there is no further ionization and the ions and electrons find each other and recombine, reducing the free-electron count. The electron densities change by one to two orders of magnitude between day and night.

There can be considerable variability from day to day, depending upon the solar activity and geomagnetic disturbances. There are significant changes with seasons and the phases of the eleven-year solar cycle. (In 2005 we are at the tail end of the solar cycle which began in 1995 and peaked in 2000–2001. The next solar max will occur around 2010.) There are also unpredictable short-term effects and localized anomalies (traveling ionospheric disturbances). In this section, we offer an overview of the ionospheric propagation effects on GPS signals. A comprehensive discussion is found in Klobuchar (1996).

The speed of propagation of radio signals in the ionosphere depends upon the number of free electrons in the path of a signal, defined as the *total electron content* (TEC): the number of electrons in a tube of 1 m² cross section extending from the receiver to the satellite.

$$TEC = \int_S^R n_e(l) dl \quad (5.20)$$

where $n_e(l)$ is the variable electron density along the signal path, and the integration is along the signal path from the satellite to the receiver. The path length through the ionosphere is shortest in the zenith direction and, therefore, TEC in the vertical direction (TECV) is the

lowest. In the previous paragraphs, when we referred to the state of the ionosphere, we really meant a map of TECV. TEC is measured in units of TEC Units (TECU), defined as 10¹⁶ electrons/m². TECV typically varies between 1 and 150 TECU. For a given place and time, vertical TEC can vary 20–25% from its monthly average! The current models of the ionosphere do not provide an adequate representation of the day-to-day variability of TEC.

The ionosphere is generally well behaved in the temperate zones but can fluctuate near the equator and the magnetic poles. The region of highest ionospheric delay is within ±20° of the magnetic equator. Solar flares and the resulting magnetic storms can create large and quickly varying electron densities, especially in the polar regions, causing rapid fluctuation in the carrier phase (called scintillation) and in amplitude (called fading) of GPS signals. This phenomenon, though transient and infrequent at mid-latitudes, can present a difficulty in tracking the signal continuously in the polar and equatorial regions.

Phase Advance and Group Delay

Ionized gas is a dispersive medium for radio waves. The refractive index for a radio wave of frequency f (to first order) is

$$n_p \approx 1 - \frac{40.3}{f^2} n_e \quad (5.21)$$

where n_p is the phase refractive index and n_e is the electron density, defined earlier. The refractive index is slightly less than one, and phase velocity of the GPS carriers in the ionosphere exceeds that of light in a vacuum by enough to be significant for precise positioning. From the expression for the refractive index we can calculate the excess phase delay $\Delta\tau_p$ (in seconds) incurred by a signal as it propagates through the ionosphere (5.12).

$$\begin{aligned} \Delta\tau_p &= \frac{1}{c} \int_S^R (n_p(l) - 1) dl \\ &= -\frac{1}{c} \int_S^R \frac{40.3 n_e(l)}{f^2} dl \\ &= -\frac{40.3 \cdot TEC}{cf^2} \end{aligned} \quad (5.22)$$

The phase delay is negative, i.e., the phase is advanced. The phase advance is in direct proportion to the number of electrons in the path of the signal. The excess phase delay in meters, which we have denoted as I_ϕ in (5.9), is

$$I_\phi = c \cdot \Delta\tau_p = -\frac{40.3 \cdot TEC}{f^2} \quad (5.23)$$

From (5.19), we can determine the group refractive index and the group delay in meters as

$$\begin{aligned} n_g &= 1 + \frac{40.3 n_e}{f^2} \\ I_\rho &= \frac{40.3 \cdot TEC}{f^2} \end{aligned} \quad (5.24)$$

The ionospheric delay terms in measurements of pseudorange (5.6) and carrier phase (5.9) are equal in magnitude but opposite in sign.

$$I_p = -I_\phi = \frac{40.3 \cdot \text{TEC}}{f^2} \quad (5.25)$$

In later sections, we will represent the ionospheric group delay simply as I , and the phase delay as $-I$. Note that a change in TEC of 1 TECU corresponds to a change in ionospheric delay at L1 of about 16 cm.

Obliquity Factor

For the purpose of simple, geometrical modeling, the ionosphere may be thought of as a thin shell surrounding the earth (Figure 5.8). The signal path length through the ionosphere varies with the satellite position in the sky: the lower the satellite, the longer the path length and higher the TEC. Assuming that there are no lateral electron gradients, we can get a simple and compact characterization of TEC along a signal path in terms of the vertical TEC (TECV) and a multiplier to account for the longer path length. The multiplier is called *obliquity factor* (OF). We use both the elevation angle (el) and its complement, zenith angle (ζ), to represent the satellite position as it suits us.

The mean height of the ionospheric shell, or the mean ionospheric height (h_f), is usually taken in the range of 300–400 km. The *ionospheric pierce point* (IP) is defined as the point of intersection of the line of sight with the spherical shell at height h_f . Now we can relate TECV to TEC at zenith angle ζ in terms of path lengths through the thin ionospheric shell as

$$\text{TEC}(\zeta) = \frac{1}{\cos \zeta'} \cdot \text{TECV} \quad (5.26)$$

where ζ and ζ' are the zenith angles of the satellite at the user position and IP, respectively. The term $(\cos \zeta')^{-1}$ defines the obliquity factor, which we can write in terms of the zenith angle of the satellite at the user location as follows. From Figure 5.8, by law of sines

$$\frac{\sin \zeta}{(R_E + h_f)} = \frac{\sin \zeta'}{R_E} \quad (5.27)$$

where R_E is the average radius of the earth. The ionospheric obliquity factor for zenith angle ζ

$$\text{OF}_f(\zeta) = \left[1 - \left(\frac{R_E \sin \zeta}{R_E + h_f} \right)^2 \right]^{-1/2} \quad (5.28)$$

The value of OF_f ranges from one for the zenith direction to about three for elevation angle of 5° . This *thin shell model* works surprisingly well.

The relationship (5.26) can be translated directly into group delay (or phase advance) for the GPS signals in accordance with (5.23) and (5.24). Denoting the ionospheric delay as a function of the zenith angle as $I(\zeta)$

$$\text{ionospheric delay}(\zeta) = \text{zenith delay} \times \text{obliquity factor}(\zeta)$$

$$I(\zeta) = I_z \cdot \text{OF}_f(\zeta)$$

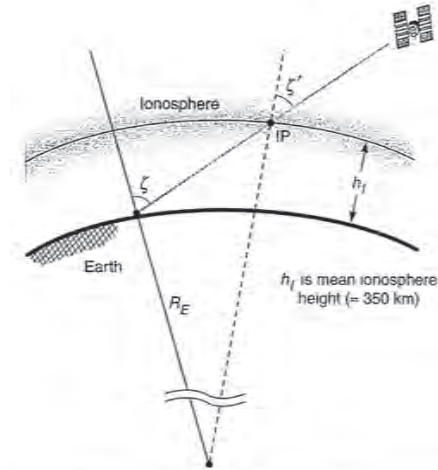


Figure 5.8 The propagation path length of a signal through the ionosphere increases with the zenith angle ζ . The increased path length is accounted for in terms of a multiplier of the zenith delay. The multiplier is called obliquity factor.

The ionospheric zenith delay I_z (i.e., the path delay in the zenith direction) typically varies at mid-latitudes from about 1–3 m at night to 5–15 m in the mid-afternoon. The zenith delay has been observed to be as high as 36 m near the equator at the peak of a solar cycle.

We can summarize our discussion so far of signal propagation through the ionosphere as follows.

- The carrier and its modulating signal (i.e., the code and navigation message) propagate at different speeds through the ionosphere: the code phase is delayed while the carrier phase is advanced by the same amount. In other words, the ionospheric refraction causes the code phase to be measured too long and the carrier phase to be measured too short by an equal amount.
- The phase advance and group delay are in direct proportion to TEC along the propagation path of a signal.
- The path length of a signal through the ionosphere depends upon the elevation angle of the satellite and is accounted for in the form of a multiplier called obliquity factor.

Delay Estimation with Dual-Frequency Measurements

A user equipped with a dual-frequency (L1-L2) GPS receiver can estimate the ionospheric group delay and phase advance from the measurements, and essentially eliminate the ionosphere as a source of measurement error. We rewrite the expression for pseudorange measurements (5.6), introducing additional notation to indicate L1 and L2 measurements.

$$\rho_q = r + c[\delta t_u - \delta t^r] + I_q + T + \varepsilon_{\rho_q}$$

where $q = L1$ or $L2$. Consolidating terms unrelated to the ionospheric effect and modeling the ionospheric delay to first order as varying inversely with the carrier frequency squared (5.25), we can rewrite the above equation simply as

$$\rho_q = \rho_{IF} + \frac{A}{f_q^2} \quad (5.29)$$

where ρ_{L1} and ρ_{L2} are the pseudorange measurements at L1 and L2, respectively; f_{L1} and f_{L2} are the corresponding carrier frequencies; ρ_{IF} is the *ionosphere-free (IF) pseudorange*, or the pseudorange measurement if the ionospheric effect were absent; and the unknown parameter $A = 40.3$ TEC. Both ρ_{IF} and A can be estimated from dual-frequency measurements (5.29). The estimate of ionospheric group delay at L1 is

$$I_{L1} = \frac{A}{f_{L1}^2} = \frac{f_{L2}^2}{(f_{L1}^2 - f_{L2}^2)} (\rho_{L2} - \rho_{L1}) \quad (5.30)$$

The estimate of ionosphere-free pseudorange measurement is

$$\begin{aligned} \rho_{IF} &= \frac{f_{L1}^2}{(f_{L1}^2 - f_{L2}^2)} \rho_{L1} - \frac{f_{L2}^2}{(f_{L1}^2 - f_{L2}^2)} \rho_{L2} \\ &= 2.546 \rho_{L1} - 1.546 \rho_{L2} \end{aligned} \quad (5.31)$$

There are two things to note about the ionosphere-free measurements (5.31). First, the errors due to satellite clock, ephemeris, and troposphere are present in these measurements in full measure, as in ρ_{L1} and ρ_{L2} . Second, the elimination of the ionospheric effect has been achieved at a price: The ionosphere-free pseudorange is significantly noisier than the pseudoranges measured at L1 and L2. If we model the multipath and receiver noise at L1 and L2 to be uncorrelated and having the same variance, the noise in the ionosphere-free pseudorange is

$$\sqrt{2.546^2 + 1.546^2} \approx 3$$

times larger than that in ρ_{L1} or ρ_{L2} . Actually, the assumption of uncorrelatedness is reasonable but the noise in the measurements at L2 is larger for the GPS receivers at present, a price for taking the forbidden measurements.

Estimates of the ionospheric delay based on L1-L2 pseudorange measurements in accordance with (5.30) are shown in Figure 5.9. The estimates appear noisy, but it is no surprise in view of the remarks above. The ionospheric delay estimates (5.30) are about twice as noisy as the individual pseudorange measurements. It's a trade-off: We get rid of a significant bias term but pick up additional noise in the process.

The carrier phase measurements are much less noisy, and we can attempt to estimate the ionospheric phase advance based on them. First, we rewrite the expression for carrier phase measurements (5.9) introducing notation to distinguish between L1 and L2 measurements.

$$\phi_q = \lambda_q^{-1} [r - I_q + T_\phi] + f_q (\delta t_u - \delta t^r) + N_q + \varepsilon_{\phi_q} \quad (5.32)$$

Equations similar to (5.30) and (5.31) can be written for the ionosphere-free carrier phase measurement, but involve integer ambiguities. It is left as an exercise to show that the phase advance at L1 is

$$I_{L1} = \frac{f_{L2}^2}{(f_{L1}^2 - f_{L2}^2)} [\lambda_{L1} (\phi_{L1} - N_{L1}) - \lambda_{L2} (\phi_{L2} - N_{L2})] \quad (5.33)$$

The estimate of ionospheric delay based on the code measurements (5.30) is unambiguous but noisy. The corresponding estimate based on carrier phase measurements (5.33) is precise but ambiguous in integer values.

We have noted previously that it's a challenge to estimate N_{L1} and N_{L2} in real time. So, what's the point of (5.33)? Well, as long as continuous carrier tracking is maintained, N_{L1} and N_{L2} remain fixed, and we can use (5.33) to estimate accurately change in the ionospheric delay between measurement epochs in real time. Such change is called *differential delay*, and its estimate obtained from (5.33) is good to centimeter level. We have plotted the estimates of differential delay in Figure 5.9 for comparison with the corresponding estimates of ionospheric group delay obtained from (5.30). The noisy code-based estimates of the ionospheric delay can be smoothed with carrier phase-based estimates of the differential delay. The basic technique is discussed in Section 5.7. The resulting smoothed estimates of the ionospheric delay would typically have decimeter-level accuracy after a satellite has risen to about 30° in elevation angle.

Dual-frequency GPS measurements have provided atmospheric physicists with a powerful tool to study the ionosphere. Actually, Transit with its dual-frequency signaling was the first satellite navigation system to be employed for monitoring the ionospheric dynamics. GPS offers a vast improvement over Transit. The ability to map TEC in real time on a global scale and

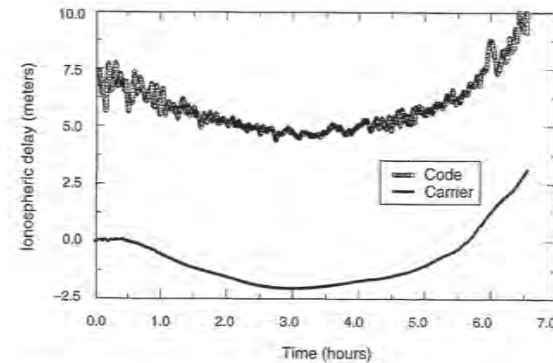


Figure 5.9 Ionospheric delay (L1) estimates obtained from code and carrier phase measurements at both L1 and L2. The code-based estimates are noisy. The carrier-based estimates are precise but ambiguous, and the plot starts arbitrarily at zero value. Changes in the ionospheric delay can be measured accurately with the carrier phase measurements while the carrier is tracked continuously.

to predict or identify disturbed regions is of great value to HF communications facilities and communication satellites, and for radar tracking of objects in space. The International GNSS Service (IGS) now offers global ionospheric maps giving TECV values over a grid, but not yet in real time.

We note in passing a calibration issue to account for differences in path lengths for the L1 and L2 signals, or inter-frequency biases, in a satellite. These parameters, also known as differential group delays or T_{GD} values, are specified in the satellite navigation message (Subframe 1). These correction terms are relevant to estimation of ionospheric delay from dual-frequency measurements, positioning with single frequency receivers, and time transfer [Wilson *et al.* (1999)].

Broadcast Model

Receivers limited to L1 measurements have recourse to an empirical model whose parameter values are broadcast by the satellites. This model, often referred to as the *Klobuchar model* in honor of its developer [Klobuchar (1996)], represents the zenith delay as a constant value at nighttime and a half-cosine function in daytime (Figure 5.10). The zenith ionospheric delay estimate $\hat{I}_{z,L1}$ at local time t is given by

$$\frac{\hat{I}_{z,L1}}{c} = \begin{cases} A_1 + A_2 \cos\left(\frac{2\pi(t - A_3)}{A_4}\right), & \text{if } |t - A_3| < A_4/4 \\ A_1 & \text{otherwise} \end{cases} \quad (5.34)$$

where A_1 : nighttime value of the zenith delay (fixed at 5×10^{-9} s),
 A_2 : amplitude of the cosine function for daytime values,
 A_3 : phase corresponding to the peak of the cosine function (fixed at 50,400 s, or 14 h, local time), and
 A_4 : period of the cosine function ($\geq 72,000$ s).

The values of parameters A_2 and A_4 are specified in the navigation message broadcast by each satellite in terms of four coefficients each of a polynomial function (Subframe 4, page 18). A set of eight coefficients is selected by the Master Control Station from a pool of 370 such sets associated with the different seasons and levels of solar activity. The parameters are typically updated every six days. The ionospheric delay is determined from the broadcast parameter values (A_2 and A_4) and the user's latitude, longitude, satellite azimuth and elevation, and local time. Step-by-step calculations of the zenith delay and obliquity factor are given in the GPS Interface Specification (IS) [IS-GPS-200D (2004)].

The Klobuchar model was developed on the basis of empirical data, but under constraints on the number of parameters to be used (eight, at most) and how often they could be updated (daily, at most). We offer a few remarks. First, the time t in (5.34) is the local time at the ionospheric pierce point (IP): The ionospheric action is at IP, not at the user location. Secondly, the position of the IP is expressed in terms of geomagnetic latitude, not geodetic. Why? Because the model fits the empirical data better when geomagnetic coordinates are used. (The computation uses the geodetic coordinates of the geomagnetic north pole: N78.3°, E291.0°.) The model represents the peak ionospheric delay as occurring at 14 h local time at all places. In fact, the time of the local peak can vary between 11 h and 17 h depending upon the season, latitude, and solar activity.

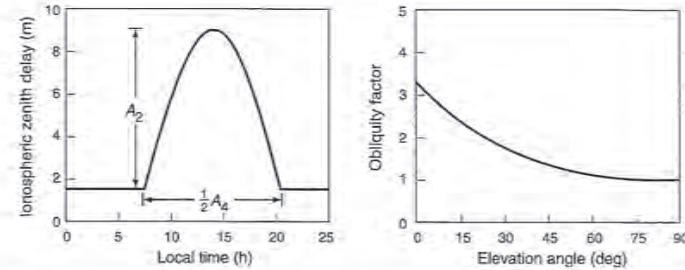


Figure 5.10 The Klobuchar ionospheric model. Parameter values A_2 and A_4 are selected by the Control Segment to reflect the prevailing ionospheric conditions and are broadcast by the satellites.

The obliquity factor for the broadcast model is expressed in terms of the satellite elevation angle el as

$$OF_1(el) = 1.0 + 16.0 \times (0.53 - el)^3 \quad (5.35)$$

where el is in units of semi-circles (1 semi-circle = 180° or π radians). This expression is an approximation of (5.28). The motivation for simplifying (5.28) is to make computations easier. In the pre-microprocessor 1970s, there was a lot of concern about keeping the computational demands on a receiver low.

The broadcast model is estimated to reduce the rms range measurement error due to uncompensated ionospheric delay by about 50% [Feess and Stephens (1987)]. At mid-latitudes, the remaining error in zenith delay can be up to 10 m during the day, and much worse during heightened solar activity.

5.3.3 Tropospheric Delay

The GPS signals are also refracted by the lower part of the earth's atmosphere composed of dry gases (mainly N_2 and O_2) and water vapor. Most of the water vapor is below 4 km, and all of it is below about 12 km, measured from sea level. The dry gases, however, are found in gradually thinning quantities at altitudes of hundreds of kilometers. About three-quarters of the gaseous mass is found in the troposphere, which stretches to about 16 km above the equator and about 9 km above the poles. The overall effect of the neutral atmosphere is, therefore, referred to as the tropospheric effect. In this section, we offer an overview of the tropospheric models. A comprehensive discussion is found in Spilker (1996).

Unlike the ionosphere, troposphere is non-dispersive for GPS frequencies (i.e., the refractive index doesn't depend upon the frequency of the signal). The refractive index of the atmospheric gases is only slightly larger than unity: $n \approx 1.0003$ at sea level, and much closer to unity at the upper end of the troposphere. The speed of propagation of GPS signals in the troposphere is lower than that in free space and, therefore, the apparent range to a satellite appears longer, typically by 2.5–25 m depending upon the satellite elevation angle. The phase and group velocities are the same, and the measurements of code and carrier at L1 and L2 frequencies all experience a common delay. This delay cannot be estimated from GPS mea-

measurements, and a user has to resort to models to correct for it. In the notation of Section 5.1, $T_p = T_\phi$, and, therefore, we will drop the subscripts p and ϕ in later sections and denote the tropospheric delay simply as T . In this section, however, we need T to denote the temperature and switch to $\tilde{T}$ to denote the tropospheric delay. (T is an overworked symbol in this book, used in Part III to denote both the period of a C/A-code and the noise temperature.)

The extent of the tropospheric delay experienced by a signal depends upon the refractive index of the air mass along its path. The refractive index of an air mass depends upon its density, which can be expressed as the sum of the densities of the dry air constituents and water vapor. These densities are functions of the pressures of the dry gases and water vapor, and the temperature. The dry and wet constituents of the atmosphere affect propagation of an RF signal differently, and each is modeled separately. The composition of the dry gases depends upon latitude, season, and altitude, and is relatively stable. The water vapor content in the troposphere, however, is much harder to model; water vapor density varies with the local weather and can change quickly. Fortunately, most of the tropospheric delay ($\approx 90\%$) is due to the more predictable dry atmosphere.

Dry and Wet Delays

It is convenient to define *refractivity*, $N = (n - 1) \times 10^6$, and express it as the sum of refractivities of the dry gases and the water vapor in the atmosphere

$$N = N_d + N_w, \quad (5.36)$$

where N_d and N_w are called the dry and wet refractivities, respectively. As in the derivation of (5.13), the excess path delay due to tropospheric refraction (in meters) can be written as

$$\begin{aligned} \tilde{T} &= 10^{-6} \int N(l) dl = 10^{-6} \int [N_d(l) + N_w(l)] dl \\ &= \tilde{T}_d + \tilde{T}_w, \end{aligned} \quad (5.37)$$

where $\tilde{T}_d$ and $\tilde{T}_w$ are called dry delay and wet delay, respectively, and the integration is along the signal path. As with signal propagation through the ionosphere, we will disregard the curvature of the signal path and focus on excess delay due to the reduced signal speed.

Models of the troposphere attempt to estimate the dry and wet refractivities along the signal path, and estimate the total delay from (5.37). The refractivity of a parcel of air depends upon its temperature and partial pressures of the dry gases and water vapor. Simple, approximate expressions (with empirically obtained coefficients) for the dry and wet refractivities of a parcel of air are

$$\begin{aligned} N_d &= 77.64 \frac{P}{T}, \\ N_w &= 3.73 \cdot 10^5 \frac{e}{T^2}, \end{aligned} \quad (5.38)$$

where P is the total pressure and e is the partial pressure of the water vapor, both in millibars; and T is the temperature in kelvin.

Knowledge of pressure, temperature, and humidity along the propagation path of a signal can determine the refractivity profile and the tropospheric delay (5.37) precisely. Such measurements are obtained for the nominally vertical direction by atmospheric researchers using

instrumented weather balloons (radiosondes). GPS users rarely have access to such measurements. A less onerous approach is to measure the meteorological conditions (pressure, temperature, and humidity) at the antenna location, and relate these measurements to P , T , and e along the signal path using gas laws or empirical models. Researchers requiring great positioning precision with GPS often record surface meteorological conditions. For example, GPS stations of a permanent tectonic monitoring network make provisions for meteorological instruments at each site.

In navigation applications, however, meteorological measurements are impractical. In fact, most GPS users base the estimation of tropospheric delay upon average meteorological conditions at their locations obtained from a model of the *standard atmosphere* for the day of the year and the user's latitude and altitude. (A standard atmosphere is basically a specification of pressure, temperature, and humidity profile with altitude, and variations in these parameters with latitude and seasons based on certain average conditions.)

We'll consider estimation of the tropospheric delay for a signal in two steps:

1. Estimation of the zenith delay ($\tilde{T}_z$) (i.e., delay associated with a signal from the zenith direction) in terms of the corresponding dry ($\tilde{T}_{z,d}$) and wet ($\tilde{T}_{z,w}$) delays

$$\tilde{T}_z = \tilde{T}_{z,d} + \tilde{T}_{z,w}$$

2. Definition of obliquity factor to scale the zenith delay as a function of the elevation angle (el) of the satellite. The atmospheric physicists refer to the tropospheric obliquity factors as mapping functions.

$$\tilde{T} = \tilde{T}_{z,d} \cdot m_d(el) + \tilde{T}_{z,w} \cdot m_w(el)$$

where we have introduced separate mapping functions m_d and m_w for the dry and wet components, respectively. Simple models often use a common mapping function ignoring differences in the atmospheric profiles of the dry gases and water vapor.

There is no dearth of tropospheric models. The models differ in their assumptions regarding changes in temperature and water vapor with altitude. We describe two simple models below. There is no shortage of mapping functions either. We discuss three below.

Tropospheric Models

The *Saastamoinen model* was derived using gas laws and simplifying assumptions regarding changes in pressure, temperature, and humidity with altitude. The zenith dry and wet delays are given as

$$\tilde{T}_{z,d} = 0.002277 (1 + 0.0026 \cos 2\phi + 0.00028 H) P_0, \quad (5.39a)$$

$$\tilde{T}_{z,w} = 0.002277 \left(\frac{1255}{T_0} + 0.05 \right) e_0 \quad (5.39b)$$

where T_0 is the temperature (kelvin), P_0 is the total pressure and e_0 is the partial pressure due to water vapor (both in millibars), all determined at the antenna location by measurements or from models of standard atmosphere; ϕ is the latitude and H is the orthometric height of the

antenna (km). This model has been refined with additional correction terms which we disregard here.

The *Hopfield model* is based on a relationship between dry refractivity at height h to that at the surface. The relationship was derived empirically on the basis of extensive measurements. This model, referred to as a quartic model of refractivity profile, is

$$N_d(h) = N_{d0} \left(1 - \frac{h}{h_d} \right)^4, \quad (5.40)$$

where h denotes the height above the antenna, N_{d0} is the dry refractivity at the surface, and h_d (≈ 43 km) is defined as the height above the antenna at which the dry refractivity is zero: $N_d(h_d) = 0$.

The Hopfield model for wet refractivity assumes a relationship similar to (5.40), though with less persuasive evidence.

$$N_w(h) = N_{w0} \left(1 - \frac{h}{h_w} \right)^2$$

where N_{w0} is the wet refractivity at the surface and $h_w = 12$ km. From (5.37)

$$\begin{aligned} \bar{T}_z &= 10^{-6} \int [N_d(h) + N_w(h)] dh \\ &= \frac{10^{-6}}{5} [N_{d0} h_d + N_{w0} h_w] \\ &= \bar{T}_{z,d} + \bar{T}_{z,w} \end{aligned}$$

Substituting expressions for the dry and wet refractivities from (5.38)

$$\bar{T}_{z,d} = 77.6 \cdot 10^{-6} \frac{P_0}{T_0} \frac{h_d}{5} \quad (5.41a)$$

$$\bar{T}_{z,w} = 0.373 \frac{e_0}{T_0^2} \frac{h_w}{5} \quad (5.41b)$$

The value of $\bar{T}_{z,d}$ is 2.3–2.6 m at sea level, and gets lower as the altitude increases: about 2 m at Denver, Colorado, the mile-high city; and about 1 m atop a Himalayan peak. The value of $\bar{T}_{z,w}$ ranges from near-zero to 80 cm (millimeters in polar region, a few centimeters in deserts, and tens of centimeters in tropical areas).

The dry delay for the zenith direction ($\bar{T}_{z,d}$) can be predicted with an accuracy of a few millimeters from accurate surface pressure measurements. The corresponding wet delay depends upon the distribution of water vapor along the signal path, and can be highly variable. (The mixing of the water vapor and dry air is a complicated process depending upon the local weather conditions, and this distribution can change quickly.) The models of wet delay ($\bar{T}_{z,w}$) based on meteorological data at the surface are less accurate, with typical error of 1–2 cm. Use of average meteorological conditions rather than actual measurements introduces additional modeling errors in both the dry and wet delays, and the total zenith delay error can be 5–10 cm.

Mapping Functions

A number of mapping functions (or, tropospheric obliquity factors) have been proposed. The simplest model for both the dry and the wet components is $1/\sin el$. This model is consistent with a flat earth, and is a poor approximation for low-elevation satellites ($el < 15^\circ$). An example of a more accurate model is

$$m(el) = \frac{1}{\sqrt{1 - (\cos el / 1.001)^2}} \quad (5.42)$$

The following is a simple example of separate mapping functions for dry and wet delays.

$$\begin{aligned} m_d(el) &= \frac{1}{\sin el + \frac{0.00143}{\tan el + 0.0445}} \\ m_w(el) &= \frac{1}{\sin el + \frac{0.00035}{\tan el + 0.017}} \end{aligned}$$

More sophisticated mapping functions are based on a truncated form of continued fraction.

$$m_i(el) = \frac{1 + \frac{a_i}{1 + \frac{b_i}{1 + c_i}}}{\sin(el) + \frac{a_i}{\sin(el) + \frac{b_i}{\sin(el) + c_i}}} \quad (5.43)$$

where subscript $i = d$ or w , and the coefficients a_i , b_i , and c_i in the different models are empirically determined constants or functions of variables such as latitude, height, surface temperature and pressure, and day of the year.

The obliquity factor increases sharply as a satellite gets lower in the sky. At low elevation angles, the typical values are about two at 30° , four at 15° , six at 10° , and ten at 5° . The obliquity factors for the troposphere are much larger at low elevation angles than those for the ionosphere. The reason for this would be clear from a comparison of the corresponding path lengths as shown in Figures 5.8 and 5.11.

A 5–10 cm residual error in tropospheric zenith delay estimates based on average meteorological conditions grows into an error of 0.5–1 m at 5° elevation. Such error would not be of serious concern for meter-level navigation but, clearly, is too large for centimeter-level positioning. For an interesting discussion of the role of tropospheric errors in DGPS for landing airplanes, see Skidmore and van Graas (2004).

The atmospheric scientists are examining the use of GPS measurements to map the amount of water vapor in the air globally and assess its role in weather prediction and climate change. The profile of water vapor content in a column of air is measured at considerable expense using radiosondes. The U.S. Weather Service actually launches radiosondes from about eighty sites in the conterminous United States twice a day for this purpose. While GPS measurements

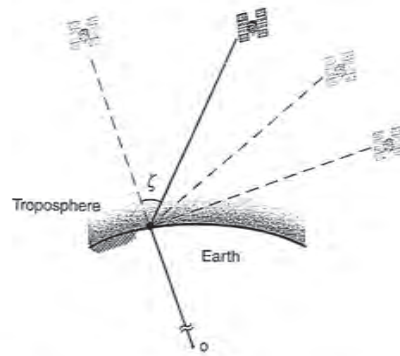


Figure 5.11 The path length of a signal through the troposphere increases significantly at low elevation angles. The obliquity factor is about two at 30° elevation angle, four at 15°, six at 10°, and ten at 5° (Symbol ζ denotes zenith angle; $\zeta = 90^\circ - \text{el.}$)

cannot give the profile of the water vapor content, a network of GPS stations can provide estimates of total water vapor content over large areas in real time continuously. It is, however, a challenge to compensate for the errors in the GPS measurements (satellite ephemeris and clock, ionosphere, tropospheric dry delay, multipath, and receiver noise) to estimate the zenith wet delay with the required accuracy in real time.

The main features of GPS signal propagation through the ionosphere and the troposphere are summarized in Table 5.1.

5.4 Measurement Errors

The errors in pseudorange measurements discussed so far are common to the code and carrier phase measurements (except for a sign change for the ionospheric delay). The signal or receiver design played no role. Now we examine a class of errors which actually depend upon signal power, code structure, and receiver/antenna design.

In this section, we address the question: How well can a receiver measure the code and carrier phase of a GPS signal reaching its antenna? We have to make a distinction between precision and accuracy of the measurements. Precision and accuracy are often used as synonyms, and we haven't been too careful about it in this book. Actually, there is a distinction. We know that accuracy is assessed in relation to the truth. Without being too pedantic about it, precision simply refers to the resolution or fineness of a measurement. A measurement or an observation can be extremely precise and completely wrong.

We look at two sources of error below. The first, receiver noise, essentially smudges the signal, affecting precision of the code or carrier phase measurement. The second, multipath, introduces interfering signals which actually change the phase being measured. Both effects are discussed in detail in Chapter 10.

Table 5.1 A summary of the features of ionosphere and troposphere relevant to GPS signal propagation

	<i>Ionosphere</i>	<i>Troposphere</i>
Variability	High (diurnal, seasonal, and solar cycles; solar flares)	Low (significant change only in the wet component, which is less than 10% of the delay)
Zenith delay	Several meters to several tens of meters	2.3–2.6 m at sea level
Obliquity factor (= 1 in zenith direction)	≈ 1.8 at 30° elevation angle; 2.5 at 15°; 3 at 5°	≈ 2 at 30° elevation angle; 4 at 15°; 10 at 5°
Modelling error for zenith delay	1–10 m, or more	5–10 cm (without meteorological data)
Dispersive for GPS frequencies?	Yes (therefore, the delay can be measured from dual-frequency measurements)	No

5.4.1 Receiver Noise

The code and carrier measurements are affected by random measurement noise, called receiver noise, which is a broad term covering the RF radiation sensed by the antenna in the band of interest unrelated to the signal; noise introduced by the antenna, amplifiers, cables, and the receiver; multi-access noise (i.e., interference from other GPS signals and GPS-like broadcasts from system augmentations); and signal quantization noise.

A receiver cannot follow changes in the signal waveform perfectly and invariably there are delays and distortions. In the absence of any interfering signals, a receiver sees a waveform which is the sum of the GPS signal and randomly fluctuating noise. The net result is that the fine structure of a signal can be masked by noise, especially if the signal-to-noise ratio is low. The measurement error due to receiver noise varies with the signal strength, which, in turn, varies with the satellite elevation angle.

According to a rule of thumb, it is not a challenge for the receiver technology to measure the phase of a reasonably strong sinusoidal signal with a precision of 1/2%–1% of a cycle. Empirical characterization of receiver noise is discussed by Nolan *et al.* (1992). The effect of receiver noise on the precision of the code and carrier phase measurements is analyzed in Section 10.6.

5.4.2 Multipath

Multipath refers to the phenomenon of a signal reaching an antenna via two or more paths. Typically, an antenna receives the direct (i.e., line-of-sight) signal and one or more of its reflections from structures in the vicinity and from the ground (see Figure 5.12). A reflected signal is a delayed and usually weaker version of the direct signal. The subsequent code and

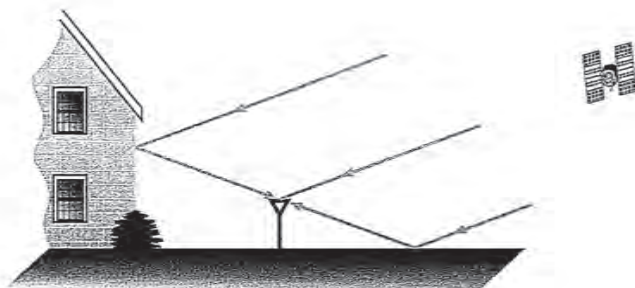


Figure 5.12 Multipath: The signal may reach an antenna via more than one path. A reflected signal is a delayed and usually weaker version of the direct signal.

carrier phase measurements are for the sum of the received signals. The range measurement error due to multipath depends upon the strength of the reflected signal and the delay between the direct and reflected signals. Multipath affects both code and carrier measurements, but the magnitudes of the error differ significantly. We offer a brief account below and an in-depth discussion in Section 10.7.

The primary defense against multipath is to locate the antenna away from reflectors, but that is not always practical. The effect of multipath can be reduced in antenna design process by lowering the contribution of some types of reflections (e.g., from the ground below the antenna). The effect can also be reduced in the signal processing step in a receiver, and several receiver manufacturers have developed and implemented proprietary techniques.

Actually, a measure of multipath immunity is built into the signal structure. A reflected signal which is delayed by more than 1.5 chips would be suppressed automatically in the correlation process in a receiver because the auto-correlation for the C/A-code is nearly zero for delays longer than 1.5 chips. Such delay corresponds to about 500 m of increased path length

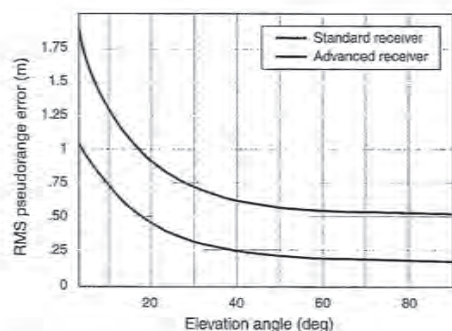


Figure 5.13 Code phase measurement error modeled as a function of elevation angle for two types of receivers.

for a C/A-code signal and 50 m for a P(Y)-code signal. One of the benefits of a higher chip-rate is greater multipath immunity. The impact of a reflected signal delayed by less than 1.5 chips would depend upon the amount of delay and the signal amplitude.

Typical multipath error in pseudorange measurements varies from 1 m in a benign environment to more than 5 m in highly reflective environment. The corresponding errors in the carrier phase measurements are typically two orders of magnitude smaller (1–5 cm). We know that the phase measurement error can be no worse than a cycle. Actually, it's no worse than a quarter cycle, as shown below.

Consider a simple model. An antenna receives two signals: a line-of-sight signal and a reflected signal with phase shift $\Delta\phi$ and amplitude attenuation α .

$$\text{received signal} = A \cos \phi + \alpha A \cos(\phi + \Delta\phi) \quad (5.44)$$

Application of straightforward but tedious trigonometric relationships allows us to determine the error in the carrier phase measurement as a result of multipath as

$$\delta\phi = \arctan\left(\frac{\sin \Delta\phi}{\alpha^{-1} + \cos \Delta\phi}\right) \quad (5.45)$$

For $\alpha < 1$, in the worst case, $\delta\phi = 90^\circ$. In other words, the error in the carrier phase measurements due to multipath does not exceed a quarter cycle if the reflected signal has a smaller amplitude than the direct signal.

5.4.3 Measurement Error Models

The measurement errors due to receiver noise and multipath depend upon the satellite elevation angle. As a satellite gets lower in the sky, the received signal power decreases and multipath increases. The combined effect on pseudorange measurements has been analyzed by McGraw *et al.* (2000) in certain 'clean' environments as a function of the satellite elevation angle. We present a summary of their results in Figure 5.13 for two categories of receivers. The first type we call 'standard' receivers. The second type, which we call 'advanced' receivers, are characterized by features such as narrow correlator, discussed in Section 10.5, and choke-ring antenna [Schupler and Clark (1991)]. Measurements from very low satellites are often more trouble than they are worth, and an elevation angle cutoff of 5° – 7.5° appears to offer a good trade-off between the loss of measurements and potential for large errors.

Typical rms errors for code and carrier phase measurements are shown in Table 5.2. Chances are that your measurements wouldn't be much better, and can be considerably worse. High-end receivers with special provision to deal with multipath would offer better measurements. Receivers intended for the consumer market can be considerably worse. Table 5.2 will serve as our simple model for measurement errors in illustrative examples in Chapters 6 and 7. For civil users, the L2 measurements are noisier than the L1 measurements, but we'll disregard this for simplicity.

5.5 User Range Error (URE)

We now return briefly to classification of a measurement error as noise or bias. Noise generally refers to a quickly varying error that averages out to zero over a 'short' time interval, where short is defined in relation to the integration time or smoothing time of a receiver [Chapters 11

Table 5.2 Typical receiver-related measurement errors (rms) in code and carrier phase measurements

	Receiver Noise	Receiver Noise and Multipath (RNRM)
Code phase	0.25–0.50 m	0.5–1.0 m
Carrier phase	0.005–0.01 cycle = 1–2 mm	0.025–0.05 cycle = 0.5–1 cm

and 12]. Receiver noise meets this requirement. The range error due to multipath is quasi-sinusoidal and may also be thought of as noise. The errors due to signal propagation delays in the ionosphere and troposphere, corrected with models or not, may persist for tens of minutes, or longer. The satellite clock and ephemeris errors also change slowly and, like the propagation errors, would be classified as biases.

We adopt a simple model for pseudorange measurements from a single-frequency (L1) receiver after corrections have been applied based on the Klobuchar model for ionospheric delay and a reasonable model for tropospheric delay. In our model, the range error (RE) attributed to the Control Segment (CS) has an rms value of about 3 m ($\sigma_{RE/CS} \approx 3$ m). This number is based on empirical data. The rms residual range error due to atmospheric propagation models at mid-latitudes is about 5 m ($\sigma_{RE/P} \approx 5$ m). The ionospheric component of the propagation error can be highly variable depending upon the user location, satellite elevation angle, and state of the medium. The rms range error due to receiver noise and multipath (RNRM), discussed above, is assumed to be 1 m ($\sigma_{RE/RNM} \approx 1$ m). The error due to multipath can easily be higher for many users. Our purpose here is to set up a simple model for later analyses.

We can now characterize the combined effect of these error sources on pseudorange measurements. The combined error is referred to as the user range error (URE), also known as the user equivalent range error (UERE). It is reasonable to model the errors due to the satellite clock and ephemeris, atmospheric propagation, multipath, and receiver noise to be uncorrelated, and the URE can be defined as the root-sum-square (rss) of these components, all expressed in units of length.

$$\sigma_{URE} = \sqrt{\sigma_{RE/CS}^2 + \sigma_{RE/P}^2 + \sigma_{RE/RNM}^2} \approx 6 \text{ m}$$

We should note in passing that while SA was active, $\sigma_{URE} = \sigma_{RE/CS} = 25$ m.

The error budget presented in Table 5.3 is noteworthy for its round numbers, reflecting our uncertainty about the actual values. A user would refine this model to take into account the knowledge of an actual measurement scenario. The atmospheric propagation error in the equatorial region or at high latitudes may double for a single-frequency receiver during high solar activity. On the other hand, $\sigma_{RE/P} \approx 0$ for a dual-frequency receiver in space. Similarly, a user with a low-end receiver located near reflectors may triple the size of the error due to multipath.

There is no standard terminology for characterization of ranging error, and this can be confusing. Sometimes total error, denoted as UERE, is divided into Signal-in-space URE

Table 5.3 Typical pseudorange measurement errors for a single-frequency (L1) receiver

Error Source	RMS Range Error (RE)
Satellite clock and ephemeris parameters	$\sigma_{RE/CS} \approx 3 \text{ m}$
Atmospheric propagation modeling	$\sigma_{RE/P} \approx 5 \text{ m}$
Receiver noise and multipath	$\sigma_{RE/RNM} \approx 1 \text{ m}$
User range error (URE)	$\sigma_{URE} \approx 6 \text{ m}$

(SIS URE or SISRE) and user equipment error (UEE). In our notation, SIS URE = $\sigma_{RE/CS}$ and $UEE = (\sigma_{RE/P}^2 + \sigma_{RE/RNM}^2)^{1/2}$.

5.6 Measurement Error: Empirical Data

To illustrate the size and character of the various errors, we have plotted in Figure 5.14 estimates of the error components in pseudorange measurements at L1. The measurements were taken in March 1997 from a satellite from the time it rose around noon until it set at about 4:30 p.m. local time. At its highest, the satellite elevation angle was 38°. In estimation of the errors, we took advantage of information generally not available to a user: We knew our antenna position precisely. We also sidestepped the receiver clock bias and drift issues by using an external cesium atomic standard. The measurements consisted of code and carrier phases at both L1 and L2 frequencies.

Figure 5.14(a) shows the dominant role of SA while it was active. The measurements are consistent with the commonly used model of the SA error as a zero-mean Gauss-Markov process with standard deviation of about 25 m. The correlation time of the error was about three to four minutes (i.e., error samples taken three to four minutes apart appear uncorrelated). SA is not missed.

Figure 5.14(b) shows the estimates of the ionospheric propagation delay obtained from the dual-frequency measurements and from the parameters of the broadcast model. The ionospheric delay profile for the entire satellite track was estimated separately from the dual frequency code and carrier measurements. The code-based estimates were then used effectively to resolve the integer ambiguity in the carrier phase measurements, and the resulting smooth estimate is shown. As noted earlier, the ionospheric effects on the code and carrier measurements are equal but in opposite directions. For simplicity, only the magnitude of the effect is shown. The ionospheric activity picks up at sunrise and peaks in the early afternoon. The plot covers the period of most intense ionospheric activity during the day, but the delays are small (< 5 m). In 1997, we were at the beginning of a new solar cycle and the ionosphere was relatively quiet. The broadcast model overcompensated for the ionospheric delay, and the remaining error is seen to be 3–6 m, depending upon the elevation angle. It is interesting to note that the propagation delay did not increase significantly at low elevation angles as the satellite set.

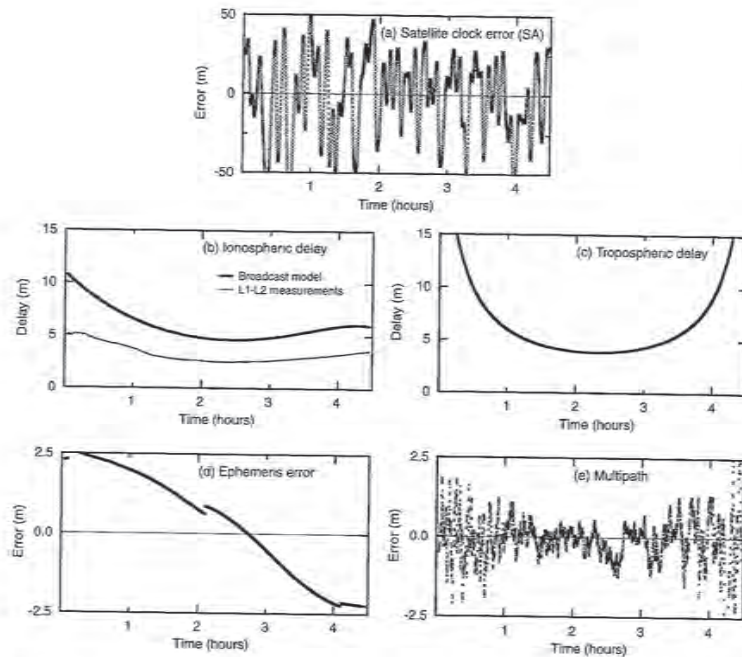


Figure 5.14 Pseudorange measurement errors due to the different error sources for a satellite from its rise time to set time (highest elevation angle: 38°). These measurements were taken in 1997 while Selective Availability (SA) was active.

The reason is simple: The satellite set in the north-east direction, away from the sun.

Figure 5.14(c) shows the tropospheric delay estimated using two different models. The first model required knowledge of the local meteorological conditions: pressure, temperature, and humidity measured at the antenna location. The second model, University of New Brunswick's UNB4 [Collins *et al.* (1996)], based the estimate on a definition of the standard atmosphere, given the season, latitude, and altitude. It is interesting that the two models gave virtually identical results. The role of the satellite elevation angle is clear: At low elevation angles, the signal traverses more of the troposphere, and the delay can be significantly larger than that at higher elevation angles. The delay ranges from about 4 m when the satellite was at its highest (38°) to about 15 m at low elevations (5°) as the satellite rose and set. If the satellite had come directly overhead, the corresponding delay would have been about 2.5 m.

Figure 5.14(d) shows the typical size of the error in broadcast ephemeris. Precise, post-fitted ephemerides are now readily available on the Internet from several sources, generally within two to four days. The broadcast ephemeris was compared with the post-fitted ephemeris obtained from the National Geodetic Survey (NGS). The pseudorange error was computed by projecting the ephemeris error vector on the line of sight. The discontinuities in the error

reflect the routine ephemeris updates at two-hour intervals. The observed ephemeris error did not exceed 2.5 m.

Figure 5.14(e) shows the remaining error in the measurements, identified as multipath and receiver noise. The antenna was mounted on the roof of MIT Lincoln laboratory in a relatively clean environment. The error is below 1 m, except for low satellite elevation angles, and shows quasi-sinusoidal oscillations with a period of several minutes, characteristic of multipath. The receiver did not have any special provisions to mitigate the multipath error. The code measurements, however, were smoothed using the carrier phase measurements, as discussed in Section 5.7, to achieve some reduction in the multipath error.

Our objective in introducing Figure 5.14 is simply to offer an appreciation for the nature and magnitude of the errors. As noted earlier, the errors show considerable variability. Uncompensated ionospheric and tropospheric delays would introduce a slowly changing bias in the measurements. We expect the receiver noise to be essentially uncorrelated from one measurement epoch to the next, but we cannot separate it from the multipath error. The combined effect has noise-like appearance, and can be mitigated to an extent by averaging, as discussed in the next section.

5.7 Combining Code and Carrier Measurements

While the carrier phase can be measured with great precision, the measurements include errors due to satellite clock and ephemeris parameter error, ionospheric and tropospheric refraction, and receiver clock bias just like the code-based pseudorange measurement. The error due to multipath and receiver noise in the carrier phase measurements, at centimeter level, is about one-hundredth of that in the pseudoranges. On the other hand, there is the extra complication due to the integer ambiguity. In this section, we examine the benefits of combining the noisy-but-unambiguous code measurements with the precise-but-ambiguous carrier phase measurements for absolute positioning in a single-receiver autonomous mode.

5.7.1 Single-Frequency Measurements

While carrier tracking is continuous, the carrier phase measurements offer precise and unambiguous measurements of change in pseudorange between two measurement epochs. In this section, we examine how to exploit the precision of these delta pseudoranges to smooth the code-based pseudoranges.

So far, we have represented the carrier phase measurement in units of cycles. When combining code and carrier measurements, it is more convenient to represent carrier phase measurements in units of length, like the code phase measurements. The measurements are

$$\rho(t) = r(t) + c[\delta t_u(t) - \delta t^i(t - \tau)] + I(t) + T(t) + \varepsilon_p(t). \quad (5.46)$$

$$\begin{aligned} \Phi(t) &= \lambda\phi(t) \\ &= r(t) + c[\delta t_u(t) - \delta t^i(t - \tau)] - I(t) + T(t) + \lambda N + \varepsilon_\phi(t), \end{aligned} \quad (5.47)$$

Both measurements are from the same satellite and a reference to the satellite number has been dropped for simplicity. We follow loosely the simple and elegant treatment of Goad (1996). Let us define ionosphere-free (IF) pseudorange [Section 5.3.2] as

$$\rho_{IF}(t) = r(t) + c[\delta t_u(t) - \delta t^r(t - \tau)] + T(t)$$

We rewrite the code and carrier phase measurement equations as

$$\begin{aligned}\rho(t) &= \rho_{IF}(t) + I(t) + \varepsilon_\rho(t) \\ \Phi(t) &= \rho_{IF}(t) - I(t) + \lambda N + \varepsilon_\Phi(t)\end{aligned}\quad (5.48)$$

We have two equations in three unknowns (ρ_{IF} , I , N), and it is not clear so far that we have gained anything aside from the obvious notational simplicity. Figure 5.15 offers a conceptual view of the two measurements, illustrating the higher code measurement noise and the ambiguity of the carrier phase measurements. In view of the ambiguity, we have set the starting value $\Phi(t_0)$ arbitrarily to zero.

Let us examine the change in the code and carrier phase measurements between two measurement epochs t_{i-1} and t_i .

$$\begin{aligned}\Delta\rho(t_i) &= \rho(t_i) - \rho(t_{i-1}) = \Delta\rho_{IF}(t_i) + \Delta I(t_i) + \Delta\varepsilon_\rho(t_i) \\ \Delta\Phi(t_i) &= \Phi(t_i) - \Phi(t_{i-1}) = \Delta\rho_{IF}(t_i) - \Delta I(t_i) + \Delta\varepsilon_\Phi(t_i)\end{aligned}\quad (5.49)$$

where $\Delta\rho_{IF}$ is the change in the ionosphere-free pseudorange between the two measurement epochs. Similarly, ΔI is the corresponding change in the ionospheric delay, and $\Delta\varepsilon$ is the change in the error term.

How to combine these measurements so as to take advantage of the low-noise carrier phase measurements and unambiguous nature of the code measurements? We have two equations in two unknowns. The noise term in the first equation is at meter level, and in the second equation at centimeter level. We could weight the two equations appropriately and solve for the two unknowns: $\Delta\rho_{IF}(t_i)$ and $\Delta I(t_i)$. An alternative is to disregard ΔI , which would be near zero if the measurement epochs are close together, and use $\Delta\Phi(t_i)$ as an accurate estimate of $\Delta\rho_{IF}(t_i)$. If we can somehow get an accurate estimate of $\rho(t_0)$, we can construct the pseudo-

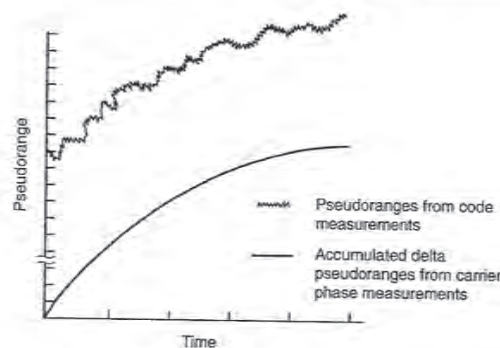


Figure 5.15 A conceptual view of the pseudoranges computed from the code and carrier phase measurements. The code-based measurements are noisy. The carrier-based estimates are precise but ambiguous, and the plot starts arbitrarily at zero value. The carrier phase measurements provide accurate estimates of delta pseudoranges, and can be used to smooth the code-based pseudoranges.

range profile using the carrier-derived delta pseudoranges (Figure 5.15).

Actually, an estimate of $\rho(t_0)$, denoted as $\hat{\rho}(t_0)$, is available from each epoch as

$$\hat{\rho}(t_0)_i = \rho(t_i) - [\Phi(t_i) - \Phi(t_0)]$$

We can simply average these estimates over n epochs to obtain a smoothed estimate

$$\bar{\rho}(t_0) = \frac{1}{n} \sum \hat{\rho}(t_0)_i$$

and reconstruct the smoothed pseudorange profile as

$$\bar{\rho}(t_i) = \bar{\rho}(t_0) + [\Phi(t_i) - \Phi(t_0)]. \quad (5.50)$$

Carrier-smoothing of the code measurements is now routine in modern receivers and offers a modest improvement. The feature of 100-second smoothing was active in the receivers used to generate the plot of the multipath error in Figure 5.14(e).

An efficient implementation of the above idea is in terms of a recursive filter of length M as

$$\begin{aligned}\bar{\rho}(t_i) &= \frac{1}{M} \rho(t_i) + \frac{(M-1)}{M} [\bar{\rho}(t_{i-1}) + (\Phi(t_i) - \Phi(t_{i-1}))], \\ \bar{\rho}(t_1) &= \rho(t_1).\end{aligned}\quad (5.51)$$

The filter weighs the carrier phase measurements more heavily than the code phase measurements. Note that in deriving (5.50), we disregarded any change in the ionosphere between measurement epochs. While such change would generally be insignificant over a few seconds, we have to be careful that the change does not add up to become significant over M epochs. Actually, we will see double the effect of this change in filter (5.51) because the change is going in one direction for the carrier phase measurement and in the opposite direction for the code phase measurements. The smoothing filter (5.51) is the reason for our earlier discussion of *code-carrier divergence* [Section 5.3.1].

We know *a priori* that such smoothing of the code phase measurements cannot compensate for the errors due to satellite clock and ephemeris, ionosphere, and troposphere. We expect the receiver noise to be smoothed, and the code multipath to be smoothed to the extent of the filter length. The effect of such smoothing applied to measurements from a stationary receiver is shown in Figure 5.16. The raw and smoothed pseudorange errors are plotted for a satellite as it rises. (These measurements were taken in 1997 from a satellite with SA inactive.)

The measured raw pseudoranges in Figure 5.16, corrected using the broadcast ionospheric model and tropospheric model UNB4, are essentially unbiased with an rms error of about 4 m. The quasi-sinusoidal oscillations are characteristic of multipath. A 100-second smoothing filter reduces the rms error to 1.7 m. The residual error retains the character of multipath, and is similar to that in Figure 5.14(e). A 15-minute smoothing filter smooths the multipath also, but a small bias becomes apparent. A 30-minute filter shows clear signs of code-carrier divergence in the middle of the time interval while the satellite is still low and rising, and the ionospheric propagation delay is changing. If the ionosphere had been more active, the divergence would have become manifest for a shorter filter length. In differential mode, however, the ionospheric propagation error is substantially eliminated, and such smoothing offers a robust technique to control measurement error, as discussed in Section 5.8.2.

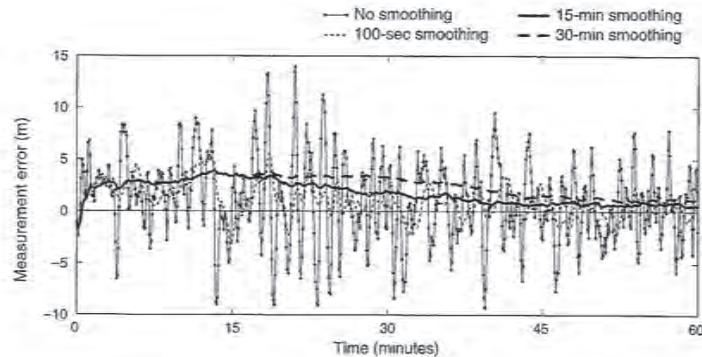


Figure 5.16 Carrier-smoothed pseudoranges with different filter lengths.

5.7.2 Dual-Frequency Measurements

Let us start by writing (5.48) for dual-frequency code and carrier phase measurements, using subscript q ($= L1$ or $L2$) to identify a measurement with $L1$ or $L2$. The ionospheric delay at $L1$ is denoted as I .

$$\begin{aligned}\rho_{L1} &= \rho_{IF} + I + \varepsilon_{\rho_{L1}} \\ \rho_{L2} &= \rho_{IF} + \left(\frac{f_{L1}^2}{f_{L2}^2} \right) I + \varepsilon_{\rho_{L2}} \\ \Phi_{L1} &= \rho_{IF} - I + \lambda_{L1} N_{L1} + \varepsilon_{\Phi_{L1}} \\ \Phi_{L2} &= \rho_{IF} - \left(\frac{f_{L1}^2}{f_{L2}^2} \right) I + \lambda_{L2} N_{L2} + \varepsilon_{\Phi_{L2}}\end{aligned}\quad (5.52)$$

Measurements from one epoch give four equations in four unknowns: $(\rho_{IF}, I, N_{L1}, N_{L2})$. If the measurements were error-free, we would be all set. Recall our model for the measurement noise: $\sigma(\varepsilon_{\rho_q}) = 1$ m, and $\sigma(\varepsilon_{\Phi_q}) \approx 1$ cm. The pseudorange measurements appear too coarse for estimation of the integers in a single snapshot of the measurements. On the bright side, N_{L1} and N_{L2} would not change from epoch to epoch as long as the carriers are tracked continuously. Each new epoch would bring four new measurements and introduce two new unknowns regardless of whether the receiver is stationary or in motion.

We can stack the measurements and process them sequentially or in batch, weighting code and carrier measurements appropriately, and expect the estimates of the integers to converge. In fact, this would be a good exercise using one of the data set on the enclosed CD. Once we have the integers, we can get precise ranges from the carrier phase measurements, and then position estimates, as discussed in Chapter 6.

Actually, the situation is trickier, but not hopeless, especially with the third civil frequency (L5) on the way. We'll pick up this discussion in Chapter 7 where we consider precise point positioning.

5.8 Error Mitigation: Differential GPS (DGPS)

The basic GPS measurements, described in Section 5.1, consist of biased and noisy estimates of ranges to the satellites. The principal source of bias is the unknown receiver clock offset relative to GPST. The remaining errors are:

- errors in modeling the satellite clocks and ephemerides,
- errors in modeling the ionospheric and tropospheric delays,
- errors in measuring the code and carrier phase due to multipath and receiver noise.

The carrier phase measurements involve an additional complication due to integer ambiguities of whole cycles. The multipath and receiver noise, however, are two orders of magnitude lower for the carrier phase than for the code measurements.

If a position is to be determined in real time using a single GPS receiver, the only option is to use code-based pseudoranges, perhaps smoothed by carrier phase. With $\sigma_{URE} \approx 6$ m, the horizontal and vertical positioning errors (95%) for an SPS user are about 10 m and 15 m, respectively. In order to obtain higher positioning accuracy, the user would have to reduce the measurement errors further. This requires a change in the mode of GPS usage from single-receiver autonomous positioning to differential GPS (DGPS), which we discuss next.

The basic idea behind DGPS is to take advantage of the fact that the errors associated with a satellite clock, ephemeris, and the atmospheric propagation are similar for users separated by tens, even hundreds of kilometers, and these errors vary 'slowly' with time. In other words, the errors exhibit spatial and temporal correlations. The closer the two users are to each other, and the closer the measurement epochs, the more similar are the errors listed above. The errors are said to get 'decorrelated' with increasing distance between the users and increasing time difference between their measurement epochs.

If the position of a GPS receiver is known, the combined effect of the errors can be estimated for each satellite. If these error estimates can be made available to the GPS users in the area, each can apply them to his measurements to mitigate the errors and improve the quality of the position estimates. That's the basic idea behind DGPS. This approach is usable with both code and carrier measurements. For navigation, such corrections will have to be made available in real time using a radio link. In practice, a user would receive and apply the corrections with some delay, called *latency*. The closer a user is to the reference station and the shorter the latency, the higher the benefit from the differential corrections.

5.8.1 Error Mitigation

Let us revisit the error sources discussed in Sections 5.2–5.4 to see how they would be mitigated in differential mode. It would be useful to review Figure 5.14 showing the nature of the errors introduced by these sources. Table 5.4 gives a summary of these errors and their mitigation in differential mode.

- *Satellite clock error.* The clock modeling error is small (2 m rms, and getting smaller), and changes slowly over hours. There is no fear of decorrelation with distance, and this error can be eliminated nearly completely in differential mode.

with correction messages every hour.

- **Satellite ephemeris error.** Another small error (2 m rms, and getting smaller) which changes slowly over minutes. Recall that damage is done only by the line-of-sight component of the ephemeris error vector (Figure 5.5). The post-differential-correction residual error, therefore, would depend upon the separation between the lines of sight from the user and reference station. The satellites are 20,000 km away, and the angular separation at the satellite of the lines of sight from two points on the earth separated by 100 km is only 0.3°. A conservative bound on the residual range error is given as [Parkinson and Enge (1996)]

$$\epsilon_R \leq \frac{d \cdot \delta r}{r}$$

where r is the range to a satellite, δr is the magnitude of the error in the ephemeris broadcast by the satellite, and d is the distance between the two receivers. For $d = 100$ km and $\delta r = 10$ m, the uncompensated range error would be less than 5 cm.

- **Ionospheric propagation delay.** Recall that the ionospheric delay depends upon the total electron content (TEC) along the signal path. Consider two receivers separated by 100 km. The corresponding ionospheric pierce points of the signals are also separated by about 100 km. The post-differential correction residual error, therefore, would depend upon the variability of TEC in the ionosphere. In fact, the ionosphere can show considerable variability both spatially and temporally resulting from solar activity and magnetic storm-induced traveling disturbances. For a satellite overhead, typical post-differential-correction residual error for two receivers separated by 100 km would be 0.1–0.2 m, but can be 1 m, or more, when the ionosphere is active.
- **Tropospheric propagation delay.** The tropospheric delay depends upon the air density profile along the signal path. Two receivers separated by several kilometers may be subject to different weather conditions. The water vapor content shows considerable variability both spatially and temporally. The post-differential-correction residual error is generally higher for the low-elevation satellites. With a 10-km separation between two receivers, the residual range error can be 0.1–0.2 m. For a longer distance or significant altitude difference, it would be preferable to correct for the tropospheric delay at the reference and user receivers separately. For a low satellite, the residual range error can be 2–7 mm per meter of altitude difference.
- **Multipath and receiver noise.** These errors are uncorrelated at the reference and user receivers, and cannot be corrected by DGPS. In fact, a user inherits the errors incurred at the reference station. Therefore, it is important to minimize these errors by careful siting and equipment selection at both the reference and user stations.

The design of the DGPS systems fielded in the 1990s was determined by SA, which introduced the largest and fastest-changing of the measurements errors. These systems typically compute new measurement corrections each five to ten seconds. In order to extend the life of a

Table 5.4 A summary of the errors in GPS measurements.
The estimates for DGPS are based on the premise that the user's distance from the reference receiver is tens of kilometers and signal latency is tens of seconds.

Source	Potential error size	Error mitigation and residual error
Satellite clock model	Clock modeling error: 2 m (rms)	DGPS: 0.0 m
Satellite ephemeris prediction	Component of the ephemeris prediction error along the line of sight: 2 m (rms)	DGPS: 0.1 m (rms)
Ionospheric delay	Effect upon the code and the carrier is equal and opposite: The code is delayed while the carrier is advanced by the same amount.	Single-frequency receiver using broadcast model: 1–5 m
	Delay in zenith direction ≈ 2 –10 m, depending upon user latitude, time of the day, and solar activity.	Dual-frequency receiver (compensates for the ionospheric delay but magnifies noise): 1 m (rms)
	Delay for a satellite at elevation angle ϵ : zenith delay $\times$ obliquity factor (ϵ)	DGPS: 0.2 m (rms)
	Obliquity factor: 1 at zenith; 1.8 at 30° elevation angle; and 3 at 5°.	
Tropospheric delay	Code and the carrier are both delayed by the same amount.	Models based on average meteorological conditions: 0.1–1 m
	Delay in zenith direction at sea level ≈ 2.3 –2.5 m; lower at higher altitudes	
	Delay for satellite at elevation angle ϵ : zenith delay $\times$ obliquity factor (ϵ)	DGPS: 0.2 m (rms) plus altitude effect
	Obliquity factor: 1 at zenith; 2 at 30° elevation angle; 4 at 15°; and 10 at 5°	
Multipath	In a 'clean' environment:	Uncorrelated between antennas.
	Code: 0.5–1 m Carrier: 0.5–1 cm	Mitigation through antenna design and siting, receiver design, and carrier-smoothing of code measurements
Receiver noise	Code: 0.25–0.5 m (rms)	Uncorrelated between receivers
	Carrier phase: 1–2 mm (rms)	Mitigation through receiver design

correction message and cut down on the data traffic, the correction messages generally transmit both the error size and its rate of change observed at the reference station at the measurement epoch. With SA gone, the correction update interval can be increased from five or ten seconds to one minute, and the correction rate can be dropped from the message.

In a common realization of local DGPS, a reference receiver is set up at a surveyed site, and the pseudorange error is computed for each satellite as the difference between the geometric range and the measured pseudorange. Such errors are computed and broadcast in the local area on a radio link as differential corrections [see Figure 2.10]. Users within a hundred kilometers can obtain one-meter-level positioning accuracy by applying these corrections to their measurements.

5.8.2 Local-Area DGPS and Relative Positioning

Let us examine pseudorange measurements at the user and reference stations. We are still dealing with measurements from a single satellite, and will use superscript s , as necessary, to identify satellite-related terms. We have to introduce additional notation to keep the measurements of the two receivers apart. Following established practice, we will use subscripts u and r to identify the user and reference receivers, respectively. With a small change in the notation of (5.6), the pseudorange measurements at the reference and user receivers are

$$\begin{aligned}\rho_u &= r_u + c(\delta t_u - \delta t^s) + I_u + T_u + \varepsilon_{pu} \\ \rho_r &= r_r + c(\delta t_r - \delta t^s) + I_r + T_r + \varepsilon_{pr}\end{aligned}\quad (5.53)$$

The geometric range to a satellite from the reference station is computed as

$$r_r = \|\mathbf{x}^s - \mathbf{x}_r\|$$

where $\mathbf{x}^s$ is the satellite position obtained from the navigation message and $\mathbf{x}_r$ is the surveyed position of the antenna at the reference station. Any error in $\mathbf{x}^s$ would generally be harmless, as we'll see below.

For simplicity, we have dropped any reference to measurement epochs in (5.53) but let us consider measurements taken at the two receivers within a minute or two of each other. The error in ρ_r , the pseudorange measurement at the reference station, is computed as

$$\begin{aligned}e_r &= r_r - \rho_r \\ &= -c(\delta t_r - \delta t^s) - I_r - T_r - \varepsilon_{pr},\end{aligned}\quad (5.54)$$

and broadcast as a differential correction. The differentially corrected pseudorange measurement of a GPS user in the area is

$$\begin{aligned}\bar{\rho}_u &= \rho_u + e_r \\ &= r_u + c(\delta t_u - \delta t^s) + (I_u - I_r) + (T_u - T_r) + \varepsilon_{pu} - \varepsilon_{pr} \\ &= r_u + c(\delta t_u - \delta t^s) + \varepsilon_{p,ur}\end{aligned}\quad (5.55)$$

The error introduced by the satellite clock bias term (δt^s) would be similar at the two receivers. If the distance between the two receivers is not 'too large,' and the corrections are not 'too old,' we can conclude that (i) the ephemeris error would affect the measurements at the two receivers similarly (see Figure 5.5), and (ii) $I_u \approx I_r$ and $T_u \approx T_r$. With two receivers separ-

ed by 25 km, the differential ionospheric delay is typically at 10–20 cm level. This difference can add up to 1 m at a distance of 100 km. A significant distance and/or difference between the altitudes of the two receivers will require that a tropospheric model be used to correct the measurements separately at each receiver. However, there is no reprieve from the errors introduced at the reference and user receivers by multipath and receiver noise, included in $\varepsilon_{p,ur}$.

In the simple treatment above, the effect of receiver clock bias at the reference station was absorbed in the corrections and transmitted to the user, resulting in a change in the apparent bias of the user receiver clock. In practice, a reference station would attempt to limit the size of the differential corrections to be transmitted by accounting for and removing its clock bias. It is important for the user and the reference receivers to coordinate how the measurements are to be processed. If the reference receiver applies the broadcast ionospheric delay model before computing the corrections, the user must do the same. Similarly, if the reference receiver uses a tropospheric delay model, so must the user. Both must also use the same ephemeris parameters. When the ephemeris parameters are updated, the reference station broadcasts corrections for a period using both the old and the new parameter sets, each identified with a unique Issue of Data (IOD) parameter. Finally, special care is required in locating a reference station to minimize multipath.

Two basic practical questions associated with DGPS are: How large an area can a reference station serve? And how often must the corrections be provided? As noted earlier, there are no clear-cut answers. Basically, the shorter the distance and the more frequent the corrections, the higher the assurance of obtaining position estimates with meter-level error. Keep in mind though that the errors due to the ionosphere and troposphere generally change slowly over minutes.

For real-time users of DGPS, the resources required are establishment of a reference station and a radio link to transmit data to the users. A standard format for data communication developed by the Radio Technical Commission for Maritime Services (RTCM) is now in wide use [RTCM SC-104 (1994)]. RTCM has defined data messages and an interface between the data link receiver and the DGPS receiver. The U.S. Coast Guard's Maritime DGPS system broadcasts differential corrections in the RTCM SC-104 format from marine radiobeacons at a modest data rate of 100–200 bps. These signals are available in coastal regions and along inland waterways of the United States for the price of a radiobeacon receiver and a differential-capable GPS receiver. Maritime DGPS is being expanded into *Nationwide DGPS* (NDGPS) to cover most of the country [Section 2.5]. Commercial services offering DGPS corrections using different radio links are available worldwide.

We have dealt so far with implementation of local DGPS in which the reference station was fixed at a surveyed location. It was a requirement for calculation of the corrections to the measurements. Actually, the concept of DGPS is useful even if the position of the reference station is not known accurately. Consider a simple case where the reference station has an unknown error in its position estimate. Clearly, the differential corrections computed at the reference station would be in error, and so would the computed position of the user based on these corrections. As we'll see in Chapter 6, the subsequent error in the user's estimated position would be approximately the same as the error in the reference station position. The relative position vector between the antennas of the two receivers would be largely unaffected. In other words, the user-computed position is relative to the position ascribed to the reference station [Hatch (1992)].

Errors in the carrier phase measurements due to satellite clock and ephemeris errors and ionospheric and tropospheric refraction can also be removed substantially in differential mode. The differentially corrected carrier phase measurements, however, differ from the similarly corrected pseudorange measurements in two important ways. On the positive side, the remaining error in the carrier phase measurements due to multipath (at centimeter level) and receiver noise (at millimeter level) is typically two orders of magnitude lower than that in the code-phase measurements. But there is a price: unlike the code phase measurements, the carrier phase measurements involve an inherent ambiguity. The basic idea of DGPS can be implemented using carrier phase measurements in the same way as with the pseudorange measurements: compute the corrections to be applied to the carrier phase measurements at the reference station, and transmit them to the users [Frodge *et al.* (1994), Lapucha *et al.* (1996)].

Relative Positioning

In an alternate implementation of DGPS, the reference station could broadcast its time-tagged measurements of pseudoranges rather than the computed differential corrections. In this case, the user receiver would form differences of its own measurements with those at the reference receiver satellite by satellite, and estimate its position relative to the reference receiver. This approach is referred to as *relative positioning*. Of course, knowledge of the reference receiver location would allow the user to determine his absolute position, if required. In some applications, for example, formation flying, the relative position vector is of main interest. Clearly, practical details of data transmission delays would have to be accounted for in generating real-time position and velocity estimates.

Differencing time-matched measurements from a satellite at the user and reference receivers results in cancellation of the same error terms as discussed earlier in this section. From (5.53), this difference can be written using a shorthand notation of a double subscript as

$$\begin{aligned}\rho_{ru} &= \rho_r - \rho_u \\ &= (r_r - r_u) + c(\delta t_r - \delta t_u) + (I_r - I_u) + (T_r - T_u) + (\varepsilon_{pr} - \varepsilon_{pu}) \\ &= r_{ru} + c(\delta t_r - \delta t_u) + \varepsilon_{\rho,ru},\end{aligned}\quad (5.56)$$

where $r_{ru} = r_r - r_u$. We have a nonlinear equation in four unknowns: three elements of the relative position vector (hidden in the r_{ru} term) and the difference in the receiver clock biases. With measurements available from four or more satellites at the two receivers, the user can solve for the relative position vector. We'll use the double-subscript notation introduced in (5.56) throughout Part II to denote the difference between measurements at two receivers and between parameters in their models: $(\bullet)_{ru} = (\bullet)_r - (\bullet)_u$.

It is instructive to compare (5.56) with (5.6). We have essentially eliminated the ionospheric and tropospheric effects and ephemeris error in our new model. Rather than estimating x_u (absolute positioning), we are now content with estimating $(x_r - x_u)$ (relative positioning), which we can accomplish with higher accuracy because we only have to contend with errors due to receiver noise and multipath. Equation (5.56) represents a re-parameterization of the problem of positioning with code phase measurements (5.6).

The errors due to receiver noise and multipath are much lower in carrier phase measurements than in code phase measurements. Of course, the problem with carrier phase measure-

ments has to do with the integer ambiguities. We can obtain much more precise estimates of relative position vector using carrier phase measurements if we can deal with the integer ambiguities. We'll develop this approach fully in Chapter 7. For now, we content ourselves with obtaining a partial benefit of the carrier phase measurements without getting tangled up in the integer ambiguity terms by smoothing the pseudorange differences (5.56) using the corresponding carrier phase differences.

The time-matched carrier phase measurements from a satellite in units of cycles measured at the user and reference receivers are

$$\begin{aligned}\phi_u &= \lambda^{-1}[r_u - I_u + T_u] + f \cdot (\delta t_u - \delta t^t) + N_u + \varepsilon_{\phi u} \\ \phi_r &= \lambda^{-1}[r_r - I_r + T_r] + f \cdot (\delta t_r - \delta t^t) + N_r + \varepsilon_{\phi r}\end{aligned}\quad (5.57)$$

The difference can be written using our double-subscript notation as

$$\begin{aligned}\phi_{ru} &= \phi_r - \phi_u \\ &\approx \lambda^{-1}r_{ru} + f \cdot (\delta t_r - \delta t_u) + N_{ru} + \varepsilon_{\phi,ru},\end{aligned}\quad (5.58)$$

where $N_{ru} = N_r - N_u$ is the integer ambiguity associated with the differenced measurements. Equation (5.58) is the counterpart of (5.56). The main differences are the presence of an ambiguity (N_{ru}) and much lower measurement error ($\varepsilon_{\phi,ru}$) due to multipath and receiver noise in (5.58).

The basic idea is similar to that of carrier-smoothed code for single-receiver measurements discussed earlier in Section 5.7.1. We can take advantage of the much higher accuracy and precision of the pseudorange differences between the user and the reference station obtained from carrier phase to smooth the corresponding differences obtained from the code measurements. The integer ambiguities drop out. The error due to multipath and receiver noise at the two receivers in the single-difference code-based pseudoranges is smoothed. The smoothing filter can be written as before (5.51)

$$\begin{aligned}\bar{\rho}_{ru}(t_i) &= \frac{1}{M} \rho_{ru}(t_i) + \frac{(M-1)}{M} (\bar{\rho}_{ru}(t_{i-1}) + \lambda(\phi_{ru}(t_i) - \phi_{ru}(t_{i-1}))), \\ \bar{\rho}_{ru}(t_1) &= \rho_{ru}(t_1),\end{aligned}\quad (5.59)$$

where $\bar{\rho}_{ru}$ is the smoothed pseudorange difference. Unlike the discussion on code smoothing in the previous subsection, code-carrier divergence poses much less of a threat to the filter above. The reason is that the ionospheric effects have been substantially eliminated in both the code and carrier single differences. The filter can be allowed to be long enough to smooth the oscillations in the pseudorange measurements due to multipath.

Figure 5.17 illustrates the effect of smoothing on L1 code measurement differences between two stations about 25 km apart. The differences of 'raw' code measurements are nearly unbiased with an rms error of 0.8 m. The 100-second smoothing reduces the rms error to 0.5 m, but the residual effect of multipath is apparent. The 15-minute smoothing filter effectively smooths out the multipath. The results for a 30-minute filter are virtually identical to those for the 15-minute filter. Each reduces the rms error to about 0.1 m, and there is no sign of code-carrier divergence.

The architecture proposed for the FAA's Local Area Augmentation System (LAAS) to be

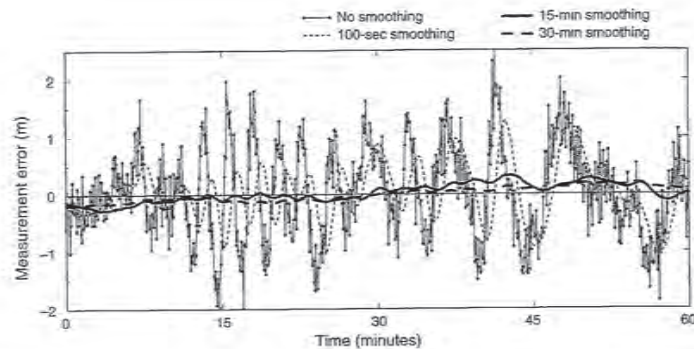


Figure 5.17 Carrier-smoothed differentially corrected pseudoranges obtained with different filter lengths.

used to guide aircraft during approach and landing operations under poor visibility is based on DGPS carrier-smoothed code measurements. The challenge lies in ensuring the accuracy of navigational guidance every time. The LAAS reference stations would be set up on airports and their measurements would be transmitted to aircraft on a VHF communications channel with a data rate of about 2 kbps. A comprehensive description of LAAS and the requirements of civil aviation is given by Enge (1999).

An idea of considerable interest in recent years has been to provide GPS-like signals from pseudo-satellites, called pseudolites, located on the ground, a ship, or aircraft. The pseudolites can be used in addition to deliver differential corrections for the GPS satellites by modulating the corrections onto the signals. Problems associated with the signal and receiver design to cope with the so-called near-far problem and the issues of multipath remain active areas of research [Elrod and Van Dierendonck (1996)].

5.8.3 Wide-Area DGPS

It would require a large number of autonomous DGPS systems to cover a continent-wide area. An alternative is wide-area DGPS (WADGPS) [Kee (1996)] implemented as a centralized system which provides differential corrections to users in a form so as to be usable over a large geographical area.

A set of reference stations is deployed over the region of interest and measurements from each are processed centrally to decompose the errors into their constituents: satellite clock, ephemeris, and ionosphere. The corrections are broadcast for each error source separately from geostationary satellites or from a network of FM stations so that each user can apply the correction terms appropriately depending upon her location. A WADGPS is said to broadcast differential correction vectors, in contrast to the scalar corrections in a local DGPS where all the errors are lumped together.

Commercial wide-area DGPS services are already operational and in wide use in offshore explorations, seismic surveys, and agriculture. A comprehensive description of the most ambitious wide-area DGPS to date, the FAA's *Wide Area Augmentation System* (WAAS), is given by

Enge *et al.* (1996). What makes WAAS special is the uncompromising requirement of safety in civil aviation. A schematic representation of WAAS is given in Figure 5.18. Dual-frequency (L1-L2) measurements from about 25 WAAS Reference Stations distributed over the conterminous United States are processed at the Master Station to estimate differential corrections and error bounds on them. The differential corrections are decomposed into three components: a fast-changing component due to the clock error and two slow-changing components due to the ephemeris error and ionospheric propagation delays for a set of points corresponding to a latitude/longitude grid.

The differential corrections are coded as a 250-bps navigation message of GPS/SPS-like signals transmitted at L1 frequency from geostationary satellites. WAAS actually generates such signals on the ground and transmits them to geostationary satellites, which essentially serve as 'bent pipes' transmitting the signals back to the earth over their area of coverage. A GPS receiver would require software modification in order to receive this additional ranging signal and to demodulate the navigation message for the differential corrections.

WAAS was commissioned in 2003. It offers positioning accuracy of 1.5–2 m in both horizontal and vertical dimensions, adequate for lateral and vertical (glide path) guidance for instrument approaches. The system also provides indications to the users where GPS or GPS/WAAS is unusable due to system errors or other effects.

WAAS-like GPS augmentations are also being developed by the European countries and Japan for deployment over regions of interest to them. These systems are called the *European Geostationary Navigation Overlay System* (EGNOS) and the *Multifunction Transportation Satellite (MTSAT)-based Satellite Augmentation System* (MSAS), respectively. Both are expected to be ready for operational use starting in 2006. The generic name for such augmentations is Satellite-Based Augmentation System (SBAS). International agreements have been negotiated so that WAAS, EGNOS, and MSAS together will provide a seamless coverage of the globe.

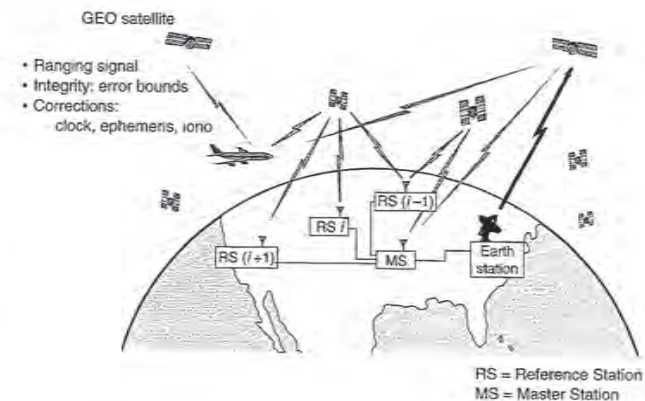


Figure 5.18 Wide Area Augmentation System (WAAS) schematic.

Another noteworthy wide-area DGPS system is NASA's Global Differential GPS (GDGPS) developed by the Jet Propulsion Laboratory (JPL) and capable of providing decimeter-level positioning accuracy to dual-frequency receiver users worldwide. The dual-frequency receivers, commonly used in low-earth orbiter missions, science applications, and high-accuracy commercial activities, have the advantage of eliminating the ionosphere as a source of error. GDGPS, therefore, focuses on providing accurate, real-time estimates of GPS ephemeris and clock parameters. Such estimates can be obtained from a relatively sparse network of global reference stations. Another innovation of GDGPS is to transmit the corrections over the Internet [http://www.gdgps.net]. (While GDGPS may offer high accuracy for gravity mapping over Greenland, it wouldn't land any airplanes in fog-bound BOS or SFO because, unlike WAAS, it wasn't designed to meet the civil aviation's requirements for integrity of each position estimates and continuity of service.)

5.9 Summary

After all kinds of preliminaries and generalities of the earlier chapters, we finally got down to brass tacks. In this chapter, we examined the GPS measurements and errors therein. Our ability to get the most out of GPS depends upon how well we understand what's being measured and how the various sources of error corrupt these measurements.

The measurement of code phases gives pseudoranges to the satellites, i.e., ranges with a common bias. Measurement of carrier phases gives pseudoranges with an additional complication of ambiguity of whole cycles. The bias and ambiguity aside, both measurements suffer from a common set of errors. The main sources of these errors are:

- Errors in satellite position and clock bias. Steady improvements in clock technology and orbital prediction algorithms, and a tight control of the constellation by the Master Control Station, have reduced these errors to 2–3 m level, an astonishing achievement.
- Errors in modeling the signal propagation delay through the ionosphere and troposphere. A signal travels through the earth's atmosphere with slightly varying speeds depending upon the composition of the medium. A dual-frequency receiver has the advantage of accounting almost fully for propagation delay through the ionosphere. When it comes to the troposphere, a dual-frequency receiver offers no advantage. Models can, however, be used to correct for most of the tropospheric delay.
- Errors due to receiver noise and multipath (i.e., antenna receiving the direct signal and one or more reflections).

The errors in satellite position and clock bias, and propagation delays are highly correlated both spatially and temporally. In other words, these errors would be substantially similar for two users located tens of kilometers apart and taking measurements within tens of seconds of each other. This observation offers a basis for augmenting the GPS signals with differential corrections to mitigate the error sources. Many differential GPS services are now in wide use worldwide offering meter-level positioning.

Homework Problems

- 5-1. Determine from Figure 5.2 if the receiver clock is running faster or slower relative to GPST, and estimate its frequency offset from the nominal value of 10 MHz.
- 5-2. Derive the relationship (5.19) between the phase and group refractive indices.
- 5-3. The problem of code-carrier divergence [Section 5.7.1] can be largely overcome with dual-frequency measurements as follows. From pseudorange measurements ρ_{L1} and ρ_{L2} , and carrier phase measurements ϕ_{L1} and ϕ_{L2} , 'create' two new measurements

$$\rho^* = \frac{f_{L1}}{f_{L1} + f_{L2}} \rho_{L1} + \frac{f_{L2}}{f_{L1} + f_{L2}} \rho_{L2}$$

$$\Phi^* = (\phi_{L1} - \phi_{L2}) \frac{c}{f_{L1} - f_{L2}}$$

- (a) Show that the wavelength of the signal $(\phi_{L1} - \phi_{L2})$ is $c/(f_{L1} - f_{L2})$, where c is the speed of light.
 - (b) Show that the ionospheric error in the two derived measurements is identical and, therefore, you can smooth the pseudorange measurements ρ^* by the carrier phase measurement Φ^* to your heart's content (to first order!).
- 5-4. Determine the ionospheric zenith delay from dual-frequency measurements over a 24-hour period, and observe the variability as a function of local time. Use the data set: (CD) *Data\Pigeon_Point\September_18_2000\Parse\pp091800.no*. The solar cycle was close to its peak on September 18, 2000, but there was no unusual solar activity on this day.
 - (a) Generate a scatter plot of ionospheric zenith delay versus local time at 15-minute intervals as follows. (i) Compute the slant delay from the dual frequency code measurements (5.30) from each satellite; (ii) determine the corresponding satellite position; (iii) use the Pigeon Point position file to determine the satellite elevation angle and the ionospheric obliquity factor; (iv) compute zenith delay.
 - (b) Compare the results from (a) with those obtained from the broadcast model. The broadcast model is described in the IS (CD) *Documents\IS-GPS200D.pdf*, pp. 125–128, and you'll find the needed coefficients in the header of file (CD) *Data\Pigeon_Point\September_18_2000\pp091800.nav*.
 - 5-5. Repeat Problem 5-4, with data from July 15, 2000, a day of intense solar storm: (CD) *Data\Pigeon_Point\July_15_2000\Parse*. Comment on the differences in the plots for the two days. (Caution: Watch for GPS week rollover.)

textbooks on probability theory [Brown and Hwang (1997), Helstrom (1991), and Stark and Woods (1994)]. The student is also expected to be comfortable with vector and matrix notation, matrix operations, and basic concepts of linear algebra. We recommend a review of the introductory sections in Strang and Borre (1997) dealing with linear algebra and least-squares solution of a system of linear equations.

In many navigation applications, it would be advantageous to base the current estimates of PVT on a time series of measurements, perhaps in combination with a model of the user dynamics. A Kalman filter implementation [Brown and Hwang (1997)] is commonly used to integrate measurements from GPS and an inertial measurement unit (IMU). Such a filter would also take into account any processing delays associated with the measurements to provide the current position and velocity estimates. In this chapter, we focus instead on estimates obtained from a snapshot of instantaneous GPS measurements. This approach simplifies the discussion of the principles of satellite navigation and allows for a cleaner characterization of the role of satellite geometry and measurement errors on the quality of the estimates.

The three main sections of this chapter deal with estimation of position, velocity, and time, in that order. The appendix presents some basic concepts from estimation theory relevant to this chapter.

6.1 Position Estimation with Pseudoranges

Let us start with the basic pseudorange measurement equation (5.6), rewritten below. We dealt in Chapter 5 with measurements from a single satellite and now need additional notation to distinguish among the measurements from K satellites in view. Following established practice, a superscript is used to identify the satellite. Parentheses are used to distinguish a superscript from an exponent. The pseudorange measurement from the k th satellite at epoch t (GPS time) can be modeled as

$$\rho^{(k)}(t) = r^{(k)}(t, t - \tau) + c[\delta t_u(t) - \delta t^{(k)}(t - \tau)] + I^{(k)}(t) + T^{(k)}(t) + \epsilon_\rho^{(k)}(t) \quad (6.1)$$

where $k = 1, 2, \dots, K$. As discussed in Section 5.1, $r^{(k)}(t, t - \tau)$ is the actual distance between the receiver antenna at signal reception time t and the satellite antenna at signal transmission time $(t - \tau)$; $\delta t_u(t)$ and $\delta t^{(k)}(t - \tau)$ are the receiver and satellite clock offsets, respectively, relative to GPST; $I^{(k)}(t)$ and $T^{(k)}(t)$ are the ionospheric and tropospheric propagation delays, respectively; $\epsilon_\rho^{(k)}(t)$ accounts for modeling errors (e.g., satellite clock modeling error and orbit prediction error) and unmodeled effects (e.g., receiver noise and multipath). Recall that we use ϵ to denote measurement errors, employing a subscript, superscript, overbar, or tilde to distinguish among the different scenarios we encounter. For simplicity, we now drop explicit reference to time in (6.1).

A user would correct each measured pseudorange for the known errors using parameter values in the navigation message from the satellite. The main corrections available to a civil user are: (i) satellite clock offset relative to GPST, (ii) relativity effect, and (iii) ionospheric delay using the parameter values for the Klobuchar model. Step-by-step implementation of these corrections is given in the GPS Interface Specification [IS-GPS-200D (2004)]. A resourceful user can go further in correcting the measured pseudoranges as follows.

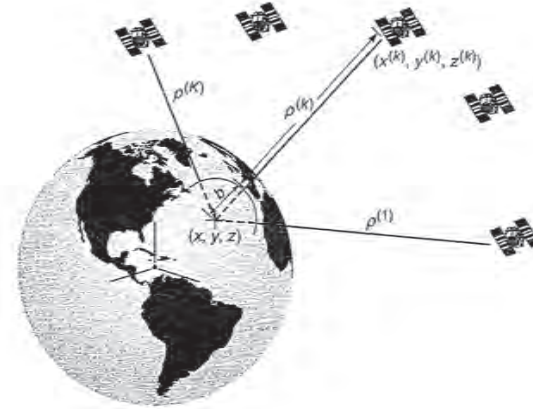


Figure 6.1 The ranges to GPS satellites measured by a receiver have a common bias and are called pseudoranges.

- The code measurements may be smoothed using the carrier phase measurements to reduce the effects of multipath [Section 5.7.1].
- The ionospheric delay can be eliminated as a source of error by a user equipped with an L1-L2 receiver [Section 5.3.2].
- The tropospheric delays can be accounted for in the software using a model [Section 5.3.3].
- The satellite position and clock errors and signal propagation delays may be mitigated if the user has access to corrections from a local- or wide-area DGPS service [Section 5.8].

We denote by $\rho_c^{(k)}$ the pseudorange obtained after accounting for the satellite clock offset and compensating for the remaining errors in the measurements to the extent it is practical for a user. We rewrite (6.1) for the ‘corrected’ pseudorange measurements simply as

$$\rho_c^{(k)} = r^{(k)} + c \cdot \delta t_u + \tilde{\epsilon}_\rho^{(k)} \quad (6.2)$$

where the error term $\tilde{\epsilon}_\rho^{(k)}$ denotes the combined effect of the residual errors. The standard deviation of $\tilde{\epsilon}_\rho^{(k)}$ can range typically from less than 1 m for differentially corrected measurements to about 6 m for measurements from a receiver in autonomous mode.

Let vectors $\mathbf{x} = (x, y, z)$ and $\mathbf{x}^{(k)} = (x^{(k)}, y^{(k)}, z^{(k)})$, for $k = 1, 2, \dots, K$, represent the position of the user at the time of the measurement and the position of the k th satellite at the time of signal transmission. The user-to-satellite geometric range (Figure 6.1) is

$$r^{(k)} = \sqrt{(x^{(k)} - x)^2 + (y^{(k)} - y)^2 + (z^{(k)} - z)^2} = \|\mathbf{x}^{(k)} - \mathbf{x}\|$$

We can write (6.2) as

$$\rho_c^{(k)} = \|\mathbf{x}^{(k)} - \mathbf{x}\| + b + \tilde{\epsilon}_\rho^{(k)} \quad (6.3)$$

where we have replaced the user clock bias term $c \cdot \delta t_u$ with a simpler b with units of meters. We use boldface, lower-case characters to denote vectors and boldface upper case to denote matrices; $\|\bullet\|$ denotes the magnitude of a vector.

Now, a subtle point. In computation of geometric range in (6.3), positions of the satellite and the user both must be expressed in the same coordinate frame, which could be either inertial or the revolving earth-centered, earth-fixed (ECEF) coordinate frame [Section 4.1]. The user position is best expressed in the ECEF frame of epoch t , the instant of signal reception and, therefore, we will express the satellite position in the ECEF frame of epoch t as well. The ephemeris parameters in the navigation message give us $\tilde{\mathbf{x}}^{(k)}$, the satellite position at $(t - \tau)$, the time of signal transmission, expressed in the ECEF frame of $(t - \tau)$. During the time it takes for the signal to propagate from the satellite to the user antenna, the earth and the ECEF coordinate frame would have rotated about the z -axis by $\omega_E \tau$, where ω_E is the rotation rate of the earth. So, we have to express the satellite position at $(t - \tau)$ in the ECEF frame of time t , and we have already discussed how this can be accomplished through a rotation matrix [Appendix 4.A].

$$\mathbf{x}^{(k)} = \begin{bmatrix} \cos \omega_E \tau & \sin \omega_E \tau & 0 \\ -\sin \omega_E \tau & \cos \omega_E \tau & 0 \\ 0 & 0 & 1 \end{bmatrix} \tilde{\mathbf{x}}^{(k)} \quad (6.4)$$

In 70–90 ms of signal transit time, the earth's rotation amounts to about 7×10^{-6} radians, and we can simplify above by using approximations: $\cos \omega_E \tau \approx 1$ and $\sin \omega_E \tau \approx \omega_E \tau$. The rotation angle may appear small but disregarding it can introduce an error of 10–20 m in the range computation in (6.3), and a corresponding east-west error in position estimate.

We are now ready to discuss position estimation.

6.1.1 Linear Model for Position Estimation

We have pseudorange measurements from K satellites, each modeled as a nonlinear equation (6.3). Each equation involves four unknowns: b and three components of $\mathbf{x}$. Clearly, four equations are required at a minimum ($K \geq 4$) to solve for the four unknowns. In other words, concurrent pseudorange measurements are required from at least four satellites in order to estimate the user's instantaneous position. The receiver clock bias is considered a nuisance parameter in the sense that in general a navigator has no need for its estimate. In fact, as noted in Section 5.1.1, the estimate is used within the receiver to schedule measurements. Once the receiver clock bias is known, the receiver can synchronize to GPST or UTC, and can generate, say, a one-pulse-per-second (1PPS) signal to control other equipment.

A simple approach to solving the K equations (6.3) is to linearize them about an approximate user position, and solve iteratively. The idea is to start with rough estimates of the user position and clock bias, and refine them in stages so that the estimates fit the measurements better. The approach described below is generally referred to as *Newton-Raphson method*. Let $\mathbf{x}_0 = (x_0, y_0, z_0)$ and b_0 be the first guesses of the user position and receiver clock bias, respectively. From (6.3), the corrected pseudorange measurement from satellite k is $\rho_c^{(k)}$. Let $\rho_0^{(k)}$ be

the corresponding approximation based on the initial guesses $\mathbf{x}_0$ and b_0 .

$$\rho_0^{(k)} = \|\mathbf{x}^{(k)} - \mathbf{x}_0\| + b_0 \quad (6.5)$$

Let the true position and true clock bias be represented as $\mathbf{x} = \mathbf{x}_0 + \delta \mathbf{x}$ and $b = b_0 + \delta b$, where $\delta \mathbf{x}$ and δb are the unknown corrections to be applied to our initial estimates. We now develop a system of linear equations in which $\delta \mathbf{x}$ and δb are the unknowns to be determined.

$$\begin{aligned} \delta \rho^{(k)} &= \rho_c^{(k)} - \rho_0^{(k)} \\ &= \|\mathbf{x}^{(k)} - \mathbf{x}_0 - \delta \mathbf{x}\| - \|\mathbf{x}^{(k)} - \mathbf{x}_0\| + (b - b_0) + \tilde{\epsilon}_\rho^{(k)} \\ &= -\frac{(\mathbf{x}^{(k)} - \mathbf{x}_0)}{\|\mathbf{x}^{(k)} - \mathbf{x}_0\|} \cdot \delta \mathbf{x} + \delta b + \tilde{\epsilon}_\rho^{(k)} \\ &= -\mathbf{1}^{(k)} \cdot \delta \mathbf{x} + \delta b + \tilde{\epsilon}_\rho^{(k)} \end{aligned} \quad (6.6)$$

where we have used a Taylor series approximation of a vector norm. In (6.6), $\mathbf{1}^{(k)}$ is the estimated line-of-sight unit vector directed from the initial estimate of the user position to satellite k , and $\mathbf{a} \cdot \mathbf{b}$ represents the dot product of vectors $\mathbf{a}$ and $\mathbf{b}$. The elements of $\mathbf{1}^{(k)}$ are direction cosines of the vector drawn from the estimated receiver location to the satellite.

$$\mathbf{1}^{(k)} = \frac{1}{\|\mathbf{x}^{(k)} - \mathbf{x}_0\|} (x^{(k)} - x_0, y^{(k)} - y_0, z^{(k)} - z_0)^T \quad (6.7)$$

A student may find it easier to derive (6.6) using long-hand notation as

$$\begin{aligned} &\sqrt{(x^{(k)} - x_0 - \delta x)^2 + (y^{(k)} - y_0 - \delta y)^2 + (z^{(k)} - z_0 - \delta z)^2} \\ &\approx \sqrt{(x^{(k)} - x_0)^2 + (y^{(k)} - y_0)^2 + (z^{(k)} - z_0)^2} \\ &\quad - \frac{(x^{(k)} - x_0)\delta x + (y^{(k)} - y_0)\delta y + (z^{(k)} - z_0)\delta z}{\sqrt{(x^{(k)} - x_0)^2 + (y^{(k)} - y_0)^2 + (z^{(k)} - z_0)^2}} \end{aligned}$$

which is written more compactly as

$$\|\mathbf{x}^{(k)} - \mathbf{x}_0 - \delta \mathbf{x}\| \approx \|\mathbf{x}^{(k)} - \mathbf{x}_0\| - \frac{(\mathbf{x}^{(k)} - \mathbf{x}_0)}{\|\mathbf{x}^{(k)} - \mathbf{x}_0\|} \cdot \delta \mathbf{x}$$

The set of K linear equations (6.6) can be written in matrix notation as

$$\delta \mathbf{p} = \begin{bmatrix} \delta \rho^{(1)} \\ \delta \rho^{(2)} \\ \vdots \\ \delta \rho^{(K)} \end{bmatrix} = \begin{bmatrix} (-\mathbf{1}^{(1)})^T & 1 \\ (-\mathbf{1}^{(2)})^T & 1 \\ \vdots & \vdots \\ (-\mathbf{1}^{(K)})^T & 1 \end{bmatrix} \begin{bmatrix} \delta \mathbf{x} \\ \delta b \end{bmatrix} + \tilde{\epsilon}_\rho \quad (6.8)$$

We now have $K (\geq 4)$ linear equations in four unknowns: $\delta \mathbf{x}$ and δb . We can write (6.8) more compactly as

$$\delta \mathbf{p} = \mathbf{G} \begin{bmatrix} \delta \mathbf{x} \\ \delta b \end{bmatrix} + \tilde{\mathbf{e}}_p \quad (6.9)$$

where

$$\mathbf{G} = \begin{bmatrix} (-\mathbf{1}^{(1)})^T & 1 \\ (-\mathbf{1}^{(2)})^T & 1 \\ \vdots & \vdots \\ (-\mathbf{1}^{(K)})^T & 1 \end{bmatrix} \quad (6.10)$$

is a $(K \times 4)$ matrix characterizing the user-satellite geometry. We will refer to $\mathbf{G}$ as the *geometry matrix*.

If $K = 4$, we can generally solve the four equations for four unknowns directly.

$$\begin{bmatrix} \delta \mathbf{x} \\ \delta b \end{bmatrix} = \mathbf{G}^{-1} \delta \mathbf{p}$$

A problem would arise if the equations are linearly dependent, making $\mathbf{G}$ rank-deficient. This could happen, for example, if the elevation angles of all satellites measured from the user position are the same. In this case the tips of the unit line-of-sight vectors would all lie in a plane, and $\text{rank}(\mathbf{G}) = 3$. Similar problem would arise if the line-of-sight vectors were all contained in a plane. In this case, the satellites would all seem to line up in the sky view plot (Figure 4.16). With the GPS constellation of 24 or more satellites, such situations are rare, and change quickly.

If the sky is not obstructed, most likely $K > 4$, we have an over-determined system of equations. $\mathbf{G}$ would be full-rank, in general. We can now look for a solution which fits the measurements best in some sense [Appendix 6.A]. It is common to use the criterion of least-squares: The best solution $(\delta \hat{\mathbf{x}}, \delta \hat{b})$ is that which minimizes the sum of squared residuals (i.e., discrepancy in the equations when we plug in the estimated values for the unknowns):

$$\min \left\| \delta \mathbf{p} - \mathbf{G} \begin{bmatrix} \delta \hat{\mathbf{x}} \\ \delta \hat{b} \end{bmatrix} \right\|^2$$

We can use tools from calculus to find the stationary point of the above expression. Alternatively, we can use concepts from linear algebra (orthogonality principle [Strang and Borre (1997)]) to find the least-squares solution for the corrections to our initial estimates from (6.9) as

$$\begin{bmatrix} \delta \hat{\mathbf{x}} \\ \delta \hat{b} \end{bmatrix} = (\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T \delta \mathbf{p} \quad (6.11)$$

The new, improved estimates of the user position and clock bias are

$$\begin{aligned} \hat{\mathbf{x}} &= \mathbf{x}_0 + \delta \hat{\mathbf{x}} \\ \hat{b} &= b_0 + \delta \hat{b} \end{aligned} \quad (6.12)$$

The observation equations may now be linearized about these new estimates of the user position and clock bias, and the solution may be iterated until the change in the estimates is

sufficiently small. Each iteration leads to new estimates of (i) GPS Time t at the time of signal reception, (ii) for each satellite, signal transit time τ , transmission time $(t - \tau)$, the corresponding satellite position, and the transformation (6.4), and (iii) geometry matrix $\mathbf{G}$. The estimates converge quickly, generally in two to four iterations, even if starting with initial estimates $\mathbf{x}_0 = \mathbf{0}$ and $b_0 = 0$.

What can we say about the quality of our least-squares estimates? The answer is: not much. A 'large' value of the sum of squared residuals would suggest that the estimates don't fit the measurements or the errors are large. 'Small' residuals might make us feel smug, but it's best to be skeptical. We expect the quality of the position estimates to depend upon the quality of the measurements, and we haven't said anything so far about the nature or the size of the errors in our measurements.

There was an implicit assumption in our least-squares solution above that all measurements are of equal quality. In practice, this assumption is almost never true. Measurements from low-elevation satellites would generally have larger errors than those from high-elevation satellites. There is no difficulty in accounting for the unequal quality of the measurements: We could simply weight the different measurement residuals appropriately, and weighting based on satellite elevation angle is often used [McGraw *et al.* (2000)]. It is left as an exercise [Homework Problem 6-1] to show that if we use weights $\{w_1, w_2, \dots, w_k\}$ on the measurements, the weighted-least-squares solution of (6.11) is

$$\begin{bmatrix} \delta \hat{\mathbf{x}} \\ \delta \hat{b} \end{bmatrix} = (\mathbf{G}^T \mathbf{W} \mathbf{G})^{-1} \mathbf{G}^T \mathbf{W} \delta \mathbf{p} \quad (6.13)$$

where $\mathbf{W}$ is a diagonal matrix with diagonal elements $\{w_1, w_2, \dots, w_k\}$.

Actually, the measured pseudoranges are not only of unequal quality, they are also correlated [Tiberius *et al.* (1999)]. If the ionospheric and tropospheric delay models are used, the error in the zenith delay enters in all measurements in a proportional way, introducing correlations. The differentially corrected measurements would also be correlated. Roughly speaking, the stronger the (positive) correlations among the pseudorange errors, the lower the position error. In the limit, an error common to all measurements shows up in the clock bias estimate and is harmless to position estimation. This should be clear from (6.8). [We saw in Figure 5.14 how the measurement errors from a satellite were correlated over time (temporal correlation). Here we are talking about measurements at an instant from different satellites. For clarity, we should refer to correlations among such measurement errors as spatial correlation.]

Ideally, we would like a characterization of the errors in the pseudorange measurements in terms of a mean vector and a covariance matrix. With such structure, we can look for a minimum mean-square error estimate of the position and determine the variance of the position error. Such characterization of the measurement error, however, is very difficult, in general. For simplicity, we treat the errors in the measurements from different satellites as zero-mean, uncorrelated, and identically distributed. (In other words, we resort to estimation techniques which would be optimal if the measurements indeed conformed to this model.)

Our model for the measurement error in (6.2) is

$$\begin{aligned} E(\tilde{\mathbf{e}}_p) &= \mathbf{0} \\ \text{Cov}(\tilde{\mathbf{e}}_p) &= E(\tilde{\mathbf{e}}_p \tilde{\mathbf{e}}_p^T) = \sigma_{URE}^2 \mathbf{I} \end{aligned} \quad (6.14)$$

where $E(\bullet)$ denotes the mean or expected value, $\text{Cov}(\bullet)$ denotes covariance, $\mathbf{I}$ is the identity matrix, and σ_{URE} is the common standard deviation of the user range error (URE) for each of the satellites, which we will often denote in the sequel simply as σ . Strictly speaking, this assumption is unjustified in most cases. (Actually, the model (6.14) fitted the pseudorange measurements while SA was active. The SA-induced errors were generated at each satellite independently and were uncorrelated across satellites. And, with $\sigma_{SA} \approx 25$ m, the SA error was also much larger than the other errors and dominated them. While deactivation of SA has made the users happy, there is a downside for the analysts.)

Model (6.14) gives useful results, in general, as we'll see below [Section 6.1.5], but we have to be careful not to push the model too far. It's important to understand that the problem is not that we don't know how to estimate position 'optimally' from pseudorange measurements with correlated errors. The problem is that an accurate characterization of the correlations is very difficult and, rather than impose an uncertain, complicated structure on the problem, we prefer to simplify the structure and use the results with appropriate caution.

6.1.2 RMS Positioning Error

Our approach to derivation of a position estimate in the previous section was algebraic, formal, and entirely devoid of insight. We attempt to make up for it in this section. Let's return to the question: How good is a GPS-based position estimate? We expect that the larger the errors in the measurements, the worse the position estimate. The set of equations (6.8) is linear in the errors, and we can say that if the measurement errors double, so would the position error. We also expect that our position estimate would improve as the number of satellites in view increases, but the exact relationship appears unclear. Given the measurement errors, how much would the position estimate improve if the number of pseudorange measurements available went up from four to eight? Actually, the number of measurements does not tell us enough because a measurement brings additional information only if the corresponding equation is linearly independent of the previous equations. This leads us to conclude that the distribution of the satellites in the sky relative to a user is important, but how? We examine this question next.

Satellite Geometry

In order to gain some insight into the role of the user-satellite geometry, let us examine a simple two-dimensional example. A user measures his distance from a pair of stations S1 and S2 at known locations (Figure 6.2). If the range measurements were perfect, the user could determine his position exactly as lying at the intersection of two circles centered at S1 and S2, with the measured ranges as their radii. The measurements, however, are imperfect, and have an uncertainty of up to, say, $\pm\epsilon$ in each case. Figure 6.2 illustrates how this measurement uncertainty translates into position uncertainty in the three cases with different geometries.

In Figure 6.2(a), the two stations are separated by an acute angle as seen from the user position. In Figure 6.2(b), this angle is 90° . In Figure 6.2(c), the angle is obtuse. While the quality of the range measurements is the same in the three cases, clearly the quality of the position estimates is not. The region of uncertainty, shown as shaded area, is smallest in case (b), and considerably larger in cases (a) and (c). The area of uncertainty is largest for case (c) because it combines the position uncertainty with position ambiguity (the two circles intersect at two points which are so close that an observer may not be able to rule one out). By the end

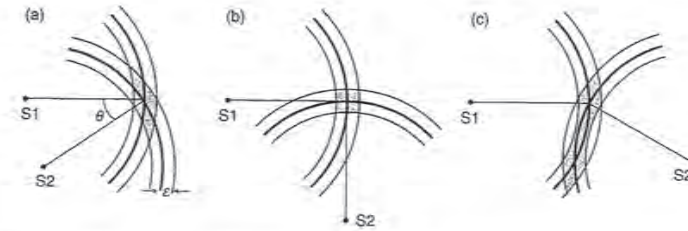


Figure 6.2 A simple 2-D example of position estimation with range measurements. The quality of the position estimate depends upon both the quality (ϵ) and geometry (θ) of the range measurements. The shaded region represents the uncertainty in the position estimate.

of this section, we will be ready to characterize the role of the measurement geometry in this problem.

Position estimation with GPS is analogous to the previous example.

We characterize the quality of the estimates in terms of the user-satellite geometry matrix $\mathbf{G}$. In our model, the pseudorange measurement errors are zero-mean and, therefore, errors in the estimates of the position coordinates and clock bias are zero-mean as well. Such estimates are said to be unbiased. We can now write down the mean and covariance of the errors in the position and clock bias estimates. Defining $\Delta\mathbf{x} = \hat{\mathbf{x}} - \mathbf{x}$ and $\Delta b = \hat{b} - b$, where $\mathbf{x}$ represents the position coordinates and $\hat{\mathbf{x}}$ its estimate from (6.12); and b is the clock bias and $\hat{b}$ its estimate,

$$\begin{aligned} E(\Delta\mathbf{x}) &= E(\hat{\mathbf{x}} - \mathbf{x}) = 0 \\ E(\Delta b) &= E(\hat{b} - b) = 0 \end{aligned} \quad (6.15)$$

It is shown in Appendix 6.A that

$$\text{Cov} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta b \end{bmatrix} = \text{Cov} \begin{bmatrix} \hat{\mathbf{x}} \\ \hat{b} \end{bmatrix} = \sigma^2 (\mathbf{G}^T \mathbf{G})^{-1} \quad (6.16)$$

We can now analyze these estimates $(\hat{\mathbf{x}}, \hat{b})$ component by component. Let σ_x^2 , σ_y^2 , and σ_z^2 denote the variances of the x -, y -, and z -component of the position error, respectively. Similarly, let σ_b^2 be the variance of the clock bias estimate. For simplicity, let

$$\mathbf{H} = (\mathbf{G}^T \mathbf{G})^{-1} \quad (6.17)$$

Denoting the i th entry on the diagonal of matrix $\mathbf{H}$ as H_{ii} , it follows from (6.16) that

$$\sigma_x^2 = \sigma^2 H_{11}, \quad \sigma_y^2 = \sigma^2 H_{22}, \quad \sigma_z^2 = \sigma^2 H_{33}, \quad \sigma_b^2 = \sigma^2 H_{44} \quad (6.18)$$

and

$$\begin{aligned} \text{RMS position error} &= \sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2} \\ &= \sigma \sqrt{H_{11} + H_{22} + H_{33}} \end{aligned} \quad (6.19)$$

Dilution of Precision (DOP)

Equations (6.18) and (6.19) show that the different components of the position estimates depend upon two factors: (i) variance of the user range error (σ^2), and (ii) a term which depends entirely on the user-satellite geometry (i.e., elements of matrix $\mathbf{H}$ (6.17) which depend entirely on matrix $\mathbf{G}$). The DOP parameters are defined on the basis of the above equations to characterize the contribution of the user-satellite geometry.

$$\text{Position dilution of precision (PDOP)} = \sqrt{H_{11} + H_{22} + H_{33}} \quad (6.20a)$$

$$\text{Time dilution of precision (TDOP)} = \sqrt{H_{44}} \quad (6.20b)$$

$$\text{Geometric dilution of precision (GDOP)} = \sqrt{H_{11} + H_{22} + H_{33} + H_{44}} \quad (6.20c)$$

The quality of the estimates obtained from a single snapshot of the measurements can be described simply (in units of length) as

$$\text{RMS (3-D position estimation error)} = \sigma \cdot \text{PDOP} \quad (6.21a)$$

$$\text{RMS (Clock bias estimation error)} = \sigma \cdot \text{TDOP} \quad (6.21b)$$

$$\text{RMS (3-D position and clock bias estimation error)} = \sigma \cdot \text{GDOP} \quad (6.21c)$$

We have now estimated the user position in the ECEF Cartesian coordinate frame and have expressions for the rms errors in its x -, y -, and z -components. The ECEF coordinates, however, are not easy for a user to relate to, and the user position is generally transformed to geodetic coordinates: latitude, longitude, and height. When it comes to assessment of a position error, it is generally more meaningful to a user to think in terms of horizontal and vertical components of the error defined relative to the local east-north-up (ENU) coordinate frame. As described in Appendix 4.A, a position vector can be transformed from ECEF to ENU with a (3×3) orthonormal matrix $\mathbf{R}_L$. The position error vector $\Delta \mathbf{x}$ defined in the ECEF coordinate frame can be represented in the ENU coordinate frame as

$$\Delta \mathbf{x}_L = \mathbf{R}_L \Delta \mathbf{x}$$

where $\Delta \mathbf{x}_L = (\Delta x_L, \Delta y_L, \Delta z_L)^T$. We use subscript L (for local) instead of ENU to keep the notation simple. We also define

$$\begin{bmatrix} \Delta \mathbf{x}_L \\ \Delta b \end{bmatrix} = \begin{bmatrix} \mathbf{R}_L & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} \Delta \mathbf{x} \\ \Delta b \end{bmatrix} = \tilde{\mathbf{R}}_L \begin{bmatrix} \Delta \mathbf{x} \\ \Delta b \end{bmatrix} \quad (6.22)$$

The covariance matrix of $(\Delta \mathbf{x}_L, \Delta b)^T$ can now be obtained from that of $(\Delta \mathbf{x}, \Delta b)^T$ from (6.16).

$$\begin{aligned} \text{Cov} \begin{bmatrix} \Delta \mathbf{x}_L \\ \Delta b \end{bmatrix} &= \tilde{\mathbf{R}}_L \text{Cov} \begin{bmatrix} \Delta \mathbf{x} \\ \Delta b \end{bmatrix} \tilde{\mathbf{R}}_L^T \\ &= \sigma^2 \tilde{\mathbf{R}}_L (\mathbf{G}^T \mathbf{G})^{-1} \tilde{\mathbf{R}}_L^T \\ &= \sigma^2 (\tilde{\mathbf{R}}_L \mathbf{G}^T \mathbf{G} \tilde{\mathbf{R}}_L^T)^{-1} \\ &= \sigma^2 [\tilde{\mathbf{G}}^T \tilde{\mathbf{G}}]^{-1} \end{aligned} \quad (6.23)$$

where $\tilde{\mathbf{G}} = \mathbf{G} \tilde{\mathbf{R}}_L^T$ has a structure similar to $\mathbf{G}$. It is a $(K \times 4)$ matrix with each row made up of

three elements of direction cosine vectors represented in the local ENU coordinate frame, and an entry of 1 in the last column. It is left as an exercise to show that the unit vector of direction cosines in the ENU coordinate frame can be written as

$$\mathbf{1}^{(k)} = \begin{pmatrix} \cos e_l^{(k)} \sin a_z^{(k)} & \cos e_l^{(k)} \cos a_z^{(k)} & \sin e_l^{(k)} \end{pmatrix}^T \quad (6.24)$$

where $a_z^{(k)}$ and $e_l^{(k)}$ are the azimuth and elevation angle of satellite k as observed from the user position [Appendix 4.A, Figure 4.A.4]. For simplicity of notation, let $\tilde{\mathbf{H}} = (\tilde{\mathbf{G}}^T \tilde{\mathbf{G}})^{-1}$, then

$$\text{Cov} \begin{bmatrix} \Delta \mathbf{x}_L \\ \Delta b \end{bmatrix} = \sigma^2 \tilde{\mathbf{H}}$$

and

$$\sigma_E^2 = \sigma^2 \tilde{H}_{11}, \quad \sigma_N^2 = \sigma^2 \tilde{H}_{22}, \quad \sigma_U^2 = \sigma^2 \tilde{H}_{33}, \quad \sigma_b^2 = \sigma^2 \tilde{H}_{44} \quad (6.25)$$

where σ_E , σ_N , σ_U are the standard deviations of the east, north, and up (vertical) components of the position error, respectively. The diagonal elements of $\tilde{\mathbf{H}}$, denoted as $\tilde{H}_{ii}$ above, correspond to DOP parameters: East DOP (EDOP), North DOP (NDOP), Vertical DOP (VDOP), and the previously defined TDOP.

$$\tilde{\mathbf{H}} = \begin{bmatrix} \text{EDOP}^2 & \cdot & \cdot & \cdot \\ \cdot & \text{NDOP}^2 & \cdot & \cdot \\ \cdot & \cdot & \text{VDOP}^2 & \cdot \\ \cdot & \cdot & \cdot & \text{TDOP}^2 \end{bmatrix} \quad (6.26)$$

where we have left the off-diagonal terms unspecified. Defining $\text{HDOP} = \sqrt{\tilde{H}_{11} + \tilde{H}_{22}}$ and $\text{VDOP} = \sqrt{\tilde{H}_{33}}$ we obtain

$$\text{RMS horizontal error} = \sqrt{\sigma_E^2 + \sigma_N^2} = \sigma \cdot \text{HDOP} \quad (6.27a)$$

$$\text{RMS vertical error} = \sigma_U = \sigma \cdot \text{VDOP} \quad (6.27b)$$

$$\text{RMS 3-D error} = \sqrt{\sigma_E^2 + \sigma_N^2 + \sigma_U^2} = \sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2} = \sigma \cdot \text{PDOP} \quad (6.27c)$$

The DOPs provide a simple characterization of the user-satellite geometry. The more favorable the geometry, the lower the DOP. The lower the DOP and σ , the better the quality of the position estimate, in general. If a receiver is limited in the number of satellites it can track simultaneously, the user pays a price with larger rms position error due to higher DOP values. The early GPS receivers were limited to tracking only four satellites at a time, and PDOP or GDOP was a common basis for the selection of the 'best' four satellites. The newer receivers, even the inexpensive ones for the consumer market, now typically track all satellites in view.

With four satellites, PDOP is lowest if three satellites are evenly distributed in azimuth near the horizon and the fourth satellite is at zenith. As a rule, a low HDOP requires good azimuthal coverage with satellites that are not too high above the horizon. A low VDOP (or, PDOP) requires a satellite near zenith and good coverage in azimuth with satellites that are close to the horizon. Actually, VDOP would be lower if a receiver could track satellites below the horizon, as would be case with space-borne receivers [Spilker (1996)].

The clean characterization of position error in (6.27) in terms of the geometry and measurement error is based on our model (6.14). The pseudorange measurement errors are uncorrelated and have a common variance. If a user had a good model to estimate the covariance matrix Σ of the measurement error, we would have formulated the problem as one of weighted least squares, with Σ^{-1} as the weighting matrix. An expression like (6.27) could still be written, but the new definition of DOP parameters, reflecting the structures of both the geometry matrix and measurement error covariance matrix, would no longer be simple or intuitive. Actually, as noted earlier, an adequate characterization of the correlation structure is difficult, and is generally sidestepped in favor of using (6.14) for broad guidance on the quality of the position estimates.

The role of DOP in GPS positioning is often misunderstood. A lower DOP value does not automatically mean a lower position error, and vice versa. To illustrate this point, Figure 6.3 gives a scatter plot of HDOP and the corresponding horizontal position error computed from measurements taken three minutes apart over a three-month period early in 1993. SA was active during this period and our data model (i.e., uncorrelated measurements with equal variance) actually fitted the data. GPS only had a partial satellite constellation during this period and the HDOP values at times were high. The main point of this figure is to illustrate that the higher the HDOP value, the bigger the scatter in position error, and you are as likely to obtain a low-error position estimate as one with high error. If we group the points by HDOP ranges and compute the rms error for each range bin, shown as diamonds, we see that the relationship between the rms position error and HDOP is in fact linear. This is an empirical validation of the relationship (6.27a).

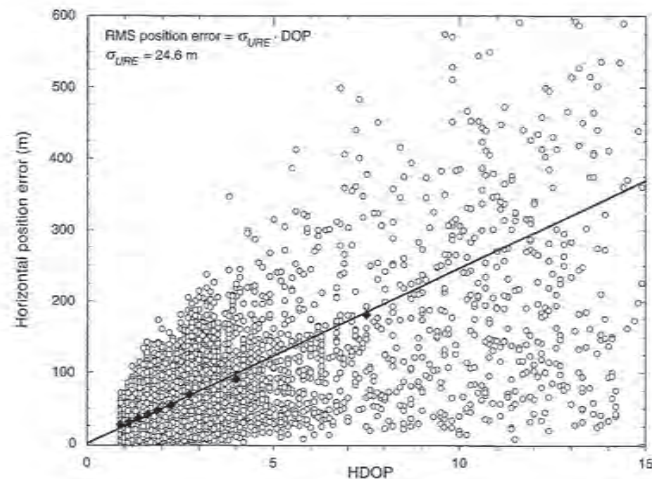


Figure 6.3 Scatter plot of the horizontal position error versus HDOP for measurements taken at three-minute intervals over a three-month period in 1993 while SA was active. The relationship between the rms horizontal position error and HDOP is linear. The slope of the line is σ_{URE} .

$$\text{RMS horizontal position error} = \sigma_{URE} \cdot \text{HDOP}$$

The slope of the line fitted to the computed rms values of position error in the different bins is σ_{URE} , the standard deviation of the user range error. This scatter plot thus gives us a way to estimate σ_{URE} . The estimate is 24.6 m, which is close to our model of $\sigma_{URE} = 25$ m while SA was active.

Now we return to the question with which we began this section: How good is a GPS-derived position estimate? The answer is that the position error depends upon the measurement geometry and pseudorange measurement errors. Given a simple model of the measurement error (6.14), the best we can do is to characterize the rms position error as in (6.27) in terms of a DOP parameter and the rms measurement error. In order to be able to say more about the position error, we would need to know more about the measurement errors.

If the pseudorange errors could be modeled as Gaussian, the error in x -, y -, and z -component of the position estimate would be Gaussian as well. We would know not just the position error statistics (i.e., mean and standard deviation), but the error distribution. We could determine the median horizontal error, and the 95th percentile of the vertical error distribution. We could even determine the 99.999th percentile of the position error, but that would be going against our dictum of pushing a simple model too far.

Distribution of DOPs

The satellites are in motion and DOP values change with the changing user-satellite geometry. To illustrate the pattern of such change, DOP values measured over a twenty-four-hour period from a roof-top antenna at MIT Lincoln Laboratory, Lexington, Massachusetts, early in 1997 are shown in Figure 6.4. The GPS constellation comprised 25 satellites, and the receiver

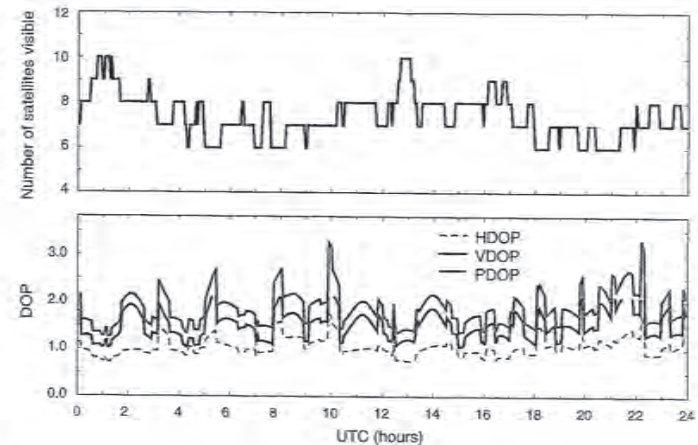


Figure 6.4 Number of satellites in view and the corresponding DOPs observed at Lexington, Massachusetts, early in 1997 with a 25-satellite constellation.

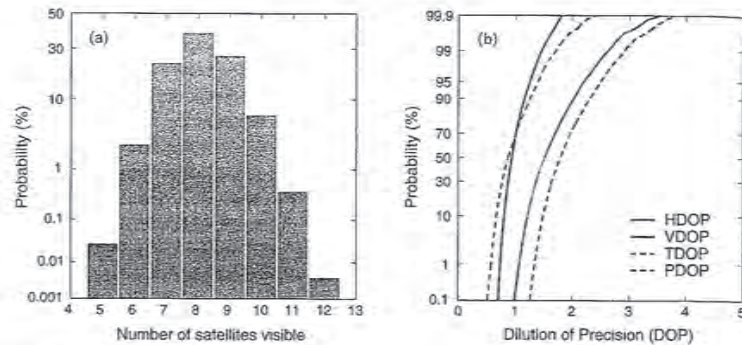


Figure 6.5 (a) Histogram of the number of satellites in view of users worldwide with an unobstructed view of the sky (elevation angle $>5^\circ$) for the 24-satellite baseline constellation. (b) The corresponding cumulative distribution functions of HDOP, VDOP, TDOP, and PDOP values.

tracked all satellites in view. As is typical, the HDOP values are consistently smaller than the VDOP values. The distribution of HDOP would be roughly similar at other locations. VDOP, however, tends to be larger at higher latitudes (north or south). The reason is that the inclination of the GPS orbits is 55° , and at higher latitudes there are no satellites overhead.

In view of (6.27), we can characterize the distribution of the positioning error in terms of probability distributions of the DOP values. Such a global characterization is offered in Figure 6.5 for the baseline 24-satellite GPS constellation on the basis of a simulation.

Figure 6.5(a) is a histogram of the number of satellites in view at elevation angles higher than 5° for users distributed worldwide. Figure 6.5(b) gives the corresponding cumulative distribution functions of HDOP, VDOP, TDOP, and PDOP. Note in Figure 6.5(a) that with the baseline 24-satellite GPS constellation, about 90% of the users with an unobstructed view of the sky see seven to nine satellites, and nearly all see five or more satellites. The quality of these position estimates would depend upon the corresponding DOPs. The HDOP and TDOP values are typically close to one, and generally do not exceed two; VDOP tends to be larger than HDOP. With the full GPS constellation, the median and 99th percentile of the HDOP distribution are 1 and 1.6, respectively; the corresponding values for VDOP are 1.6 and 3.7, respectively. We should note in passing that (i) the GPS constellation has had more than 25 working satellites continuously since 1995, and (ii) an unobstructed view of the sky is a luxury for most navigators while on land.

What's DOP Good For?

In the 1980s, there was only a partial satellite constellation and GPS coverage was spotty. The receivers were also less capable, generally limited to tracking four satellites. The early researchers, therefore, had to schedule their data collection experiments on the basis of predicted DOPs in order to get useful measurements. The best time for GPS observations was invariably between 2 a.m. and 4 a.m., regardless of the user location, or that's how the old-timers remember it. In any case, that's how DOP came to acquire its considerable importance in the GPS user community.

Now with 24-plus satellites in the constellation and most receivers tracking all satellites in view, HDOP is seldom greater than one, unless parts of the sky are obstructed from view. The urban users and hikers who commonly face such situations have to be advised of poor satellite geometry and its consequences for positioning, but it seems pointless to quote DOP to them. The concept of DOP, however, is a valuable tool in design and analysis of a satellite constellation, and as a diagnostic tool to indicate receiver malfunction.

6.1.3 Price of an Inexpensive Receiver Clock

Clocks are at the heart of GPS. As discussed in Section 4.2.4, the satellites carry on board atomic frequency standards which are extraordinarily stable. The widespread use of GPS, however, is made possible by the fact that the receivers can work with inexpensive quartz oscillators not much better than those common in wristwatches. Geodetic-quality receivers may use temperature-compensated crystal oscillators (TCXO), which are a bit larger and use more power. The use of atomic standards in GPS receivers is rare, generally limited to monitoring stations and a few reference stations used for research purposes.

In our discussion so far, it has been understood implicitly that the price of using an inexpensive receiver clock is simply that we cannot estimate 3-D position from three satellites, and need a minimum of four satellites in view. Actually, there is more to it. There is a price in terms of the quality of the position estimates. If the clock bias were known and we could measure ranges rather than pseudoranges to the satellites, we would obtain better position estimates.

The vertical error in GPS-based position estimates tends to be larger than the horizontal error. This can be inferred from the plots of HDOP and VDOP values in Figures 6.4 and 6.5. The discrepancy between the quality of the horizontal and vertical position estimates is often explained by an appealing intuitive argument as follows. The horizontal position estimates are better because the satellites are on all sides of a user. The vertical position estimates suffer by comparison because all the satellites are above the user, and none below.

Indeed, if the earth were a transparent bubble allowing each user to see all GPS satellites, a simple and rough argument based on symmetry would lead us to conclude that the error in each of the north, east, and vertical directions would be similar. The blockage of satellites below a user's horizon introduces asymmetry which puts the estimation of vertical position at a disadvantage. In fact, the consequence of this geometrical asymmetry would be relatively small if the receiver clock bias were known *a priori*. It is left as an exercise to show that the VDOPs would be reduced significantly, and would be generally lower than the corresponding HDOPs, if the receiver could measure true ranges to the satellites rather than pseudoranges [Misra (1996)]. Our earlier definition of DOPs was for a 4-D estimation problem dealing with three unknowns of the user position and the receiver clock bias. If the receiver clock bias is known, the problem reduces to 3-D estimation, and the definition of DOPs would be amended appropriately, following the same lines of treatment as given earlier in this section.

It may appear at first that the idea of a receiver clock with a known bias is strictly academic. Not true. Note that we didn't require that the bias remain constant, only that it be known accurately. It would be enough if the clock behavior could be predicted, given the recent past. This would require that the clock bias change smoothly, without any corners or jumps. Actually, this is a rather stringent requirement as the timekeeping characteristics of a clock can change with temperature, pressure, and acceleration. An ingenious, alternate approach to 'clock aiding' to improve DGPS navigation accuracy is described by Hwang *et al.* (2005).

6.1.4 Other Performance Measures and Specifications

In the previous section, we characterized the rms position error for a user in terms of the standard deviation of the pseudorange measurement error (σ), and a DOP parameter (which depends upon the satellite geometry). The measurement errors change over time. The rms value of the measurement errors would depend upon the mode of usage (autonomous, differential, or aided). The satellite geometry changes as the satellites rise, move across the sky, and set. Since the GPS constellation essentially repeats its ground tracks from one day to the next (actually, after 23h 56m), the user would see the same geometries repeat. The performance of GPS, therefore, changes with place and time, and is said to be dynamic.

Our discussion so far has focused on *absolute accuracy*: How well a position estimate matches what we know to be the right answer. If we didn't know the right answer, we may want to evaluate the position estimates in terms of their repeatability. We could call this *repeatable accuracy*. (A fisherman may simply want to be led back to the same spot without regard to the position coordinates). There are still other ways to think of accuracy, but we'll limit ourselves to absolute accuracy.

In order to characterize the global performance of GPS, we have to take into account the position errors of users at all places at all times. Different users employ receivers of different capabilities and some users operate in areas with significant sky blockage or are hampered by severe multipath. The satellite constellation changes, too, as satellites are added and removed or temporarily sidelined for maintenance. For global specifications, it is necessary to limit the scenarios and define average performance predicated upon a number of conditions [SPS (2001)]: 24-satellite constellation; clear view of the sky for elevation angles higher than 5°; at least four satellites in view with PDOP ≤ 6 ; and performance averaged over twenty-four hours. The error sources can also be limited only to those for which the Control Segment can be held responsible. The resulting so-called signal-in-space (SIS) accuracy specifications, cited in Table 2.1, are to be understood in the context of these conditions.

There are several ways to characterize positioning accuracy. Given a mathematical model of the measurements (6.9) and a characterization of the measurement error (6.14), we determined mean error, standard deviation, and rms error in 1-D, 2-D, and 3-D earlier (6.27). We can also determine the various percentiles of the error distributions if we assume further that the errors follow a Gaussian distribution. Alternatively, the error statistics can be computed empirically on the basis of a sample of position estimates. Let Δx_j , Δy_j , and Δz_j , $j = 1, 2, \dots, n$, be the errors in the east, north, and up components of the j th position estimate sample. RMS horizontal (2-D), vertical, and 3-D errors are defined as

$$\text{RMS vertical error} = \sqrt{\frac{1}{n} \sum_{j=1}^n \Delta z_j^2} \quad (6.28a)$$

$$\text{2-D rms error} = \sqrt{\frac{1}{n} \sum_{j=1}^n (\Delta x_j^2 + \Delta y_j^2)} \quad (6.28b)$$

$$\text{3-D rms error} = \sqrt{\frac{1}{n} \sum_{j=1}^n (\Delta x_j^2 + \Delta y_j^2 + \Delta z_j^2)} \quad (6.28c)$$

The more extensive the sample, the more faith we can have in the measure of the quality of a position estimate.

The rms error is sometimes represented as synonymous with standard deviation of the error. That's incorrect. The two terms are interchangeable only if the mean position error along each component is zero (i.e., the estimates of position coordinates are unbiased). For GPS position estimates, mean error equals zero if averaged over a long time interval. Position estimates averaged over seconds or minutes are generally biased because the pseudorange measurement errors due to satellite clock, orbit, and propagation effects change slowly.

If the position estimates are modeled as unbiased, the above expressions for rms error can be written in terms of standard deviations of the east, north, and vertical components of position, introduced earlier as σ_E , σ_N , and σ_U , respectively, as follows.

$$\text{2-D rms error} = \sqrt{\sigma_E^2 + \sigma_N^2} \quad (6.29a)$$

$$\text{RMS vertical error} = \sigma_U \quad (6.29b)$$

$$\text{3-D rms error} = \sqrt{\sigma_E^2 + \sigma_N^2 + \sigma_U^2} \quad (6.29c)$$

The 50th and 95th percentiles of the horizontal, vertical, and 3-D errors are used commonly, and are easily computed. The median (50th percentile) horizontal error, called *circular error probable* (CEP), defines the radius of a circle centered at the true position which would contain the position estimate with probability of 0.5. *Spherical error probable* (SEP) is defined similarly as median 3-D error. The 95th percentiles are often written as horizontal error (95%) and vertical error (95%) as shorthand. They are also sometimes referred to as error values at 95% confidence level. (Actually, confidence level has a different, well-defined meaning in statistics, and we don't recommend this usage)

Another measure of the quality of the position estimates in common use is *2drms*, defined as twice distance rms in 2-D (or, twice the rms horizontal error, $2 \times 2\text{-D rms}$). It's a confusing nomenclature which now appears entrenched. To make matters worse, *2drms* is often used incorrectly as equivalent to the 95th percentile of the horizontal error.

The system requirements and performance of positioning systems are often expressed by different, often inconsistent, parameters. For example, the original GPS positioning specifications were: 100 m (2drms) for the SPS users, and 16 m (SEP) for the PPS users. That's all. Different user groups choose to specify their requirements in terms of 2drms, CEP, SEP, and 95th percentile. Given one measure of performance, we can determine another. Such 'translation' of the performance measures often requires simplifying assumptions which, though not strictly true, serve as a useful model if not pushed too far. We follow van Diggelen (1998).

Our model for GPS position estimates is:

- The position estimates follow a multivariate normal distribution centered at the true position.
- Errors in the east, north, and up (E, N, and U) components are uncorrelated. (It's a reasonable assumption for low DOPs.) The standard deviation of the north and east errors are roughly equal: $\sigma_E = \sigma_N$. The standard deviation of the vertical error is roughly twice the rms horizontal error: $\sigma_U = 2(\sigma_E^2 + \sigma_N^2)^{1/2}$.

With the above assumptions, the various performance measures can be related as follows using the basic properties of the normal and chi-squared distributions.

(i)	RMS vertical error	$\approx 0.5 \times \text{vertical error (95\%)}$	(6.30a)
		$\approx 2\text{drms}$	(6.30b)
		$\approx 0.9 \times 3\text{-D rms error}$	(6.30c)
(ii)	2-D rms error	$\approx 0.5 \times 2\text{drms error}$	(6.31a)
		$\approx 0.6 \times \text{horizontal error (95\%)}$	(6.31b)
		$\approx 1.2 \times \text{CEP}$	(6.31c)
(iii)	3-D rms error	$\approx 2.2 \times 2\text{-D rms error}$	(6.32a)
		$\approx 1.2 \times \text{horizontal error (95\%)}$	(6.32b)
		$\approx 1.3 \times \text{SEP}$	(6.32c)

6.1.5 Empirical Positioning Results

In this section, we examine the typical positioning results obtained from GPS. The objective is to offer a broad appreciation for the variability in the position error in different modes (autonomous and differential). The measurements were taken with stationary rooftop antennas at surveyed locations at Stanford University and MIT Lincoln Laboratory and the receiver tracking all satellites in view. Each of the error plots of Figures 6.6 and 6.7 below presents horizontal position error computed from pseudorange measurements taken one minute apart over a 24-hour period early in 2004.

SPS Position Estimates

Representative positioning results from GPS Standard Positioning Service (SPS) are given as scatter plots in Figure 6.6. Figure 6.6(a) presents positioning results from an inexpensive and popular receiver designed for outdoorsmen. As expected, the user's manual isn't too specific about how the receiver treats the ionospheric or tropospheric errors. The receiver computes and displays 3-D position estimates as latitude, longitude, and height, and can record these data internally. The results are plotted in Figure 6.6(a) as north error and east error. The pixellated look of the scatter plot comes from the limited precision of the internal recording: 0.00001 degrees for latitude and longitude (and 0.1 feet in height). The median error is about 4 meters and the 95th percentile point is below 7.5 meters. The performance is impressive, and more than adequate for a hiker. The performance far exceeds the specifications [Table 2.1].

Figure 6.6(b) shows SPS results obtained with a more capable receiver which recorded code and carrier phase measurements and the navigation messages. The C/A-code pseudorange measurements were corrected for propagation errors using the models discussed previously. The ionospheric delay was estimated from the broadcast parameter values of the Klobuchar model [Section 5.3.2]. The tropospheric delay was estimated from the model UNB4 [Collins *et al.* (1996)]. We didn't, however, use the carrier phase measurements to smooth the pseudoranges [Section 5.7.1]. The performance is significantly better than that for the pocket receiver discussed above. The median error is 2 meters, and the 95th percentile point is 4 meters.

It is interesting to note in Figure 6.6 that the east and north position errors show no obvi-

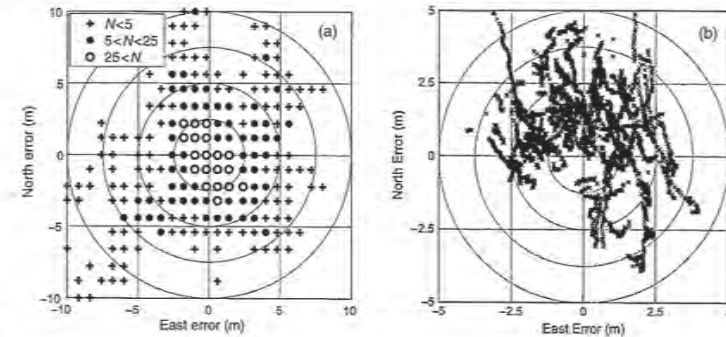


Figure 6.6 Scatter plots of horizontal position error in position estimates obtained from GPS Standard Positioning Service (SPS). (a) Quantized position estimates from an inexpensive receiver for hikers (one-minute samples over 24 hours, N represents the number of outcomes corresponding to each bin). (b) A more capable receiver which gave access to the pseudorange measurements and navigation data (one-second samples over 24 hours), courtesy of Dr. Todd Walter, Stanford University.

ous correlation, thus validating our model in Section 6.1.4. Though not shown, there is no correlation apparent between the vertical and horizontal position errors either.

DGPS Position Estimates

Representative positioning results from wide-area and local-area differential GPS are provided in Figures 6.7. The wide-area corrections were obtained from the Wide Area Augmentation System (WAAS) [Section 5.8.3] for the data set used in Figure 6.6(b). The corrections were applied to the pseudorange measurements, and the resulting positioning errors are shown. The median and 95th percentile errors are now 0.7 meters and 1.7 meters, a significant improvement over SPS.

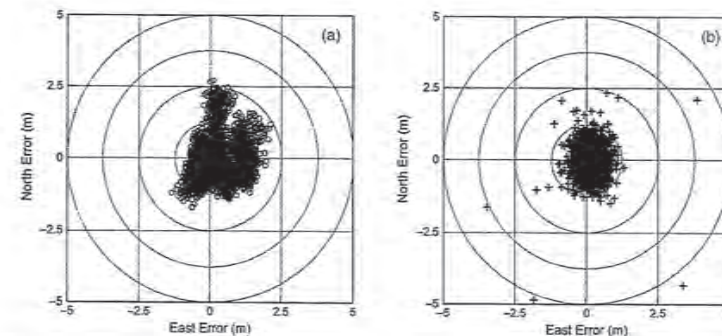


Figure 6.7 Scatter plots of horizontal position error in DGPS position estimates using pseudorange corrections obtained from (a) WAAS (1-second samples over 24 hours), courtesy of Dr. Todd Walter, Stanford University. (b) Local-area DGPS with the reference station 25 km away (1-minute samples over 24 hours).

A local-area differential GPS can do better yet. Figure 6.7(b) shows the positioning results using corrections obtained from a CORS site [Section 2.5] located about 25 km away. In this experiment, the pseudorange measurements at both ends were smoothed using carrier phase measurements [Section 4.7.1]. The 95th percentile point is less than 1 meter.

6.2 Position and Velocity from Pseudorange Rates

6.2.1 Velocity Estimation

The relative motion of a satellite and the user results in changes in the observed frequency of the satellite signal. This Doppler shift is measured routinely in the carrier tracking loop of a GPS receiver [Chapter 12]. Given the satellite velocity, the Doppler shift can be used to estimate the user velocity. The Doppler shift, or equivalently, the range rate [Section 1.3.2], can be written as a projection of the relative velocity vector on the satellite line-of-sight vector. The measurement, however, is biased by the receiver clock bias rate (i.e., frequency offset), and what we actually measure is the pseudorange rate.

The delta pseudoranges obtained from carrier phase measurements are proportional to the average pseudorange rates or the line-of-sight velocity of the user relative to the satellite over the time interval. The model for pseudorange rates can be obtained by differentiating (6.1). It is left as an exercise to show that

$$\begin{aligned}\dot{\rho}^{(k)} &= \dot{r}^{(k)} + (\dot{b} - \dot{b}^{(k)}) + \dot{I}^{(k)} + \dot{J}^{(k)} \\ &= (\mathbf{v}^{(k)} - \mathbf{v}) \cdot \mathbf{1}^{(k)} + \dot{b} + \varepsilon_{\phi}^{(k)},\end{aligned}\quad (6.33)$$

where $\mathbf{v}^{(k)}$ is the satellite velocity vector, obtained from the navigation message broadcast by the satellite (Homework Problem 4-12); $\mathbf{v}$ is the user velocity vector, to be estimated. Both $\mathbf{v}^{(k)}$ and $\mathbf{v}$ are expressed in the ECEF coordinate frame. The user-to-satellite line-of-sight unit vector $\mathbf{1}^{(k)}$ is determined from an estimate of the user position; $\dot{b}$ and $\dot{b}^{(k)}$ are the rates of change in the receiver and satellite clocks (m/s), respectively, and $\varepsilon_{\phi}^{(k)}$ denotes the combined error due to changes during the measurement interval in the satellite clock, ionosphere, and troposphere. Note that the velocity of an object attached to the earth is zero in the ECEF coordinate frame.

The principal source of error in (6.33) throughout the 1990s was the satellite clock frequency dithering ($\dot{b}^{(k)}$) due to SA. Now with SA gone, this term is negligible. The errors due to changes during the measurement interval in the ionospheric and tropospheric delays and in multipath are also generally small. For moderate user speeds, the main source of error in velocity estimation is the error in the predicted satellite position and velocity $\mathbf{v}^{(k)}$. Any error in the predicted satellite position (especially the along-track and cross-track components) shows up in the line-of-sight vectors, and is magnified in the dot product with the satellite velocity vector.

Problems can arise if the user dynamics are high. The delta ranges give only an average velocity over a time interval. High accelerations and jerks would clearly be problematic. Such dynamics can also create an additional error source by raising the receiver clock phase noise [Section 4.2.3]. For these reasons, the DoD has avoided specifying the velocity estimation performance. (According to an old document [JPO (1991)], the PPS performance specifications for a constant-velocity scenario are: 0.1 m/s rms in any direction; 0.2 m/s 2σrms. These are

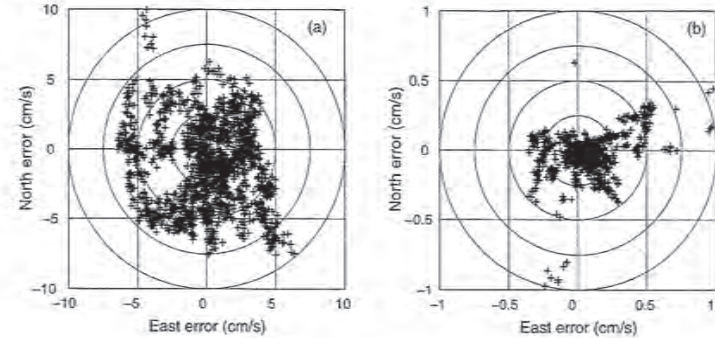


Figure 6.8. Scatter plots of horizontal error in GPS velocity estimates computed from pseudorange rate (Doppler frequency) measurements taken one minute apart over a 24-hour period from a pair of stationary antennas about 25 km apart. (a) Error in velocity estimates; (b) error in relative velocity estimates. (Note: 1 m/s = 2 knots.)

overly conservative, as we'll see below.)

Equation (6.28) is linear in user velocity components, and can be rewritten as

$$(\dot{\rho}^{(k)} - \mathbf{v}^{(k)} \cdot \mathbf{1}^{(k)}) = -\mathbf{1}^{(k)} \cdot \mathbf{v} + \dot{b} + \varepsilon_{\phi}^{(k)}$$

Denoting $(\dot{\rho}^{(k)} - \mathbf{v}^{(k)} \cdot \mathbf{1}^{(k)})$ as $\tilde{\rho}^{(k)}$, the combined set of measurements from the K satellites can be written as a set of equations compactly in matrix notation as

$$\tilde{\mathbf{p}} = \mathbf{G} \begin{bmatrix} \mathbf{v} \\ \dot{b} \end{bmatrix} + \tilde{\mathbf{e}}_{\phi}, \quad (6.34)$$

where matrix $\mathbf{G}$ characterizes the user-satellite geometry, as defined previously (6.10). The problem of estimation of user velocity based on pseudorange rates is identical in structure to that of estimation of user position from pseudoranges (6.9). A least-squares solution and the DOP parameters can be defined, as before, and related to the rms error in these estimates.

Velocity estimates based on pseudorange rates are presented in Figure 6.8. We recorded pseudorange rates measured one minute apart over a day at a pair of receivers with stationary antennas about 25 km apart. Figure 6.8(a) gives a scatter plot of the horizontal velocity estimates for one of the receivers. The 50th and the 95th percentiles of the horizontal velocity error are 0.03 m/s and 0.08 m/s, respectively. The effect of errors in the predicted satellite position and velocity, discussed above, is mitigated in relative velocity computations, and results in a dramatic improvement. This is seen in Figure 6.8(b), which presents the corresponding estimates of relative velocity between the two receivers.

The estimates of the vertical component of velocity are not as good as those for horizontal velocity. The reasons are similar to those given earlier in this section in the discussion of the accuracy of the horizontal and vertical position estimates from conventional processing.

6.2.2 Position Estimation

As an aside of some historical interest, we discuss Doppler positioning [Section 1.2.3] with GPS. If a user is stationary at an unknown location, we can rewrite (6.33) as

$$\begin{aligned}\dot{\rho}^{(k)} &= \mathbf{v}^{(k)} \cdot \mathbf{1}^{(k)} + b + \varepsilon_{\phi}^{(k)}, \\ &= \mathbf{v}^{(k)} \cdot \frac{(\mathbf{x}^{(k)} - \mathbf{x})}{\|\mathbf{x}^{(k)} - \mathbf{x}\|} + b + \varepsilon_{\phi}^{(k)}\end{aligned}\quad (6.35)$$

In principle, this set of K equations ($K \geq 4$) can be solved for the four unknowns: $(\mathbf{x}, b)$. As before, we could resort to linearization about the initial estimates of position and receiver clock rate, and set up a linear system of equations to solve iteratively. Actually, the information content in the above formulation regarding the user position is weak: the line-of-sight vector is not very sensitive to small changes in user position. As noted previously, the angle at a satellite between the lines of sight from two users separated by 100 km is less than 0.3° .

The DOP parameters defined for the system of equations (6.35) would be large (in thousands), in general. This is another way to characterize the information content in the system of equations (6.35) for position estimation. The DOPs can be brought down by stacking the measurements taken over a period of time. It would take measurements at well-separated epochs to get a good position estimate.

While Doppler positioning has little to recommend to the GPS users, it was the basis for positioning with Transit, the first operational satellite navigation system [Section 1.3.2]. A Transit user saw only one satellite at a time for ten to twenty minutes, took multiple measurements, and solved for his 2-D position and the difference of the clock bias rates of the satellite and receiver clocks (i.e., frequency offset between the satellite and receiver clock). Unlike GPS, the clocks in Transit satellites did not require precise synchronization, but the frequencies of the Transit satellites and user clocks were required to be stable over ten to twenty minutes. In other words, any frequency instability during a satellite pass contributed to position error.

6.3 Time Transfer

The problem of keeping precise time and synchronizing clocks that are separated by considerable distances is an old one. What's changed over time is that the requirements of new applications for accuracy and precision have become more and more demanding. In Section 4.2, we discussed time scales and the evolution of timekeeping devices. In this section, we discuss the distribution of time, or time transfer, in which GPS now plays an increasingly important role. In particular, we discuss two applications: direct time distribution from GPS and comparison of remote clocks.

The coming of the telegraph in the 19th century and its use in distribution of time was vital to smooth operation of railroads across a continent. Early in the 20th century, telegraph was superseded by radio for time transfer. Radio broadcast time over land and oceans and in the air with a millisecond-level accuracy. Then came radionavigation systems with transmission of their signals carefully synchronized with reference to a precise time standard. Omega, Loran-C, and Transit were all capable of time transfer with an accuracy level of about 1 ms [Klepczynski (1983)]. With the current techniques, GPS can distribute time with an accuracy

of about 30 ns, and can compare remote clocks with an accuracy of about 5 ns.

Precise synchronization of the clocks aboard the satellites is at the heart of GPS. Synchronization of a satellite clock with UTC(USNO) at nanosecond level is made possible by use of atomic frequency standards aboard the satellites and in the monitor stations of the Control Segment. GPS provides access to this network of synchronized atomic clocks for dissemination of precise time and frequency data worldwide.

A user interested in obtaining precise time would typically require a one-pulse-per-second (1PPS) signal synchronized with UTC, the civil standard. Actually, we'll look for a 1PPS signal synchronized to UTC(USNO), introduced earlier in Section 4.2 as UTC realized in real time at the U.S. Naval Observatory (USNO). In order to generate such a signal from a GPS receiver, we have to deal with four time scales:

- time kept by a satellite clock (t^s),
- GPS time (GPST) (t_{GPS}), defined by the Control Segment on the basis of a set of atomic standards aboard the satellites and in monitor stations,
- UTC(USNO) (t_{UTC}), the U.S. national standard defined by the U.S. Naval Observatory,
- time kept by a user's receiver clock (t_u).

Starting with t_u , we want to generate t_{UTC} . We do so by estimating and accounting for the offset between these two time scales in steps as follows.

Recall that the Control Segment determines the bias in each satellite clock (δt^s) on the basis of measurements from the monitor stations.

$$\delta t^s = t^s - t_{GPS} \quad (6.36)$$

The bias is modeled as a quadratic function over a time interval. The parameters $\{a_{f0}, a_{f1}, a_{f2}\}$ of this model are computed and uploaded to the satellite, which broadcasts them in the navigation message (Subframe 1) to the users. At time t_{GPS}

$$\delta t^s = a_{f0} + a_{f1}(t_{GPS} - t_{0c}) + a_{f2}(t_{GPS} - t_{0c})^2 + \Delta t_r \quad (6.37)$$

where t_{0c} is the reference time for the model in GPST, and Δt_r is the relativistic term [Section 4.2.4]. The rms error in estimation of δt^s is currently estimated to be about 5 ns [$\sigma(\delta t^s) = 5$ ns].

The U.S. Naval Observatory monitors the offset of GPST relative to UTC(USNO) as

$$\delta t_{UTC} = t_{GPS} - t_{UTC} \quad (6.38)$$

and provides this information to the Control Segment. The bias between GPST and UTC(USNO) can be computed at any instant (defined in GPST) from

$$\delta t_{UTC} = A_0 + A_1(t_{GPS} - t_{0U}) + \Delta t_{LS} \quad (6.39)$$

where the first two terms specify the bias modulo 1 second, and Δt_{LS} is the number of leap seconds added to UTC since 1980 ($\Delta t_{LS} = 14$ seconds as of 1 January 2006). Values of parameters $\{A_0, A_1, t_{0U}, \Delta t_{LS}\}$ are broadcast by each satellite in its navigation message (page 18 of Subframe 4). The rms error in estimation of δt_{UTC} is currently estimated to be about 10 ns [$\sigma(\delta t_{UTC}) = 10$ ns].