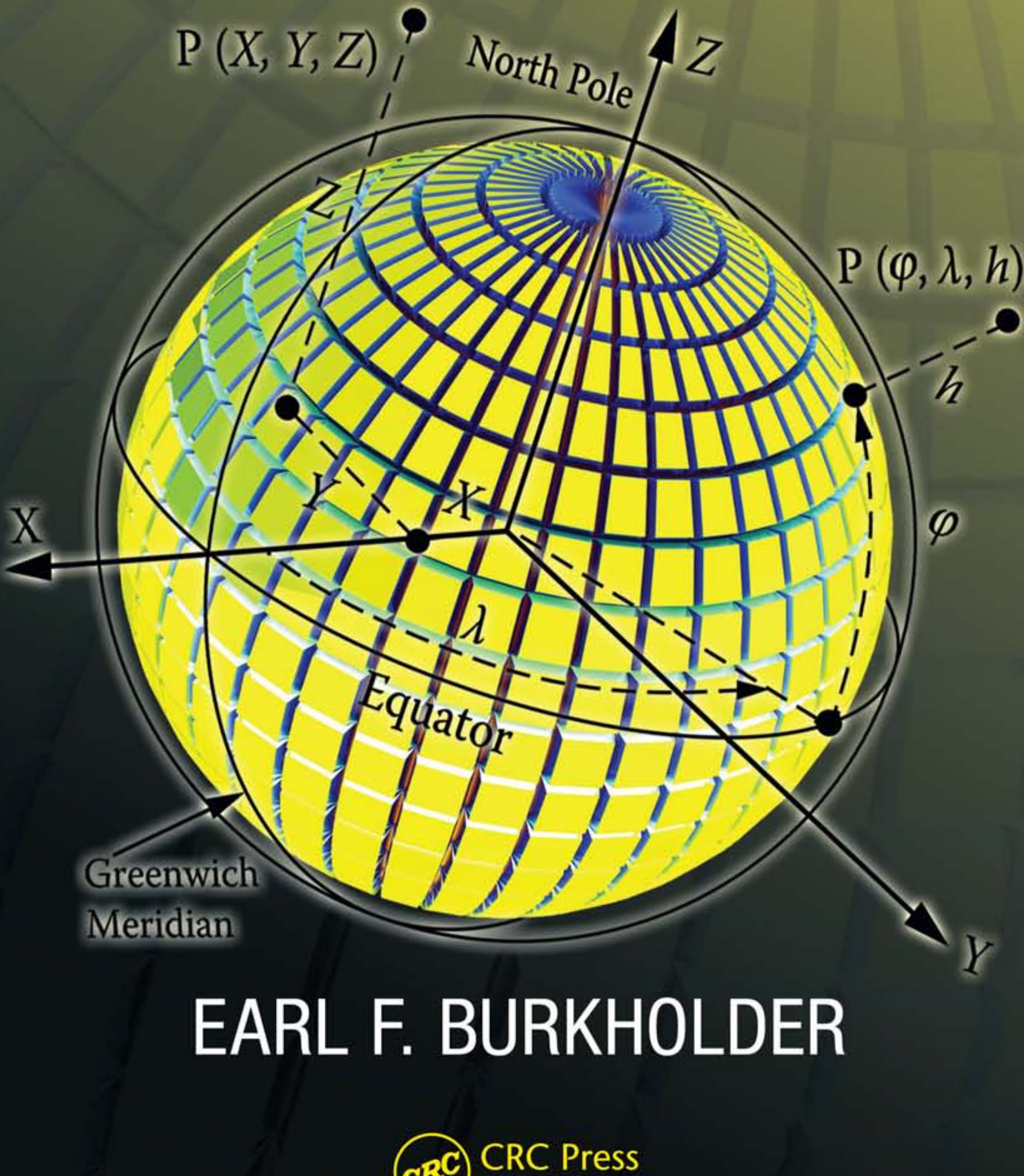
THE 3-D GLOBAL SPATIAL DATA MODEL

Foundation of the Spatial Data Infrastructure



EARL F. BURKHOLDER

CRC Press
Taylor & Francis Group

THE 3-D GLOBAL SPATIAL DATA MODEL

Foundation of the Spatial Data Infrastructure

THE 3-D GLOBAL SPATIAL DATA MODEL

Foundation of the Spatial Data Infrastructure

EARL F. BURKHOLDER

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-4200-6301-1 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Burkholder, Earl F.
The 3-D global spatial data model : foundation of the spatial data infrastructure / Earl F. Burkholder.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4200-6301-1 (hardback : alk. paper)
1. Spatial data infrastructures--Mathematics. 2. Spatial systems. 3. Three-dimensional imaging. I. Title.

QA402.B866 2008
516--dc22 2008008413

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

Foreword

The global spatial data model (GSDM) is an arrangement of time-honored solid geometry equations and proven mathematical procedures. In that respect, it contains nothing new. But, the GSDM is built on the assumption of a single origin for three-dimensional (3-D) geospatial data and formally defines procedures for handling spatial data that are consistent with digital technology and modern practice. In that respect, the GSDM is a new model.

Why another spatial data model? Compatibility is essential when persons from disparate disciplines share data. Success is assured to the extent fundamental concepts are clearly defined and basic procedures for data exchange are formalized. Such standardization and centralization provide economies of scale to the user community. Subsequently, additional benefits are derived through decentralization in which innovative applications expand upon the capabilities supported by the underlying standard. Such benefits, in turn, spawn new markets and applications. As the cycle continues, the underlying standards and the assumptions upon which they were established need to be reexamined and, if appropriate, updated. The telephone is an example. A centralized regulated monopoly was largely responsible for placing a telephone in most homes in the United States. With the underlying infrastructure in place, additional benefits were realized as the industry was deregulated and competition between providers brought the consumer more options relative to telecommunications equipment and services. The twisted-pair analog standard upon which the telecommunications network was built is no longer adequate, and digital technology has been implemented to support significantly greater levels of service. Consumers can now select various (even wireless) network connections for data, voice, fax, and cable.

Geospatial data are another example. In the past, analog storage of geospatial data on a map was standard practice. Geographic coordinates provide global standardization, and derivative uses such as map projection (or state plane) coordinates are commonplace. The benefits of such centralization were a driving force in the early stages of building geographic information systems (GIS's) as users and agencies needed to pool resources to achieve desired economies of scale. But spatial data users are experiencing the same analog-digital transition as the telecommunications industry, and the underlying model needs to be reexamined. With the advent of affordable digital technologies (i.e., global positioning systems [GPS] and related computer resources), the demand for spatial data products is growing rapidly. Enormous gains in productivity have been achieved by automating procedures for handling spatial data and by switching from analog to digital spatial data storage. However, traditional (horizontal and vertical) spatial data models fail to exploit fully the wealth of data available. In a sense, the spatial data user community continues to "put new (digital) wine into old bottles." The GSDM is a new bottle model that preserves the integrity of 3-D spatial data while providing additional benefits (i.e., simpler equations, worldwide standardization, and the ability to track spatial data accuracy with greater specificity and convenience).

Computer databases are digital. Analog maps are still used, but, increasingly, maps and data visualizations are generated upon demand from a digital database. Rarely is a map now used for primary spatial data storage. Spatial data are 3-D, and maps are 2-D. Modern measurement systems collect 3-D data, yet many computer databases store spatial data as 2-D horizontal data and 1-D vertical (elevation) data. An improved practice is to build and share a 3-D database that supports both 2- and 3-D applications in either analog or digital mode. Separately, the importance of spatial data accuracy has come to the fore as evidenced by efforts to develop meta-data standards and specifications. These issues and others are addressed by reexamining the underlying spatial data model and by designing spatial data collection, storage, manipulation, adjustment, and visualization procedures based upon the 3-D GSDM.

But, perhaps even more compelling arguments favoring adoption of the GSDM can be derived from two documents prepared by the U.S. National Academy of Public Administration (NAPA):

- "The Global Positioning System: Charting the Future" was prepared for the U.S. Congress and the U.S. Department of Defense (DOD), and was published in 1995. It describes the history, performance, and future of GPS. This document is particularly important to those who build, operate, and utilize the systems that generate reliable geospatial data. The executive summary states, in part, "GPS is much more than a satellite system for positioning and navigation. It represents a stunning technological achievement that is becoming a global utility with immense benefits for the U.S. military, civil government, and commercial users and consumers worldwide."
- "Geographic Information for the 21st Century" was prepared for the U.S. Bureau of Land Management, U.S. Forest Service, U.S. Geological Survey, and National Ocean Service, and was published in 1998. It describes the instrumental roles that agencies of the U.S. government have played in "surveying, mapping, and other geographic information functions since the beginning of the Republic." The report includes various excellent recommendations based upon the use of GIS's under the conceptual umbrella of the National Spatial Data Infrastructure (NSDI).

Many persons in various professions are comfortable with both reports. But, by and large, the GPS group includes highly technical specialists such as aerospace engineers, electrical engineers, geodesists, physicists, and photogrammetrists. The GIS group probably involves a greater number of people and includes spatial data users whose professional-technical focus tends toward administration, local government services, information technology (IT), civil engineering, surveying and mapping, planning, and business. In addition to those with professional interests, the number of people using GPS and/or GIS on a personal level is growing exponentially, a trend reasonably expected to continue.

Interoperability is the key. The GSDM builds a conceptual bridge between the two NAPA reports by providing a consistent 3-D geometrical framework for both GPS and GIS. The GSDM serves the scientific end of the spectrum without sacrificing technical rigor while simultaneously providing local spatial data users the opportunities to work with local flat-Earth coordinate differences and to view the (virtual) world from any location. Examples of this interoperability bridge are highlighted at various places throughout the book.

Preface

"The right tool for the job" is a simple phrase with profound implications. The proliferation of tools for handling spatial data is somewhat daunting, as benefits associated with their use spawn the development of even better tools. Although we are where we are because of where we came from, the path to the future should be viewed in terms of the analog-to-digital revolution. Given the many specializations associated with developing technology, it is difficult to write a comprehensive book about an umbrella topic like spatial data. Therefore, acknowledging that others will add details to illuminate the path ahead even better, this book is written to define and describe a global spatial data model (GSDM) that

- is easy to use because it is based upon rules of solid geometry.
- is standard between disciplines and can be used all over the world.
- accommodates modern measurement and digital data storage technologies.
- supports both analog map plots and computer visualization of digital data.
- preserves geometrical integrity and does not distort physical measurements.
- combines horizontal and vertical data into a single 3-D database.
- facilitates rigorous error propagation and standard deviation computations.
- provides (and defines assumptions associated with) various choices with respect to spatial data accuracy.

In a way, this book is organized backward. Chapter 1 contains the results, and chapter 2 justifies chapter 1. Fundamental geometrical concepts are developed in terms of more traditional material in subsequent chapters. That is done to accommodate readers with various backgrounds. Managers and those with a strong technical background might concentrate only on the beginning chapters. Spatial data professionals at various levels who wish to gain a better understanding of geometrical relationships should start with the beginning chapters so they know where the rest of the book is going. Given that chapters 1 and 2 are not easy reading, they should be read first as an overview. It is expected, then, as the reader progresses through subsequent chapters, that chapters 1 and 2 will be revisited as required to help refresh the focus on the overall objective of defining an appropriate spatial data model. For those just beginning to work with spatial data, serious reading and study should begin in chapter 3. With that said, the plan for building a comprehensive spatial data model is to present fundamental mathematical concepts in chapter 3 and to add concepts from surveying, geodesy, and cartography in subsequent chapters.

The material is presented as simply as possible without compromising technical rigor. Some readers will find the review of mathematical concepts redundant, and some readers may never have occasion to use linear algebra, matrix manipulation, or error propagation. Acknowledging the certain diversity of readers, the goals are to provide a logical development of concepts for those who wish to follow the theory and to provide all readers a collection of tools that can be used to handle spatial data more efficiently.

Whether the reader is involved in technical applications, is making managerial and administrative decisions with regard to spatial data, or is a programmer writing software for handling spatial data, all should agree that the most appropriate tools for handling spatial data are those that are, at the same time, both simple and appropriate. The GSDM is simple because it uses existing practice and rules of solid geometry for manipulating spatial data. And, the GSDM is appropriate because it is built on local coordinate differences, preserves true 3-D geometrical integrity on a global scale, accommodates modern digital technology, handles error propagation with aplomb, and supports subsequent computation of complex geometrical relationships in geodesy, cartography, and other sciences. In the past, spatial data models were selected by default as people (rightfully) focused on impressive gains in utility and productivity made possible by automating existing processes. But the GSDM is a result of examining those processes in terms of digital technology and fundamental geometrical concepts. With features of the various models described and compared, it is anticipated that spatial data analysts in various fields will, as a matter of conscious choice, begin using the GSDM because it establishes a common geometrical link between spatial data sets, applications, and disciplines, and because it provides an efficient method of defining, tracking, and evaluating the accuracy of spatial data.

What does it mean to "think outside the box"? Is thinking outside the box something beneficial and desirable? Or is thinking outside the box to be avoided? What do elephant jokes have to do with boxes? Without answering those questions, consider the following (to whom does one credit elephant jokes?):

1. How does one determine the number of elephants in the refrigerator? Answer: Count their tracks in the butter.
2. How does one kill a blue elephant? Answer: Shoot it with a blue elephant gun.

OK, now the pattern is established, and the reader is ready for whatever else comes along.

3. How does one kill a red elephant? No, you don't shoot it with a red elephant gun, because you don't have one. The correct answer is "Choke it until it turns blue, then shoot it with your blue elephant gun."

Elephant jokes may have no place in a rigorous technical book (except maybe in the preface), but these illustrate a very important point. Humans are very good at using whatever tools are available to do what needs to be done. Without being critical, many wonderful accomplishments have involved (figuratively) choking the elephants. But, everyone should be aware that sometimes it is better, easier, and more appropriate to look for a red elephant gun than it is to keep choking those red elephants. Most red elephant guns are found outside the box.

The GSDM is viewed as a red elephant gun for handling 3-D digital geospatial data.

Acknowledgments

I am indebted to many persons for the motivation to write this book. Writing took far longer than first anticipated, but rushing to publication was not my goal. First, I need to acknowledge my mother, Brownie, who encouraged me to be curious, and my father, Irvin, who, unlike his father to him, avoided prejudging my aspirations. I must also recognize my undergraduate mentor at the University of Michigan, Professor Ralph Moore Berry, who helped me gain a profound appreciation of the surveying profession; Professor Edward Mikhail at Purdue University, who emphasized the importance of rigor in the thought process; Professor Alfred Leick at the University of Maine, who, in a single 5-minute conversation during my 1990–1991 sabbatical, identified the 3-D geodetic model as an appropriate model for spatial data; and Dr. Kurt W. Bauer, emeritus executive director of the Southeastern Wisconsin Regional Planning Commission, who commissioned preparation of the 1997 report "Definition of a Three-Dimensional Spatial Data Model for Southeastern Wisconsin." But I owe the largest debt of gratitude to my wife, Donna, and our three (now adult) children, Franklin, Valorie, and Barbara. None of them ever once suggested during the last 10 years that I should give up writing the book, even though none of us knew (or knows) where this series of many small steps might lead. The journey has been long and challenging at times, but the book portion of my life journey has been good because innumerable persons, especially students, have made it so. To all those persons, both named and unnamed, I offer my sincere appreciation—thank you.

Earl F. Burkholder, PS, PE

List of Abbreviations

2-D two-dimensional
3-D three-dimensional
AASHTO American Association of State Highway and Transportation Officials
ACSM American Congress on Surveying and Mapping
A-S anti-spoofing
ASCE American Society of Civil Engineers
ASCII American Standard Code for Information Interchange
ASPRS American Society of Photogrammetry and Remote Sensing
BK1 through BK22 labels assigned to routine procedures listed in Table 1.1 and identified in Figure 1.4
BLM U.S. Bureau of Land Management
BURKORD™ trademark for 3-D software and database
C/A code course acquisition code
COGO coordinate geometry
CORPSCON coordinate conversion software by the U.S. Army Corps of Engineers
CORS continuously operating reference station
c+r correction for curvature and refraction
CTP Conventional Terrestrial Pole
D° degree of curve
DGPS differential GPS
DMA Defense Mapping Agency (no longer used—see NIMA)
DMD double-meridian-distance
DOD U.S. Department of Defense
DOT U.S. Department of Transportation
ECEF Earth-centered Earth-fixed
EDM electronic distance meter
e/n/u local perspective right-handed rectangular coordinates
FAA Federal Aviation Administration
FOC full operational capability
GALILEO European satellite positioning system (similar to GPS, which is owned by the United States)
GIS Geographic Information System
GLONASS Russian Global Navigation Satellite System (similar to GPS, which is owned by the United States)
GNSS global navigation satellite systems
GPS Global Positioning System
GRS80 Geodetic Reference System of 1980
GSDI Global Spatial Data Infrastructure
GSDM global spatial data model

HARN high-accuracy reference network
HD(1) horizontal distance used in plane surveying
HPGN high-precision geodetic network
HTDP horizontal time-dependent positioning
IAG International Association of Geodesy
IERS International Earth Rotation Service
IGLD, IGLD(xx) International Great Lakes Datum (year realized)
IOC initial operational capability
IT information technology
ITRF International Terrestrial Reference Frame
LDP low distortion projection
LLR lunar laser ranging
MSL mean sea level
NAD27 North American Datum of 1927
NAD83(CORS) North American Datum of 1983 based upon NGS CORS stations
NAD83(2007) North American Datum of 1983 based upon 2007 adjustment of the NSRS
NAD83(xx) North American Datum of 1983 (realized in 19xx)
NADCON North American Datum conversion software
NAPA U.S. National Academy of Public Administration
NAVD88 North American Vertical Datum of 1988
NAVSTAR navigation satellite timing and ranging
NGA National Geo-spatial Intelligence Agency (formerly NIMA)
NGS National Geodetic Survey
NGVD29 National Geodetic Vertical Datum of 1929
NIMA National Imagery and Mapping Agency (formerly DMA, now NGA)
NOAA National Oceanic and Atmospheric Administration
NRC National Research Council
NSDI National Spatial Data Infrastructure
NSRS National Spatial Reference System
OPUS On-line Positioning User Service
OPUS-RS OPUS—rapid static
PC personal computer
P-code precision code
P.O.B. point of beginning
PPS precise positioning service
RINEX Receiver Independent Exchange format
RTK real-time-kinematic
RTN real-time network
SA selective availability (discontinued on May 1, 2000)
SI international system for units of measurement
SLR satellite laser ranging
SPC(S) state plane coordinate (system)
SPS standard positioning service
TAI International Atomic Time

TCT transcontinental traverse
UAV unmanned aerial vehicle
USC&GS U.S. Coast and Geodetic Survey (now NGS)
USPLSS U.S. Public Land Survey System
UTC Coordinated Universal Time
UTM universal transverse Mercator
VERTCON vertical datum conversion software
VLBI very long baseline interferometry
WAAS wide area augmentation system
WADGPS wide area differential GPS
WGS84 (Gxxxx) World Geodetic System of 1984 (specifies epoch for GPS operations)
WGSxx World Geodetic System (realized in 19xx)
WVDXX World Vertical Datum of 20xx
X/Y/Z geocentric rectangular coordinates

1 The Global Spatial Data Model (GSDM) Defined

INTRODUCTION

Geospatial data representing real-world locations are three-dimensional (3-D), and modern measurement systems collect data in a physical 3-D environment. Time as the fourth dimension is acknowledged, but this book focuses on 3-D data. This chapter defines and describes the global spatial data model (GSDM) as a collection of mathematical concepts and procedures that can be used to collect, organize, store, process, manipulate, evaluate, and use 3-D spatial data. Measurements of quantities such as angles, length, time, current, mass, and temperature are used with known physical and geometrical relationships to compute spatial data components that are stored for subsequent use and reuse.

In the past, records of such measurements were written in field books, logs, or journals, and the spatial information was compiled into an analog map that typically served two purposes. The map was simultaneously the primary storage medium for the spatial information and the end product of the data collection process. Spatial data are now collected, stored, and manipulated digitally in an electronic environment, and the primary storage medium is rarely the end product. Instead, the same digital data file can be duplicated repeatedly and used to generate and/or support many different spatial data products. In either case, whether developing an analog or digital spatial data product, algorithms are the mathematical rules used to manipulate measurements and spatial data to obtain meaningful spatial information. In addition, the quality of spatial information is dependent upon the quality of the original measurement, completeness of the required information, and appropriateness of the algorithms used to manipulate the data.

The GSDM includes both the algorithms for processing spatial data and the procedures that can be used to provide a defensible statistical description of spatial data quality. That means measurement professionals can focus on building and/or using systems that generate reliable spatial data components and spatial data users in various disciplines can devote attention to using and interpreting the data with the assurance that all parties generating and/or using the data are "on the same page" (i.e., using a common spatial data model).

This first chapter is a summary of the defining document for the GSDM (Burkholder 1997b). The intent is to cite primary works because other people developed most of the concepts described herein. For example, appendix C in Bomford (1971) is titled "Cartesian Coordinates in Three Dimensions." Leick (2004) defines the 3-D geodetic model of which the GSDM is a part, Mikhail (1976) provides a comprehensive discussion of functional and stochastic models, and, when discussing models, Moritz (1978) comments on the simplicity of using the basic global rectangular *X*/*Y*/*Z* system without an ellipsoid. When the aforementioned concepts are combined in a systematic way with particular attention to the manner in which spatial data are used, the synergistic whole—the GSDM—appears to be greater than the sum of the parts.

Neither is the GSDM concept a new one. Seeber (1993) states that H. Burns proposed the concept of a global three-dimensional polyhedron network as early as 1878. The differences now are that the Global Positioning System (GPS) and other modern technologies have made a global network practical and that the polyhedron need not be limited to Earth-based points. The GSDM might also be an appropriate model for describing the "best" instantaneous positions of a global network of continuously operating reference stations (CORS) computed in real time with respect to the International Terrestrial Reference Frame (ITRF). An adopted mean position for each CORS may serve the needs of most users, but corrections for short-term variations caused by the Earth's tides, long-term continental drift velocities, and even catastrophic events such as earthquakes should be available to those needing them. It is readily acknowledged that such policies are already being used in the scientific community and that a space-fixed inertial reference system is more appropriate for describing the motion of Earth satellites. The GSDM should not be viewed as a prescriptive model, but as an inclusive model that accommodates the diverse practice of many spatial data users and provides an efficient bridge between local "flat-Earth" uses and rigorous scientific applications.

THE GSDM

The GSDM is a collection of mathematical concepts and procedures that can be used to manage spatial data both locally and globally. It consists of a functional model that describes the geometrical relationships and a stochastic model that describes the probabilistic characteristics—statistical qualities—of spatial data. The functional part of the model includes equations of geometrical geodesy and rules of solid geometry as related to various coordinate systems and is intended to be consistent with the 3-D geodetic model described by Leick (2004) with the following exception: the GSDM, being strictly spatial, does not accommodate gravity measurements but presumes gravity effects are appropriately accommodated before data are entered into the spatial model. The stochastic portion of the GSDM is an application of concepts described by Mikhail (1976), Leick (2004), and Wolf and Ghilani (1997).

Although the GSDM makes no attempt to accommodate non-Euclidean space or concepts, it does provide a simple universal foundation for many disparate coordinate systems used in various parts of the world and offers advantages of standardization for spatial data users in disciplines such as those listed in Figure 1.1. As such, the GSDM should be viewed as the geometrical portion of a larger concept being promoted as the Global Spatial Data Infrastructure (GSDI) described by Holland et al. (1999) as "[t]he policies, organizational remits, data, technologies, standards, delivery mechanisms, and financial and human resources necessary to ensure that those working at the global and regional scales are not impeded in meeting their objective." For more information on the GSDI, see http://www.gsdi.org.

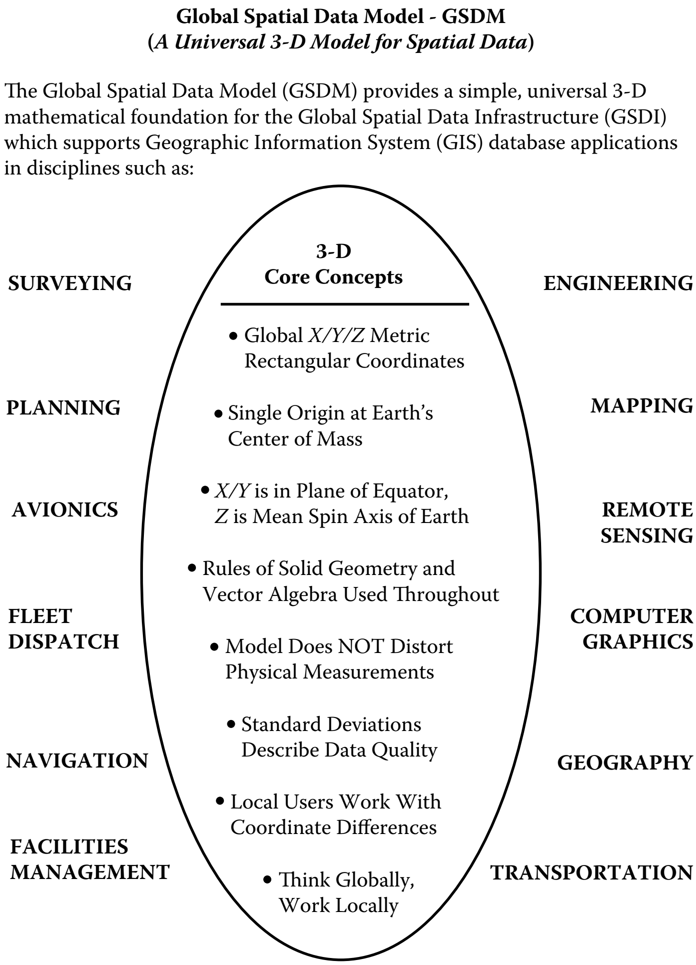


FIGURE 1.1 The Global Spatial Data Model

The Functional Model Component

The functional model component of the GSDM is based upon a three-dimensional, right-handed rectangular Cartesian coordinate system with the origin located at the Earth's center of mass. The *X*/*Y* plane lies in the equatorial plane, with the X-axis at the 0° (Greenwich) meridian. The Z-axis coincides with the mean spin axis of the Earth as defined by the Conventional Terrestrial Pole (Leick 2004). This geocentric coordinate system is called an Earth-centered Earth-fixed (ECEF) coordinate system by the United States' National Imagery and Mapping Agency (1997) and is widely used by many who work with GPS and related data. Rules of solid geometry and vector algebra are universally applicable when working with ECEF coordinates and coordinate differences.

As shown in Figure 1.2, the unique 3-D position of any point on Earth or near space is equivalently defined by traditional *latitude*/*longitude*/*ellipsoid height*

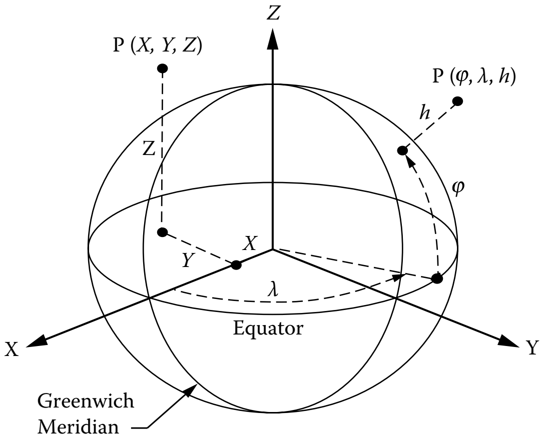


FIGURE 1.2 Geocentric *X*/*Y*/*Z* and Geodetic φ/λ/*h* Coordinates

coordinates or by a triplet of *X*/*Y*/*Z* coordinates expressed in meters. Due to the large distances involved, the *X*/*Y*/*Z* coordinate values can be quite large, but personal computers (PC's) operating in double precision routinely handle fifteen significant digits. Twelve significant digits will accommodate all ECEF coordinate values within the "birdcage" of GPS satellites down to 0.1 mm. Some users may object to working with such large coordinate values, but, as shown in Figure 1.3, such objections will likely become inconsequential to the extent that end-user applications are designed to utilize coordinate differences (much smaller numbers and fewer digits).

Figure 1.4 is a schematic that illustrates relationships between the ECEF coordinate system and various other coordinate systems commonly used in connection with

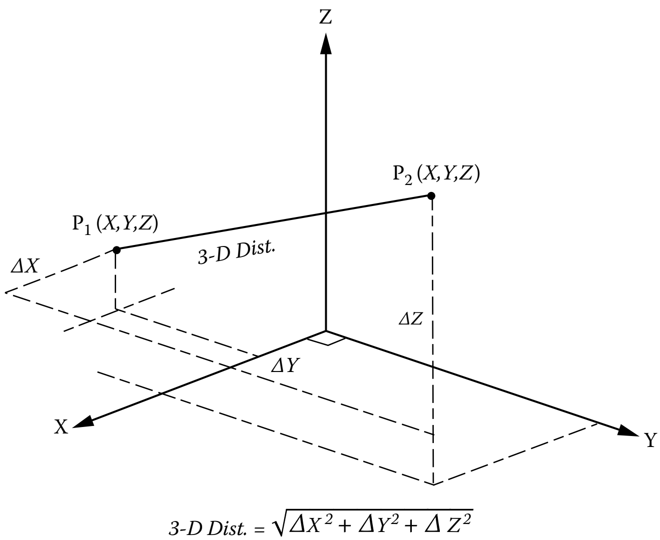


FIGURE 1.3 GPS Technology Provides Precise *ΔX*/*ΔY*/*ΔZ* Components

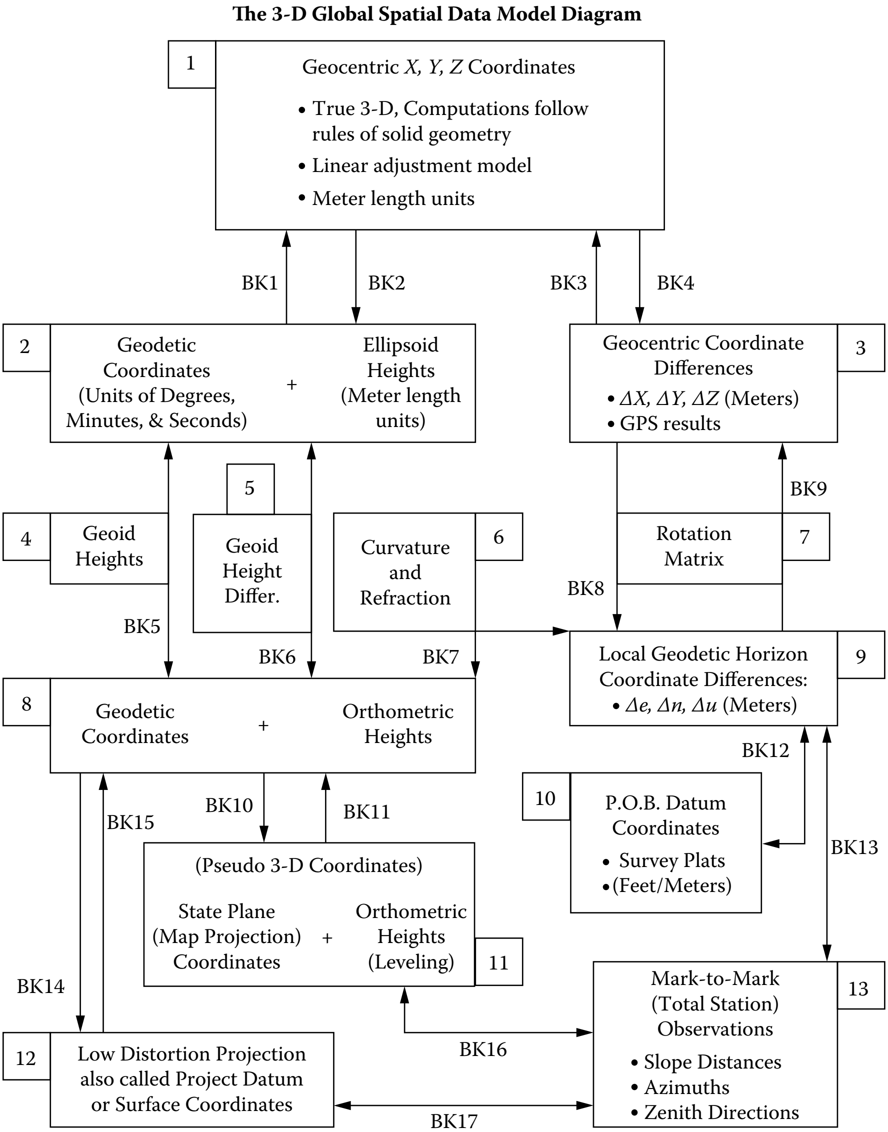


FIGURE 1.4 Diagram Showing Relationship of Coordinate Systems

spatial data. A key feature on the diagram is a rotation matrix used to convert $\Delta X/\Delta Y/\Delta Z$ coordinate differences to local $\Delta e/\Delta n/\Delta u$ coordinate differences at any point (local origin) specified by the user. Since a vector in 3-D space is not altered by moving the origin or by changing the orientation of the reference coordinate system, a vector defined by its geocentric $\Delta X/\Delta Y/\Delta Z$ components is equivalently defined by local components, and the rotation matrix is the mechanism that efficiently transforms a global perspective into a local one. The transpose of the rotation matrix is used to transform local components of a space vector to corresponding geocentric components.

Computational Designations

With regard to Figure 1.4, the functional model includes equations for transforming spatial data described by coordinates in one numbered box to equivalent expressions in a different coordinate system. The contents of the numbered boxes are as follows:

Box 1: Geocentric *X*/*Y*/*Z* coordinates are the basis for all other coordinate values obtained from the GSDM. These are the primary defining values stored for each point in a digital spatial data file. Coordinate values in other coordinate systems are derived from the stored ECEF coordinates using algorithms that have been tested and proven for mathematical "exactness" and computational precision. This part of the GSDM features meter units, a linear adjustment model, and vector algebra along with universal rules of solid geometry.

Box 2: Geodetic coordinates of latitude and longitude combined with ellipsoid height can define a three-dimensional position with the same precision and exactness as geocentric *X*/*Y*/*Z* coordinates. Equations are listed in a subsequent section by which coordinate values in one box can be converted to equivalent values in another. Using both angular sexagesimal units (degrees, minutes, and seconds) on the ellipsoid and length units of meters for height makes traditional 3-D geodetic computations more complicated than when using ECEF rectangular coordinates.

Box 3: GPS technology has been a driving force behind the use of 3-D spatial data and helps create demand for the GSDM. Although practice includes displaying coordinates in a defined system, the primary output of a GPS survey historically has been a 3-D vector defined by its $\Delta X/\Delta Y/\Delta Z$ components. Because existing control stations were defined with geodetic coordinates of latitude and longitude (and other reasons), it was natural to continue building a 2-D network using 3-D measurements. And there certainly are cases where that practice can still be justified. But, the GSDM defines an environment in which the full value of 3-D data can be used to build high-quality 3-D networks without being encumbered by 2-D assumptions and complex equations found in classical geometrical geodesy. Another benefit of the GSDM is that the associated stochastic model lends itself to implementation in the rectangular 3-D environment more readily than in the *latitude/longitude/height* system.

Box 4: In practice, geoid height is taken to be the difference between ellipsoid height and elevation. With any two of the three elements known, the third can be found. If a reliable ellipsoid height for a point (from GPS data) is combined with an appropriate geoid height (from geoid modeling), it is possible to obtain high-quality orthometric height (elevations). Appropriate use of standard deviations for the constituent components will provide a statistical assessment of the quality of such elevations.

Box 5: Box 5 is the same as Box 4 except that the computations are performed using differences. As will be explained in chapter 7, modeled geoid height differences are often more reliable than modeled absolute geoid heights.

That means better elevations can be computed by starting with a known high-quality benchmark elevation and combining observed ellipsoid height differences with modeled geoid height differences to compute the orthometric height difference.

Box 6: The Δu component from Box 9 is the perpendicular distance of the forepoint from the tangent plane through the standpoint. An elevation difference from the standpoint to the forepoint includes the Δu component plus the curvature and refraction (c+r) correction. This c+r procedure is based on the modeled distance between the horizontal plane and a level surface and does not include any geoid modeling between the standpoint and forepoint.

Box 7: Given that the statistical qualities of a vector in space are independent of the perspective from which it is viewed, the rotation matrix is a very efficient method for changing a global perspective (geocentric coordinate differences) into a local perspective (local "flat-Earth" components). Similarly, the transposed rotation matrix converts the local perspective into a global one.

Box 8: Historically, horizontal latitude and longitude coordinates have been combined with vertical elevations when mapping features on or near the Earth's surface. The generic zero reference surface for elevation has been the geoid (or mean sea level), which admits to a physical definition but, as it turns out, is very difficult, if not impossible, to find. As a result, an arbitrary reference surface that approximates, but does not define, mean sea level was selected for the North American Vertical Datum of 1988 (Zilkoski, Richards, and Young 1992).

Box 9: The local geodetic horizon (Trimble 1990) is essentially the same as the local geodetic frame defined more precisely by Soler and Hothem (1988) and shares many similarities with local plane surveying practice. The primary difference is that "up" is defined by the ellipsoid normal instead of the plumb line. That difference is largely inconsequential except in cases where very high precision is required or where the slope of the geoid (with respect to the normal) is severe. Another difference with the GSDM is that the origin moves with the observer because one is working with local coordinate differences with respect to the user-specified standpoint. See Box 10 for working with a traditional (or fixed) origin.

The Δu component in Box 9 can be used as an approximate elevation difference because it does not include the slope of the geoid, Earth curvature, or refraction (c+r)—all inconsequential for short lines. Although suggested as a secondary means for obtaining elevation differences, the standard c+r correction can be combined with the Δu component to obtain elevation differences between standpoint and forepoint. Understandably, the primary method for obtaining elevation differences still relies on differential leveling or accurate geoid heights and ellipsoid heights or their differences. See chapter 8 for more details.

Box 10: Point-of-beginning (P.O.B.) datum coordinates are a feature within the GSDM that accommodates long-established local plane surveying practices without compromising geometrical integrity. P.O.B. coordinates permit the user to select any point in the database as an origin. The 3-D location of each additional point selected is listed with respect to the P.O.B. Admittedly, this practice makes little sense for large distances, but these local coordinate differences can be treated in the same manner as local plane coordinates and used on survey plats. Horizontal distances are in the tangent plane through the P.O.B., and azimuths are with respect to the meridian through the P.O.B. If surveys of adjacent tracts do not use the same P.O.B., there will be two azimuths for a common line (the difference is convergence of the meridians between the two P.O.B.'s). However, if the P.O.B. is the same for both tracts, they will share a common basis of bearing—the geodetic meridian through the P.O.B.

Box 11: Map projections were invented to address the challenge of representing a curved Earth on a flat map. In particular, conformal projections have been used in surveying and mapping to define precisely a 2-D relationship between latitude and longitude positions on the Earth and equivalent plane coordinate positions on a flat map. Systematic use of map projections includes state plane coordinate systems as implemented in the United States and worldwide use of universal transverse Mercator (UTM) coordinates. However, it is important to note that elevations combined with map projection *x*/*y* (or *east*/*north*) plane coordinates are not an appropriate 3-D rectangular model for two reasons:

A. Conformal projections are well defined in two dimensions only. There is no mathematical definition of elevation in conformal mapping.
B. The zero reference surface for elevation (approximated by sea level) is a nonregular curved surface. Full 3-D integrity is preserved only to the extent that a flat Earth can be safely assumed. Therefore, map projection coordinates combined with elevations are referred to as "pseudo 3-D."

Box 12: An important consideration when using state plane coordinates is the relationship of the grid inverse distance to actual ground-level horizontal distance. In applications such as highway centerline stationing, the difference between grid and ground distance quickly becomes too great to ignore. Project datum coordinate systems were invented to accommodate that difference. Lack of standardization is an issue when considering project datum coordinates. For a summary of comments from forty-six out of fifty state departments of transportation (DOT's) on the grid-ground distance difference, see appendix 3 of Burkholder (1993a). On the other hand, states such as Wisconsin (1995) and Minnesota (Whitehorn 1997) have formally defined countywide coordinate systems for local use. A concise formulation of local coordinate system algorithms is found in chapter 10.

When working with the $\Delta e/\Delta n$ components, the horizontal distance is in the tangent plane through the standpoint and is the same horizontal distance that plane surveyors have been using for generations. It is also the same as HD(1) (i.e., ground-level horizontal difference) as described in Burkholder (1991). Understandably, with a unique tangent plane at each standpoint, the tangent plane from point A to point B is slightly different than the tangent plane from point B to point A. But, geometrical integrity in three dimensions is preserved by the GSDM and underlying *X*/*Y*/*Z* coordinates.

The 3-D azimuth from standpoint to forepoint obtained from arctan ($\Delta e/\Delta n$) gives the correct azimuth between each pair of points. The forward azimuth of a line differs from the back azimuth of the same line due to convergence of the meridians between the two endpoints. The GSDM competently provides the correct answer in each case. The 3-D azimuth is defined simply and is easy to use. The azimuth of a geodetic line has a more complex definition and differs only slightly from the 3-D azimuth. The geodetic azimuth is "better" than the 3-D azimuth only in the most demanding cases. See Burkholder (1997a) and Chapter 6 for more details.

Box 13: Spatial data measurements with conventional total station surveying instruments include slope distances, vertical (or zenith) angles, and determinations of bearings or azimuths. These measurements are used to compute local geodetic horizon coordinate differences of $\Delta e/\Delta n/\Delta u$. In reality, those measurements are referenced to the plumb line while the GSDM presumes the results are normal based. The difference is small, but important. Current procedures for making Laplace corrections are still applicable and should be used as will be described in chapter 8.

Equations for moving spatial data from one box to another have been given various names over the years. When used in context, there may be little confusion over what is a "forward" and what is an "inverse" computation. But, when brought together in a common collection, the duplication of conventional names can be confusing and misleading. Therefore, as a matter of convenience and in the interest of promoting unambiguous communication, the designations shown in Table 1.1 are used to describe the various computations and transformations. Many of them are illustrated in Figure 1.4.

Algorithm for Functional Model

A more complete set of equations and derivations is provided in chapter 6, but the following symbols are defined and used in this summary as follows:

$X/Y/Z$: Geocentric right-handed rectangular coordinates
$\Delta X/\Delta Y/\Delta Z$: Geocentric coordinate differences
$\Delta e/\Delta n/\Delta u$: Local coordinate differences
$\varphi/\lambda/h$: Geodetic latitude/longitude (east) and ellipsoid height
a and b: Semimajor and semiminor axes of reference ellipsoid
f: Flattening of reference ellipsoid
e^2: Eccentricity squared of reference ellipsoid; $e^2 = 2f - f^2$
N: Length of ellipsoid normal; also used for geoid height
S: Spatial slope distance between standpoint and forepoint
α: Geodetic azimuth at standpoint to forepoint
z or V: Zenith direction or vertical angle to forepoint
H: Orthometric height (elevation)
$HD(1)$ or D: Ground-level horizontal distance

TABLE 1.1
Designations for Spatial Data Computations and Transformations

Name	Conventional Description	See Page
BK1	Converting geodetic latitude/longitude/height coordinates to geocentric *X/Y/Z* coordinates	11
BK2	Converting geocentric *X/Y/Z* coordinates to geodetic latitude/longitude/height coordinates	11
BK3	3-D geodetic forward computation using $\Delta X/\Delta Y/\Delta Z$	12
BK4	3-D geodetic inverse computation using $\Delta X/\Delta Y/\Delta Z$	12
BK5[a]	Any combination of using ellipsoid height, orthometric height (elevation), and geoid height	12
BK6	Any combination of using differences for ellipsoid height, orthometric height (elevation), and geoid height	12
BK7	Using curvature and refraction corrections to refine elevation difference computations—especially trig heights	12
BK8	Converting geocentric coordinate differences to local coordinate differences	13
BK9	Converting local coordinate differences to geocentric coordinate differences	13
BK10	Converting geodetic latitude/longitude to state plane or map projection coordinates (also known as “forward computation”)	266
BK11	Converting state plane or map projection coordinates to latitude/longitude (also known as “inverse computation”)	266
BK12	Using local coordinate differences to compute P.O.B. datum coordinates and vice versa	13
BK13	Using conventional total station observations to compute local coordinate differences (1-D, 2-D, or 3-D) and vice versa	13
BK14	Converting geodetic latitude/longitude to low-distortion, project datum, or surface coordinates	13
BK15	Converting low distortion, project datum, or surface coordinates to latitude/longitude	13
BK16	2-D COGO computations based upon named state plane coordinate system zone	—
BK17	2-D COGO low-distortion, surface, or project datum computations of designated (countywide) system	—
BK18	2-D geodetic forward computation—not shown on Figure 1.4	162
BK19	2-D geodetic inverse computation—not shown on Figure 1.4	164
BK20	Generic 2-D COGO computations—not shown on Figure 1.4	—
BK21	Generic differential leveling—not shown on Figure 1.4	—
BK22	Generic trig height leveling—not shown on Figure 1.4	—

[a] *BK5, BK6, and BK7 are quite similar, but having different designations will help avoid problems caused by the subtle differences.*

Notes:

1. All distances are in units of meters.
2. Where two points are concerned, the standpoint is indicated by the subscript 1, while the forepoint is indicated by the subscript 2.

The BK1 equations are as follows:

$$N = \frac{a}{\sqrt{1 - e^2 \sin^2 \phi}} \tag{1.1}$$

$$X = \left(N + h \right) \cos \phi \cos \lambda \tag{1.2}$$

$$Y = \left(N + h \right) \cos \phi \sin \lambda \tag{1.3}$$

$$Z = \left(N\,[1 - e^2] + h \right) \sin \phi \tag{1.4}$$

The BK2 equations are more difficult to use because iteration is normally required to solve them. Equation 1.5 is quite straightforward, but equations 1.6 and 1.7 need to be iterated, as will be explained in chapter 6. An alternate (noniterative) method for performing the BK2 transformation is also given in chapter 6.

$$\lambda = \tan^{-1}\left(\frac{Y}{X} \right) \tag{1.5}$$

$$\phi = \tan^{-1}\left[\frac{Z}{\sqrt{X^2 + Y^2}} \left(1 + \frac{e^2 N \sin \phi}{Z} \right) \right] \tag{1.6}$$

$$h = \frac{\sqrt{X^2 + Y^2}}{\cos \phi} - N \tag{1.7}$$

The BK3 and BK4 equations are also called the 3-D “forward” and “inverse” as shown here:

BK3—Forward **BK4—Inverse**

$X_2 = X_1 + \Delta X$ $\quad\quad$ $\Delta X = X_2 - X_1$ (1.8) and (1.11)

$Y_2 = Y_1 + \Delta Y \qquad \Delta Y = Y_2 - Y_1$ (1.9) and (1.12)

$Z_2 = Z_1 + \Delta Z \qquad \Delta Z = Z_2 - Z_1$ (1.10) and (1.13)

The BK5 computation handles any combination of orthometric height (*H*), ellipsoid height (*h*), and geoid height (*N*) as follows:

$$N = h - H \tag{1.14}$$

$$H = h - N \tag{1.15}$$

$$h = H + N \tag{1.16}$$

The BK6 computation is the same as BK5 except that *differences* are used as follows:

$$\Delta N = \Delta h - \Delta H \tag{1.17}$$

$$\Delta H = \Delta h - \Delta N \tag{1.18}$$

$$\Delta h = \Delta H + \Delta N \tag{1.19}$$

Differences are important because geoid modeling provides better answers when using relative geoid height differences rather than absolute geoid heights. See chapter 8 for more details.

The BK7 computation relies on the combined curvature and refraction (c+r) correction for the difference between a level surface and tangent plane surface. For modest precision over short distances, the c+r correction can be used beneficially as follows (Davis et al. 1981, equation 5.7):

$$H_2 = H_1 + \Delta u + 0.0675 \frac{D^2}{1{,}000{,}000} = H_1 + \Delta u + 0.0675 \frac{\Delta e^2 + \Delta n^2}{1{,}000{,}000} \tag{1.20}$$

The BK8 and BK9 transformations involve using a rotation matrix to convert geocentric differences to local differences and local differences to geocentric differences. See appendix A at the end of this book for more details.

The BK8 transformation of geocentric differences to local differences is:

$$\begin{bmatrix} \Delta e \\ \Delta n \\ \Delta u \end{bmatrix} = \begin{bmatrix} -\sin\lambda & \cos\lambda & 0 \\ -\sin\phi\cos\lambda & -\sin\phi\sin\lambda & \cos\phi \\ \cos\phi\cos\lambda & \cos\phi\sin\lambda & \sin\phi \end{bmatrix} \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix} \tag{1.21}$$

The BK9 transformation of local differences to geocentric differences is as follows:

$$\begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix} = \begin{bmatrix} -\sin\lambda & -\sin\phi\cos\lambda & \cos\phi\cos\lambda \\ \cos\lambda & -\sin\phi\sin\lambda & \cos\phi\sin\lambda \\ 0 & \cos\phi & \sin\phi \end{bmatrix} \begin{bmatrix} \Delta e \\ \Delta n \\ \Delta u \end{bmatrix} \tag{1.22}$$

Equations 1.23 and 1.24 are not BKX transformations, but they are used to obtain the local tangent plane horizontal distance and the true direction from the standpoint (PT_1) to the forepoint (PT_2):

$$Distance = \sqrt{\Delta e^2 + \Delta n^2} \tag{1.23}$$

$$\tan\alpha = \left(\frac{\Delta e}{\Delta n}\right) \text{ with due regard to quadrant} \tag{1.24}$$

The BK10 and BK11 transformations are used to handle state plane coordinate transformations and are discussed in chapter 10.

BK12 computations are used to develop local tangent plane coordinates with respect to any P.O.B. selected by the user. See chapter 12 for more details.

BK13 transformations are used to convert terrestrial observations into local coordinate differences that can then be converted to geocentric differences using the BK9 transformation. Such observations may need to be corrected for instrument calibration, atmospheric conditions, polar motion, and local deflection-of-the-vertical.

$$\Delta e = S\sin z\sin\alpha = HD(1)\ \sin\alpha \tag{1.25}$$

$$\Delta n = S\sin z\cos\alpha = HD(1)\cos\alpha \tag{1.26}$$

$$\Delta u = S\cos z \tag{1.27}$$

There is no one correct set of equations for BK14 and BK15 computations. The primary force behind the use of project datum (or surface) coordinates is that an inverse between grid coordinates (grid distance) is not the same as the horizontal ground distance. Various methods for handling the grid-ground distance difference are described in Burkholder (1993a). Concise rigorous procedures for using local coordinate systems are provided in Burkholder (1993b), and some applications of project datum coordinates may continue to be justified. More recently, the concept of a low distortion projection (LDP) has been proposed, and the advantages of central administration for LDPs certainly have merit. But the issue of using project datum coordinates becomes moot when using the GSDM.

Computations BK16 and BK17 are the same traditional coordinate geometry computation with the following exception: BK16 computations involve formal use of state plane coordinates, and BK17 computations involve use of project datum coordinates. Being specific between those two is very prudent and can help avoid frustration and wasted efforts as a result of unwittingly using one for the other.

BK18 through BK22 computations consist of traditional surveying practices and are not shown in Figure 1.4.

The Stochastic Model Component

The stochastic component of the GSDM is devoted to answering the question "Accuracy with respect to what?" The stochastic model is based upon storing the covariance matrix associated with the geocentric *X/Y/Z* rectangular coordinates that define the location of each stored point. A user can compute the local *east/north/up* covariance matrix of any point on an "as needed" basis using the standard covariance error propagation (this minimizes storage requirements). The same basic procedure is extended to other functional model computations and provides a statistically defensible method for tracking the influence of random errors to any derived quantity. In particular, the user can look at the standard deviation of a coordinate position (by individual component) in either the geocentric or local reference frame. The standard deviation of other derived quantities such as distance, azimuth, slope, area, or volume can be obtained using the same error propagation procedure with the appropriate functional model equations.

The GSDM Covariance Matrices

The functional component of the GSDM consists of geometrical equations that are used to manipulate *X/Y/Z* geocentric coordinates defining the spatial position of each point. The stochastic component of the GSDM is an application of the laws of variance-covariance error propagation and utilizes the following matrix formulation (Mikhail 1976; Burkholder 1999, 2004):

$$\boldsymbol{\Sigma}_{YY} = \boldsymbol{J}_{YX}\,\boldsymbol{\Sigma}_{XX}\,\boldsymbol{J}^{t}_{X} \tag{1.28}$$

where

$\boldsymbol{\Sigma}_{YY}$ = covariance matrix of computed result,
$\boldsymbol{\Sigma}_{XX}$ = tcovariance matrix of variables used in computation, and
$\boldsymbol{J}_{YX}$ = Jacobian matrix of partial derivatives of the result with respect to the variables.

The GSDM uses two covariance matrices for each point—the geocentric covariance matrix and the local covariance matrix. The geocentric covariance matrix is stored and the local covariance matrix is computed on an "as needed" basis. In particular, the following symbols and matrices are used in the stochastic model:

σ_X^2 σ_Y^2 σ_Z^2: Variances of geocentric coordinates for a point
σ_{XY} σ_{XZ} σ_{YZ}: Covariances of geocentric coordinates for a point
σ_e^2 σ_n^2 σ_u^2: Variances of a point in the local reference frame
σ_{en} σ_{eu} σ_{nu}: Covariances of a point in the local reference frame
$\sigma_{\Delta X}^2$ $\sigma_{\Delta Y}^2$ $\sigma_{\Delta Z}^2$: Variances of geocentric coordinate differences
$\sigma_{\Delta X\Delta Y}$ $\sigma_{\Delta X\Delta Z}$ $\sigma_{\Delta Y\Delta Z}$: Covariances of geocentric coordinate differences
$\sigma_{\Delta e}^2$ $\sigma_{\Delta n}^2$ $\sigma_{\Delta u}^2$: Variances of coordinate differences in the local frame
$\sigma_{\Delta e\Delta n}$ $\sigma_{\Delta e\Delta u}$ $\sigma_{\Delta n\Delta u}$: Covariances of coordinate differences in the local frame
σ_S^2 σ_α^2: Variances of local horizontal distance and azimuth
$\sigma_{S\alpha}$: Covariance of local horizontal distance with azimuth
$\sigma_{X_1X_2}$ $\sigma_{Y_1Y_2}$: Elements of a point 1–point 2 submatrix

Geocentric Covariance Matrix ***Local Covariance Matrix***

$$\Sigma_{XYZ} = \begin{bmatrix} \sigma_{X^2} & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & \sigma_{Y^2} & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & \sigma_{Z^2} \end{bmatrix}; \qquad \Sigma_{enu} = \begin{bmatrix} \sigma_{e^2} & \sigma_{en} & \sigma_{eu} \\ \sigma_{en} & \sigma_{n^2} & \sigma_{nu} \\ \sigma_{eu} & \sigma_{nu} & \sigma_{u^2} \end{bmatrix}$$

(1.29 and 1.30)

Notes about the individual point covariance matrices:

1. Each covariance matrix is 3 × 3 and symmetric. Six numbers are required to store upper (or lower) triangular values.
2. The unit for each covariance matrix element is meters squared, the off-diagonal elements represent correlations, diagonal elements are called variances, and standard deviations are computed as the square root of the variances.
3. Each covariance matrix (with its unique orientation) represents the accuracy of a point with respect to a defined reference frame (or to whatever control is held fixed by the user) and is designated **datum accuracy**.

The local covariance matrix and the geocentric covariance matrix are related to each other mathematically by a rotation matrix for the latitude/longitude position of a point computed from its *X*/*Y*/*Z* coordinates (Burkholder 1999).

$$\boldsymbol{R} = \begin{bmatrix} -\sin\lambda & \cos\lambda & 0 \\ -\sin\phi\cos\lambda & -\sin\phi\sin\lambda & \cos\phi \\ \cos\phi\cos\lambda & \cos\phi\sin\lambda & \sin\phi \end{bmatrix} \tag{1.31}$$

The relationship between the covariance matrices is as follows:

$$\Sigma_{e/n/u} = \boldsymbol{R}\,\Sigma_{X/Y/Z}\,\boldsymbol{R}^t; \qquad \Sigma_{X/Y/Z} = \boldsymbol{R}^t\,\Sigma_{e/n/u}\,\boldsymbol{R} \quad (1.32 \text{ and } 1.33)$$

With regard to the rotation matrix in equation 1.31, longitude is counted 0° to 360° east from the Greenwich meridian, west longitude is a negative value, and latitude is counted positive north of the equator and negative south of the equator.

The GSDM 3-D Inverse

Given that point 1 is defined by $X_1/Y_1/Z_1$ and point 2 by $X_2/Y_2/Z_2$, matrix formulations of the 3-D geocentric coordinate inverse and covariance error propagation are as follows:

$$\begin{bmatrix}\Delta X\\ \Delta Y\\ \Delta Z\end{bmatrix} = \begin{bmatrix}-1 & 0 & 0 & 1 & 0 & 0\\ 0 & -1 & 0 & 0 & 1 & 0\\ 0 & 0 & -1 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}X_1\\ Y_1\\ Z_1\\ X_2\\ Y_2\\ Z_2\end{bmatrix}; \quad \Sigma_\Delta = \boldsymbol{J}_1\,\Sigma_{1\text{-}2}\,\boldsymbol{J}^t$$

(1.34 and 1.35)

The Jacobian matrix from equation 1.34 and the general covariance error propagation procedure (equation 1.35) are used to find the overall geocentric inverse covariance matrix as follows:

$$\Sigma_\Delta = \begin{bmatrix}-1 & 0 & 0 & 1 & 0 & 0\\ 0 & -1 & 0 & 0 & 1 & 0\\ 0 & 0 & -1 & 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\begin{bmatrix}\sigma^2_{X_1} & \sigma_{X_1Y_1} & \sigma_{X_1Z_1}\\ \sigma_{X_1Y_1} & \sigma^2_{Y_1} & \sigma_{Y_1Z_1}\\ \sigma_{X_1Z_1} & \sigma_{Y_1Z_1} & \sigma^2_{Z_1}\end{bmatrix} & \begin{bmatrix}\sigma_{X_1X_2} & \sigma_{X_1Y_2} & \sigma_{X_1Z_2}\\ \sigma_{Y_1X_2} & \sigma_{Y_1Y_2} & \sigma_{Y_1Z_2}\\ \sigma_{Z_1X_2} & \sigma_{Z_1Y_2} & \sigma_{Z_1Z_2}\end{bmatrix}\\ \begin{bmatrix}\sigma_{X_1X_2} & \sigma_{Y_1X_2} & \sigma_{Z_1X_2}\\ \sigma_{X_1Y_2} & \sigma_{Y_1Y_2} & \sigma_{Z_1Y_2}\\ \sigma_{X_1Z_2} & \sigma_{Y_1Z_2} & \sigma_{Z_1Z_2}\end{bmatrix} & \begin{bmatrix}\sigma^2_{X_2} & \sigma_{X_2Y_2} & \sigma_{X_2Z_2}\\ \sigma_{Y_2X_2} & \sigma^2_{Y_2} & \sigma_{Y_2Z_2}\\ \sigma_{Z_2X_2} & \sigma_{Z_2Y_2} & \sigma^2_{Z_2}\end{bmatrix}\end{bmatrix}\begin{bmatrix}-1 & 0 & 0\\ 0 & -1 & 0\\ 0 & 0 & -1\\ 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$$

(1.36)

Correlation between points 1 and 2 is described by the off-diagonal submatrices. Various accuracies are defined by a choice with regard to the use of the covariance matrix in equation 1.36. The matrix operation in equation 1.36 can be used to compute the following:

1. **Local accuracy** if the full covariance matrix is employed (relative accuracy based upon the quality of measurements connecting adjacent points)
2. **Network accuracy** if the correlation between points 1 and 2 is zero (relative accuracy based upon the combined quality of each point with respect to the network)

3. **P.O.B. accuracy** if the covariance matrix of point 2 is the only one used (relative accuracy based solely on the network quality of point 2)

Equation 1.36 is really the heart of the stochastic model portion of the GSDM. Yes, equations 1.32 and 1.33 can be used to convert absolute datum accuracy from one reference frame to another (geocentric and local), but, when looking at one point with respect to another, equation 1.36 offers important choices in answering the question "Accuracy with respect to what?" If a connecting measurement between two points has a smaller standard deviation than would be computed given no correlation between them, then the local accuracy of one point with respect to the other can be computed with statistical reliability. These tools give the spatial data user a number of choices and the option of computing the standard deviation of all subsequently derived quantities such as distance, direction, height, volume, and area. More details on spatial data accuracy are included in chapter 11.

BURKORD™: SOFTWARE AND DATABASE

The mathematical concepts and equations described and used in formulating the global spatial data model are all in the public domain. The phrase "global spatial data model (GSDM)" is generic. The term "BURKORD" has been trademarked (1) as the name of a software package that performs 3-D coordinate geometry and error propagation computations as described in this chapter, and (2) as the name of a 3-D database used by the BURKORD software. The end user is free to use the term "BURKORD™" as applied to the underlying database or to software obtained from Global COGO, Inc. However, anyone offering a product or services to others whose value relies upon or is enhanced by reference to or use of the BURKORD trademark will be expected to pay an appropriate licensing fee. Inquiries related to using the term "BURKORD™" should be directed to Global COGO, Inc., P.O. Box 3162, Las Cruces, NM 88003.

SUMMARY

The GSDM gives each user both control and responsibility. If good or bad information is used inappropriately, unreliable answers can be obtained. However, the opposite case is the important one. The GSDM defines a model and computational environment that can be used to manage spatial data efficiently. Each user has the option of establishing criteria that must be met before spatial data can be used for a given purpose. The concept of meta data is important in establishing and preserving the credibility of spatial data, but standard deviation (in any or all components) is a very efficient method for evaluating the quality of spatial data. Once the *X*/*Y*/*Z* position of a point is defined along with its variance-covariance matrix, the spatial data can be exchanged in a very compact format. The same solid geometry and error propagation equations for using such shared spatial data are equally applicable worldwide, and the mathematical procedures are already proven and individually implemented. Using the GSDM is primarily a matter of choosing to do so.

REFERENCES

Bomford, G. 1971. Cartesian coordinates in three dimensions. In *Geodesy*, 3rd ed., appendix C. Oxford: Oxford University Press.

Burkholder, E. 1991. Computation of horizontal/level distances. *Journal of Surveying Engineering* 117 (3): 104–16.

———. 1993a. Using GPS results in true 3-D coordinate system. *Journal of Surveying Engineering* 119 (1): 1–21.

———. 1993b. Design of a local coordinate system for surveying, engineering, and LIS/GIS. *Surveying & Land Information Systems* 53 (1): 29–40.

———. 1997a. The 3-D azimuth of GPS vector. *Journal of Surveying Engineering* 123 (4): 139–46.

———. 1997b. *Definition and description of a global spatial data model (GSDM).* Washington, DC: U.S. Copyright Office. http://www.globalcogo.com/gsdmdefn.pdf.

———. 1999. Spatial data accuracy as defined by the GSDM. *Surveying and Land Information Systems* 59 (1): 26–30.

———. 2004. *Fundamentals of spatial data accuracy and the global spatial data model (GSDM).* Washington, DC: U.S. Copyright Office.

Davis, R., F. Foote, J. Anderson, and E. Mikhail. 1981. *Surveying: Theory and practice*. New York: McGraw-Hill.

Holland, P., M. Reichardt, D. Nebert, S. Blake, and D. Robertson. 1999. The Global Spatial Data Infrastructure initiative and its relationship to the vision of a digital Earth. Paper presented by Peter Holland at the International Symposium on Digital Earth, Beijing, China, 29 November–2 December.

Leick, A. 2004. *GPS satellite surveying*, 3rd ed. New York: John Wiley.

Mikhail, E. 1976. *Observations and least squares*. New York: Harper & Row.

Moritz, Helmut. 1978. Definition of a geodetic datum. In *Proceedings of the Second International Symposium on Problems Related to the Redefinition of North American Geodetic Networks*, Arlington, VA, hosted by the U.S. Department of Commerce; Department of Energy, Mines & Resources (Canada); and Danish Geodetic Institute. Washington, DC: U.S. Department of Commerce

National Imagery and Mapping Agency. 1997. Department of Defense World Geodetic System 1984: Its definition and relationships with local geodetic systems. Technical Report 8350.2. Fairfax, VA: National Imagery and Mapping Agency.

Seeber, G. 1993. *Satellite geodesy*. Berlin: Walter de Gruyter.

Soler, T., and L. Hothem. 1988. Coordinate systems used in geodesy: Basic definitions and concepts. *Journal of Surveying Engineering* 114 (2): 84–97.

Trimble. 1990. *Trimnet: Survey network software user's manual*. Sunnyvale, CA: Trimble Navigation.

Whitehorn, K. L. 1997. *The Minnesota county coordinate system: A handbook for users*. St. Cloud, MN: Precision Measuring Systems.

Wisconsin. 1995. *Wisconsin coordinate system*. Madison: Wisconsin State Cartographer's Office.

Wolf, P. R., and C. D. Ghilani, 1997. *Adjustment computations: Statistics and least squares in surveying and GIS*. New York: John Wiley.

Zilkoski, D., J. Richards, and G. Young. 1992. Results of the general adjustment of the North American Vertical Datum of 1988. *Surveying and Land Information Systems* 52 (3): 133–49.

2 Spatial Data and the Science of Measurement

INTRODUCTION

Many disciplines work with spatial data, and many people use a GIS to reference geospatial data. Starting with a concise definition of spatial data, this chapter describes how spatial data and their accuracy are related to the measurement process and one's choice of a measurement system. The goal is to describe how 3-D spatial data can be manipulated more efficiently and how spatial data accuracy can be established without ambiguity using the GSDM as the foundation for GIS's and the GSDI.

Modern practice and instruments are used to collect and record spatial measurements. These data are processed electronically, and digital results are stored in computer files. Paper maps are inherently two-dimensional (they flatten the Earth), and humans traditionally view spatial relationships in terms of "horizontal" and "vertical." Computer graphics and data visualization procedures offer an endless array of display options. Although not exhaustive, this chapter summarizes characteristics of pertinent coordinate systems, defines spatial data, and looks at measurement processes by which spatial data are generated. Today, 3-D digital spatial data are more appropriately stored in a database that combines horizontal and vertical into a single database. And, as discussed later in this chapter, differences between how spatial and geospatial data are used may become significant. The accuracy of spatial data is also considered, and an important distinction is made between primary and derived spatial data.

SPATIAL DATA DEFINED

Use of the GSDM can foster greater insight into the relationships between coordinate systems and how they are used to handle spatial data. Spatial data are **described** as those numerical values that represent the location, size, and shape of objects found in the physical world. Examples include points, lines, directions, planes, surfaces, and objects. For the purposes of this book, spatial data are **defined** as the distance between endpoints of a line in Euclidean space (see the definition of a point in chapter 3). The endpoints may be nonphysical entities such as an origin or a specific location on the axis of a coordinate system. An endpoint may also represent the location of some physical feature such as a survey monument, building corner, benchmark, or other object. Geometrical elements such as planes, surfaces, and other objects formed by the movement and aggregation of distances also qualify as spatial data. Although straight-line distances are generally presumed, the measure of a distance can also be along a curved line, in either linear or angular units, without violating the definition. As used here, the definition of spatial data also includes, but is not limited to, geospatial data.

COORDINATE SYSTEMS GIVE MEANING TO SPATIAL DATA

When working with spatial data, assumptions are made about the underlying coordinate system. Since each reader deserves to know at all times "with respect to what," an attempt is made to be very specific about the underlying coordinate system and whether the spatial data are absolute or relative. As a matter of convention, **absolute** spatial data are taken to be data with respect to a defined coordinate system, while **relative** spatial data are taken to be the difference between two absolute values in the same system. A coordinate is an absolute distance with respect to the defined coordinate system, and an azimuth is an absolute direction with respect to the zero reference. Spatial data components are coordinate differences (in the same system) and are used as relative values. An angle, defined as the difference between two directions, is also a relative value. Absolute data are often used to store spatial information, while relative data are more often associated with measurements.

Admitting the use of undefined terms, relying upon prior knowledge, and acknowledging a difference between a reference system and a reference frame, the information presented in this chapter is intended to be consistent with current definitions of coordinate systems, such as those described by Soler and Hothem (1988). Three coordinate systems are an integral part of the GSDM.

1. **ECEF**: A functional 3-D geocentric coordinate system for spatial data is called the Earth-centered Earth-fixed (ECEF) rectangular Cartesian coordinate system and defined by the National Imagery and Mapping Agency (NIMA; 1997). (See Figure 2.1.) With its origin at the Earth's center of mass, the *X/Y* plane is coincident with the Earth's equator, and the Z-axis is defined by the location of the Conventional Terrestrial Pole (CTP). The X-axis is defined by the arbitrarily fixed location of the Greenwich meridian, and the Y-axis is at longitude 90° east, giving a right-handed coordinate system.
2. **Geodetic**: A geodetic coordinate system (Figure 2.2) is used to reference spatial data by geodetic positions on the ellipsoid, a mathematical approximation of the Earth's surface. Position is defined in the north-south direction by angular units (degrees, minutes, and seconds) of latitude and in the east-west direction by angular units of longitude. Lines of equal latitude are called parallels, and lines of equal longitude are called meridians. The sign convention for latitude is positive north of the equator and negative south of the equator. The sign convention for longitude is positive eastward for a full circle from 0° on the Greenwich meridian to 360° (arriving again on the Greenwich meridian). A west longitude, as commonly used in the western hemisphere, is acceptable and mathematically compatible if used as a negative value.

Geodetic latitude and longitude are 2-D curvilinear coordinates given in angular units. The third dimension, ellipsoid height, in this worldwide coordinate system is the distance above or below the mathematical ellipsoid and is measured in length units, meters being the international standard. With the conceptual separation of horizontal and vertical, this system of geodetic coordinates more closely matches physical reality in a global sense than

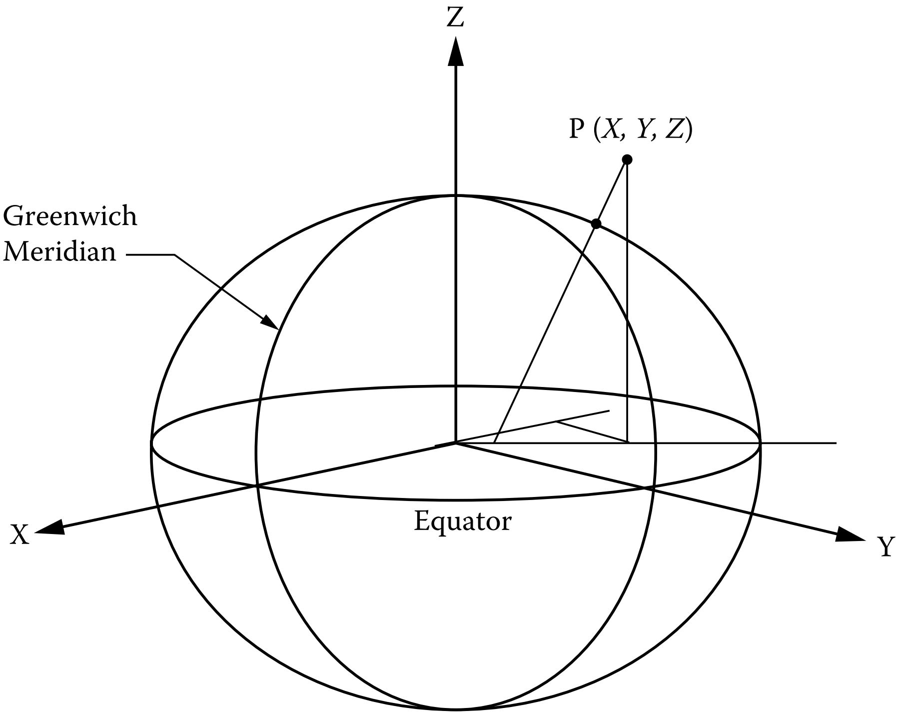


FIGURE 2.1 Geocentric ECEF Coordinate System

does the ECEF system and remains very useful for cartographic visualizations. But, the geodetic coordinate system is computationally more complex and more cumbersome to use than rectangular components when working with 3-D spatial data.

3. **Local**: Local coordinate systems (Figure 2.3) portray the location of spatial data with respect to some user-specified reference and/or origin. A local coordinate system can be defined such that horizontal and vertical relationships are both accurately portrayed and 3-D relationships are preserved. However, many local coordinate systems enjoy true 3-D geometrical integrity only to the extent that a flat Earth can be assumed. If spatial data issues are addressed strictly on a local basis, the error caused by such flat-Earth assumptions can be negligible. However, as one works over larger areas, needs greater precision in small areas, or needs to establish compatibility between local coordinate systems, the flat-Earth model is not adequate for referencing spatial data. But, when used as a component of the GSDM, the local flat-Earth model can support visualization and use of 3-D data without being adversely affected by the underlying curved-Earth distortions. That means local rectangular (flat-Earth) relationships can be utilized in a global environment without compromising the geometrical integrity of spatial data.

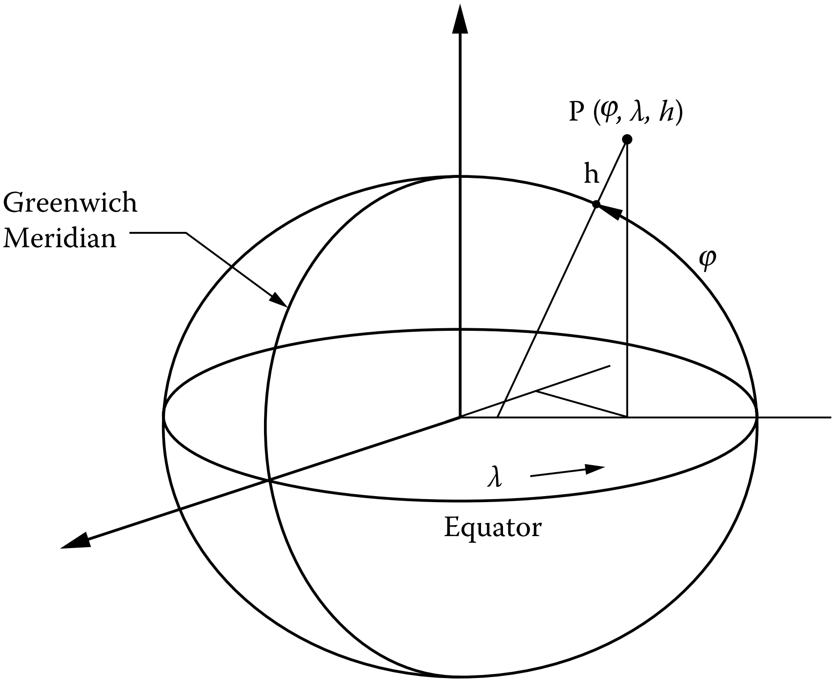


FIGURE 2.2 Geodetic Coordinate System

Spatial Data Types

Given descriptions of the geocentric ECEF coordinate system, the geodetic coordinate system, and a local coordinate system, the following spatial data types are listed:

1. Absolute geocentric *X*/*Y*/*Z* coordinates are perpendicular distances in meter units from the respective axes of an ECEF reference system.
2. Absolute geodetic coordinates of *latitude*/*longitude*/*height* are derived and computed from ECEF coordinates with respect to some named model (geodetic datum).
3. Relative geocentric coordinate differences, $\Delta X/\Delta Y/\Delta Z$, are obtained by differencing compatible geocentric *X*/*Y*/*Z* coordinate values.
4. Relative geodetic coordinate differences, $\Delta\varphi/\Delta\lambda/\Delta h$, are obtained as the difference of compatible (common datum) geodetic coordinates.
5. Relative local coordinate differences, $\Delta e/\Delta n/\Delta u$, are local components of a space vector defined by relative geocentric coordinate differences.
6. Absolute local coordinates, *e*/*n*/*u*, are distances from some origin whose definition may be mathematically sufficient in 3-D, 2-D, or 1-D. Examples are as follows:
   - Point-of-beginning (P.O.B.) datum coordinates, as defined in chapter 1. These derived coordinates enjoy full mathematical definition in 3-D,

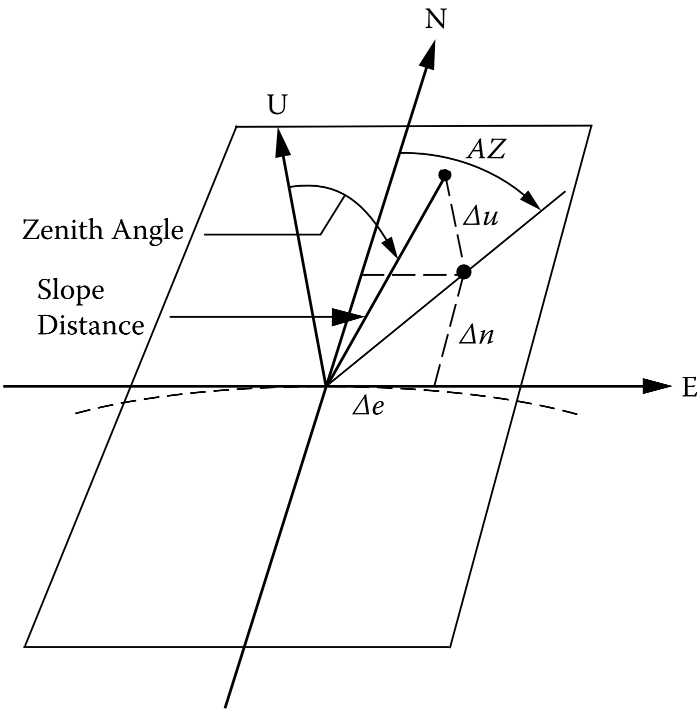


FIGURE 2.3 Local Coordinate System

suffer no loss of geometrical integrity in the GSDM, and serve the local needs of many spatial data users.

- Map projection (state plane) coordinates, which are well defined in 2-D with respect to some named origin and geodetic datum.
- Elevations, which are 1-D distances above or below some named reference equipotential surface. In the past, mean sea level was assumed to be acceptable as a vertical reference, but, due to the difficulty of finding mean sea level precisely, modern vertical datums are referenced to an arbitrary equipotential reference surface (Zilkoski, Richards, and Young 1992).

7. Arbitrary local coordinates may be 1-D (assumed elevations), 2-D (assumed plane coordinates), or 3-D (spatial objects, rectangular coordinates, or assumed elevations and plane coordinates). Although useful in some applications, arbitrary local coordinates are generally not compatible with other local systems and have limited value in the broader context of georeferencing. Many computer graphics and data visualization programs use arbitrary local coordinates.

The GSDM efficiently handles spatial data that fall into categories 1, 3, and 5 (absolute geocentric coordinates, relative geocentric coordinate differences, and relative local coordinate differences). Spatial information is stored most efficiently using digital geocentric coordinates, manipulated most readily using geocentric coordinate differences, and displayed for human visualization and analysis using relative local coordinate differences. Spatial data consisting of geodetic coordinates and geodetic coordinate differences (categories 2 and 4) are useful for cartographic portrayal and, to the extent they can be competently related to category 1, generally are not a problem. Category 6 spatial data (local coordinate differences) can be incorporated into the GSDM if and only if they enjoy full 3-D mathematical definition. Without additional survey measurements, attempts to incorporate category 7 data into the GSDM are not viewed as fruitful. This is where the difference between spatial and geospatial data definitions may become significant.

SPATIAL DATA VISUALIZATION IS WELL DEFINED

Spatial data are used extensively in computer graphics, visualization programs, computer-aided design and drafting, and the manipulation of spatial objects. The GSDM provides for the connection of spatial data to the physical Earth, but otherwise makes no attempt to impose conditions on the use of spatial data. It is anticipated that the scope, utility, and value of many spatial data manipulation, visualization, and 3-D coordinate geometry (COGO) programs may be enhanced by taking advantage of the physical Earth connection as defined by the GSDM.

DIRECT AND INDIRECT MEASUREMENTS CONTAIN UNCERTAINTY

Spatial data are created by measurement, and no measurement is perfect. In a simple case, a distance is determined by direct comparison of some unknown length with a standard such as a ruler, steel tape, or wavelength. Whether the distance is horizontal or vertical is a condition noted by the person recording the observation. More often, however, spatial data are obtained as the result of an indirect measurement in which one or more spatial data components are computed from the observations, as is the case when a slope distance is resolved into its horizontal and vertical components. In other cases, some physical quantity is observed and a distance is computed using known mathematical relationships (a model). An example is computing a distance from a voltage, which represents the phase shift of a sine wave signal in an electronic distance-measuring (EDM) instrument. Restating, spatial data measurements may be the result of a direct comparison, or, more often, they are computed indirectly from observations of various fundamental physical quantities.

Fundamental Physical Constants Are Held Exact

Fundamental physical quantities as expressed in the International System (SI) are as follows:

Length: meter
Time: second
Mass: kilogram
Current: ampere

Temperature: Kelvin
Luminous intensity: candela
Amount of substance: mole

Derived physical quantities include the following (there are others—see Nelson 1999):

Frequency: hertz
Force: newton
Pressure: pascal
Energy: joule
Power: watt
Electric charge: coulomb
Electric potential: volt
Plane angle: radian
Solid angle: steradian

Measurements Contain Errors

Spatial data are created by measurement of some combination of physical quantities, and those measurements are used in models that relate the observed quantity to a physical distance (spatial data) relative to one of the three coordinate systems listed earlier. The accuracy of such spatial data is dependent upon (1) the quality and sufficiency of the measurements, (2) the appropriateness of the models used to compute the spatial data components, and (3) error propagation computations. The GSDM accommodates all three considerations.

MEASUREMENTS USED TO CREATE SPATIAL DATA INCLUDE ...

Taping

A calibrated tape is laid flat on a horizontal surface at some specified tension and temperature. The measurement involves a visual comparison of the unknown length with uniform markings on the tape (a fundamental physical quantity). The observation is recorded as a measurement. If the temperature (another physical quantity) is different than the specified calibration temperature or if the tension (a derived physical quantity) is not what it should be, these other measurement conditions must also be noted. Using this additional information and appropriate equations, corrections to the taped distance are computed and applied to this otherwise direct measurement. Whether the computed distance is a direct or an indirect measurement is left to the reader.

Leveling

A level rod with graduations marked on it is held erect in the field of view of an observer looking through the telescope of an automatic (or tilting) level, and the distance from the bottom of the rod to the cross-hair intercept is read and recorded. Separate readings are made with the rod resting on other objects. In this case, the difference of two direct readings provides an indirect determination of the relative heights of the two objects. Among others, the accuracy of such an indirect measurement is affected by (1) whether the line of sight is perpendicular to the plumb line, (2) the presence of parallax, (3) whether or not a vernier was used to refine the reading, (4) the plumbness of the level rod when the readings were made, and (5) the distance from the instrument to the rod (curvature and refraction correction). Modern bar-scale reading instruments are becoming commonplace.

Electronic Distance Measurement

An electronic distance-measuring (EDM) instrument emits electromagnetic radiation, which is modulated with a known frequency (giving a known wavelength). The signal is returned by a retro-reflector from the forepoint end of a line, and the phase of the returned waveform is electronically compared to that of the transmitted signal. The measurement of phase differences on several modulated frequencies provides information used to compute the distance between the EDM and a reflector. Other quantities such as temperature and barometric pressure are also measured to determine corrections that account for the signal traveling through the atmosphere between the standpoint and forepoint at some speed slower than it would have traveled through a vacuum. The point is that several physical quantities are measured and the physical environment is modeled with equations before a collection of observations can be converted into spatial data.

With later-generation pulse laser instruments and scanners, the physical distance between EDM and object is determined using the time interval required for a pulse to travel from EDM to object and back. Of course, atmospheric delay must be modeled and direction to the target must be known before a slope distance can be resolved into rectangular components.

Angles

Although not a fundamental physical quantity, angles are commonly measured and used in computing spatial data components. Two examples are (1) using a vertical angle to resolve a slope distance into horizontal and vertical rectangular components, and (2) using the bearing of a line to find the latitudes and departures of a traverse course. Looking beyond the obvious where an angle is measured directly with a protractor on paper or on the ground using a total station surveying instrument, angles are also measured indirectly as the difference of two directions such as might be observed with a compass, a gyroscope, or GPS. Whether an angle was measured in the horizontal, vertical, or some other plane is also an important consideration, especially when using angles to resolve the hypotenuse of a triangle into its rectangular components. Two examples are resolving slope distances into horizontal and vertical and resolving traverse courses into latitudes and departures.

GPS

GPS is very versatile in that several kinds of fundamental observations can be used to determine spatial data quantities. An oversimplified view of GPS measurements includes three concepts: (1) distance is the product of rate and a time interval obtained from code phase observations, (2) the Doppler shift of a frequency recorded on the ground as compared to the frequency transmitted by a satellite, and (3) interferometric interaction of the carrier phase signal as recorded simultaneously at two different antennas (carrier phase observations). Without going into detail, the point is that portable handheld code phase GPS equipment routinely determines *absolute* geocentric ECEF coordinates and converts them to absolute geodetic coordinates before displaying them. The accuracy of code phase instruments is typically less than that obtained using carrier phase instruments. GPS carrier phase data must be collected at two points simultaneously and can be processed to give a very precise 3-D space vector between two antennas in terms of *relative* geocentric coordinate differences. The relative accuracy of such GPS observed vectors (with operator care and diligence) routinely exceeds one part in a million. If such a vector is attached to a known control point, a precise 3-D position of the second point can be easily computed.

This brief description implies that GPS positions determined using portable handheld equipment are not as accurate as those collected using a receiver mounted on a tripod. That may be, but is not necessarily, true because whether or not a GPS receiver collects code phase or carrier phase data is not determined by whether or not it is portable. For example, differential corrections (from a base station) may be used to improve upon the accuracy of code phase solutions, and radio connections between portable carrier phase receivers mean that relative differences may be obtained in real time (such as real-time kinematic, or RTK). Even though the two GPS technologies are quite different, in either case, once a GPS position is determined, answers can be viewed in a coordinate system of the user's choice.

Photogrammetric Mapping

Relative spatial data, both local and geocentric, can be determined efficiently and precisely using geometrical relationships reconstructed from stereoscopic photographs of a common image. A photogrammetric measurement is the relative location of an identifiable feature on a photographic plate with respect to fiducial marks on the same plate as determined with a comparator. A more complex measurement of 3-D spatial relationships based upon principles of photogrammetry requires mechanical reconstruction of the stereoscopic image by achieving the proper relative and absolute orientation of the stereo photographs in a mechanical stereo plotter. A 3-D contour map of the ground surface is the end result of the plotting operation. That traditional photogrammetric mapping process has been computerized and automated and now comes under the banner of softcopy photogrammetry. The end result of the modern computerized process is a 3-D digital model of the terrain. Hardcopy maps, computer displays, and other products, both digital and analog, are made from a common digital spatial data file.

Remote Sensing

The American Society of Photogrammetry and Remote Sensing (1984) describes remote sensing as the process of gathering information about an object without touching or disturbing it. Photogrammetry is an example of remote sensing, and ray tracing based upon stereo photographs of a common image is very geometrical. Bethel (1995) also discusses remote sensing and describes interpretative (less metrical) applications of remote sensing, which include use of nonvisible portions of the electromagnetic spectrum to record the response of an object or organism to stimuli from a distant source. Sensors include infrared film, digital cameras, radar, satellite imagery, and so on, and information is stored pixel by pixel in a raster format. Determining the unique spatial location represented by each pixel can be a daunting challenge and typically requires enormous storage capacity.

Other measurement methods are also used to create spatial data. But, regardless of the technology used to measure fundamental physical quantities, the GSDM provides a common universal foundation for expressing fundamental spatial relationships. Various equations (models) are used to convert observations into spatial data components, which are then used as measurements in subsequent operations, for example traverse computations, network adjustments, and plotting maps. The GSDM also accommodates fundamental error propagation in all cases, and that information, if available, is stored in the covariance matrix for each point. From there, each user can make informed decisions about whether or not the spatial data accuracy is sufficient to support a given application.

Errorless Spatial Data Must Also Be Accommodated

Several cases exist in which spatial data are considered errorless. Examples include (1) spatial data created during the design process, (2) physical dimensions (such as the width of a street right-of-way) defined by ordinance or statute, and (3) spatial data whose standard deviations are small enough to be judged insignificant for a given application. In the case of a proposed development such as a highway, bridge, skyscraper, or other civil works project, the planned location of a feature and the numbers representing the size and shape of each feature qualify as spatial data. But, they are the result of a design decision instead of a measurement process. Such design dimensions are without error until they are laid out. After being laid out and constructed, the location of the feature or object is determined by measurement and typically recorded on as-built drawings or in project files. Considered that way, the perfection of a design dimension is transitory and ceases to exist when laid out during construction.

An exception to the transitory nature of an errorless design dimension exists when a dimension is established by ordinance or statute. Such a dimension may be fixed by law, but the physical realization of that dimension is still subject to the procedures and quality of measurements used to create it. Under ideal conditions there will be no conflict between a statutory dimension and its subsequent remeasurement if the layout process was more accurate or reliable than the measurements made to document its location. For example, a 100-foot right-of-way may have been monumented very carefully and current measurements between the monuments are all 100.00 feet, plus or minus 0.005 feet. In that case, the right-of-way width can be shown as 100.00 feet, measured and recorded. Under less-than-ideal conditions, several possible dilemmas are as follows:

1. The right-of-way monuments really are separated by 100.00 feet, but the survey is based upon a state plane coordinate grid and the measured grid distance is 99.97 feet, plus or minus 0.005 feet (possible at elevations over 4,200 feet). Understandably, the monuments are not to be moved, so they are separated by a grid distance of 100.00 feet, but some users may not be willing to accept the implication that a foot is not really a foot. The apparent discrepancy arises from the use of two different definitions for horizontal distance, local tangent plane distance or grid distance (Burkholder 1991).
2. The right-of-way monuments appear stable and undisturbed, but the measurement between them is consistently 99.96 feet, plus or minus 0.005 feet (it could happen if the monument locations were staked during cold weather and no temperature corrections were applied to the steel tape measurements at the time). The conflicting principles are that the statutory dimension (of 100.00 feet) must be honored and that the original undisturbed monument controls, even if originally located with faulty measurements.

The intent here is not to solve those problems, but to acknowledge the potential for conflicting principles when working with so-called errorless spatial data. Other authors have written entire textbooks devoted to survey law, evaluation of evidence, and analysis of survey measurements. The point here is that the GSDM offers a consistent standard environment in which to make comparisons between conflicting data. The GSDM does not distort horizontal distances as does the use of map projections and/or state plane coordinates.

When the coordinates of any point are held "fixed," the result is the same as assuming the standard deviations associated with the coordinates are very small or are zero. In many cases, such an assumption is reasonable and defensible because the standard deviations of a point are small enough to be insignificant and the implied statement "with respect to existing control" is acceptable. However, each spatial data user making decisions about which control points are held "fixed" should document such decisions specifically so that subsequent users may always know "with respect to what." With the accuracy of spatial data becoming ever more important, criteria for judging the quality of spatial data should be unambiguous and easy to understand. The stochastic model portion of the GSDM uses 3-D standard deviations to describe the accuracy of spatial data and accommodates errorless spatial data as those data having zero standard deviations (Burkholder 1999). An answer to the question "Accuracy with respect to what?" is determined by (1) the control points used as primary data, (2) the 3-D standard deviations of those control points, (3) the quality of measurements used to establish new positions, (4) competent determination of the covariance matrix of each new point, and (5) the manner in which equation 1.36 is used in subsequent computations.

PRIMARY SPATIAL DATA ARE BASED UPON MEASUREMENTS AND ERRORLESS QUANTITIES

Earlier, spatial data were defined as distances. Spatial data types were also listed as distances represented by coordinates or coordinate differences in one of several coordinate systems. And, unless attempting to convert from one datum to another (see chapter 7), it should be understood that equations for converting spatial data from one coordinate system to another should have little or no mathematical uncertainty associated with them. Any uncertainty should be the result of an imperfect measurement, not a defective model, equation, or algorithm. With that said, primary spatial data are defined as geocentric *X*/*Y*/*Z* coordinates, their associated covariance matrices, and point-pair correlation matrices. Primary spatial data are created by a specific measurement process or determined on the basis of some prescribed geometry. Measurements have standard deviations and covariances associated with them, while errorless quantities have zero standard deviations. The GSDM accommodates both measurements and errorless quantities by using standard deviations of all three components at each point, covariances between components, and correlations between points.

Observations and Measurements

Mikhail (1976) describes how measurement and observation are very similar and, in fact, are used interchangeably. A mathematical distinction made here is that observations are independent, while measurements may be correlated. Stated differently, an observation (whether it is the process or the numerical outcome) is taken to be the actual comparison of some quantity with a standard, while a measurement is taken to be the same as either an observation or a subsequently computed quantity after corrections are applied as dictated by observation conditions. For example, a horizontal distance is said to be measured by an EDM. Actually, an EDM uses (1) the observed phase difference of two electromagnetic signals on several frequencies (the transmitted and received signal, and an internal reference signal), (2) the estimation (observation) of air temperature and barometric pressure for the atmospheric correction, and (3) the measurement of the vertical (or zenith) angle. These observations are used to compute the horizontal distance, which is called a measurement, when, in fact, physical quantities other than length were observed. Also note that the same observations are used to compute the vertical component of the slope distance. If one of the observations is changed, it may affect both computed values. Hence, the horizontal and vertical measurements are correlated and not independent. Slope distance and zenith directions are the independent observations.

Having made a distinction between observations and measurements, several other points also need to be made:

- In the strictest sense, primary spatial data should include only errorless quantities and independent observations. However, given the multitude of sensors used to make observations and the number of steps often needed to convert observations into spatial data components (measurements), it would be onerous indeed for each spatial data user to assume responsibility for the integrity of his or her data all the way back to the observation. It is hereby suggested the GSDM will conveniently serve two distinct groups: those responsible for generating quality spatial data and those who use spatial data. The work of scientists, physicists, electrical engineers, and programmers is completed upon delivering a measurement system that can be used to generate quality spatial data. If cartographers, geographers, planners, and other spatial data users know they can rely on the quality of data provided, they need not be so concerned with the science of measurement, computations, and adjustments but are free to focus their energies on spatial analysis and other chosen applications. The geomatician (geodesist, surveying engineer, photogrammetrist, etc.) provides a valuable service to society by interacting with and serving both groups.

- All primary spatial data have covariance matrices associated with them. In the case of errorless quantities, the covariance matrix is filled with zeros. Otherwise, the covariance matrices are obtained by formal error propagation from basic observations through a competent network adjustment.
- Computation of measurements often results in correlation between computed spatial data components. That correlation is defined and determined by the error propagation computation procedure applied to independent observations and the mathematical equations used to obtain the spatial data components. For that reason, it is necessary to store the full (3×3 symmetric, six unique values) covariance matrix along with the $X/Y/Z$ coordinates of each point.
- As used here, errorless quantities and unadjusted measurements are the basis of primary spatial data. But, in reality, primary spatial data are the $X/Y/Z$ coordinates and associated covariance values stored following rigorous network adjustments and successful application of appropriate quality control measures.

A statement of the obvious is that primary spatial data having small standard deviations are more valuable than primary spatial data with large standard deviations. Whether a standard deviation is large or small is dependent upon the measurements made and the correct propagation of the measurement errors to the spatial data components. The GSDM handles 3-D spatial data the same way, component by component, regardless of the magnitude of the standard deviations, and each user has the option of deciding what level of uncertainty is acceptable for a given application. Additionally, the GSDM is strictly 3-D and makes no mathematical distinction between horizontal and vertical data. But, the GSDM readily provides local $\Delta e/\Delta n/\Delta u$ components that can be used locally as flat-Earth (local tangent plane) distances.

DERIVED SPATIAL DATA ARE COMPUTED FROM PRIMARY SPATIAL DATA

Spatial data that owe their existence to mathematical manipulation of existing primary spatial data are considered to be derived spatial data. Derived spatial data include geodetic coordinates, UTM coordinates, state plane coordinates, project datum coordinates, and coordinates in other mathematically defined systems. Derived spatial data also include inversed bearings and distances (as shown on survey plats and subdivision maps), areas, volumes, and elevations. In each case, the accuracy of each derived quantity is computed from the standard deviations (and covariances) of the underlying primary spatial data on which they are based.

A clear distinction between primary spatial data and derived spatial data is critical to the efficient collection, storage, management, and use of spatial data. Primary spatial data require measurement of physical quantities and computation of spatial data components according to very specific procedures. For example, taping corrections are needed to determine a precise horizontal distance measured with a steel tape, and an EDM measurement needs to be corrected for reflector offset, atmospheric conditions, and slope if the endpoints are at different elevations. The cost of acquiring primary spatial data is still prohibitive in some cases, but, due to automation, computerization, and other technical developments (such as GPS), spatial data are much less expensive to obtain now than in the past. Even so, by comparison, derived spatial data are generally much less expensive than primary spatial data. Derived data can be computed, used, and discarded without detrimental economic consequence. Other than the effort required to assemble the needed primary spatial data and to make the computations, derived spatial data can be generated as often as needed based upon a prescribed algorithm. The challenge is to archive primary spatial data efficiently and to make sure other users know specifically what algorithms were used in generating the derived quantities. The practice of storing derived spatial data is potentially wasteful.

ESTABLISHING AND PRESERVING THE VALUE OF SPATIAL DATA

Establishing the value and integrity of spatial data is not a trivial undertaking. An oversimplified statement is that the right measurement needs to be made with the correct equipment under well-documented conditions and that appropriate equations must be used to compute primary spatial data components. Once the components are run through an appropriate least squares network adjustment, they are attached to the chosen 3-D datum and the resulting geocentric *X*/*Y*/*Z* coordinates (and covariances) become the primary spatial data. A thorough treatise on the science of measurement and subsequent computation of spatial data components would fill an entire book. The goal in this chapter is to establish a connection between our measurements and spatial data components with the idea of showing how both can be handled more efficiently in the context of the GSDM.

But, in addition to the points made here about measurements and computations, another question also needs to be asked: "What makes spatial data lose their value?" Often attention is focused on doing whatever is necessary to get the most data for the least cost, but avoiding the cost of replacement or the inconvenience of not having needed data also needs to be considered. Therefore, efforts made to preserve the value of spatial data may be efforts well spent. Although less directly, the GSDM also facilitates those efforts by providing a simple standard model that can be used worldwide by all spatial data disciplines and users. The key concepts are data compatibility and interoperability.

Specifically, spatial data lose their value if

- potential users do not know that they exist or are available.
- these data are incomplete, incompatible, or of dubious quality.
- they are in the wrong format or stored in the wrong location.
- a user does not know what to do with them.
- they are replaced by data having smaller standard deviations.

SUMMARY

This chapter:

- Defines spatial data as the distance between points on a line.
- Describes three coordinate systems used to reference spatial data.
- Lists spatial data types as coordinates and coordinate differences.
- Acknowledges spatial data visualization is a solved problem.
- Admits that spatial data created by measurements contain uncertainties.
- Gives examples of nontrivial spatial data measurements.
- Describes the role of errorless spatial data.
- Defines primary spatial data and derived spatial data.
- Suggests ways to preserve the value of spatial data.

The remainder of this book:

- Reviews mathematical concepts needed for working with spatial data.
- Describes existing geometrical models used for spatial data computations.
- Develops geometrical relationships in the context of
    - geometrical geodesy.
    - physical geodesy.
    - satellite geodesy and GPS surveying.
    - geodetic datums.
    - map projections and state plane coordinate systems.
    - how spatial data are used.
- Shows how the GSDM accommodates 3-D coordinate geometry and error propagation.

REFERENCES

American Society of Photogrammetry and Remote Sensing. 1984. *Multilingual dictionary of remote sensing and photogrammetry*. Bethesda, MD: American Society of Photogrammetry and Remote Sensing.

Bethel, J. S. 1995. Surveying engineering. In *The civil engineering handbook*, ch. 53. Boca Raton, FL: CRC Press.

Burkholder, E. F. 1991. Computation of level/horizontal distance. *Journal of Surveying Engineering* 117 (3): 104–16.

———. 1999. Spatial data accuracy as defined by the GSDM. *Journal of Surveying & Land Information Systems* 59 (1): 26–30.

Mikhail, E. M. 1976. *Observations and least squares*. New York: Harper & Row.
Nelson, R. A. 1999. Guide for metric practice. *Physics Today: 1999 Buyer's Guide*, August, BG11–BG12.
National Imagery and Mapping Agency. 1997. Department of defense world geodetic system 1984: Its definition and relationships with local geodetic systems, 3rd ed., Technical Report 8350.2. Fairfax, VA: National Imagery and Mapping Agency (now the National Geo-spatial Intelligence Agency).
Soler, T., and Hothem, L. 1988. Coordinate systems used in geodesy: Basic definitions and concepts. *Journal of Surveying Engineering* 114 (2): 84–97.
Zilkoski, D., Richards, J., and Young, G. 1992. Results of the general adjustment of the North American Datum of 1988. *Surveying & Land Information Systems* 52 (3): 133–49.

3 Summary of Mathematical Concepts

INTRODUCTION

The term "mathematics" is difficult to define, in part because it includes so many concepts. Even so, the primary goal of this book is to organize mathematical concepts and geometrical relationships for the convenience of spatial data users. The approach is to start with simple, well-defined ideas and add understandable pieces as needed to develop tools for handling spatial data more efficiently. Mathematics has been concisely defined as the study of quantities and relationships through the use of numbers and symbols. The terms "quantities" and "relationships" may be somewhat abstract, but their meaning should become clearer with use. "Numbers" includes the set of all real values from negative infinity to positive infinity, and "symbols" includes letters of the alphabet (English, Greek, or otherwise) used to represent numerical values. Symbols also include other markings to indicate mathematical operations such as addition, subtraction, multiplication, division, and square roots. As illustrated by the definition and the two following examples, the goals of presenting simple well-defined concepts and keeping a focus on practical applications for spatial data will not be easy to achieve. Understandably, some readers will be distracted by temptations to pursue peripheral interests. Although that is acceptable and encouraged, space and print limitations do not permit joint excursions. The reader is always welcome back.

- With respect to numbers, few people (certainly not the author) can comprehend the vastness of infinity, yet reference is made to negative infinity and positive infinity with the implication that they might somehow be the same size (but in opposite directions on the real number line). The point is not whether that implication is true, but to note instead that there are just as many numbers between zero and one (0 and 1) as there are numbers greater than one. That statement is proved by taking the reciprocal of any number greater than 1.
- Symbols for mathematical operations such as +, –, ×, ÷, and √ are simple and are used the world over. Symbols also include letters that are used to represent certain numerical values. Perhaps the most common mathematical symbol is the Greek letter pi (π), used to represent the ratio of the circumference of a circle divided by its diameter. The definition is simple, and that ratio finds many applications when working with spatial data. However, mathematicians (in what could be called esoteric pursuits) have spent years chasing an increasing number of digits for pi (Beckman 1971). Access to millions of digits for π is now as simple as typing "pi" into a World Wide Web search engine.

CONVENTIONS

If you went looking for π, welcome back. In an effort to be specific about concepts and to communicate clearly using unambiguous symbols, the following conventions are identified and used throughout the book as consistently as possible even though they may differ from one discipline to the next or from one culture to another.

Numbers

Whole numbers are integers, and numerical values between integers are called real numbers. A line such as that shown in Figure 3.1 can be used to represent all numbers—both integer and real. If some point on the line is assigned the value "zero," points to the left of zero are negative numbers, and points to the right of zero are positive numbers. Numbers are composed of the Arabic digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9, and the Hindu-Arabic number system used in modern mathematics is decimally constructed by decades where the column of a digit implies its value times 1, 10, 100, 1,000, and so on.

Fractions

A ratio of one integer divided by any other (except 0) is known as a fraction. Many examples exist, but successive division by 2 is an intuitive example that serves very well when cutting a pie or a pizza. Fractions of 1/2, 1/4, 1/8, and so on are familiar to everyone and are used, for example, when driving a car and judging the amount of fuel remaining in the tank. Successive division by 2 is also appropriate in other cases, but when carried too far it becomes cumbersome and using decimal equivalents is easier. In the case of the U.S. Public Land Survey System (USPLSS), it is no coincidence that 1 square mile nominally contains 640 acres. Successive division of area by quarters (two divisions of length by two) is convenient down to a parcel of 10 acres. Although more could be said about repetitive division, it is noted that some disciplines in the United States still use fractions (e.g., architects, carpenters, millwrights, and ironworkers).

Decimal

Another prevalent practice for counting objects utilizes decades of 10, presumably based upon prehistoric humans having ten fingers. With the invention of "zero" by mathematicians in India about 600 a.d., the decimal system was developed in its

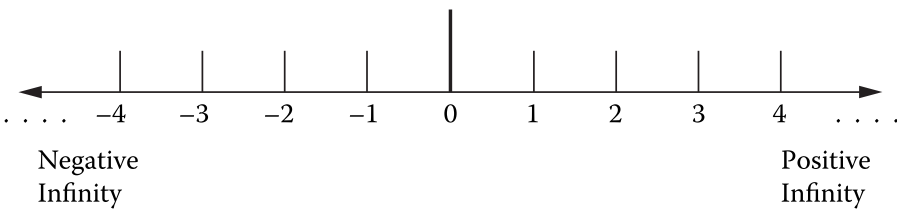


FIGURE 3.1 The Real Number Line

present form, borrowed by the Arabs about 700 a.d., and subsequently adopted by European merchants. The decimal form conveniently handles numbers (both positive and negative, and both integer and real) of any size from the very large to the very small, and is used worldwide.

Spatial data users are probably more interested in the development of the meter as a decimal standard of length. The goal of the French Academy of Science in the 1790s was to devise a standard of length that could be duplicated in nature and that was decimally divided. The arc distance from the Earth's equator to the North Pole was determined as accurately as possible by means of a geodetic survey, and the result—5,130,766 toises—was defined to be exactly 10 million meters (Smith 1986; Alder 2002).

The decimal system is used in the International System of units (SI) adopted by the Eleventh General Conference on Weights and Measures in 1960 (Chen 1995, 2455). The SI system is a coherent system of units included in, or derived from, the seven independent SI base units of the meter, kilogram, second, ampere, degrees Kelvin, mole, and candela. One advantage of the decimal system is names for units that differ by a magnitude of 1,000, as shown in Table 3.1. Perhaps the names are best recognized when referring to computer speeds (megahertz), data storage (gigabytes), or very short periods of time (nanoseconds). While the SI system defines decimal subdivision for length (meters), time (seconds), and angles (radian), standard practice in many parts of the world still uses the sexagesimal division of angles (degrees, minutes, and seconds) and time (hours, minutes, and seconds).

Sexagesimal System

About 5,000 years ago, the Babylonians related 360° in a circle to 365 days in a year. Given that six circles will fit exactly around a seventh, a subdivision of 360° into six sectors of 60° each is plausible. Tooley ([1949] 1990) credits the Babylonians with subdividing both the sky into degrees and the day into hours. The sexagesimal system of minutes and seconds was applied to each, allowing stars of the night sky to be plotted in a consistent proportional manner. Wilford (1981) credits Ptolemy with subdividing the degree into 60 minutes and each minute into 60 seconds, while

TABLE 3.1
Prefixes for Numbers

1,000,000,000,000.	tera-	one trillion
1,000,000,000.	giga-	one billion
1,000,000.	mega-	one million
1,000.	kilo-	one thousand
1.	unit	one
0.001	milli-	one thousandth
0.000 001	micro-	one millionth
0.000 000 001	nano-	one billionth
0.000 000 000 001	pico-	one trillionth

Smith (1986) credits the Chinese with first using zero and a circle sexagesimally divided into degrees, minutes, and seconds. Regardless of the origin of the practice, the second is now the defining unit of time, and the sexagesimal system (60 seconds = 1 minute, 60 minutes = 1 hour, and 24 hours = 1 day) is used worldwide. The sexagesimal system is also widely used in angular measurement, but the radian (defined as an angle whose arc length equals its radius) is the defining unit for rotation in the SI system.

Binary System

Computer science professionals use combinations of zeros and ones (0's and 1's) to represent numbers, letters, and other symbols within a computer's memory. A string of eight bits (0's and 1's) is called a byte and can be used to represent up to 256 different items. The American Standard Code for Information Interchange (ASCII) uses combinations of seven of the eight bits to represent uppercase and lowercase letters of the English alphabet, digits 0–9, and other symbols as text characters. If ASCII characters were used to represent real numbers in a computer, it would be quite costly in terms of memory requirements. Therefore, numeric data are stored using a base 2 (binary) system that accommodates real numbers and integers, both positive and negative. Rather than exploring the details of the binary system, the point here is to recognize that numerical values, whether integer or real, are coded differently than text. In most cases, the user need not be concerned with computer binary operations because numeric input in decimal form is immediately converted to binary, and numeric output, unless specified otherwise by the user, is displayed or printed in conventional decimal format. It is interesting to note, however, that there are similarities between fractions (dividing by two) and binary arithmetic. For an interesting tongue-in-cheek discussion of the advantages of binary computations for survey measurements, see Stanfel (1994).

Conversions

The ability to use numbers is important, but the real meaning of any mathematical operation comes from knowing what physical quantities (i.e., units) are associated with the numbers. And, it is important to understand the use of ratios where the units cancel out and the relative size of the number is really the issue. The value of π is one example; trigonometric ratios comprise another. The solution of most problems involves a reasonable number of something (units) that can be understood. Conversion is the process whereby equivalency is established between seemingly unrelated physical quantities. Mathematical operations also apply to the units. An easy example might be area (m^2) = length (m) × width (m). Not as obvious, but not really that unusual, the volume of concrete in a hypothetical sidewalk is length (30 meters) × width (5 feet) × thickness (4 inches). Unit conversions are "exact" and are used as ratios to find the desired answer. Note how units cancel in the separate computations and make it easy to find the volume of concrete in either cubic yards or cubic meters.

$$Vol\ yd^3 = (30\ m)\left(\frac{39.37\ ft}{12.00\ m}\right)(5\ ft)(4\ inches)\left(\frac{1\ ft}{12\ inches}\right)\left(\frac{1\ yd^3}{27\ ft^3}\right) = 6.1\ yd^3$$

$$Vol\ m^3 = (30\ m)(5\ ft)\left(\frac{12.00\ m}{39.37\ ft}\right)(4\ inches)\left(\frac{1\ m}{39.37\ inches}\right) = 4.6\ m^3$$

Coordinate Systems

An origin and three mutually perpendicular axes that intersect at the origin define a generic Cartesian coordinate system, as shown in Figure 3.2. When studying 2-D phenomena, the X-axis and Y-axis are the ones most commonly used. In a three-dimensional system, the direction of the positive Z-axis for a right-handed coordinate system is given by the direction of the thumb on one's right hand when rotating the positive X-axis (one's index finger) toward the Y-axis (one's palm). An example of a left-handed coordinate system is given as *north/east/up*. The convention in this book is to use a right-handed rectangular Cartesian coordinate system whenever possible. That includes both *X/Y/Z* and *east/north/up*. Note that while labeling the axes of a coordinate system is really the prerogative of the user, the GSDM uses *X/Y/Z* for geocentric ECEF coordinates and uses *e/n/u* for local perspective coordinates.

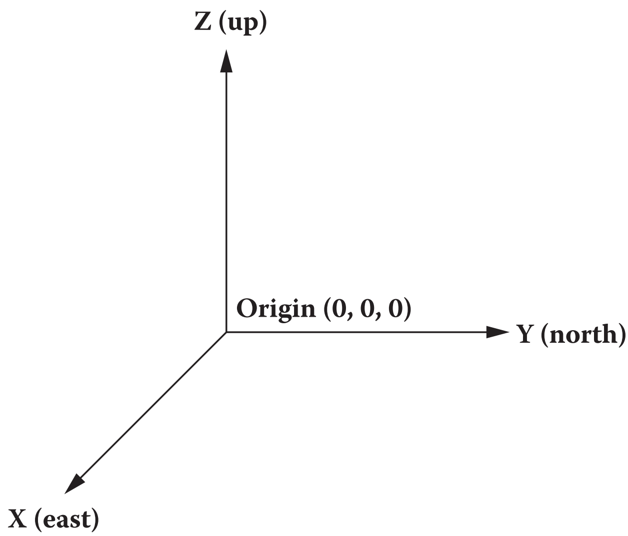


FIGURE 3.2 Three-Dimensional Rectangular Coordinate System

Significant Figures

The rules of significant digits are not arbitrary, but are based upon the principles of error propagation. If the validity of an answer obtained using significant figures is in doubt, the uncertainty of any answer can be verified using the standard deviations of measured quantities and error propagation computations.

- Integer values may have an infinite number of significant figures. For example, when dividing the area of a rectangle in half, 2 is an exact number. But, physical operations are not so precise. If one of two siblings cuts a piece of candy in half, the astute parent permits the second sibling to have first choice of the two pieces (in case one piece is larger).
- Conversions that are exact (12" = 1' or 27 ft^3 = 1 yd^3) contain an infinite number of significant figures when used as a ratio.
- With regard to controlling round-off error, it is recommended to carry at least one more digit in the computations than can be justified by the original data. The final answer of any computation should reflect the user's judgment with respect to significant figures.

Addition and Subtraction

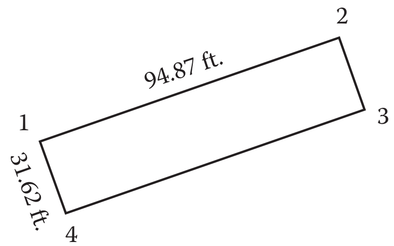
The column (decade) of the least accurate number in a sum (addition) or in a difference (subtraction) determines the number of significant digits in the answer. A zero listed after the decimal point is usually counted as significant. But, zeros that serve only to position the decimal point are generally not significant. Table 3.2 shows examples for addition and subtraction. For an exception, see the area computation in Figure 3.3 where $3{,}00\bar{0}$ ft^2 has four significant figures, not one. In such cases, some authors place a bar over the last significant digit.

Multiplication and Division

The number of significant digits in the product or quotient is determined by the term used in the operation containing the least number of significant digits. A product or quotient does not contain more significant digits than either term used to compute it. A simple example is area computed as length (L = 94.87') × width (W = 31.62'), which gives area = 2,999.7894 ft^2. An appropriate answer to four significant digits is $3{,}00\bar{0}$ ft^2.

TABLE 3.2
Significant Figure Examples Using Addition and Subtraction

Addition	Addition	Subtraction	Subtraction
120.	0.0023 435	1.00000000	5.44536724 * 109
+ 13.42	+ 0.101	– 0.00676865	– 5.43407673 * 109
130.	0.103	0.99323135	11,290,510.
2 s.f.	3 s.f.	8 s.f.	7 s.f.



PT #	X	Y
1	2,160,107.36 ft.	507,032.16 ft.
2	2,160,197.36 ft.	507,062.16 ft.
3	2,160,207.36 ft.	507,032.16 ft.
4	2,160,117.36 ft.	507,002.16 ft.

FIGURE 3.3 Rectangle for Area Computation

Area for the same rectangle (Figure 3.3) can be computed using the area-by-coordinates (cross multiplication) equation and the listed state plane coordinates of its corners. A common form of the area-by-coordinates equation is

$$A = \frac{1}{2}\left[\begin{array}{l}(Y_1 X_2 + Y_2 X_3 + Y_3 X_4 + Y_4 X_1) \\ -(X_1 Y_2 + X_2 Y_3 + X_3 Y_4 + X_4 Y_1)\end{array}\right] \tag{3.1}$$

Sum of Negative Products	Sum of Positive Products
$X_1Y_2 = 1.095308704 * 10^{12}$ ft^2	$Y_1X_2 = 1.095289533 * 10^{12}$ ft^2
$X_2Y_3 = 1.095289533$ "	$Y_2X_3 = 1.095359410$ "
$X_3Y_2 = 1.095229798$ "	$Y_3X_4 = 1.095248971$ "
$X_4Y_3 = 1.095248971$ "	$Y_4X_1 = 1.095179097$ "
$\text{Sum}_1 = 4.381077006 * 10^{12}$ ft^2	$\text{Sum}_2 = 4.381077011 * 10^{12}$ ft^2

Area is one-half the difference of the two sums:

$Sum_2 = 4{,}381{,}077{,}011{,}000$ ft^2
$Sum_1 = 4{,}381{,}077{,}006{,}000$ ft^2
$Difference = 5{,}000$ ft^2
$Difference/2 = 2{,}500$ ft^2 (Not Good! The correct answer is $3{,}00\bar{0}$ ft^2.)

What happened? The area equation is correct. The coordinates are good. A ten-digit calculator was used to compute the answer. But, the answer is obviously in error by 500 ft^2. The problem is one of significant figures. Two issues (innocent mistakes) are as follows:

- The *Y* coordinate contains only eight significant figures, yet each *XY* product lists ten significant figures. This is mistake number one and invalidates the computation.
- But, a second mistake (and separate issue) is that significant figures are lost in taking the difference of two large numbers of similar magnitudes. In this

case, only one significant figure remains after computing the difference of the two sums of products.

Both mistakes can be avoided by working with coordinate differences. By finding a coordinate difference first, the problem of working with large coordinate values is avoided because the cross products are formed using much smaller numbers. A derivation of the improved area equation 3.2 is given in Burkholder (1982), but it can also be obtained from equation 3.1 by moving the coordinate system origin to the first point of the figure (a fair amount of algebraic manipulation is required). Another advantage of equation 3.2 is that it can be extended easily for any number of points. A minimum of three points is required in the first line of equation 3.2. Beyond that, point 4 is brought into line 2, point 5 is brought into line 3, and so on. Note that with the orderly progression of subscripts, a program need only be written for the first line and used in a loop with updated subscripts until all points around the figure are used. Points used must be in graphical sequence to form a closed figure, and no line crossovers are permitted.

$$2A = (X_2 - X_1)(Y_3 - Y_1) - (Y_2 - Y_1)(X_3 - X_1) +$$

$$(X_3 - X_1)(Y_4 - Y_1) - (Y_3 - Y_1)(X_4 - X_1) +$$

$$(X_4 - X_1)(Y_5 - Y_1) - (Y_4 - Y_1)(X_5 - X_1) + \ldots$$

$$\text{for any number of points.} \qquad (3.2)$$

LOGIC

The following thirteen points on logic are adapted from Bumby and Klutch (1982, ch. 1).

1. A **statement**, also called an assertion, is any sentence that is true or false, but not both. The truth or falsity of a statement is called its "truth value."
2. Placeholders in mathematical sentences are called **variables**. An open sentence contains one or more placeholders. Variables are selected from a collection of possibilities called the **domain**, and the collection of all variables from the domain that make an open sentence true is called the **solution set**.
3. If a statement is represented by *p*, then *not p* is the **negation** of that statement. A negation of a negation is the same as the original: that is, *p* equals *not (not p)*.
4. A compound statement formed by joining two statements with the word "and" is called a **conjunction**. Each of the statements is called a **conjunct**.
5. A compound statement formed by joining two statements with the word "or" is called a **disjunction**. Each statement is called a **disjunct**.
6. A compound statement formed by joining two statements with the words "if … then" is called a **conditional**.

7. In logic, the symbol → is used to represent a conditional. Therefore, the conditional “if *p* then *q*” can be written as “$p \rightarrow q$.” Statement *p* is called the **antecedent** of the conditional, and statement *q* is called the **consequent**.
8. A **tautology** is a compound statement that is true regardless of the truth values of the statements of which it is composed.
9. The **converse** of a conditional is formed by switching the order of *p* and *q* of the original conditional.
10. The **inverse** of a conditional is formed by negating both the antecedent and the consequence.
11. The **contrapositive** of a conditional is formed by the inverse of the converse, that is, by negating both *p* and *q* and reversing their order.
12. A statement formed by the conjunction of the conditionals $p \rightarrow q$ and $q \rightarrow p$ is a **biconditional**. The phrases “necessary and sufficient” and “if and only if” are biconditionals.
13. Whenever two statements are always either both true or both false, the two statements are **equivalent**.

The following rules of logic, still applicable in problem solving and computer programming, were adopted by the French mathematician and philosopher René Descartes (Russell 1959):

- Never accept anything but clear distinct ideas.
- Divide each problem into as many parts as are required to solve it.
- Thoughts must follow an order from the simple to the complex. Where there is no order, one must be assumed.
- One should check thoroughly to assure no detail has been overlooked.

ARITHMETIC

Arithmetic consists of the manipulation of numbers by addition, subtraction, multiplication, division, and square roots to solve problems. Simple examples are as follows:

Addition: $8 + 4 = 12$ and $8 + (-4) = 4$
Subtraction: $8 - 4 = 4$ and $8 - (-4) = 12$
Multiplication: $8 \times 4 = 32$ and $8 \times (-4) = -32$
Division: $8 \div 4 = 2$ and $8 \div (-4) = -2$
Square root: $\sqrt{(64)} = 8$ and $\sqrt{(-64)}$ = undefined

ALGEBRA

Algebra is an extension of arithmetic that includes the use of letters to represent unknown numerical values. This permits a problem to be solved in a general form and for the solution of a specific problem to be obtained more efficiently by substituting the variables into an algebraic solution as opposed to re-solving the problem with different numbers on every occasion.

Rules associated with the properties and manipulation of algebraic quantities (Dolciani et al. 1967) are given below.

Axioms of Equality (for real numbers A, B, and C)

Reflexive property: $A = A$.
Symmetric property: if $A = B$, then $B = A$.
Transitive property: if $A = B$ and $B = C$, then $A = C$.

Axioms of Addition (for real numbers A, B, and C)

Associative Rule: $(A + B) + C = A + (B + C)$.
Existence of Identity: There is a unique number, zero, such that $0 + A = A$ and $A + 0 = A$.
Existence of Inverses: For each real number A, there is a number $-A$ such that $A + (-A) = 0$ and $(-A) + A = 0$.
Commutativity: For all real numbers A and B, $A + B = B + A$.

Axioms of Multiplication (for real numbers A, B, and C)

Associative Rule: $(A * B) * C = A * (B * C)$
Existence of Identity: There exists a unique number, 1, such that for every real number A, $1 * A = A$ and $A * 1 = A$.
Existence of Inverse: For every real number A (except 0), there is an element $1/A$ such that $A * (1 / A) = 1$ and $(1 / A) * A = 1$.
Note: This property is the basis of the rule that division by zero is undefined.
Commutativity: $A * B = B * A$.
Distributive: $A * (B + C) = (A * B) + (B * C)$ and $(B + C) * A = (B * A) + (C * A)$.

Boolean Algebra

Boolean algebra involves the use of logic conditions in which the result of some operation is assigned only one of two values, true or false. In terms of binary computer code, the conditions equate to 1 for true and 0 for false. Computer programmers make extensive use of Boolean algebra in writing computer programs and testing for differing conditions that may exist at the time a given part of the program is executed. Boolean algebra is beyond the scope of this book.

GEOMETRY

This book is written to define a model for spatial data on a global scale. Geometry is fundamental to that mission and includes the study of points, lines, circles, curves, planes, triangles, rectangles, cubes, spheres, and other objects. In chapter 2, spatial data were defined as the distance between endpoints of a line in Euclidean space. In order to provide additional clarification, the following elements are described.

Point

A point is a dimensionless quantity that occupies a unique location. The irony is that location cannot be defined without reference to (distance from) some other point. And, what does it mean? "A point has no dimension."

Distance

Distance is defined as the spatial separation of two objects (points) and has length as an attribute—dimension. Note that this is an example of circular logic because the concept of a point is used to define distance and the concept of distance (reference to axes) is used to define a point.

Dimension

Dimension, or the units of the quantity being measured, is another intuitive concept that seems to defy definition. A point has no dimension. A line has one dimension, surface area has two dimensions, volume has three dimensions, and space-time is generally considered to have four dimensions. According to Hawking (1988), anything more than four dimensions is a part of science fiction. But, he also describes "string theories" that accommodate many dimensions.

Line

As stated earlier, a line has length and can be described as the path of a moving point. Strictly speaking, a straight line is infinitely long but a line segment has two endpoints. A line can also be straight or curved. A straight line is the shortest distance between two points in Euclidian space, while a curved line has a radius associated with it. Understandably, a straight line can also be defined as a curve with an infinite radius. The slope-intercept form for the equation of a line is $y = ax + b$, where a is the slope and b is the intercept on the Y-axis (that place where the x value is zero).

Plane

A plane is two-dimensional, contains something called area, and is formed by the lateral movement of a straight line. More precisely, a plane is a flat surface defined by three noncollinear points. In three-dimensional space, a plane is also a flat surface that is perpendicular to a given straight line. Spatial data users should relate to the last definition because a horizontal plane is defined as perpendicular to the plumb line at a point. That is, humans stand erect and the perception is that we walk on a flat Earth. That is important because, in the GSDM, the origin moves with the observer and the model provides a view of all other points from that occupied or specified by the user.

Angle

Given that straight lines are infinitely long, it is said for Euclidean geometry that two lines lying in the same plane are parallel if and only if they never intersect (i.e., the distance between them never changes). If two lines in a plane are not parallel, they will intersect. An angle is defined as the geometrical shape formed by the intersection of two lines in a plane. Later, an angle will also be defined as the difference between two directions. If the four angles formed by the intersection of two lines are all the same size, it is said the two lines are perpendicular to each other and the four angles are all called right angles (i.e., one-fourth of a circle).

Circle

A circle is a closed figure formed by a uniformly curved line lying in a two-dimensional plane. Being uniformly curved, all points on the circle are the same radius distance from a common center point. The maximum distance from one side of the circle to the other is the diameter, or two radii. The angular measure of a full circle is one revolution, four right angles, 2π radians, or 360° (arbitrary units).

Ellipse

An ellipse is a continuous closed plane curve having a major axis dimension longer than its minor axis dimension. A unique characteristic of an ellipse is that the sum of distances from any point on the ellipse to each of the two foci located on the major axis is a constant whose value is twice the length of the semimajor axis. The ellipse becomes more nearly a circle as separation of the two foci becomes smaller and smaller.

Radian

One radian is the angle formed by an arc whose length is the same as the radius of the circle. Since there are 2π radians in a full circle of 360°, it is also popular to say 1 radian = $180°/\pi$ or 57° 17'44."8062… Another useful relationship is derived from the fact that there are 1,296,000 seconds of arc in a full circle. Dividing that value by 2π, there are 206,264.806247096355… seconds of arc per radian.

Triangle

A triangle is a three-sided figure in a two-dimensional plane. Said differently, a triangle is a closed figure in a plane formed by three line segments. The sum of the interior angles of any plane triangle is always 180°.

1. A right triangle has one right (90°) angle. In all seriousness, a student once asked, "What is a left triangle?" How should such a question be answered?
2. An equilateral triangle has three 60° angles.
3. An isosceles triangle has two equal angles. Sides opposite those angles are also equal. The third angle makes the sum of 180°.
4. An acute triangle is a triangle having three angles each less than 90°.
5. An obtuse triangle is a three-sided figure having one angle greater than 90°.

Quadrilateral

A quadrilateral is any four-sided figure in a plane bounded by straight lines. If additional conditions are imposed (as noted following), special names can be used.

Rectangle

A rectangle is a quadrilateral having four right (90°) angles. That means opposite sides are parallel and the same length. Any rectangle divided by a line through opposite corners gives two right triangles.

Square

A square is a rectangle with all sides being the same length.

Trapezoid

A trapezoid is any quadrilateral having at least two parallel sides.

Polygon

A polygon is a closed figure in a plane having many (i.e., any number of) sides. There is no restriction on lengths or angles except that the figure must be closed. A regular polygon is one in which all sides (regardless of the number) are of equal length and all deflection angles are equal. A regular polygon having an infinite number of sides is also the same as a circle.

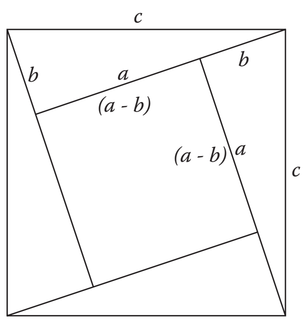


FIGURE 3.4 Theorem of Pythagoras

Pythagorean Theorem

With regard to Figure 3.4, the area of the total figure is the product of two sides, $c \times c$, or c^2. If the area is also computed as the sum of four triangles plus the smaller square inside the figure, proof of the Pythagorean theorem can be written as shown in equation 3.3:

$$c^2 = 4\frac{a\,b}{2} + (a-b)(a-b) = 2\,a\,b + a^2 - 2\,a\,b + b^2 = a^2 + b^2 \tag{3.3}$$

SOLID GEOMETRY

The rules of solid geometry apply to 3-D objects and to the position(s) of other geometrical elements in 3-D space. Although an entire textbook could be written about solid geometry, only a brief summary is included here.

Sphere

A sphere, also known as a ball, is a closed, uniformly curving, three-dimensional surface; all points on its surface are the same distance from an interior point, the center.

Ellipsoid

An ellipsoid is the solid figure formed by rotating an ellipse about one of its two axes. An ellipse rotated about its major axis looks rather like a football. An ellipse rotated about its minor axis is used to approximate the size and shape of the Earth.

Polyhedron

A polyhedron is a solid figure bounded by plane surfaces—usually more than six.

Tetrahedron

A tetrahedron is a regular solid whose sides consist of four equilateral triangles. It is a figure formed by the minimum number of plane sides.

Pyramid

A pyramid is a solid whose base is a polygon and whose triangular sides meet at a common point. Most pyramids are five-sided, having a square base and four triangular sides.

Cube

A cube is a polyhedron having six identically square faces (sides).

Equation of a Plane in Space

A plane in space is described by a first-order (containing only powers of 1 for each variable) equation using variables $X/Y/Z$ as

$$aX + bY + cZ + d = 0 \tag{3.4}$$

The letters a, b, and c are coefficients of the variables and d is a constant; all of them are real numbers. Note that the entire equation could be divided by d, giving equation 3.5 having only three independent coefficients. That makes sense, because it takes three points in space to determine a plane.

$$(a/d)X + (b/d)Y + (c/d)Z + 1 = 0 = a'X + b'Y + c'Z + 1 \tag{3.5}$$

Equation of a Sphere in Space

A sphere in space is described by a second-order (contains powers of 2 on one or more variables) equation using variables $X/Y/Z$ as

$$(X - a)^2 + (Y - b)^2 + (Z - c)^2 = R^2 \tag{3.6}$$

The letters a, b, and c are the $X/Y/Z$ coordinates of the center of the sphere, and R is the radius of the sphere. If a, b, and c are all zero, the center of the sphere lies at the origin of the coordinate system.

Equation of an Ellipsoid Centered on the Origin

A two-dimensional ellipse rotated about its major axis forms a football-shaped figure. A two-dimensional ellipse rotated about its minor axis is used to approximate the Earth's size and shape. The North Pole and South Pole lie on the Earth's spin axis, which is the minor axis of the ellipsoid.

$$\frac{X^2}{a^2} + \frac{Y^2}{a^2} + \frac{Z^2}{b^2} = 1 \tag{3.7}$$

The letter a is used as the semimajor axis, and b is the semiminor axis of the ellipse. The equatorial plane goes through the origin and is 90° from each pole. The equator forms a circle having a as its radius. Each meridian section is perpendicular to the equator and forms an ellipse defined by the letters a and b.

Conic Sections

A cone is a triangular-shaped solid whose base is a closed curve. A right circular cone is one whose base is a circle that is perpendicular to the cone axis. Conic sections are two-dimensional shapes obtained by intersecting a cone with a plane at different orientations.

1. A circle is formed by the intersection of a cone with a plane perpendicular to the axis of the cone. If the intersection occurs at the vertex, the circle is reduced to a point.
2. An ellipse is formed by the intersection of a cone with a plane that is not perpendicular to the cone axis. The intersection is a closed figure.
3. A parabola is formed by the intersection of a cone with a plane that is parallel with the opposite side of the cone. The intersection is not a closed figure.
4. A hyperbola is formed by the intersection of a cone with a plane that is parallel with the axis of the cone. The intersection is not a closed figure.

Conic sections can all be derived from the general second-degree polynomial equation by the appropriate selection of coefficients, A, B, C, D, E, and F. It is to be understood that the X/Y coordinate system used in equation 3.8 lies in the plane intersecting the cone.

$$AX^2 + BXY + CY^2 + DX + EY + F = 0 \tag{3.8}$$

1. For a line: $A = B = C = 0$.
2. For a circle: $A = C$ and $B = D = E = 0$.
3. For an ellipse: $A \neq C$ and $B = D = E = 0$.

4. For a parabola: $B = C = 0$
5. For a hyperbola: $A = -C$ and $B = D = E = 0$.

Vectors

A vector is a directed line segment in space. In terms of a right-handed coordinate system, a vector is composed of signed components in each of the *i*/*j*/*k* directions. The length of a vector is called its magnitude and is computed as the square root of the sum of the components squared (a three-dimensional hypotenuse). A unit vector has a length of 1.0 and is obtained by dividing each component by the vector magnitude. The resulting rectangular components of a unit vector are its direction cosines. The underlying *i*/*j*/*k* orientation of a vector can be rotated to any *X*/*Y*/*Z* coordinate system without changing the length or statistical qualities of the vector. The direction cosines do change.

Vectors can be added and subtracted component by component (subtraction is the same as addition given that the sign is changed for each component of the second vector).

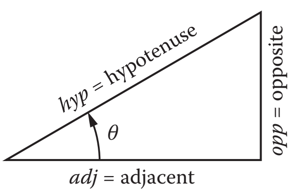
A vector can be multiplied by a scalar that has the effect of changing only the magnitude of the vector. The direction cosines remain the same.

Vector multiplication takes two forms: a dot (inner) product of two vectors is a number (scalar) that can be used to find the angle between two vectors in a plane and the cross product of two vectors that is a third vector perpendicular to the plane common to the first two.

TRIGONOMETRY

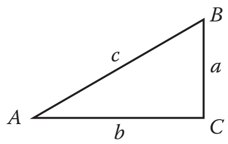
Trigonometry is the study of triangles and relationships between the various sides and angles. The trigonometric functions are defined as ratios of the sides of a right triangle. Once an angle is identified, the defining trigonometric ratio for angle θ are as given in Figure 3.5 (*opp* = opposite, *adj* = adjacent, *hyp* = hypotenuse).

Standard abbreviations for the six ratios are sin, cos, tan, sec, csc, and cot. The first three ratios are used extensively, and the second three, while just as valid, are used much less. Generic nomenclature for a triangle is to label each of the vertex angles with capital letters *A*, *B*, and *C*. The sides opposite the vertex angles are labeled with lowercase letters *a*, *b*, and *c*. Using the generic labeling and noting that the sum of angles *A* and *B* is 90°, the following relationships in Figure 3.6 are fundamental.



Full Function Name	Symbolic Name		Definition
sine θ	sin θ	=	*opp/hyp*
cosine θ	cos θ	=	*adj/hyp*
tangent θ	tan θ	=	*opp/adj*
secant θ	sec θ	=	*hyp/opp*
cosecant θ	csc θ	=	*hyp/adj*
cotangent θ	cot θ	=	*adj/opp*

FIGURE 3.5 Trigonometric Definitions



$\sin A = a/c \qquad \sin B = b/c$
$\cos A = b/c \qquad \cos B = a/c$
$\tan A = a/b \qquad \tan B = b/a$

FIGURE 3.6 Right Triangle Relationships

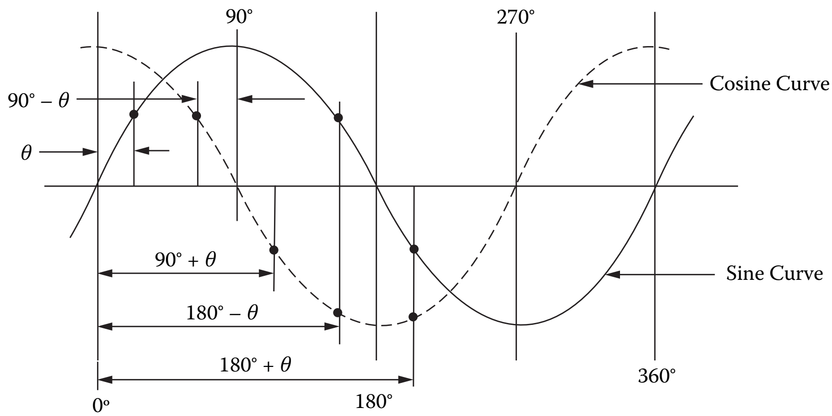


FIGURE 3.7 Plot of Sine and Cosine Functions

Trigonometric Identities

Referring to Figures 3.6, 3.7, and 3.8, the following trigonometric identities can be proved using the Pythagorean theorem and fundamental relationships.

$\sin^2 \theta + \cos^2 \theta = 1 \Rightarrow \sin^2 \theta = 1 - \cos^2 \theta$ and $\cos^2 \theta = 1 - \sin^2 \theta$

$\sin \theta = -\cos(\theta + 90°) = \cos(90° - \theta) = \cos(\theta - 90°) = -\sin(\theta + 180°) = \sin(180° - \theta)$

$\cos \theta = \sin(\theta + 90°) = \sin(90° - \theta) = -\sin(\theta - 90°) = -\cos(\theta + 180°) = -\cos(180° - \theta)$

Law of Sines

Figure 3.9 is a standard generic triangle with extension lines added to make two right triangles; note that sin A = y/c and that sin (180°– C) = y/a, from which y = a sin (180° – C). Next, use the trigonometric identity, sin (180° – C) = sin C, to get y = a sin C. Substituting those in the original equation, sin A = (a sin C)/c, from which (sin A)/a = (sin C)/c. Similarly, with regard to the same Figure 3.9, sin B = z/c and sin (180° – C) = z/b from which z = b sin (180° – C). Using the same

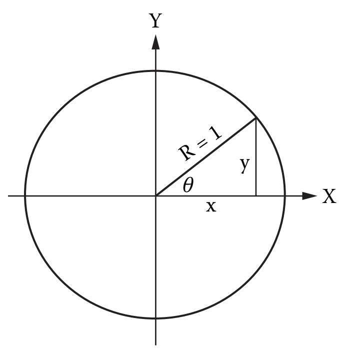


FIGURE 3.8 Circular Trigonometric Functions

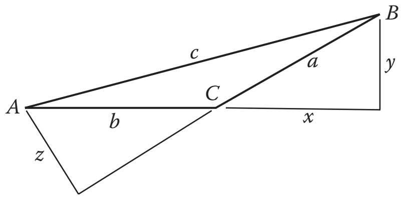


FIGURE 3.9 Generic Triangle

trigonometric identity again, sin (180° $-C$) = sin C. Making substitution for these two equalities, the original equation now becomes sin $B = (b \sin C)/c$, from which $(\sin B)/b = (\sin C)/c$. Since it's been shown the two quantities are both equal to $(\sin C)/c$, they are equal to each other, and the law of sines is

$$\frac{\sin A}{a} = \frac{\sin B}{b} = \frac{\sin C}{c} \tag{3.9}$$

Law of Cosines

The same Figure 3.9 is also used to write a relationship for the law of cosines. Using the Pythagorean theorem on the two right triangles, write $c^2 = (b + x)^2 + y^2$ and $a^2 = x^2 + y^2$ (or $y^2 = a^2 - x^2$). Also note cos (180° – C) = x/a from which $x = a$ cos (180° – C). But a useful trigonometric identity is cos (180° – C) = –cos C. Putting those together,

$$c^2 = (b + x)^2 + a^2 - x^2 = b^2 + 2bx + x^2 + a^2 - x^2$$

$$= a^2 + b^2 + 2ab\cos(180° - C)$$

$$= a^2 + b^2 - 2ab\cos C \tag{3.10}$$

An alternate version of the law of cosines giving an angle in terms of the sides is

$$\cos C = \frac{a^2 + b^2 - c^2}{2ab} \tag{3.11}$$

SPHERICAL TRIGONOMETRY

The rules of spherical trigonometry can be used to solve for the great circle arc distance between latitude/longitude points on the Earth or to solve triangles on the celestial sphere when determining the astronomical azimuth of a line on the ground. Two important considerations are as follows:

- The Earth is slightly flattened at the poles so the spherical great circle distance will be somewhat longer than the actual ellipsoidal distance. For example, the ellipsoidal (GRS80) distance between points in New Orleans and Chicago is 1,354 kilometers, and the great circle distance between the same two points is 1,359 kilometers. If accuracy to two or three significant figures is sufficient, the spherical triangle solution may be appropriate. The spherical triangle solution is certainly simpler and easier to find than the ellipsoidal distance using a geodetic inverse (BK19) computation.

- The radius of the celestial sphere is considered to be infinitely large, which means there is no similar spherical/ellipsoidal approximation when spherical trigonometry is applied to solving the pole-zenith-star (*PZS*) triangle. The equations and procedures for determining an astronomical azimuth can be found in texts such as Moffitt and Bossler (1998), Davis et al. (1981), or Wolf and Ghilani (2005), and are not covered here.

As shown in Figure 3.10, a spherical triangle has three vertex angles and three sides. All six elements are given in terms of angles because changing the size of the sphere does not change the angular relationships. The three vertex angles are on the surface of the sphere and are labeled with capital letters *A*, *B*, and *C*, while the sides opposite each vertex angle are labeled with lowercase letters *a*, *b*, and *c*. Each side is the arc of an angle subtended at the center of the sphere, and the value listed for each side is really the value of the corresponding subtended angle.

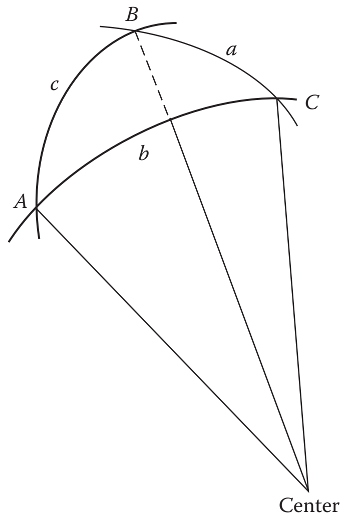


FIGURE 3.10 Spherical Triangle

Equations for solving a spherical triangle include the spherical law of sines and two forms of the spherical law of cosines. In order to solve a spherical triangle, three of the six angles must be known. If two vertex angles and the side opposite one of them are known, or if two sides and the vertex angle opposite one of them are known, the spherical law of sines can be used. The spherical law of sines is

$$\frac{\sin A}{\sin a} = \frac{\sin B}{\sin b} = \frac{\sin C}{\sin c}. \tag{3.12}$$

The spherical law of cosines has two forms: the first form solves for a vertex angle if the sides are all known, and the second form solves for a side if all the vertex angles are known. They are

$$\cos A = \frac{\cos a - \cos b \cos c}{\sin b \sin c}$$

$$\cos a = \frac{\cos A + \cos B \cos C}{\sin B \sin C} \qquad \text{(3.13) and (3.14)}$$

Note that equations 3.13 and 3.14 are cyclic in that the symbols for the vertices and the sides can be relabeled on the diagram or that equations for vertex angles *B* and *C* can be found by rearranging the sides in equation 3.13. Similarly, equations for sides *b* and *c* can be found by rearranging vertex angles in equation 3.14. For example:

$$\cos B = \frac{\cos b - \cos c \cos a}{\sin c \sin a} \quad \text{while} \quad \cos b = \frac{\cos B + \cos C \cos A}{\sin C \sin A}$$

Equation 3.13 is the form used to solve for great circle arc distances. Given that the Earth's radius is 6,372 km and that the latitude and longitude for the two cities are as follows:

New Orleans: latitude = 30° 02' 17", longitude = 90° 09' 56" W
Chicago: latitude = 42° 07' 39", longitude = 87° 55' 12" W

Note that vertex angle *A* is the longitude difference at the North Pole, vertex angle *B* is at Chicago, and vertex angle *C* is at New Orleans.

Find the great circle distance between the two cities.

Side *b* = 90° – 30° 02' 17" = 59° 57' 43"
Side *c* = 90° – 42° 07' 39" = 47° 52' 21"
Angle *A* = 90° 09' 56" – 87° 55' 12" = 2° 14' 44"

Rewriting equation 3.13 to solve for side *a* and using the values for *b*, *c*, and *A*,

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

$$\cos a = \sin(59°57'43'') \sin(47°52'21'') \cos(2°14'44'') + \cos(59°57'43'') \cos(47°52'21'')$$

cos *a* = 0.9773288; *a* = 12° 13' 25" = 0.21334 radians (only 5 s.f.)

where distance = *R a* (in radians) = 1,359.4 km = 844.70 miles.

Remember: this computation assumes the Earth is a sphere. Computing the ellipsoid arc distance between the same two points will be covered in chapter 6.

CALCULUS

Calculus is a valuable mathematical tool that can be described as a study of rates of change or, said differently, cause and effect. Calculus has an undeserved reputation of being difficult to learn. That may be true for some, but consider that most people who receive a paycheck for wages are already experts at calculus. That is, upon being notified of a 5 percent increase in wages, they will quickly determine how the different pay rate affects their take-home pay. In a more formalized manner, the procedures of *differential* calculus apply known rules to smaller and smaller increments so that the instantaneous rate of change at any point can be computed. Called a derivative and given the symbol dy/dx, the graphical representation of dy/dx is the slope of a line in the X/Y plane.

Convention

Three different ways of showing similar values are as follows:

Deltas: Δx, Δy, when using small or specific numerical values.
Derivative: dx, dy, infinitesimally small components of calculus.
Partial: ∂x, ∂y, same as a derivative, but the derivative of the computed result is taken with respect to one variable at a time.

In another application of calculus, small pieces of a quantity (such as arc length or area) are computed according to some equation. *Integral* calculus uses the antiderivative of dy/dx to perform an infinite number of summations in order to find the grand total result (see computation of the meridian arc in chapter 6). In cases where a given mathematical equation cannot be integrated, a fallback procedure is to use numerical integration to compute an approximation based upon intervals (ΔX's) selected by the user. With a computer programmed to do the repetitive calculations, a numerical integration can be made to be as accurate as desired (within reason) by adding up more and more ever smaller pieces. Computing area under the standard error bell curve is an example of numerical integration.

Many students learn calculus by memorizing the rules of manipulation, and, with continued use, the concepts become more understandable. Therefore, the goal here is to include several simple examples along with a summary of the fundamental rules of manipulation. It is hoped, however, that the examples will lead toward greater understanding of specific applications involving spatial data manipulation.

EXAMPLE

The volume of a cylindrical storage tank (see Figure 3.11) is computed from the radius (R) and height (h) of the tank; see equation 3.15. Two questions are as follows:

1. If the height of the tank is changed, how will that affect the total volume?
2. If the radius of the tank is changed, how will that affect the total volume?

$$V = \pi R^2 h \qquad (3.15)$$

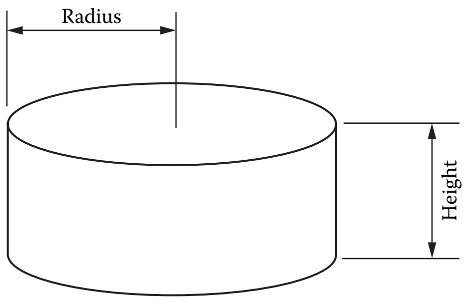


FIGURE 3.11 Volume of Tank

Certainly, one could compute the volume for specific incremental values of radius and height. The change is then found by taking the difference of various answers. Calculus is not needed for that. But, by looking at the rate of change given by the derivative of the equation used to compute the volume, extra calculations can be avoided. And, different answers are obtained depending upon whether the height changes or the radius changes. Using calculus, height and radius variables are handled separately and the user has the option of using either, neither, or both depending upon the question to be answered.

The relationship between tank volume, height, and radius is illustrated in Figure 3.12. Note that volume increases linearly with height, but that the volume of the tank increases exponentially with radius. These changes are handled with partial derivatives, one variable at a time. Note too that when taking the volume partial derivative with respect to height, the answer is the slope of the line in the volume/height plane. The radius is treated as a constant, and the slope (derivative) is constant. In the radius/volume plane, the slope of the curve increases exponentially with radius and is not linear—the larger the radius, the larger the slope.

Partial derivative of volume with respect to height: $\partial V / \partial h = \pi R^2$
Partial derivative of volume with respect to radius: $\partial V / \partial R = 2 \pi R h$

In order to find the combined impact of changing both radius and height, the two expressions are combined. The symbols are changed to the "delta" nomenclature to indicate that one is expected to use actual differences to compute a change.

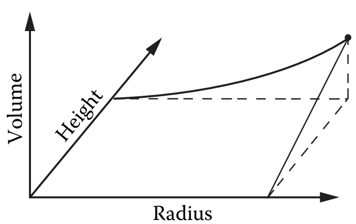


FIGURE 3.12 Plot of Volume

$$\Delta V = \pi R^2 \Delta h + 2 \pi R h \Delta R \qquad (3.16)$$

Equation 3.16 is quite useful, but, because R is nonlinear, equation 3.16 is accurate only for "small" values of ΔR. As ΔR grows larger, the accuracy of equation 3.16 becomes unacceptable. The problem will be addressed in the section on error propagation where Δh and ΔR are taken to be standard deviations of those dimensions. Using standard deviations of R and h, partial derivatives of the volume with respect to R and h, and formal error propagation procedures, the standard deviation of the volume can be determined with statistical reliability.

Differential Calculus Equations

Following is a brief list of derivatives. These rules presume x is the independent variable, u and v are intermediate variables, y is the computed result, and the derivative is dy/dx. Appropriate substitutions for y are listed in each case.

y = constant, a	$d(constant)/dx = 0$
y = constant * variable, a * u	$d(au)/dx = a\ du/dx$
y = sum of variables u and v	$d(u + v)/dx = du/dx + dv/dx$
y = product of variables u and v	$d(uv)/dx = u\ dv/dx + v\ du/dx$
y = quotient of variables u and v	$d(u/v)/dx = (v\ du/dx - u\ dv/dx) / v^2$
y = variable raised to power n	$d(un)/dx = n\ un^{-1}\ du/dx$
$y = \sin u$	$d(\sin u)/dx = \cos u\ du/dx$
$y = \cos u$	$d(\cos u)/dx = -\sin u\ du/dx$
$y = \tan u = \sin u / \cos u$	$d(\tan u)/dx = \sec^2 u\ du/dx$

Integral Calculus Equations

Using the same conventions as in the previous section, the following is a brief summary of integrals. When evaluating integrals between stipulated limits, the constant of integration (C in the following equations) cancels out.

$$\int du = u + C$$
$$\int a\ du = a \int du = a\ u + C$$
$$\int (du + dv) = \int du + \int dv = u + v + C$$
$$\int u^n\ du = [\ u^{n+1} / (n + 1)\ ] + C$$
$$\int du/u = \ln |u| + C$$
$$\int \cos u\ du = \sin u + C$$
$$\int \sin u\ du = -\cos u + C$$
$$\int \sec^2 u\ du = \tan u + C$$

PROBABILITY AND STATISTICS

Introduction

The fields of probability and statistics are distinct disciplines each deserving more coverage than given here. Since knowledge of underlying mathematical principles is essential to understanding the importance of each discipline's contribution to spatial data, the reader is referred to a variety of sources for more information. Some books are written purposefully with a focus on mathematical theory and principles (e.g., Dwass [1970] was an undergraduate text for math majors), others were written to highlight applications in various fields (e.g., Wine [1964] contains a focus on mathematical, not behavioral, sciences), and others are written to explore specific applications to spatial data (e.g., (Cressie [1993] focuses specifically on spatial data applications). The goal in this book is to utilize fundamental principles from each discipline and to organize them as needed to describe efficient procedures by which 3-D spatial data accuracy (standard deviations) can be reliably established, stored, tracked, and used—component by component. Other more sophisticated tools such as krigging and hypothesis testing are also useful when analyzing spatial data. Although not included here, such tools are viewed as being compatible with the underlying GSDM.

Concepts of probability and statistics overlap each other when applied to spatial data, and, beyond the definitions, little effort is made here to preserve their distinction. Wolf and Ghilani (1997) say that probability is the ratio of the number of times that an event should occur to the total number of possibilities. Mikhail (1976) defines probability as the limit of the relative frequency of occurrences of a random event. A simple example of probability is the results one would expect when tossing a coin. It should come up heads half the time and tails the other half. If the coin is tossed seven times, the ratio of heads to the number of tosses will not be 0.5 for several reasons: (1) because of the odd number of tosses, and (2) because the outcome of a random event is never certain. It would be possible, but not probable, for eight persons to toss a coin seven times each and for each person to obtain different results ranging from zero heads to seven heads. However, if a very large number of tosses is used, the limit of relative frequency should be the same as the (expected) ratio of "heads up" to the total number of tosses. Two important characteristics of probability are that a zero probability is associated with an event that will not happen, while a 100 percent probability is associated with an event that is certain to happen. Therefore, a number between 0 and 1 gives an estimate of uncertainty associated with some event or statement.

Wine (1964) calls statistics the science of decision making in the face of uncertainty. He goes on to say statistics should be thought of as "both a pure and an applied science which is involved in creating, developing, and applying procedures in such a way that the uncertainty of inferences may be evaluated in terms of probability." Cressie (1993) begins his book by saying that statistics, the science of uncertainty, attempts to model order in disorder. Incorporating the definitions into the goal of this book, the GSDM defines an environment and computational procedures by which the standard deviation of each 3-D spatial data component can be determined, enabling better decisions to be made regarding use of the data.

Standard Deviation

The standard deviation of a distance is used to describe its uncertainty at some level of confidence. Procedures for computing standard deviation have been formalized, are well documented under the umbrella of error propagation, and are summarized later in this chapter. The confidence level at which decisions are made regarding the use of standard deviations is selected by the user. One standard deviation is associated with a confidence level of 68 percent, two standard deviations correspond to a 95 percent confidence level, and virtual certainty (99.7 percent) is achieved at three standard deviations.

Standard deviations are not all that exact. Although they can be computed by very specific equations, they are acknowledged to be approximations rarely having more than two significant figures. That means standard deviation can, at times, also be assigned on the basis of seasoned judgment. For example, the standard deviation of a GPS vector might be assigned a standard deviation of 5 mm, a distance measured by an electronic distance meter (EDM) may have a standard deviation of 0.01 feet, and a single code phase GPS position may have a standard deviation of 100 meters, 10 meters, 5 meters, 1 meter, or even less based upon circumstances of the measurement. Such assignments, while subjective, are based upon prior experience and can be quite valid. But, more specifically, standard deviations are determined from repetitious measurements made under known conditions and/or computed via error propagation of such measurements.

Measurement

In a sense, measurement and observation are both the result of comparing some unknown quantity with a standard. Accepting a graduated meter (yard) stick as a standard, the distance between points is measured by aligning graduations on the scale with the endpoints of the distance being measured. The process utilizes both the human eye and judgment. The measurement is the observation, and the observation is the measurement. If the same distance is measured by GPS, the result may be more precise, but the concept of observation is not so clear because the comparison was computed rather than viewed by the human eye. Admittedly, care and judgment are involved in positioning the GPS antenna precisely over the marks, but the point remains that the measurement involved computations rather than a direct scale comparison. Chapter 2 describes the difference between a direct measurement and an indirect one. Chapter 2 also notes that observation and measurement are essentially the same with one subtle distinction: observations are taken to be independent, while measurements may be correlated.

Errors

Another fundamental principle is that no measurement is perfect. Here a distinction is made between a count and a measurement. An integer count of pencils can be exact, but repeated measurements of the length of a pencil, when compared carefully to a fine scale, will invariably yield different results. The standard deviation of a measurement is determined from those variations. Admitting that a true length can never be found, the goal in making a measurement is to find an acceptable estimate and to have some knowledge of the uncertainty of the estimate. The mean (or average) of a group of measurements is taken to be the estimate, and the standard deviation of the mean provides a measure of confidence. Results of measuring the length of a pencil could be reported as 181.3 mm +/– 0.4 mm.

The word "error" is used when referring to the variability of results within a set of measurements. When associated with a mistake or blunder, the word "error" has a justifiably bad connotation. But, given the impossibility of making a perfect measurement, errors are not necessarily bad but regularly occur with predictable behavior and are categorized as systematic errors or random errors.

Blunders

Blunders are mistakes and are not considered legitimate observations. Any measurement containing a blunder should be discarded and not used. A blunder is the responsibility of the person making the measurement and is eliminated by exercising care, checking one's work, and making redundant observations. There is no mathematical magic for accommodating blunders in a set of data.

Systematic Errors

Systematic errors arise from a mismatch between ideal (assumed) conditions and actual conditions of a measurement. A systematic error is characterized by its predictability and the fact that it always occurs with a given set of measurement conditions. Bias is another word sometimes used to describe systematic errors. Two examples of systematic error are (1) measuring a distance with a steel tape that has been foreshortened by cold temperature, and (2) measuring a distance with an EDM without specifying the correct parts-per-million correction for given temperature and pressure conditions.

Random Errors

Random errors are the result of imperfect observations. Even though the goal is to use well-calibrated equipment according to proper procedures in order that random errors are kept as small as practical, they do occur and their predictable behavior has been well documented. The normal distribution curve shown in Figure 3.13 illustrates the collective characteristics of random errors.

The characteristics of random errors are:

1. Small random errors occur more frequently than large ones.
2. Positive and negative random errors occur with similar frequency.
3. Very large random errors do not occur. If they do, they are taken to be blunders and discarded.

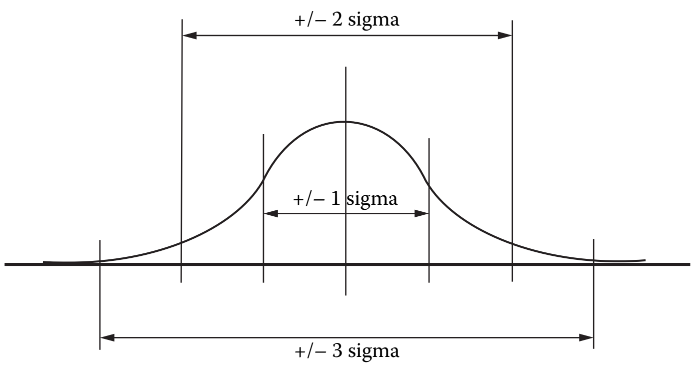


FIGURE 3.13 Normal Distribution Curve

Error Sources

Knowledge of possible error sources is important, both to those making measurements (collecting spatial data) and to those responsible for determining the circumstances under which data are collected (buying equipment and/or writing specifications). Three general error sources are:

Personal

All three types of errors can be attributed to a person making a measurement. A careless person will make mistakes and blunders. Even the most conscientious person makes random errors and systematic errors can arise from choices made about how computations are made or how data are used. For example, choosing to show state plane grid area for a parcel instead of horizontal ground area is a systematic error. If the difference is small, the error may be inconsequential.

Environmental

Environmental conditions give rise to both systematic errors and random errors. If one determines the temperature of a steel tape, it is possible to compute a temperature correction for a measured distance. By applying the temperature correction to the measurement, it is possible to eliminate that systematic error source. If a cold front moves through during GPS data collection, the results of a GPS survey may be affected by the changing weather conditions. Depending upon the severity of the weather changes and upon the availability of reliable meteorological data, the differences of a computed GPS position could be a combination of systematic and random errors.

Instrumental

If a piece of equipment malfunctions, one could call it a blunder. More often, instrumental errors are systematic and due to the physical construction of the instrument. An example is a steel tape whose length at standard temperature and tension differs from its nominal length. Random error for the same tape could be represented by scale graduations that are not perfectly spaced. And, gradual change of calibration parameters with respect to the performance of an EDM instrument gives rise to random errors until such time as the EDM is calibrated. With the calibration parameters known, that part of the error becomes systematic and an appropriate correction should be applied.

Accuracy and Precision

Accuracy and precision are also related to errors. Blunders are mistakes related to one's level of diligence and professionalism. Systematic errors are related to the concept of accuracy, while random errors are related to the concept of precision.

Accuracy is a measure of absolute nearness of a measurement to the true value (the true value is never known, but an estimate is used in its place). A data set containing little or no systematic error is said to be accurate. Due to random error, there may be significant variation among different measurements of the same quantity, but the mean of the data set will be quite close to the true value.

Precision is a measure of consistency (or repeatability) within a given data set. A data set containing small random errors is said to be precise. However, if systematic error (a bias) is also present, it is possible to be precisely wrong.

Examples of accuracy and precision are often given as the result of shooting a gun at a target. Figure 3.14 shows four different results:

1. Both accurate and precise: There is a small grouping of holes located near the center of the target. This represents the desired result.
2. Precise, but not accurate: There is a small grouping of holes, but they are obviously not located near the center of the target. Such a result indicates the presence of systematic error and/or the need to remove a bias by adjusting the crosshairs in the telescope of the rifle.
3. Accurate, but not precise: There is a wide dispersion of holes over the target with no obvious grouping. Taken as a whole, the results can be said to be accurate, but the randomness in the grouping indicates the need for more practice, a steady rest, or some other factor to assure greater consistency.
4. Neither accurate nor precise: When some of the shots obviously miss the target, the marksman is neither accurate nor precise. The first step to improvement might be to eliminate the blunders (misses), after which decisions can be made as to the need for greater improvement.

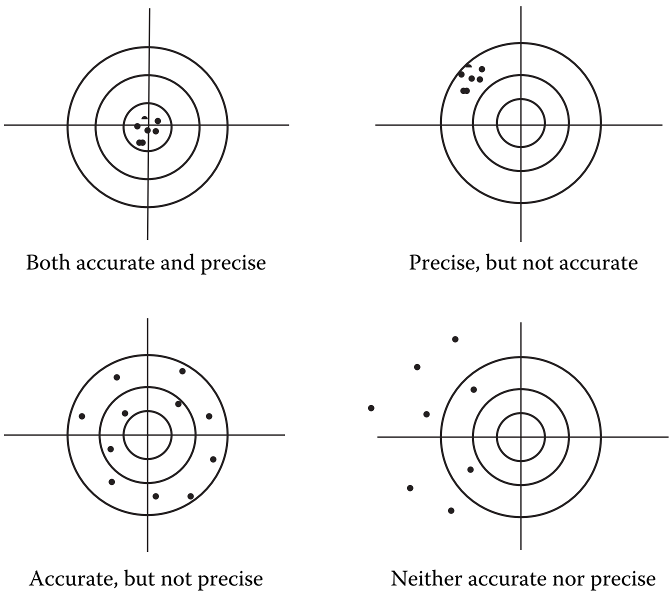


FIGURE 3.14 Examples of Accuracy and Precision

Computing Standard Deviations

What is the definition of "small" when dealing with random errors? The relative magnitude of a random error is defined mathematically by its standard deviation. Whether the standard deviation is large or small, the shape of the distribution is that given in Figure 3.13. However, whether the range covered by the random error is large or small is illustrated in Figure 3.14. In either case, the Greek letter sigma, σ, is used to denote standard deviation, which is computed using equation 3.17.

Standard Deviation of the Mean

Equation 3.17 gives the standard deviation associated with an individual measurement in the data set. That is an appropriate value to use when making comparisons between different methods of measurement or evaluating equipment. But, when using the mean of a set of observations in subsequent computations, the computed mean has a greater chance of being close to the true value than do any of the individual measurements. Specifically, in subsequent computations, the standard deviation of the mean (equation 3.18) is needed rather than the standard deviation of a measurement. The standard deviation of the mean is the standard deviation of the measurement divided by the square root of the number of measurements. In the example, the standard deviation of a single 1,000 meter measurement is 0.032 meters, but the standard deviation of the mean is 0.014 meters.

$$\sigma_i = \sqrt{\frac{\sum_{i=1}^{n}(mean - x_i)^2}{n-1}} \tag{3.17}$$

$$\sigma_{mean} = \sqrt{\frac{\sum_{i=1}^{n}(mean - x_i)^2}{n(n-1)}} = \frac{\sigma_i}{\sqrt{n}} \tag{3.18}$$

where

σ_i = Greek letter, sigma, for standard deviation of data set;
σ_{mean} = standard deviation of the mean;
mean = average of data set;
x_i = single measurement; and
n = number of observations.

Number	Observation	Mean	Difference	Difference Squared
1.	999.98 m	999.994 m	0.014 m	0.000196
2.	1,000.02 m	999.994 m	–0.026 m	0.000676
3.	999.95 m	999.994 m	0.044 m	0.001936
4.	1,000.03 m	999.994 m	–0.036 m	0.001296
5.	999.99 m	999.994 m	0.004 m	0.000016
Totals	4,999.97 m			0.004120

Results: σ_i = 0.032 meters and $\sigma_{mean} = \sigma_i / \sqrt{n}$ = 0.014 meters.

Confidence Intervals

The standard deviation as computed in the example corresponds to a confidence level of 68 percent. As applied to the computed mean, there is a 68 percent chance the true value of the distance, whatever it is, lies between 999.980 meters and 1,000.008 meters. If one is not comfortable at the 68 percent level of confidence, it is routine to quote results at the 95 percent confidence level. For 95 percent confidence, the range quoted is the mean plus or minus two standard deviations. In this case, there is a 95 percent probability that the true value lies between 999.996 meters and 1,000.022 meters. The level of confidence is greater, but the quoted interval of uncertainty is also larger. Detail: two standard deviations correspond to 95.5 percent, and 95 percent corresponds to 1.96 standard deviations. Common practice is to use two standard deviations with 95 percent.

A limit of three standard deviations is associated with a confidence level of 99.7 percent. This criterion is often used to judge whether or not a given observation should be discarded as a blunder. As applied to the standard deviation of each observation, if the difference of an individual observation from the mean is greater than three standard deviations of the data set, standard practice is to eliminate that observation from the data set and *re-compute* the mean and standard deviations with a smaller data set. Strictly speaking, a difference could exceed three standard deviations three times out of 1,000 observations and not be a blunder, but, generally, little harm is done by rejecting an observation that lies more than three standard deviations from the mean of the data set.

Hypothesis Testing

Extending the concept of confidence intervals further gets into hypothesis testing. When comparing the results (statistics) of one data set with another, one question to ask is whether the data sets are compatible. Is it really proper to make the comparisons between them, and what is the likelihood of drawing conclusions based upon comparisons of unlike data (apples and oranges)? A discussion of hypothesis testing is beyond the scope of this book, but the principles are well formulated in books such as Wolf and Ghilani (1997) and constitute additional valuable tools for the spatial data analyst.

Matrix Algebra

Matrix algebra is a compact way of representing and manipulating systems of linear equations. Compact representation helps humans grasp and discuss overall concepts without getting bogged down in details, and compact manipulation makes it possible to utilize standard programming procedures more efficiently in computerized solutions. Rules of matrix manipulation are developed in many math books and included, often as an appendix, in many texts devoted to survey computations or data adjustment.

A matrix is a rectangular array of rows and columns and is denoted by boldfaced type when written in text or in equations. Each subscripted position in the array is assigned a real number (as opposed to the cell of a spreadsheet, which may contain text or equations in addition to numbers). The number of rows and columns are the dimensions of a matrix and are, at times, written as subscripts to the boldfaced matrix symbol. Individual elements within a matrix are often given by lowercase subscripted variables. An example of a simple 2 rows by 3 columns matrix is

$$\boldsymbol{A}_{2,3} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \end{bmatrix} = \begin{bmatrix} 2 & 1 & -1 \\ 4 & 0 & -2 \end{bmatrix} \tag{3.19}$$

Several important matrix concepts are:

1. A square matrix has the same number of rows and columns.
2. The diagonal of a square matrix contains elements in those positions having identical row and column numbers. A diagonal matrix is one in which any nondiagonal element is zero.
3. An identity matrix is a diagonal matrix with 1's on the diagonal.
4. A symmetrical matrix has a mirror image with respect to the diagonal.
5. The transpose of a matrix (indicated by the superscript t) is obtained by switching the rows and columns of the parent matrix. The dimensions are switched accordingly. The transpose of $\boldsymbol{A}_{2,4}$ is $\boldsymbol{A}^t_{4,2}$.
6. A vector is a matrix having only one row or one column. A matrix with only one row and one column contains a single element, a real number.
7. Matrix addition is defined, for matrices having compatible dimensions, as a matrix containing the sum of the corresponding elements in the two parent matrices.
8. Likewise, matrix subtraction is defined, for matrices having compatible dimensions, as the difference of corresponding elements in the stipulated order.
9. Matrix multiplication is defined for matrices having compatible dimensions. That is, the number of rows in the second matrix must be the same as the number of columns in the first matrix. Each element in the product matrix is obtained as the sum of the products of row/column elements of the matrices being multiplied, as shown in equation 3.20.

$$\begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} x \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix} = \begin{bmatrix} a_{1,1}b_{1,1} + a_{1,2}b_{2,1} & a_{1,1}b_{1,2} + a_{1,2}b_{2,2} \\ a_{2,1}b_{1,1} + a_{2,2}b_{2,1} & a_{2,1}b_{1,2} + a_{2,2}b_{2,2} \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} x \begin{bmatrix} -2 & 4 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} -4 & 8 \\ -8 & 14 \end{bmatrix} \tag{3.20}$$

10. Matrix division is not defined. Instead, the matrix inverse is an alternate procedure that is used and produces the same result. Given matrix $\boldsymbol{A}$, the inverse is $\boldsymbol{A}^{-1}$. In regular algebra, $A * 1/A = 1$. In matrix algebra. $\boldsymbol{A} * \boldsymbol{A}^{-1}$ gives the identity matrix, $\boldsymbol{I}$. Details for computing a matrix inverse are somewhat involved and not included here. Wolf and Ghilani (1997) is an appropriate reference containing material on the theory and use of matrices.

MODELS

Models provide a connection between abstract concepts and human experience. In the context of spatial data, models give relevance and meaning to the concepts of location and geometrical relationships. Two kinds of models are used in this book: functional models and stochastic models.

Functional Models

Functional models consist of physical, geometrical, mechanical, electrical, and other relationships that exist with respect to the cause and effect between observed fundamental quantities such as length, time, temperature, and current and computed results such as spatial data components. Using a model consists of writing and solving equations that describe the problem being considered and interpreting the results in terms of how well they agree with the original assumptions and observations.

Stochastic Models

A stochastic model describes the probabilistic characteristics of various elements of the functional model. Whether a quantity is fixed by law, determined by repeated measurements, or assigned on the basis of personal judgment, the stochastic model represents the "totality of the assumptions on the statistical properties of the variables involved" (Mikhail 1976). The standard deviation of any quantity is a statistical measure of its quality. Statistical interaction between variables is known as correlation and, along with standard deviations, is captured in the appropriate variance/covariance matrix. With regard to 3-D coordinates representing spatial data, the following variance/covariance matrix represents the probabilistic characteristics of the defined point.

$$\Sigma_{XYZ} = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{YX} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{ZX} & \sigma_{ZY} & \sigma_Z^2 \end{bmatrix} \quad (3.21)$$

Notes:

1. The diagonal elements are called variances, and the off-diagonal elements are called covariances. It is proper to refer to the entire matrix as a covariance matrix even though it contains both variances and covariances.
2. The standard deviation of each respective *X*/*Y*/*Z* coordinate is the square root of the variance.
3. Correlation between variables (coordinates is a number between –1.0 and 1.0 and is obtained from elements in the covariance matrix. Since the correlation of X with respect to Y is the same as teh correlation of Y with respect to X, the covariance matrix is symmetric. Quantities that are statistically independent have zero correlation. Correlation is mathematically defined as:

$$\rho X, Y = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (3.22)$$

where:
cov(X,Y) = covariance value for element X,Y.
σ_X and σ_Y = standard deviations of X and X.

ERROR PROPAGATION

The theory of error propagation is derived in books such as Mikhail (1976) and Wolf and Ghilani (1997), and is presented concisely in matrix form as

$$\Sigma_{YY} = \boldsymbol{J}_{YX} \Sigma_{XX} \boldsymbol{J}^t{}_{YX} \quad (3.23)$$

where

Σ_{YY} = covariance matrix of computed result,
Σ_{XX} = covariance matrix of variables used in computation, and
$\boldsymbol{J}_{YX}$ = Jacobian matrix of partial derivatives of the result with respect to the variables.

Error propagation involves calculus and is used to answer the question "If something is computed on the basis of a measurement and the measurement contains uncertainty, how is the computed quantity affected?" In a trivial case ($Y = X$), there is a direct correspondence between the two, and the error in the result is the same as the error in the measurement. In another simple case—the volume of a tank and its height—the relationship between the measurement and the computed result is linear: volume = (area of base) * (height). Other cases—the volume of a tank and its radius—are more complex: volume = (πR^2) * (height). Here the relationship between the volume and radius is exponential. However, even more complexity arises when several measurements contribute simultaneously to the quantity being computed. Equation 3.23 handles all cases from the trivial to the most complex.

Equation 3.16 could be used to compute approximate changes in volume based upon changes in radius and height, but that approach is somewhat limited. Equation 3.23 is used to answer the specific question "What is the standard deviation of the volume if the standard deviation of the radius and the standard deviation of the height are both known?" Admitting that standard deviations are estimates, the answer will still be an estimate. But, unlike equation 3.16 (which is accurate only for "small" values of Δh and ΔR), equation 3.23 is a definitive procedure that is statistically reliable, and the approximation is in the standard deviation of the measurements (the user's responsibility) and not in the equation. A simplified list of steps for performing error propagation is as follows:

1. Identify the variables (i.e., the measurements), and determine their standard deviations on the basis of repeated measurements, computations, or professional judgment.

Measurements and standard deviations in the tank example are as follows:

R = 50.0 meters +/– 0.07 meters

h = 10.00 meters +/– 0.02 meters

2. Formulate the equations that will be used to compute the result.

$$V = \pi R^2 h = 78{,}539.82 \text{ meters}^3 \tag{3.24}$$

3. Take the partial derivatives, one variable at a time.

$$\partial V/\partial h = \pi R^2 = 7{,}853.98 \text{ meters}^2 \tag{3.25}$$

$$\partial V/\partial R = 2\pi Rh = 3{,}141.59 \text{ meters}^2 \tag{3.26}$$

4. Build the matrices as shown in equation 3.23.

$$\boldsymbol{\Sigma}_{XX} = \boldsymbol{\Sigma}_{RH} = \begin{bmatrix} \sigma_R^2 & \sigma_{Rh} \\ \sigma_{Rh} & \sigma_h^2 \end{bmatrix} = \begin{bmatrix} 0.0049 & 0.0000 \\ 0.0000 & 0.0004 \end{bmatrix} \tag{3.27}$$

$$\boldsymbol{J}_{YX} = \begin{bmatrix} \dfrac{\partial V}{\partial R} & \dfrac{\partial V}{\partial h} \end{bmatrix} = \begin{bmatrix} 7{,}853.98 & 3{,}141.59 \end{bmatrix} \tag{3.28}$$

5. Perform the matrix operations. Computers make this task much easier.

$$\Sigma_{YY} = \begin{bmatrix} 7,853.98 & 3,141.59 \end{bmatrix} \begin{bmatrix} 0.0049 & 0.0000 \\ 0.0000 & 0.0004 \end{bmatrix} \begin{bmatrix} 7,853.98 \\ 3,141.59 \end{bmatrix}$$

$$\Sigma_{YY} = \textit{Variance of volume} = 306,204.34 \;\; m^6 \qquad (3.28)$$

6. Interpret the results.
   A. The standard deviation of the volume = square root of the variance = 553.36 meters^3. Realistically, this answer has no more than two significant figures. At the 68 percent confidence level the standard deviation of the volume is 550 m^3, and at the 95 percent confidence level the standard deviation of the tank volume is 1,106.7 meters^3 (1,100 meters^3).
   B. The answer in step two really has only three significant figures and could be reported as 78,500 meters^3 +/– 550 meters^3 (or, at two sigma, +/– 1,100 meters^3).
   C. In this case, there is no correlation between the measurements of height and radius. Had there been, the Σ_{XX} matrix off-diagonals would be nonzero.

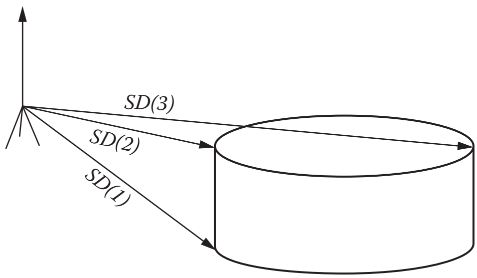


FIGURE 3.15 Survey Measurements of a Tank

What happens if correlation is present? The following example may help. A total station surveying instrument was used to measure the size of the tank, as shown in Figure 3.15. Admittedly, this might not be the best way to measure a tank, but this procedure was chosen to show how correlation is included. Measurement of the radius and height is derived from independent observations of slope distances and zenith directions. This example assumes a standard deviation of 0.10 meters for each slope distance and 20 seconds of arc for each zenith direction. It could also be appropriate to make other assumptions about standard deviations for the observations.

SD_1 = 101.119 m +/– 0.10 m $\quad$ Z_1 = 98° 31’ 51” +/– 20”
SD_2 = 100.125 m +/– 0.10 m $\quad$ Z_2 = 92° 51’ 45” +/– 20”
SD_3 = 200.062 m +/– 0.10 m $\quad$ Z_3 = 91° 25’ 55” +/– 20”

To find radius and height from the field measurements, the functional model equations are

$$R = 0.5\,(SD_3 \sin Z_3 - SD_3 \sin Z_3) \qquad (3.30)$$

$$H = SD_1 \cos Z_1 - SD_2 \cos Z_2 \tag{3.31}$$

The partial derivatives of the radius are

$$\partial R/\partial SD_1 = 0 = 0.0 \tag{3.32}$$

$$\partial R/\partial SD_2 = -0.5 \sin Z_2 = -0.499376165 \tag{3.33}$$

$$\partial R/\partial SD_3 = 0.5 \sin Z_3 = 0.499843856 \tag{3.34}$$

$$\partial \mathrm{R}/\partial \mathrm{Z}_1 = 0 = 0.0 \tag{3.35}$$

$$\partial R/\partial Z_2 = -0.5 \cos Z_2 = 2.500011935 \tag{3.36}$$

$$\partial R/\partial Z_3 = 0.5 \cos Z_3 = -2.499971439 \tag{3.37}$$

The partial derivatives of the height are

$$\partial H/\partial SD_1 = \cos Z_1 = -0.148340663 \tag{3.38}$$

$$\partial H/\partial SD_2 = -\cos Z_2 = 0.049937816 \tag{3.39}$$

$$\partial \mathrm{H}/\partial \mathrm{SD}_3 = 0.0 = 0.0 \tag{3.40}$$

$$\partial H/\partial Z_1 = -SD_1 \sin Z_1 = -100.0002519 \tag{3.41}$$

$$\partial H/\partial Z_2 = SD_2 \sin Z_2 = 100.0000769 \tag{3.42}$$

$$\partial H/\partial Z_3 = 0.0 = 0.0 \tag{3.43}$$

The Jacobian matrix (transposed for ease of printing) of partial derivatives is

$$\boldsymbol{J}^{\boldsymbol{T}} = \begin{bmatrix} 0 & -0.148340663 \\ -0.499376165 & 0.049937816 \\ 0.499843856 & 0 \\ 0 & -100.0002519 \\ 2.500011935 & 100.0000769 \\ -2.499971439 & 0 \end{bmatrix} \tag{3.44}$$

The covariance matrix of the observations is

$$\Sigma_{observations} = \begin{bmatrix} \sigma^2_{SD_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_{SD_2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2_{SD_3} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{Z_1} & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_{Z_2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{Z_3} \end{bmatrix} \quad (3.45)$$

Note that each slope distance was assumed to have a standard deviation of 0.10 meters and that each zenith direction has a standard deviation of 20 seconds of arc. Given that 20 seconds squared in radians is 9.401755*10^{-9}, the elements of the observation covariance matrix are

$$\Sigma_{obs} = \begin{bmatrix} 0.01 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.01 & 0 & 0 & 0 \\ 0 & 0 & 0 & 9.401755E\text{-}9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 9.401755E\text{-}9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 9.401755E\text{-}9 \end{bmatrix} \quad (3.46)$$

The covariance matrix of the derived radius and height is then computed as

$$\Sigma_{RH} = J\,\Sigma_{obs}\,J^T = \begin{bmatrix} 0.0049923218667 & -0.000247027986 \\ -0.000247027986 & 0.00043302309592 \end{bmatrix} \quad (3.47)$$

Note that the covariance matrix in equation 3.47 is almost the same as in equation 3.27 except that equation 3.47 contains covariance data for the derived measurements of radius and height. The standard deviation of the radius is 0.0707 m (not 0.07 m), and the standard deviation of the height is 0.0208 m (not 0.02 m). In the first example, radius and height were independent measurements (observations). In the second example, radius and height were both computed from the independent slope distance and zenith direction observations. Radius and height are not independent, and the covariance matrix contains correlation. The variance of the computed volume is now computed (the same procedure as in equation 3.28) as

$$\Sigma_{YY} = \begin{bmatrix} 7{,}853.98 & 3{,}141.59 \end{bmatrix} \begin{bmatrix} 0.0049923219 & -0.000247027 \\ -0.000247027 & 0.0004330231 \end{bmatrix} \begin{bmatrix} 7{,}853.98 \\ 3{,}141.59 \end{bmatrix}$$

$$\Sigma_{YY} = \textit{Variance of volume} = 300{,}034.86 \text{ m}^6 \quad (3.48)$$

The correlated standard deviation of the computed volume is 547.75 m^3.

When the results in equation 3.47 are compared to those in equation 3.28, the correlated results are somewhat smaller. However, in this case, when significant figures are taken into account, the overall reported answer is the same in each case. The tank volume is 78,500 m^3 +/– 550 m^3.

When does correlation make a significant difference? Each user must answer that question for him or herself. As technology permits observations to be made with greater and greater precision and as smaller tolerances are imposed upon the computed result, correlated measurements will need to be considered. The important point here is that the independent observations must be identified and that the equations (models) used to compute spatial data components will need to be used properly to compute the correlated covariance matrices.

As shown in the tank example, equation 3.23 is very powerful. Specifically, matrix tools were used to illustrate using both correlated and uncorrelated measurements. Many surveying measurements are uncorrelated, and, as shown here, even correlated measurements may give the same answer. In the past, the nonmatrix form of equation 3.23 has been quoted as the special law of propagation of variance, which is used without correlations as

$$\sigma_U^2 = \left(\frac{\partial U}{\partial X}\right)^2 \sigma_X^2 + \left(\frac{\partial U}{\partial Y}\right)^2 \sigma_Y^2 + \left(\frac{\partial U}{\partial Z}\right)^2 \sigma_Z^2 + \ldots \tag{3.49}$$

where $U = f(X,Y,Z\ldots)$ and $X/Y/Z$ are independent variables.

If the variables really are independent, equation 3.49 can be applied to simple equations to give error propagation equations listed in various textbooks and memorized by many as

$$U = \text{sum} = A + B, \ \ \sigma_{A+B} = \sqrt{\sigma_A^2 + \sigma_B^2} \tag{3.50}$$

$$U = \text{difference} = A - B, \ \ \sigma_{A-B} = \sqrt{\sigma_A^2 + \sigma_B^2} \tag{3.51}$$

$$U = \text{product} = A * B, \ \ \sigma_{A*B} = \sqrt{A^2\sigma_B^2 + B^2\sigma_A^2} \tag{3.52}$$

$$U = \text{quotient} = A / B, \ \ \sigma_{A/B} = \sqrt{\frac{\sigma_A^2}{B^2} + \left(\frac{A}{B}\right)^2 \frac{\sigma_B^2}{B^2}} \tag{3.53}$$

Even when used without correlations, the error propagation equation is a powerful tool that has been underutilized. But, the matrix form of the error propagation, equation 3.23, is even more powerful in that it utilizes the power of matrices to handle systems of complex equations and it handles any and all correlations that may be part of a problem, simple or complex.

ERROR ELLIPSES

Error ellipses are a graphical tool used to illustrate the pair-wise correlation that exists between computed values. Using 2-D plane coordinates as an example, if the correlation between the computed coordinates is zero, then the orientation of the error ellipse corresponds to that of the host coordinate system. Special case: if the correlation is zero and the standard deviations are the same for both coordinates, then the error ellipse is a circle (and orientation is immaterial). However, if the standard deviation is the same for both coordinates and the correlation is not zero, then the maximum and minimum standard deviations will occur with some other orientation.

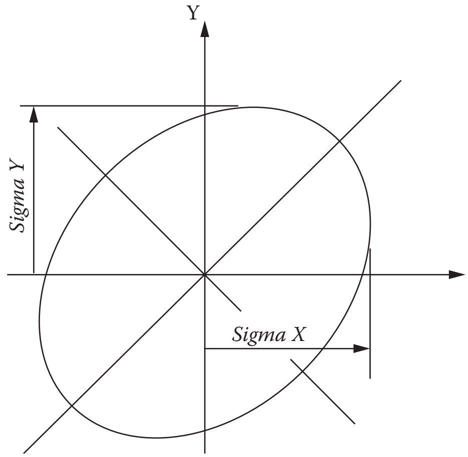


FIGURE 3.16 General Error Ellipse

The general case, illustrated in Figure 3.16, accommodates standard deviations of coordinates that are *not* the same with respect to the X and Y axes and in which correlation *does* exist. The respective *x*/*y* standard deviations and orientation of the ellipse major axis are obtained from the point covariance matrix. Additional material on error ellipses is given in Wolf and Ghilani (1997) and should be studied carefully in order to fully understand their usefulness. An incidental point is that error ellipses provide excellent visualization of correlation in two dimensions, typically in the horizontal plane. Visualization of 3-D error ellipsoids needs more study and discussion.

LEAST SQUARES

The principle of least squares states that the sum of the squares of the residuals—multiplied by their appropriate weights—will be a minimum for that set of answers (parameters) that has the greatest probability of being correct. The concept is simultaneously simple and complex because it applies with equal validity to computing a simple mean of two equally weighted measurements as well as adjustment of the most complex problem that can be described with functional model equations. Although it is correct to say that there is no known method proven to be better than a least squares solution, it is also true to say, within reason, that least squares can be used to obtain any desired answer. The difference lies in the selection of weights, and that is the responsibility of the user. The least squares procedure is specific and proven, but least squares can also be abused, sometimes unwittingly, to the point where a solution has questionable, marginal, little, or no value. The challenge in using least squares is to select the appropriate model, to write the equations (observation and/or condition) correctly, and to assign legitimate weights to the observations. Generating the solution could be a challenge (if done longhand), but computers are programmed to handle the matrices and to crunch the numbers needed to find the most probable solution to the problem. Having done all that, one could still argue that the most challenging part of using least squares is interpreting the results—such as using the covariance matrices to track three-dimensional spatial data accuracy.

Given one measurement of a distance, there is no basis for adjustment. In order to use least squares, there must be "extra" measurements. When tying in sideshots to a survey traverse, there is no "extra" measurement to the radial point, and no adjustment of that point is possible. But when computing around a closed loop traverse, the coordinates of the endpoint must be the same as those used for the beginning point. Once coordinates of each traverse point are computed, least squares is an appropriate procedure by which to find the most probable coordinate values of the surveyed points. Of course, if the sideshot points are tied in a second time from a separate survey point, redundancy does then exist and an adjustment of such redundant positions is possible.

In order to use least squares competently, the user must decide upon the appropriate model (is the coordinate system local, state plane, UTM, or geodetic) and write equations for the computations that utilize the measurements. As stated above, if there is no redundancy in the measurements, no adjustment is possible. But, given a loop traverse and redundant measurements, the equations used in the computations must be consistent with the model being used. In most cases, a slope distance must be reduced to horizontal, and the horizontal distance must be reduced to the ellipsoid if using geodetic coordinates or to the state plane coordinate grid (in the appropriate zone) if using state plane coordinates. Least squares cannot be used to correct errors caused by using the wrong model. Often, the model is implicit and defined by the context. For example, when measuring a distance with a plumb bob and steel tape, the implied model is horizontal distance, and everyone knows what that is. But, if the distance is measured with EDM or with GPS, it becomes more important to be specific about the definition of horizontal. For example, Burkholder (1991) gives six different definitions of horizontal, each being more precise (specific) than the previous one. It is the user's responsibility to assure compatibility between the measurements, the model, and the solution obtained from a least squares adjustment.

In addition to providing the best possible geometrical answer to a network of redundant observations, a least squares adjustment can also be used to determine the statistical properties (standard deviations) of the answers obtained from the adjustment. This provides the spatial data analyst with valuable tools (error propagation) for making decisions about how the answers are used or interpreted. For example, elevations of points on a building are determined very carefully and compared with elevations determined earlier (say, six months). Did the building move during that time interval? If the building did move, how much did it move? If the differences are small, does that mean the wall really did move? Or, is it possible that the observed difference is the result of accumulation of random errors in the measurements? Least squares, error propagation, error ellipses, and hypothesis testing are tools that can be used to make statements based upon inferences having a rigorous statistical foundation at a level of confidence chosen by the user.

Linearization

Taken by itself, the concept of least squares represents too much computational effort to be practical. However, if the least squares process is combined with matrices (for computational compactness) and with computers (for processing speed), a least squares solution becomes feasible for a greater variety of problems. A second drawback to using least squares is that matrices are valid only for systems of linear equations and many spatial data computations involve nonlinear geometrical relationships. That obstacle is overcome using a process called linearization in which nonlinear equations are replaced by their Taylor series approximation.

When using least squares to solve a linear problem, the solution is obtained on the basis of a single iteration. However, successive iteration is required when using least squares to solve nonlinear problems. Since only the first two terms of the Taylor series are used in the matrix formulation (point-slope form of a line), the solution (a set of corrections to the previously adopted values) of a set of equations will be only an approximation. Stated differently, a consequence of linearization is that the answer being sought changes from "the actual numerical value" to "What correction(s) to the previous approximate value(s) will provide a better solution?" (For linear problems, the correction is the value between zero and the correct answer.) The most important question is "What is an acceptable answer?" An acceptable answer is one that fulfills the original (nonlinear) conditions within some tolerance selected by the user. That puts the user in control.

A corollary to the previous question might be "How small must the corrections be to be acceptable?" Answering this question also keeps the user in control. But, achieving the goal of finding the right answer or making the corrections go to zero (within some tolerance) typically requires an enormous amount of number crunching—feasible only when done on a computer. Very briefly, the overall process for solving a nonlinear problem is as follows:

1. Identify the geometry of the problem and write the appropriate equations.
2. Linearize the equations and take partial derivatives to be used in the matrix formulation.
3. Establish some initial value for each unknown parameter as being reasonably close to the final answer.
4. Run the least squares adjustment to find corrections to the initial estimates.
5. Look at the results. Are the corrections small enough to quit? If so, do. If not, update the previous estimate using current corrections and run the adjustment again. Are the corrections smaller, and are they small enough?

Two important concepts described above are iteration and convergence. Iteration is the process of using results from a previous solution to solve the problem again. Convergence is the desirable condition realized when each successive correction is smaller than the previous one. If a solution converges slowly, it may take many iterations to solve a problem. Given that computers are programmed to do the number crunching, the time and effort required may or may not be an issue. If solutions are being generated in real time, rapid convergence is preferable and linear models that do not require iteration are even more desirable. When using the GSDM, network adjustments can be formulated as a linear model.

APPLICATIONS TO THE GLOBAL SPATIAL DATA MODEL (GSDM)

Given that this book is devoted to describing the global spatial data model (GSDM), it will be shown in later chapters how the various mathematical concepts in this chapter are combined in a single comprehensive model, the GSDM. Very briefly:

- Observations are independent measurements of fundamental physical quantities.
- Observations are manipulated in conformance with physical laws and geometrical relationships to obtain spatial data components. For example, carrier phase GPS observations are processed to obtain geocentric components, while total station survey measurements provide local components. Each can be used and/or combined with the other when using the GSDM.
- The standard deviation of each observation is propagated to the corresponding spatial data component. Correlation data are tracked in the associated covariance matrix. Correlation is largely responsible for the difference between network accuracy and local accuracy from one point to another (Burkholder 1999, 2004).
- Spatial data components (indirect measurements) are combined into networks according to existing geometrical configurations and evaluated for statistical reliability.
- Once spatial data components pass rigorous quality control criteria, they are combined with existing primary database points in a least squares adjustment.
- Stored values include the geocentric *X*/*Y*/*Z* values of each point, the associated covariance values, and the point-correlation values.
- When drawn from the 3-D database, the spatial accuracy of each point is given by its standard deviation, component by component. The value or utility of each database point is determined by whether it passes a tolerance filter (for each component) as selected by the user.
- Standard deviations of any or all derived quantities are available using proven standard error propagation computations.
- The local accuracy between (especially directly connected) points can be determined using the full covariance matrix in Burkholder (1999, equation 9).

REFERENCES

Alder, K. 2002. The measure of all things: The seven-year odyssey and hidden error that transformed the world. New York. The Free Press.

Beckman, P. 1971. *A history of Pi*. New York: St. Martin's.

Bumby, D. R., and R. J. Klutch. 1982. *Mathematics: A topical approach*. Columbus, OH: Merrell.

Burkholder, E. F. 1982. Algorithm for accurate area computations. *Surveyor Calculations Journal,* 92 (7):9–10. (Joe Bell, publisher, San Bernardino, CA 92412).

———. 1991. Level/horizontal distances. *Journal of Surveying Engineering* 117 (3): 104–16.
———. 1999. Spatial data accuracy as defined by the GSDM. *Surveying & Land Information Systems* 59 (1): 26–30.
———. 2004. *Fundamentals of spatial data accuracy and the global spatial data model (GSDM)*. Washington, DC: Filed with the U.S. Copyright Office.
Chen, W. F., ed. 1995. *The civil engineering handbook*. Boca Raton, FL: CRC Press.
Cressie, N. A. C. 1993. *Statistics for spatial data*. New York: John Wiley.
Davis, R., F. Foote, J. Anderson, and E. Mikhail. 1981. *Surveying: Theory and practice*. New York: McGraw-Hill.
Dolciani, M. P., E. F. Beckenbach, A. Donnelly, R. Jurgensen, and W. Wooton. 1967. *Modern introductory analysis*. Boston: Houghton Mifflin.
Dwass, Meyer. 1970. *Probability: Theory and applications*. New York: W. A. Benjamin.
Hawking, Stephen W. 1988. *A brief history of time*. New York: Bantam.
Mikhail, Edward. 1976. *Observations and least squares*. New York: Donnelly.
Moffitt, F., and J. Bossler. 1998. *Surveying*, 10th ed. New York: Addison-Wesley.
Russell, Bertrand. 1959. *Wisdom of the West*. London: Crescent Book.
Smith, J. R. 1986. *From plane to spheroid: Determining the figure of the Earth from 3000 b.c. to the 18th century Lapland and Peruvian survey expeditions*. Rancho Cordova, CA: Landmark Enterprises.
Stanfel, L. 1994. Survey distance units: A better way. *Journal of Surveying Engineering* 120 (3): 130–32. (See also subsequent discussion in vol. 122 [1]: 41–44.)
Tooley, R. V. [1949] 1990. *Maps and mapmakers*. New York: Dorset.
Wilford, J. N. 1981. *The mapmakers*. New York: Vintage.
Wine, R. L. 1964. *Statistics for scientists and engineers*. Englewood Cliffs, NJ: Prentice-Hall.
Wolf, P., and C. Ghilani. 1997. *Surveying adjustments and statistics for GIS and surveying*. New York: John Wiley.
———. 2005. *Elementary surveying*, 11th ed. New York: HarperCollins.

4 Geometrical Models for Spatial Data Computations

INTRODUCTION

As previously defined, a mathematical model is a set of rules used to make a conceptual connection between abstract concepts and human experience. A model is judged "good" to the extent it is both simple and appropriate. When working with spatial data, the simplest model is a one-dimensional (1-D) distance. Models of increasing complexity include a 2-D plane coordinate system formed by the perpendicular intersection of the X and Y axes, a generic 3-D *X*/*Y*/*Z* rectangular Cartesian coordinate system having three mutually perpendicular axes, a spherical Earth model, and, finally, the ellipsoidal Earth model. Figure 4.1a shows the standard 2-D *X*/*Y* coordinate system, Figure 4.1b shows a right-handed *X*/*Y*/*Z* coordinate system, and Figure 4.1c illustrates the sexagesimal coordinate system of latitude and longitude used to describe the geodetic location of points on the Earth's ellipsoidal surface. Other choices will be discussed later, but the goal at this point is to identify a variety of geometrical model choices. With regard to working with spatial data, considerations include, but are not necessarily limited to, the following:

- Are the observations or subsequently computed measurements 1-D, 2-D, or 3-D?
- Is a 1-D or 2-D model sufficient, or is a 3-D model required?
- Is the extent of a given project small enough to use "flat-Earth" relationships, or is a different model needed to accommodate the Earth's curvature? Is a spherical-Earth model appropriate, or is the ellipsoidal Earth model required?
- Is the project of such a nature that a local coordinate system is sufficient, or should the data be referenced to the National Spatial Reference System (NSRS)?
- What issues of compatibility (e.g., units of feet or meters) must be addressed? What is required for new measurements to be compatible with and/or add to the value of existing data?
- Is there a spatial data model that accommodates all computational concerns? If so, what is it? That decision should be documented specifically for each project. Otherwise subsequent users are forced to infer the model from the way spatial data are used. For example, project datum coordinates (also called surface coordinates) often resemble state plane coordinates, and serious problems may result if project datum coordinates are used as if they were state plane coordinates.

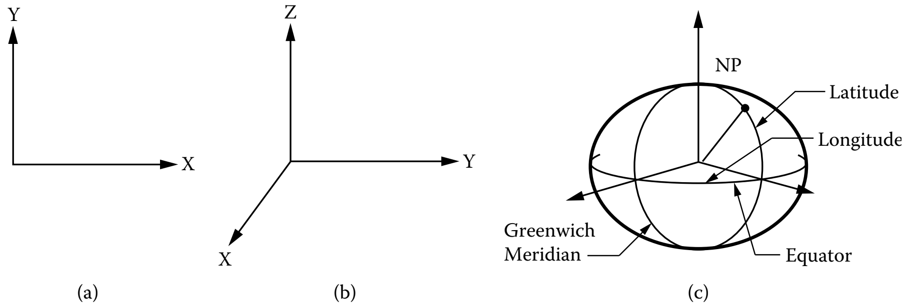


FIGURE 4.1 2-D, 3-D, and the Geodetic Coordinate Systems

CONVENTIONS

With regard to the use of various models for handling spatial data, the following conventions are, although contradictory at times, commonly used.

- Standard computational practice in the scientific, engineering, and mathematical disciplines employs the right-handed rectangular 3-D coordinate system described earlier and shown in Figure 4.1b. The conventional 2-D *X*/*Y* rectangular coordinate system is a subset of the right-handed 3-D *X*/*Y*/*Z* system.
- Conventional surveying and mapping practice includes use of latitudes and departures, northings and eastings, and *x*/*y* coordinates in the 2-D plane. The goal here is to accommodate existing practice to the extent possible. However, the convention of east/north/up (not north/east/up) will be used for local perspective coordinates because it is right-handed and compatible with the underlying geocentric ECEF system described in chapter 2. Regretfully, clockwise azimuth is a left-handed convention.
- Coordinate differences between points are denoted as "Δ" and computed as forepoint (point 2) minus standpoint (point 1). Given that computations are performed with respect to the standpoint and that another point of interest is the forepoint, rectangular spatial data (vector) components are computed as

$$\Delta x = x_2 - x_1 \tag{4.1}$$

$$\Delta y = y_2 - y_1 \tag{4.2}$$

$$\Delta z = z_2 - z_1 \tag{4.3}$$

- Directions are given in sexagesimal units as bearings or azimuths. An azimuth is counted from north clockwise through a complete circle of 360°. Negative azimuths imply a counterclockwise (angle left) rotation. A negative azimuth can be changed to a positive azimuth by adding some multiple of 360°. Bearings are quadrant based and related to azimuths as follows:
    - Quadrant I: a northeast (NE) bearing has the same value as an azimuth.
    - Quadrant II: a southeast (SE) bearing is 180°– the azimuth of the line.

- Quadrant III: a southwest (SW) bearing is the azimuth of the line minus 180°.
- Quadrant IV: a northwest (NW) bearing is 360°minus the azimuth of the line.

- Bearings and azimuths are distinguished by the meridian to which they are referenced. Common examples are as follows:
  - Magnetic bearings are referenced to lines of the Earth's magnetic field that converge at the magnetic poles. Given that the magnetic pole is not coincident with the true pole (as defined by the Earth's spin axis), magnetic declination is the difference between magnetic north and true north.
  - An astronomic azimuth is determined by observing the sun and/or stars using a transit or theodolite leveled with respect to the local plumb line (vertical). An astronomical meridian goes to the Earth's instantaneous spin axis. This pole position changes slowly and differs slightly from the Conventional Terrestrial Pole (CTP), the mathematical North Pole, adopted by international agreement.
  - Geodetic azimuth is referenced to the meridian to the CTP with respect to the ellipsoid normal (instead of the vertical) and is used in geodetic position computations. A geodetic azimuth differs from an astronomic azimuth due to (1) the difference between the direction of the ellipsoid normal and the direction of the vertical (deflection-of-the-vertical), and (2) the difference between the instantaneous North Pole and the CTP (polar wandering). See chapter 6 for more detail on each difference.
  - Grid azimuths are commonly encountered when working with state plane (or map projection) coordinates. All geodetic meridians converge to the CTP, but grid meridians are generally parallel on the projection surface. At the center of each map projection zone, the central meridian coincides with the true geodetic meridian. Convergence is the difference between a geodetic azimuth and its corresponding grid azimuth from one point to another. Convergence is zero on the central meridian of a map projection.
  - A 3-D azimuth is very nearly the same as a geodetic azimuth but is much easier to compute. The 3-D azimuth is computed as $\tan^{-1}(\Delta e/\Delta n)$, where Δe and Δn are the local tangent plane components of a 3-D (GPS) vector. See Burkholder (1997) and/or chapter 11 for more details.
  - An assumed azimuth, encountered in many places, is whatever the user declares it to be.
- Most scientific calculators operate in the decimal degree mode when computing trigonometric functions. It is important to convert sexagesimal degrees/minutes/seconds to decimal degrees before computing a trigonometric function. Time is also expressed in sexagesimal units, and some calculators provide a hardwired function for converting hours/minutes/seconds to decimal hours. It is the same sexagesimal conversion. Look for D/M/S to DD or H/M/S to HR for converting sexagesimal to decimal. Or, it can be done easily on the keyboard as

D.D (decimal degrees) = *degrees* + *minutes*/60 + *seconds*/3600

The opposite computation is also important. When computing inverse trigonometric functions such as $\tan^{-1}(\Delta x/\Delta y)$ for azimuth, the answer is displayed by many calculators as decimal degrees. The function to convert decimal degrees to degrees/minutes/seconds is hardwired in many calculators as *HR* to *H*/*M*/*S* (or *DD* to *D*/*M*/*S*). If not hardwired in the calculator, the conversion can also be done on the keyboard.

The keyboard procedure for converting decimal degrees to degrees/minutes/seconds is as follows:

Degrees: integer portion of *DD.DDDDDDDDD*
Minutes: integer portion of ([*DD.DDDDDDD* – *degrees*] * 60)
Seconds: DD.DDDDDDDD*3600 – degrees*3600 – minutes*60

Some calculators will show symbols for degrees/minutes/seconds. An alternative formatting convention for showing angular units in a calculator display is as follows:

Decimal degrees: *DD.DDDDDDDDDDD*
Degrees/minutes/seconds: *DD.MMSSSSSSS* (with an implied decimal point in the seconds as SS.SSSS)

Another practice used by some software packages includes the follow:

Decimal degrees: *DD.DDDDDDDD*
Degrees/minutes/seconds: *DDMMSS.SSSSSSS*

Standard degree/minute/second symbols, the use of zeros for placeholders, and the placement of the decimal point are illustrated as follows:

Azimuth = 038° 00’ 03.”44563, where the seconds symbol (") is either over or follows the decimal point in the seconds. Otherwise, the (") symbol might be interpreted as meaning inches.

Of course, a calculator may also be set to operate in radian units, in which case the conversions are (using π or *SPR* = 206,264.”806247096 per radian)

radians = *DD* * (π/180°) and *DD* = *radians* * (180°/π),
radians = *DD* * 3600/*SPR* and *DD* = *radians* * *SPR*/3600, or
radians = seconds/SPR and seconds = radians * SPR.

- When geodetic latitude or longitude is given in sexagesimal units, it is customary to show five decimal places of seconds for control point positions. If the Earth were a sphere with a radius of 6,372,000 meters, 1 second of arc would correspond to a distance of 30.89 meters in the north-south direction. So, 0.”00001 seconds of arc in the north-south direction correspond to 0.000309 meters (submillimeter accuracy). It is also understood that

latitude is used as a positive value (0° to 90°) in the northern hemisphere and as a negative value (0° to –90°) in the southern hemisphere; and
- longitude starts with 0° at the Greenwich meridian and is used 0° to 360° eastward. However, at the October 1884 International Meridian Conference held in Washington, D.C., it was agreed that longitude would be counted both east and west from the Greenwich meridian up to 180° (Clarke 2000). Therefore, west longitude is the standard practice for geographic locations in the western hemisphere. Using a west longitude as a negative number is compatible with the mathematically unambiguous practice of using 0° to 360° eastward.
- For many spatial data applications, horizontal distance is computed from plane surveying latitudes and departures. As such, it is the hypotenuse of a plane right triangle. But, when working with different elevations, horizontal distance is also taken to be the right-triangle component of a slope distance. Depending upon the coordinate system being used, other definitions of horizontal distance include the geodetic distance on the ellipsoid or the grid distance as used in the state plane coordinate systems.

Conventions used by the GSDM include horizontal distance as the local tangent-plane right-triangle component of a GPS vector (the same as used in plane surveying) and the 3-D azimuth (in degrees/minutes/seconds) as computed from the latitude/departure components of the same horizontal distance. To the extent possible, the GSDM includes relevant plane-surveying practices without sacrificing the advantages of a rigorous connection to a National Spatial Reference System.

TWO-DIMENSIONAL CARTESIAN MODELS

The standard 2-D rectangular coordinate system has an origin formed by the perpendicular intersection of the abscissa and the ordinate. Two systems, called the Math/Science Reference System and the Engineering/Surveying Reference System, are quite similar, but, from one system to the other, the reference axes are different and the direction of positive rotation is reversed (see Figure 4.2a and Figure 4.2b).

Math/Science Reference System

The 2-D coordinate system commonly used by mathematicians and scientists labels the abscissa as the X-axis and the ordinate as the Y-axis. The positive X-axis is considered to be the reference for angles, and rotation is counted positive counterclockwise. In this system, the Δx and Δy components of any directed line segment (vector) are

$$\Delta x = d \cos \theta \tag{4.4}$$

$$\Delta y = d \sin \theta \tag{4.5}$$

where

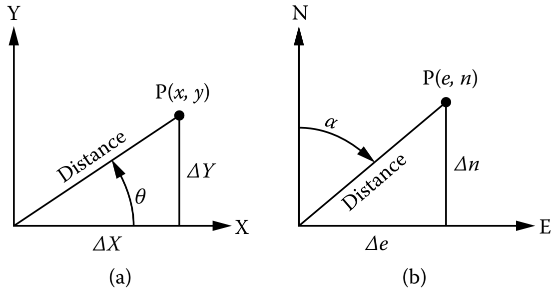


FIGURE 4.2 Math/Science and Engineering/Surveying Coordinate Systems

d = length (distance) of the directed line segment, and
θ = direction of the line segment as measured counterclockwise with respect to positive X-axis (consistent with right-handed rule).

Engineering/Surveying Reference System

Surveyors, engineers, and others who work with mapping data often use a 2-D rectangular coordinate system, which is similar to the math/science system except that cardinal directions of north-south and east-west are superimposed upon the two axes and rotation is counted clockwise from north. In this system, the Δx (*easting*) and Δy (*northing*) components of any directed line segment are

$$\Delta x = \Delta e = d \sin \alpha \tag{4.6}$$

$$\Delta y = \Delta n = d \cos \alpha \tag{4.7}$$

where

d = length (distance) of the directed line segment, and
α = azimuth of the line segment as measured clockwise with respect to north, the positive Y-axis (contrary to right-handed rule).

The specific relationship between the math/science system and the engineering/surveying system is that, in all cases, $\alpha + \theta = 90°$, which also means $\alpha = 90° - \theta$ or $\theta = 90° - \alpha$. Being aware of these similarities and differences helps spatial data users make greater use of the polar/rectangular conversions hardwired into many calculators.

For example, most scientific calculators are hardwired according to the math/science system. To compute rectangular components of a line 100.00 meters long having an azimuth of 30°, the calculator will show 50.000 meters as being the north-south component when it is really the east-west component. Similarly, because it is using the math/science convention, the calculator will show 86.602 meters as being the east-west component when it really is the north-south component (in the engineering/surveying system). The only thing the user needs to do is switch the label of the computed components. A word of caution: each user should practice and work with known quantities to make sure the procedure being used is giving correct answers in the intended system.

In each case, the 2-D math/science reference system and the 2-D engineering/surveying reference system can be expanded into a 3-D system by adding a Z-axis. The convention adopted in this book is to use *e*/*n*/*u* for the local reference frame because it is right-handed and $\Delta e/\Delta n/\Delta u$ can be conveniently rotated into the ECEF right-handed *X*/*Y*/*Z* reference frame. Regretfully, azimuth in the engineering/surveying system is not consistent with the right-hand rule.

COORDINATE GEOMETRY

Once rectangular components are obtained in either the math/science or engineering/surveying system, standard coordinate geometry (often referred to as COGO) operations are the same. Admittedly, it becomes a challenge to keep track of various conventions employing *x*/*y* coordinates and *eastings*/*northings* in the same rectangular system, but the COGO procedures are the same in each case. Burkholder (1984) gives a derivation for the following 2-D COGO operations listed as follows:

1. Forward (traverse to new point)
2. Inverse (find direction and distance from standpoint to forepoint)
3. Line-line (also called bearing-bearing) intersection
4. Line-circle (also called bearing-distance) intersection
5. Circle-circle (also called distance-distance) intersection
6. Perpendicular offset distance from a line to a point

Forward

Given:

e_1 and n_1 = coordinates at standpoint
d = distance from standpoint to forepoint
α = azimuth from standpoint to forepoint

Compute:

Δe and Δn = rectangular components
e_2 and n_2 = coordinates of forepoint

Solution:

$$e_2 = e_1 + \Delta e = e_1 + d \sin \alpha \tag{4.8}$$

$$n_2 = n_1 + \Delta n = n_1 + d \cos \alpha \tag{4.9}$$

Inverse

Given:

e_1 and n_1 = coordinates at standpoint
e_2 and n_2 = coordinates at forepoint

Compute:

Δe and Δn = rectangular components
d = distance between standpoint and forepoint
α = azimuth from standpoint to forepoint

Solution:

$$\Delta e = e_2 - e_1 \text{ and } \Delta n = n_2 - n_1 \tag{4.1 and 4.2}$$

$$d = \sqrt{(\Delta e^2 + \Delta n^2)} \tag{4.10}$$

$$\tan \alpha = \Delta \text{e}/\Delta n$$

(Use signs of Δe and Δn to determine the proper quadrant.)

$$\text{If } \Delta n \text{ is } -,\ \alpha = 180° + \arctan(\Delta e/\Delta n). \tag{4.11}$$

$$\text{If } \Delta n \text{ is } + \text{ and } \Delta e \text{ is } -,\ \alpha = 360° + \arctan(\Delta e/\Delta n). \tag{4.12}$$

$$\text{If } \Delta n \text{ and } \Delta e \text{ are both } +,\ \alpha = \arctan(\Delta e/\Delta n). \tag{4.13}$$

Intersections

When performing design or other COGO computations, it is often necessary to determine where lines intersect (see Figure 4.3). Three methods are as follows:

1. Line-line: in this case, points 1 and 2 are given. The directions from point 1 and the direction to point 2 are also given. The problem is to compute the coordinates of the intersection point. Line-line is also called a bearing-bearing intersection.
2. Line-circle: in this case, points 1 and 2 are given. The direction from point 1 and the distance from the intersection point to point 2 are given. The problem is to compute the coordinates of the intersection point. Line-circle is also called a bearing-distance intersection.

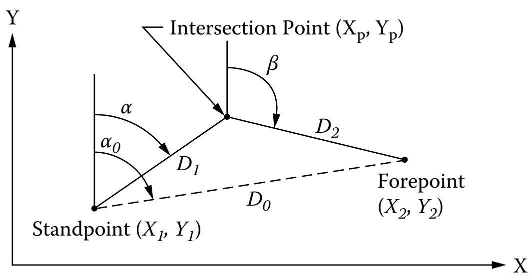


FIGURE 4.3 Geometry of Intersections

3. Circle-circle: in this case, points 1 and 2 are given. The distance from point 1 to the intersection point and the distance from the intersection point to point 2 are given. The problem is to compute the coordinates for the intersection point. Circle-circle is also called a distance-distance intersection.

In each case, the computation starts at the standpoint and ends at the forepoint while establishing the location of an intermediate intersection point. The rules to be used in establishing the computed intersection point vary according to the type of intersection, but they fall into one of the three categories listed. Symbols and conventions for the computations are follows:

(e_1,n_1) = coordinates of standpoint (given)
(e_2,n_2) = coordinates of forepoint (given)
(e_p,n_p) = coordinates of intersection point (computed result)
d_0 = distance from standpoint to forepoint
d_1 = distance from standpoint to intersection point
d_2 = distance from intersection point to forepoint
α_0 = azimuth from standpoint to forepoint
α = azimuth from standpoint to intersection point
β = azimuth from intersection point to forepoint

In each case, the solution starts by computing $\Delta e = e_2 - e_1$ and $\Delta n = n_2 - n_1$. Also note that no solution exists if, in the first case, the lines are parallel; in the second case, the line does not intersect the circle; and, last, the circles do not intersect. A different mathematical impossibility is encountered in each case.

Line-line: if the lines are parallel, azimuths α and β are the same and the denominator of equation 4.14 goes to zero. As a result, the distance d_1 is undefined. There is no intersection. See Figure 4.4a.

Line-circle: if the line does not intersect the circle, the perpendicular offset distance from the line to the circle is greater than the radius of the circle. If that happens, the quantity under the radical in equation 4.15 is negative. Since it is not possible to take the square root of a negative number, d_1 is undefined and there is no intersection. See Figure 4.4b.

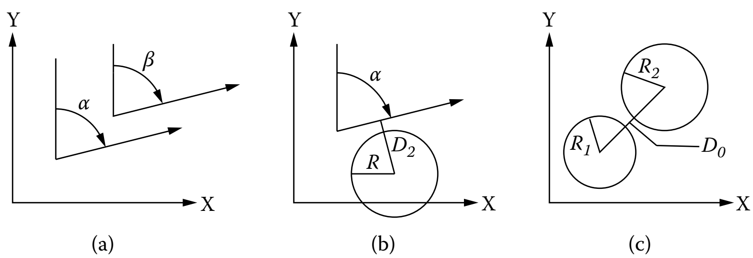


FIGURE 4.4 Examples of Intersections That Fail

Circle-circle: if the two circles do not intersect, it is not possible to find the angle γ using equation 4.16 because, in one case, the value of cos γ is greater than 1.0 and, in the other case (one circle entirely within the other), the value of cos γ is less than –1.0. See Figure 4.4c.

Other than the nonintersection cases described, the following formulas provide very efficient procedures for computing intersections in a two-dimensional plane.

Line-Line: One Solution or No Solution if Lines Are Parallel

Given:

$$e_1, n_1, e_2, n_2, \alpha, \text{ and } \beta \ (\Delta e = e_2 - e_1 \text{ and } \Delta n = n_2 - n_1)$$

Compute:

$$d_1 = (\Delta e \cos \beta - \Delta n \sin \beta) / \sin (\alpha - \beta) \tag{4.14}$$

(d_1 may be either a positive or negative value.)

Solution:

$$e_p = e_1 + d_1 \sin \alpha \tag{4.8}$$

$$n_p = n_1 + d_1 \cos \alpha \tag{4.9}$$

Check: inverse intersection point to forepoint, and compare computed direction β with given direction β. If they are not the same, a mistake was made.

Line-Circle: May Have Two Solutions, One Solution, or No Solution

Given:

$$e_1, n_1, e_2, n_2, \alpha, \text{ and } d_2$$

Compute:

$$d_1 = \Delta e \sin \alpha + \Delta n \cos \alpha \pm \sqrt{d2 - (\Delta ex \cos \alpha - \Delta n \sin \alpha)^2} \tag{4.15}$$

(d_1 normally has two values, one for each solution.)

Solution:

$$e_p = e_1 + d_1 \sin \alpha \tag{4.8}$$

$$n_p = n_1 + d_1 \cos \alpha \tag{4.9}$$

Check: inverse intersection point to forepoint, and compare computed distance d_2 with given distance d_2. They should be the same. Also make sure the solution obtained is the one desired. Depending upon where the line intersects the circle, the values of d_1 could both be positive, could be one positive and one negative, or could both be negative.

Circle-Circle: May Have Two Solutions, One Solution, or No Solution

Given:

$$e_1, n_1, e_2, n_2, d_1, \text{ and } d_2$$

Compute:

$$d_0 = \sqrt{(\Delta e^2 + \Delta n^2)} \quad (4.10)$$

$$\alpha_0 = \tan^{-1} (\Delta e/\Delta n) \quad (4.11, 4.12, \text{ or } 4.13)$$

$$\gamma = \cos^{-1} (d_1^2 + d_0^2 - d_2^2) / (2\, d_1\, d_0) \quad (4.16)$$

$$\alpha = \alpha_0 + \gamma \text{ and } \alpha = \alpha_0 - \gamma \quad (\text{two solutions; 4.17 and 4.18})$$

Solution:

$$e_p = e_1 + d_1 \sin \alpha \quad (4.8)$$

$$n_p = n_1 + d_1 \cos \alpha \quad (4.9)$$

Check: inverse intersection point to forepoint, and compare computed distance d_2 with given distance d_2. They should be the same. Also make sure the solution obtained is the one desired. Several solutions may exist. Also be aware that equation 4.16 has no solution if the two circles do not intersect.

Perpendicular Offset

In many situations, it is desirable to find the perpendicular distance from a line to a point. Given one is on a standpoint $P(e_1,n_1)$ and looking in the direction of a line, α, the problem (as illustrated in Figure 4.5) is to find the perpendicular offset distance right (positive) or left (negative) from the line to the point specified as the forepoint $P(e_2,n_2)$. Note that the azimuth, α, may be any azimuth from 0° to 360°. Using the

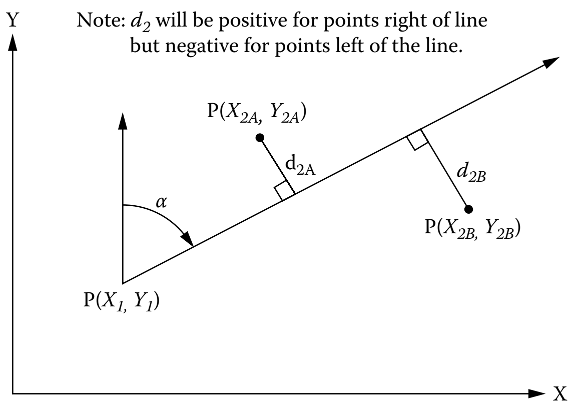


FIGURE 4.5 Perpendicular Offset to a Line

conventions given earlier, the perpendicular offset distance is the distance d_2 computed as

$$\Delta e = e_2 - e_1 \text{ and } \Delta n = n_2 - n_1$$

$$d_2 = \Delta e \cos \alpha - \Delta n \sin \alpha \tag{4.19}$$

Area by Coordinates

Area is the product of length times width and is generally presumed to be computed on a flat surface. Computing area on a spherical or ellipsoidal surface is more of a challenge and is addressed in chapter 6. The method generally used for computing the area of an irregular tract is area-by-coordinates, as used in chapter 3. Area by double-meridian-distance (DMD), useful if working with latitudes and departures, is described in many surveying texts and is not presented here. Development of the area (equation 3.1) involves adding and subtracting trapezoids, as shown in Figure 4.6.

Area = Trapezoid I + Trapezoid II – Trapezoid III – Trapezoid IV

Trapezoid I: A-1-2-B-A	Area = $0.5(e_1 + e_2)(n_1 - n_2)$
Trapezoid II: B-2-3-D-B	Area = $0.5(e_2 + e_3)(n_2 - n_3)$
Trapezoid III: C-4-3-D-C	Area = $0.5(e_3 + e_4)(n_3 - n_4)$
Trapezoid IV: A-1-4-C-A	Area = $0.5(e_4 + e_1)(n_4 - n_1)$

Combining the four trapezoids into one equation and multiplying by two gives

$$2A = (e_1 + e_2)(n_1 - n_2) + (e_2 + e_3)(n_2 - n_3) - (e_3 + e_4)(n_4 - n_3) - (e_4 + e_1)(n_1 - n_4)$$

Considerable algebraic manipulation and combination of terms are needed to get

$$2A = n_1e_2 + n_2e_3 + n_3e_4 + n_4e_1 - (e_1n_2 + e_2n_3 + e_3n_4 + e_4n_1) \tag{4.20}$$

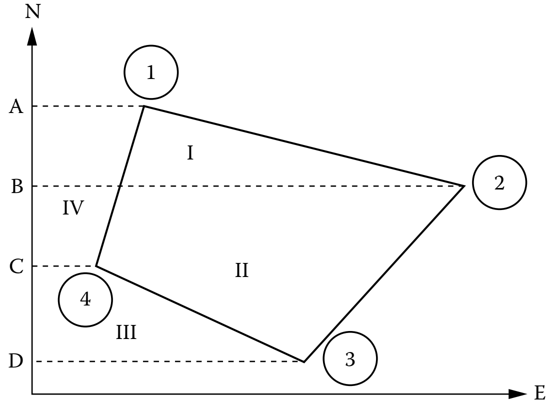


FIGURE 4.6 Area by Coordinates

Comments about equation 4.20 are as follows:

- The coordinates are often arranged in tabular form with the beginning values listed again at the end. The computation is then illustrated by showing accumulation of cross products, and area is computed as half the difference of the two accumulated sums. See the example in chapter 3 using equation 3.1.
- If points around a figure are chosen in a clockwise sequence, as was done in the derivation, a positive area is computed. If a counterclockwise sequence is used, a negative area will be the result. Although each user is free to "traverse" a parcel in either direction, the convention is to use positive area.
- Care must be exercised to assure that area is being computed for a legitimate figure. If one line of the figure crosses another (e.g., due to erroneous point sequence input), the equation will provide an answer that reflects removal of the crossover area.
- Significant figures may become an issue if equation 4.20 is used with large (i.e., state plane) coordinates. See the example in chapter 3. Equation 3.2 is recommended as the preferred area-by-coordinates formula.

CIRCULAR CURVES

Circular logic defines a curve as a line that is not straight and a straight line as one that is not curved. Separating circular geometry from circular logic, a circle is a curve that has completed one cycle, or, said differently, a curve is a portion of a circle. There are two fundamental definitions that relate a circle to the value of pi (π) and to the measure of an angle.

Definitions

$$\text{Definition 1: } \pi \equiv \text{circumference / diameter} = C/2R \tag{4.21}$$

$$\text{Definition 2: angle in radians} \equiv \text{arc length / radius} = L/R \tag{4.22}$$

Using the two definitions, it is quickly established that there are 2π radians in a complete circle (360°) and that any arc length is the product of the radius times the subtended angle in radians. One of the simplest, most powerful equations available to the spatial data user is

$$L = R\,\theta = \mathrm{R}\,\Delta^\circ\,(\pi / 180^\circ) \tag{4.23}$$

where

L = arc length,
R = radius of curve,
θ = a subtended angle in radians,
Δ° = a subtended angle in decimal degrees, and
π = the value of pi.

Degree of Curve

Circular curves are frequently encountered when working with road, street, or highway data. Distance along a road centerline is often measured in stationing (increments of 100 feet or other units), and straight centerline segments between the angle points are called tangents. The angle points in a centerline are called points of intersection (PI's). A curve is used at each *Pi* to provide for a gradual, rather than an abrupt, change in direction. Historically, the sharpness of a curve has been defined by degree-of-curvature ($D°$) according to one of two definitions: one definition for railroad work, and one for highway applications. The $D°$ definitions and the equations for computing radius from each of them are as follows:

degree of curve (**highway**) ≡ central angle subtended by an arc of 100 feet

$$radius = 100 / [D° (\pi / 180°)] = 5{,}729.57795 / D° \quad (4.24)$$

$$D° = 5{,}729.57795 / radius \quad (4.24a)$$

degree of curve (**railroad**) ≡ central angle subtended by a chord of 100 feet

$$radius = (100 / 2) / \sin (D° / 2) = 50 / \sin 0.5D° \quad (4.25)$$

$$D° = 2 \sin^{-1} (50 / radius) \quad (4.25a)$$

Table 4.1 compares values of radii for several values of $D°$ by each of the two definitions. Note that values of radius decrease linearly for the highway definition, but not for the railroad definition.

TABLE 4.1
Comparison of Degree of Curve Radii

Degree of Curve	Highway Definition	Railroad Definition
1°	5,729.578'	5,729.651'
2°	2,864.789'	2,864.934'
5°	1,145.915'	1,146.279'
10°	572.958'	573.686'
20°	286.479'	287.939'
30°	190.986'	193.185'
45°	127.324'	130.656'
90°	63.662'	70.711'

Elements and Equations

The degree-of-curve definitions are not compatible with the metric system (SI). For that and other reasons, modern practice tends to use the radius and another element to define a circular curve. As shown in Figure 4.7, there are five primary circular curve elements routinely encountered when working with spatial data involving circular curves. Of the ten possible combinations of five elements taken two at a time, the combinations of tangent/length and tangent/long chord are rarely used because of their weak geometry and computational complexity (see, for example, Thompson 1974). Given any other pair of primary circular curve elements, the solution can be found using some combination of equations 4.26 to 4.31.

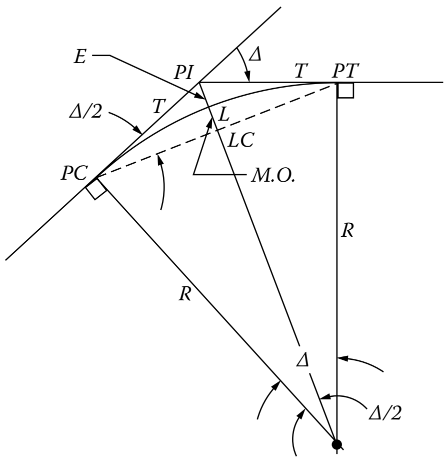


FIGURE 4.7 Elements of a Circular Curve

radius = R
tangent = T
long chord = LC
length = L
central angle = $\Delta°$

Secondary curve elements also labeled in Figure 4.7 include the following:

point of intersection = *PI* = intersection of straight-line tangents
point of curvature = *PC* = centerline station at beginning of curve
point of tangency = *PT* = centerline station at end of curve
intersection angle = $\Delta°$ = same as the central angle of the curve
deflection angle = $\Delta°/2$ = half the central angle
middle ordinate = *M.O.* = radial distance from chord to arc
external = E = radial distance from arc of curve to *PI*

With reference to Figure 4.7, note the following:

- The word “tangent” is used twice. The straight centerline segment between *PI*’s is called a tangent. And, one of the curve elements, the distance from the *PC* to the *PI*, is called the tangent (of the curve). Added to those two uses, the word “tangent” is also associated with the use of a trigonometric function. The context of usage generally dictates which meaning is intended.
- The radius of the curve meets the centerline (both tangents) at a right angle.

- The angle of intersection is the same as the central angle of the curve.
- The deflection angle is half the central angle.
- The diagram is symmetric. The distance *PC* to *PI* equals the distance *PI* to *PT*, and a radial line to the *PI* is perpendicular to the long chord. This right angle statement should not be accepted at face value, but each reader should prove it to him or herself.

Using $L = R\theta$, the definitions of trigonometric ratios, the labels assigned to the curve elements, and the diagram in Figure 4.7, it is possible to write the following relationships (⇒ means "implies that"):

$$L = R\Delta^\circ\left(\frac{\pi}{180^\circ}\right) \rightarrow R = \left(\frac{L}{\Delta^\circ}\right)\left(\frac{180^\circ}{\pi}\right) \tag{4.26}$$

$$\Delta^\circ = \left(\frac{L}{R}\right)\left(\frac{180^\circ}{\pi}\right) \tag{4.27}$$

$$\tan\left(\frac{\Delta}{2}\right) = \frac{T}{R} \rightarrow T = R\tan\left(\frac{\Delta}{2}\right) \tag{4.28}$$

$$\mathrm{R} = \frac{T}{\tan\left(\frac{\Delta}{2}\right)} \tag{4.29}$$

$$\sin\left(\frac{\Delta}{2}\right) = \frac{LC}{2R} \rightarrow LC = 2R\sin\left(\frac{\Delta}{2}\right) \tag{4.30}$$

$$R = \frac{LC}{2\sin\left(\frac{\Delta}{2}\right)} \tag{4.31}$$

$$M.O. = R - R\cos\left(\frac{\Delta}{2}\right) = \mathrm{R}\left(1\text{-}\cos\left(\frac{\Delta}{2}\right)\right) \tag{4.32}$$

$$E = \frac{R}{\cos(\Delta/2)} - R = R\left(\frac{1}{\cos(\Delta/2)} - 1\right) \tag{4.33}$$

STATIONING

When designing and building a road (or other centerline-referenced project), horizontal distance along the centerline is the basis of stationing. An arbitrary value such as 0+00 or 100+00 is assigned to a point near the beginning of a project, and points on the centerline are stationed as xx+xx.xx (the value of xx+xx.xx being the accumulated centerline distance to the point). Difference in stationing is the horizontal distance along the centerline between stationing values—even along curved portions of the centerline. For example, the distance from station 132+16.58 to station 163+45.32 is 3,128.74 feet. Discontinuities in stationing (either gaps or overlaps) are handled with a station equation assigned to some centerline point. A common format for a station equation is "XXX+XX.XX Back = YYY+YY.YY Ahead."

Station equation policies vary from one organization to another, but if changes are made to the centerline alignment or if curves are added after centerline stationing is assigned, stationing for the entire project might be reassigned. Since restationing is often not practical, a station equation is used to account for gaps and overlaps. In the case of adding a curve, a station equation is used at the end of the curve (*PT*). The "ahead" station is found by adding the curve tangent distance to the *PI* station, and the "back" station for the same point is found by adding the curve length to the *PC* station. Used in reverse, distance along centerline is the difference in stationing—except when a station equation is encountered. Then, distance along centerline is computed separately on two sides of the station equation. For example, a station equation (175+48.92 BK = 175+50 AH) is used at the end of the curve shown in Figure 4.8. What is the centerline distance between station 170+00 on the curve and station 180+00 on the tangent?

Distance along curved portion = 175+48.92 – 170+00 = 548.92 feet
Distance along tangent portion = 180 + 00 – 175+50 = 450.00 feet
Total distance = 998.92 feet

Of course, other changes to the centerline alignment also require a station equation to accommodate the change in length along the revised route. Some changes will lengthen the centerline, causing an overlap at the station equation. Other changes will shorten the centerline and cause a gap between the "back" and "ahead" stations at a point.

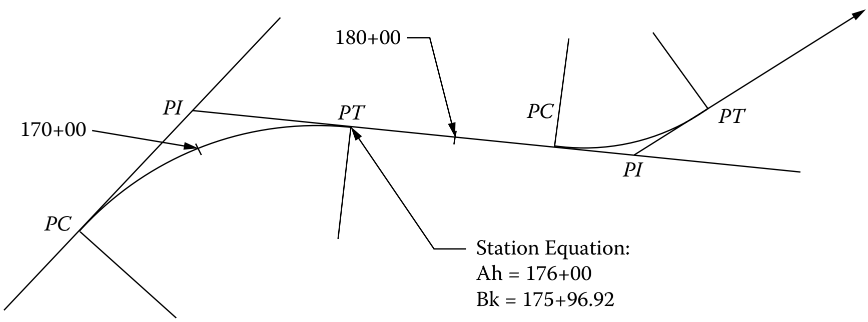


FIGURE 4.8 Station Equation Example

Metric Considerations

Stationing is used to identify cross-section locations on road or similar projects. Cross-section spacing of each station or half-station is common. There appears to be no set standard when using stationing on metric-based projects. Some use a 100 meter station with cross-section intervals of 20 or 30 meters. Others use 1,000 meters (1 kilometer) as the station with cross-section intervals every 100 meters. Therefore, each user is encouraged to confirm and/or be specific about stationing policies on any metric project. Possible sources of information include any state department of transportation (DOT) office or the American Association of State Highway Transportation Officials (AASHTO).

Area Formed by Curves

The area of a rectangular-shaped figure is length times width. The area of a circle is developed using small rectangles and tools of calculus. As shown in Figure 4.9, a differential element of area is the length (circumference at that radius distance from the center) times an infinitesimally small width (*dR*). Written in calculus as a summation of an infinite number of small rings, the area of a circle with radius *R* is computed as

$$dA = 2\pi R\, dR$$

$$Area = \int_0^R 2\pi R\, dR = \pi R^2 \tag{4.34}$$

A **sector** of a curve is defined as the "pie-shaped portion" of a circle, and the **segment** of a curve is defined as that area between the arc and chord of the curve. Both are shown in Figure 4.10. Area of a curve sector is linearly proportional to the total curve area in the same manner as a central angle is proportional to a complete circle. If the central angle is 1/4 of 360°, then the area of the sector is 1/4 the area of the circle; or if the central angle is 0.4475 of 360°, then the area of the sector is 0.4475 πR^2. For any sector, the area of the sector is

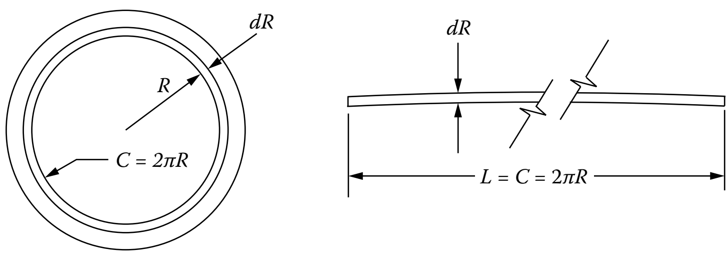


FIGURE 4.9 Area of a Circle

$$Sector\ area = \left(\frac{\Delta}{360}\right)\pi R^2 \qquad (4.35)$$

where Δ = central angle in decimal degrees.

The area of a segment is computed as the remainder left when the area of the inscribed triangle is subtracted from the area of the corresponding sector, as shown in Figure 4.10. The area of a triangle is 1/2 base times height. Using the radius as the base and the radius times sin of central angle as the height of the inscribed triangle, the segment area is computed as

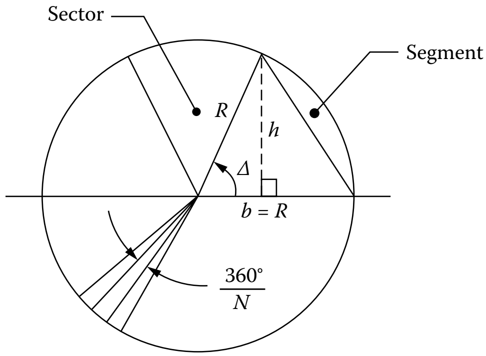


FIGURE 4.10 Area of a Sector and Area of a Segment

$$Area\ of\ segment = \frac{\Delta}{360}\pi R^2 - \frac{1}{2}R^2 \sin\Delta$$

And, with a bit of algebraic manipulation,

$$Area\ of\ segment = \frac{R^2}{2}\left(\frac{\pi}{180}\Delta - \sin\Delta\right) \qquad (4.36)$$

Note that the last expression in equation 4.36 is really Δ in radians minus the sin of Δ.

Area of Unit Circle

What is the area of a unit circle ($R = 1$)? The answer is π. If the area of a unit circle is computed as the sum of a large number of small triangles, the answer will be in error by the accumulated segment area between the arc and the chord. As a larger number of triangles is used, the error becomes very small. As illustrated in Figure 4.10, equation 4.37 approximates the area of a unit circle and permits the user to choose any large value of *N*, the number of triangles.

$$Area = \pi \approx \frac{N}{2}\left(\sin\frac{360}{N}\right) \quad for\ large\ values\ of\ N. \qquad (4.37)$$

Table 4.2 shows a summary of answers for various values of *N*. The last two lines are the value of pi computed using a spreadsheet, $\pi = 4\arctan(1.0)$, and the first twenty digits of π (Beckmann 1971).

TABLE 4.2
Approximations for π Based on Area of Unit Circle

Value of N	Area of Corresponding Unit Circle (to 15 significant digits)
100	3.13952597646567
1,000	3.14157198277948
10,000	3.14159244688129
100,000	3.14159265152271
1,000,000	3.14159265356913
10,000,000	3.14159265358959

Note: Spreadsheet computation: 4 * atan(1) = 3.14159265358979.
First 20 digits of π: 3.1415926535897932385.

Comment: mathematicians have devised better ways of computing π, but this technique shows an interesting connection between the value of π and the area of a unit circle while illustrating the concept of limits.

SPIRAL CURVES

A spiral is defined as a curve whose radius is inversely proportional to its length. Various kinds of spirals can be developed from the same definition. Mathematicians often view spirals from the perspective of an origin located at that point where the radius approaches zero. Spirals used in geomatics applications share the definition, but are viewed from a different perspective (i.e., beginning at that point where the spiral length is zero and the radius is infinitely long). This origin is selected because spirals are used to provide a rigorous gradual transition from traveling in a straight line along the tangent to traveling along a circular curve having a constant radius. Spirals are used extensively on railroad layout, but they are also used in some highway applications.

Typically, spirals are used in pairs. Starting on a straight-line tangent, a spiral is used to make the transition from traveling in a straight line to traversing a circular curve. At the end of the circular curve, another spiral is used to transition back to traveling in a straight line on the next tangent. Although the entrance spiral and exit spiral could have different lengths, standard practice is for them to be symmetrical. Therefore, the discussion here will focus only on the geometry of the entrance spiral.

SPIRAL GEOMETRY

In some cases, spirals have been avoided because of their computational complexity. But, even though the equations have not changed, modern computers, coordinate geometry routines, and radial surveying techniques have eased the burden of using spirals. It is intended for the following to provide a comprehensive computational algorithm that can be completed using only a simple scientific calculator.

Admittedly, the formulas given here do not lend themselves to easy longhand solutions but will find ready applications in a spreadsheet solution or computer program. The equations presented here will likely be most useful to those writing coordinate geometry routines (for computers and/or data collectors).

Symbols for the spiral illustrated in Figure 4.11 are as follows:

α = azimuth of beginning tangent.
TS = station at transition tangent to spiral.
SC = station at transition spiral to circular curve.
L = total length of spiral (*L* also = *SC-TS*).
R = radius of circular curve at end of the spiral.
K = spiral constant = 1 / (2*RL*).
s = distance along spiral from *TS* to specified point on spiral.
r = instantaneous radius of spiral at location defined by *s*.
θ = angular difference between spiral radius vectors at two points. Typically one point is at the beginning of the spiral, and the second is any point on the spiral at a distance *s* from the beginning.
x = tangent component of the distance from *TS* to any point on the spiral.
y = perpendicular distance from the tangent to any point on the spiral. An appropriate sign convention is for *y* to be positive if the spiral curves to the right and negative if the spiral curves to the left when standing at *TS* and looking along the tangent adjacent to the spiral.

Equations for computing coordinates of a point on a spiral are given below.
Given:

e_1,n_1 = local plane coordinates of tangent to spiral, *TS*

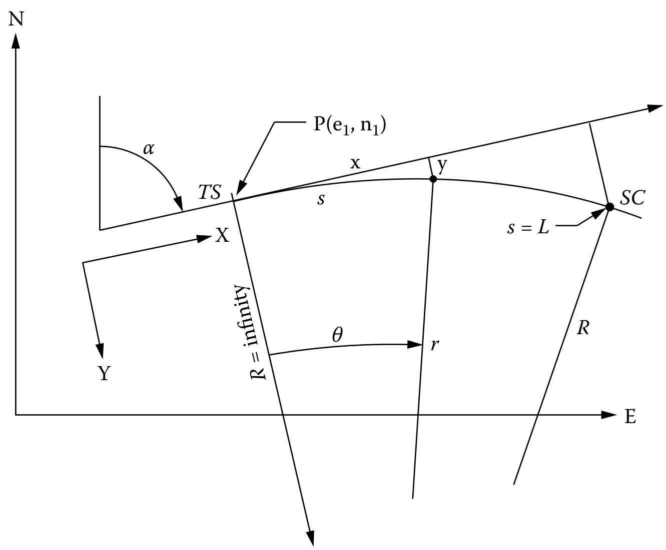


FIGURE 4.11 Spiral Elements and Geometry

α = local azimuth of straight-line tangent at TS
R = radius of circular curve at the end of the spiral
L = total length of the spiral
s = centerline distance from TS to point on the spiral

Store: constants common to spiral solutions (Libby and Booth, 1973, 8–9).

$C_1 = 1/3$
$C_2 = -1/10$
$C_3 = -1/42$
$C_4 = 1/216$
$C_5 = 1/1{,}320$
$C_6 = -1/9{,}360$
$C_7 = -1/75{,}600$
$C_8 = 1/685{,}440$

Compute:

$$\theta = s^2 / (2RL) = K\, s^2 \tag{4.38}$$

$$x = s\,(1 + C_2\,\theta^2 + C_4\,\theta^4 + C_6\,\theta^6 + C_8\,\theta^8 + \ldots)$$

$$= s\,(1 + \theta^2\,(C_2 + \theta^2\,(C_4 + \theta^2\,(C_6 + C_8\theta^2\,)))) \tag{4.39}$$

$$Y = s\,(C_1\,\theta + C_3\,\theta^3 + C_5\,\theta^5 + C_7\,\theta^7 + \ldots)$$

$$= s\,\theta\,(C_1 + \theta^2\,(C_3 + \theta^2\,(C_5 + C_7\,\theta^2\,))) \tag{4.40}$$

Solution: coordinates for a point on a spiral are computed using the COGO forward computation equations 4.8 and 4.9 as follows (sign convention: y is negative for spiral to left):

$$e_2 = e_1 + x \sin \alpha + y \sin (\alpha + 90°)$$

$$= e_1 + x \sin \alpha + y \cos \alpha \tag{4.41}$$

$$n_2 = n_1 + x \cos \alpha + y \cos (\alpha + 90°)$$

$$= n_1 + x \cos \alpha - y \sin \alpha \tag{4.42}$$

A line parallel to a spiral is not a spiral. But, points on an offset to a spiral, (e_3, n_3), can be computed using equations 4.43 and 4.44. To compute coordinates of points lying to the right of the spiral, use a plus offset distance. Points to the left of the spiral are computed using the offset distance as a negative value. The azimuth of the line (radius vector) perpendicular to the line tangent to the spiral at a point is (α + 90° + θ).

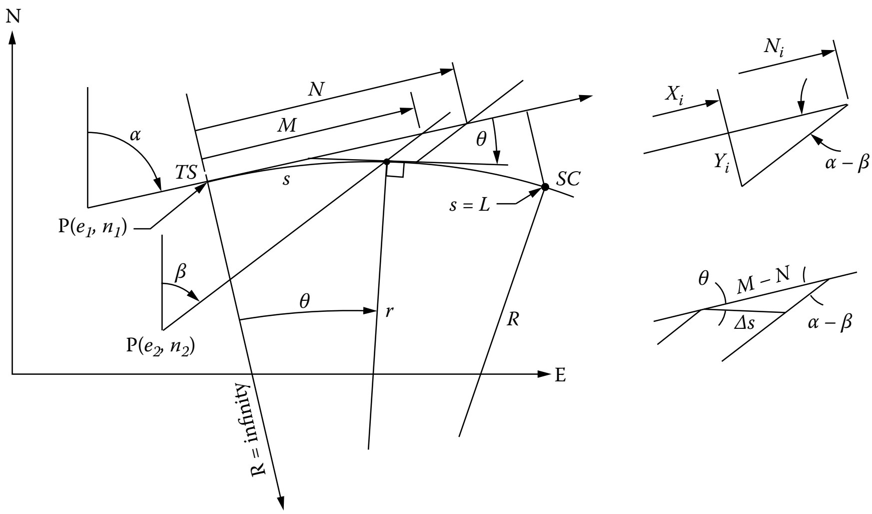


FIGURE 4.12 Spiral Intersection Elements and Geometry

$$e_3 = e_2 \pm \text{offset distance} * \sin(\alpha + 90° + \theta) \tag{4.43}$$

$$n_3 = n_2 \pm \text{offset distance} * \cos(\alpha + 90° + \theta) \tag{4.44}$$

With spiral points and offset line points computed according to equations 4.41, 4.42, 4.43, and 4.44, those points can be used along with other project points according to radial surveying techniques and standard coordinate geometry procedures available in most field computers and/or data collectors.

Intersecting a Line with a Spiral

Computing the intersection of a straight line with a spiral is not encountered very often, but there are times it needs to be done. Possible combinations include no intersection (trivial), one intersection (covered here), or, in extremely rare cases, two intersections. This section, as mentioned, looks at the single intersection case. As shown in Figure 4.12, the key to finding the solution is determining a correct value of *s*, the distance from the *TS* along the spiral centerline to the intersection. Once the value of *s* is known, the *x* and *y* spiral components are computed according to equations 4.39 and 4.40. Then the *e* and *n* coordinates of the intersection point are computed using equations 4.41 and 4.42. The value of *s* is determined using an iterative process.

Given:

e_1,n_1 = local plane coordinates of tangent to spiral, *TS*
α = local azimuth of straight-line tangent at *TS*
e_2,n_2 = local plane coordinates of any point on line
β = local azimuth from point on line to spiral intersection
R = radius of circular curve at the end of the spiral
L = total length of the spiral

Compute:

$K = 1 / (2RL)$.
M = line-line intersection distance along original tangent. Use equation 4.14 and distance d_1 as M. M does not change.
s_1 = value of M. The value of s will be improved.

Iterate: Start with $i = 1$, and continue incrementing i by 1 until tolerance is met.

θ_i = angle (in radians) to initial trial point on spiral = Ks_i^2.
x_i = trial distance along spiral tangent—see equation 4.30.
y_i = trial distance perpendicular to spiral tangent—see equation 4.31.
N_i = check distance; $N_i = x_i + y_i \cot(\alpha - \beta)$.
Tol = absolute value $|M - N|$. Is it small enough? If yes, quit.

If not:

$$\Delta s_i = \text{correction to } \Delta s \qquad s = \frac{(M - N)\sin(\alpha - \beta)}{\sin(\alpha - \beta + \theta_i)}$$

$$s_{i+1} = s_i + \Delta s_i \text{ } (\Delta s_i \text{ should get smaller and smaller.})$$

Increment i by 1.
Return to beginning of iteration using new value of i.

Solution: When the tolerance is met, s_i is the correct value.

The values of x_i and y_i have already been computed.
Use equations 4.41 and 4.42 to compute e and n at intersection.

As a final check, inverse from e_2 and n_2 to e and y. The computed direction should be the same as β.

Computing Area Adjacent to a Spiral

If a spiral is part of a boundary, computing the area adjacent to a spiral becomes an issue. The method presented here allows the user to compute the area between the original tangent and the spiral to any precision desired by choosing smaller and smaller increments of Δs and accumulating the area by numerical integration. Without a computer programmed to perform the repetitive calculations, the method loses its practicality. As shown in Figure 4.13, the area is accumulated from numerous trapezoids formed by the original tangent and perpendicular lines to the spiral. The separation of construction lines perpendicular to the tangent is not constant, but is determined by equal values of Δs on the spiral centerline.

Given:

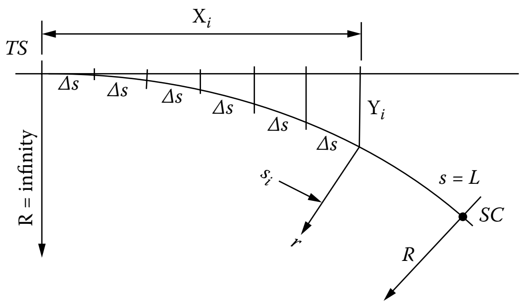


FIGURE 4.13 Area Adjacent to a Spiral

R = radius of circular curve at end of spiral
L = total length of spiral
$s = L$ or some portion thereof

Find: area bounded by original tangent and spiral up to distance s.

Solution:

$K = 1/(2RL)$. Spiral constant.
Δs = increment chosen by user. $\Delta s = s/n$. User chooses n.
$s_0 = 0.0$.
$x_0 = 0.0$.
$y_0 = 0.0$.

Loop: for $i = 1$ to n:

s_i = distance to point i. $s_i = s_{i-1} + \Delta s$.
$\theta_i = Ks_i^2$. Angle in radians to point on spiral.
x_i = tangent component distance to point on spiral. See equation 4.39.
y_i = perpendicular distance tangent to spiral. See equation 4.40.

End of loop.
Area by trapezoids:

$$2 \text{ area} = (x_1 - x_0)(y_1 + y_0) + (x_2 - x_1)(y_2 + y_1) \ldots$$

$$+ (x_n - x_{n-1})(y_n + y_{n-1})$$

which, after considerable algebraic manipulation, reduces to

$$2 \text{ area} = x_2y_1 - x_1y_2 + x_3y_2 - x_2y_3 \ldots. \quad + x_ny_{n-1} - x_{n-1}y_n + x_ny_n \tag{4.45}$$

Or, the area adjacent to a spiral can also be written as

$$2A = \sum_{i=2}^{n} x_iy_{i-1} - \sum_{i=2}^{n} x_{i-1}y_i + x_ny_n \tag{4.46}$$

RADIAL SURVEYING

The following discussion relates specifically to conventional total station surveying procedures, but it is also somewhat applicable to using GPS equipment and procedures. Radial surveying is a term used to describe the practice of measuring an angle and distance from a known point to tie in another point (determine coordinates for its location) or for setting out a point at a predetermined location. The procedures are well defined and simple to perform, but they have the disadvantage of no redundancy. If a blunder is made in measuring the angle, reading the distance, or setting up over the wrong point, there is no built-in check. One way to check the position of such open sideshots is to set up over a different point, tie in all sideshot points a second time, and compare the results. Another method is to measure (say, with a steel tape) distances between the surveyed points. The inverse distance in each case should compare favorably with the taped distance. Lack of expected consistency in the results using either technique is an indication of a blunder or uncontrolled errors. A high degree of consistency in such checks is an indication, but not a guarantee, that there are no blunders and that random and systematic errors have been controlled at an acceptable level. Given that many production measurements may be checked in this manner, it must be understood that such checks are a poor substitute for carefully designed redundant measurements.

Radial surveying methods can be accomplished using various equipment combinations, but they are ideally suited for the modern total station instrument connected with a data collector. Radial surveying techniques are also suited for the one-person crew using a robotic total station. Each point is identified by number in the data file, and instructions are given to the instrument in terms of commands, point numbers, and, in some cases, attributes. If collecting data for a topographic site plan, the point numbers may be assigned sequentially by default. In the layout mode, the operator specifies the points to be used or staked in any order desired. In either case (collecting data or laying out points), both the point occupied by the instrument and the backsight point must be specified by the user.

Radial surveying has two primary advantages:

- The geometry of points to be staked may involve curves, spirals, offsets, or other complex geometrical relationships, but, in the field, the solution boils down to an angle from a known backsight and a distance from the instrument.
- Decisions about which point to occupy with the instrument and which point to use as a backsight can be deferred to the responsible person in the field. Intervisibility between points is essential for line-of-sight equipment, but, whether tying in points or staking locations, the logistical operations are made easier by the flexibility of the method.

When performing radial stakeout, the computations are performed by computer, and details are rarely of concern to the end user. However, should it be necessary, for whatever reason, to perform the angle-right/distance computations by hand, the following procedure as illustrated in Figure 4.14 may be helpful. The conventional procedure is to perform two separate inverse computations and to use the two com-

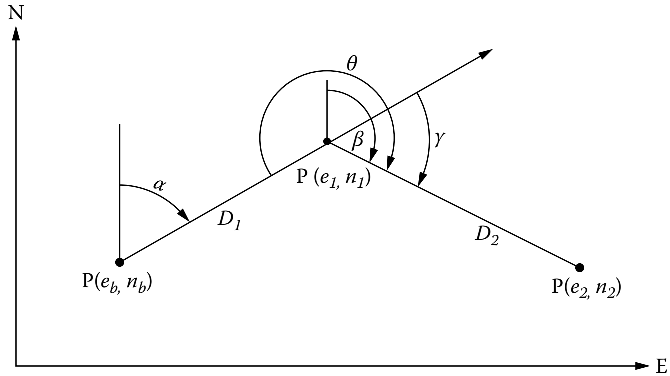


FIGURE 4.14 Angle-Right on a Radial Stakeout

puted azimuths to find the appropriate angle-right. Using a trigonometric identity, the procedure can be shortened and done directly on any scientific calculator.

The following procedure is moot if using GPS to stake out points. If using a total station for layout, three points are required to perform an angle-right/distance layout computation: the standpoint, a backsight, and the forepoint. The solution is the angle-right (at the standpoint) from the backsight to the forepoint and the distance from the standpoint to the forepoint.

Symbols:

e_B, n_B = backsight coordinates
e_1, n_1 = standpoint coordinates
e_2, n_2 = foresight coordinates
α = azimuth from backsight to standpoint
β = azimuth from standpoint to forepoint
γ = deflection angle, $\beta - \alpha$
θ = angle-right, $\gamma + 180°$
D_1 = distance backsight to standpoint
D_2 = distance standpoint to forepoint

Trigonometric identity:

$$\tan(B - A) = \frac{\tan B - \tan A}{1 + \tan A \tan B} \tag{4.47}$$

Compute:

$$\Delta e_1 = e_1 - e_B,\ \Delta e_2 = e_2 - e_1$$

$$\Delta n_1 = n_1 - n_B,\ \Delta n_2 = n_2 - n_1$$

$$\tan \alpha = \Delta e_1/\Delta n_1,\ \tan \beta = \Delta e_2/\Delta n_2$$

$$D_1 = \sqrt{(\Delta e_1^2 + \Delta n_1^2)},\ D_2 = \sqrt{(\Delta e_2^2 + \Delta n_2^2)}$$

Solution (first find deflection angle, then angle-right):

$$\tan\gamma = \tan(\beta - \alpha) = \frac{\tan\beta - \tan\alpha}{1 + \tan\alpha \tan\beta}$$

Substitute values for tan α and tan β from above:

$$\tan\gamma = \frac{\left(\frac{\Delta e_2}{\Delta n_2}\right) - \left(\frac{\Delta e_1}{\Delta n_1}\right)}{1 + \left(\frac{\Delta e_1}{\Delta n_1}\right)\left(\frac{\Delta e_2}{\Delta n_2}\right)} = \frac{\Delta e_2\,\Delta n_1 - \Delta n_2\,\Delta e_1}{\Delta n_1\,\Delta n_2 + \Delta e_1\,\Delta e_2} \tag{4.48}$$

Angle right = $\theta = \gamma + 180°$ (subtract 360° if required) (4.49)

Notes:

1. The distance D_1 is not needed, but can be used to check distance to backsight.
2. Equation 4.48 can be solved using the rectangular/polar key of a scientific calculator by inputting the numerator and denominator separately. The result will be the deflection angle (add 180° for angle-right) and distance D_2.
3. If programming equation 4.48, include provision for a zero denominator. (Most arctan2 functions will accommodate a zero denominator.) If the denominator is zero, the angle-right is 90° or –90° (270°).
4. Angle-right can also be computed by taking the difference of two azimuths as obtained from separate coordinate inverse computations; see equations 4.11, 4.12, and 4.13.

VERTICAL CURVES

The coordinate geometry tools discussed so far have looked at lines, curves, and spirals in the horizontal plane. This section looks at the use of a parabolic equation in the vertical plane (profile) as used to provide gradual changes in grade on a road, street, or highway. Commonly called a vertical curve, the parabolic equation is one of the conic sections and is obtained from the general polynomial equation (equation 3.8) by setting the *B* and *C* coefficients equal to zero. The result is an equation of the general form

$$y = ax^2 + bx + c \tag{4.50}$$

The grade (slope) of most highways is gradual and expressed in percent. A 2 percent grade means a uniform rise of 2 feet vertically for each station (100 feet). A –4 percent slope means a uniform drop of 4 feet per station. Sight distances across the crest of a hill and passenger comfort for those traveling on high-speed interstate roadways are related to the rate of change of grade when making a transition from

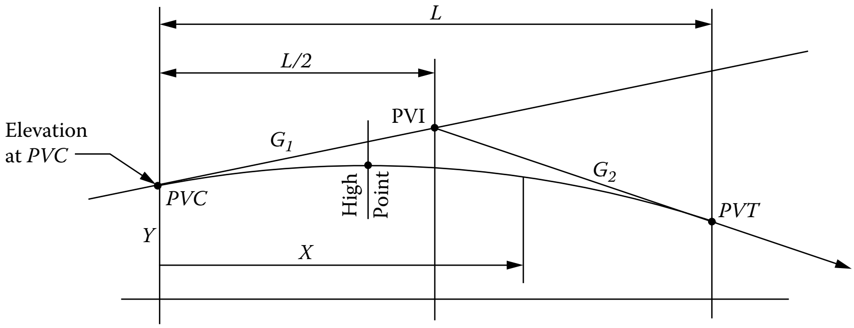


FIGURE 4.15 Vertical Curve Geometry

one grade to another. This rate of change of grade per station is a design parameter that varies according to intended use of the travel way.

Symbols used in equation 4.52 and Figure 4.15 are as follows:

PVI = point of vertical intersection.
PVC = point of vertical curvature (also called the *BVC*, or beginning of vertical curve).
PVT = point of vertical tangency (also called the *EVC*, or end of vertical curve).
L = length of vertical curve computed as difference in stationing (station at *EVC* – station at *BVC*; *L* is a horizontal distance).
G_1 = grade of roadway along the centerline coming into the vertical curve.
G_2 = grade of roadway along the centerline leaving the vertical curve.
y = elevation of point on the vertical curve at station XXX+XX.
x = distance of station XXX+XX from *PVC*. $0 \leq x \leq L$.
r = rate of grade per station = $(G_2 - G_1) / L$, grade in percentage and *L* in stations.

Notes about derivation and the solution of vertical curve problems:

1. The first derivative of the parabolic equation is the slope of the tangent to the curve at that point.
2. The slope of the curve is zero at the high point (summit) or low point (sag) of the curve.
3. The slope of the curve is G_1 at *BVC* (@ $x = 0$) and G_2 at *EVC* ($x = L$).
4. The vertical curve is symmetrical with respect to length. That is, the *PVI* is halfway between the *BVC* and the *EVC*. That's why it is called an equal tangent length vertical curve. Note: unequal tangent length vertical curves are solved by establishing secondary *PVI*'s at the midpoints of the *BVC* to *PVI* and *PVI* to *EVC* and treating the result as two abutting equal tangent length vertical curves.
5. Units: two consistent practices will each give good results. Mixing the two conventions may result in bad answers and lots of frustration. And, be careful with using metric stationing.

A. Grade expressed as slope (2 percent = 0.02) is consistent with the x distance used as feet or meters.
B. Grade expressed as a percentage is consistent with x used in stations.

The a, b, and c coefficients must be found before equation 4.50 can be used to solve for elevations at each station. Note that at $x = 0$, the first two terms of equation 4.50 drop out, leaving c = elevation of vertical curve at BVC ($x = 0$). To find coefficients a and b, we need to take the derivative of equation 4.50 as

$$dy/dx = 2ax + b$$

At $x = 0$, the slope is G_1 and $dy/dx = b$. Therefore, $b = G_1$.
At $x = L$, the slope is G_2 and $dy/dx = 2ax + G_1$. Therefore, $a = (G_2 - G_1)/(2L)$.

And, finally, if the first derivative is set to 0, we have an expression that can be solved for x, the distance from the BVC to the high point or low point of the vertical curve. Using the values of a and b, as found above.

$$0 = 2 * x * \frac{G_2 - G_1}{2L} + G_1 \rightarrow x = \frac{G_1 L}{G_1 - G_2} \tag{4.51}$$

Note that if equation 4.51 gives a value of x less than zero or greater than L, it means there is no sag or summit between the BVC and EVC. That will happen if both grades are positive or if both grades are negative. In that case, the highest point or the lowest point (not the sag or the summit) will be at either the BVC or EVC.

The vertical curve equation is

$$y = \frac{(G_2 - G_1)}{2L} x^2 + G_1 x + Elev@BVC \tag{4.52}$$

Example: compute elevations at each station and half-station (every 50 feet) for the following vertical curve. Find the station and elevation at the low point (sag).

PVI station = 172+00
elevation @ *PVI* = 2,300.00 feet
length of curve = 400.00 feet
G_1 and G_2 = –4 percent and 1 percent, respectively

Solution:

1. Find *BVC* station and elevation:

$$\text{station at } BVC = \text{station at } PVI - L/2 = 170{+}00$$

$$\text{elevation at } BVC = \text{elevation @ } PVI - L/2 * G_1 = 2{,}308.00 \text{ feet}$$

2. Find EVC station and elevation:

station at EVC = station at $PVI + L/2$ = 174+00 (also same as $BVC + L$)

elevation at EVC = elevation at $PVI + L/2 * G_2$ = 2,302.00 feet

3. Find station of low point:

x (low) = $G_1 * L/(G_1 - G_2)$ = 320 feet; station = 173+20

4. Find coefficients a and b:

$$a = (G_{22} - G_1) / (2L) = (0.01 - 0.04) / 800 = 0.000062500$$

$b = G_1 = -0.04$

5. Use equation 4.43 to find each elevation as follows:

Station	*x*	*Elevation*	Station	*x*	*Elevation*
170+00	0	2,308.00’	170+50	50’	2,306.16’
171+00	100’	2,304.63’	171+50	150’	2,303.41’
172+00	200’	2,302.50’	172+50	250’	2,301.91’
173+00	300’	2,301.63’	173+50	350’	2,301.66’
174+00	400’	2,302.00’			
173+20	320’	2,301.60’			
low point of vertical curve					

THREE-DIMENSIONAL MODELS FOR SPATIAL DATA

So far, this chapter has looked specifically at COGO models used for routine 2-D computations. More complex models include three dimensions (3-D) and are needed to compute volumes. Several fundamental examples are given below.

Volume of Rectangular Solid

Perhaps the easiest volume to find is that of a rectangular solid computed as length * width * height. The same concept will be used along with tools of integral calculus to compute the volume of other shapes.

Volume of a Sphere

To compute the volume of a sphere, an elemental surface area (length * width) is multiplied by an elemental increment of radius (height). The elemental surface area, as shown in Figure 4.16, is $R\,d\varphi$ in the north-south direction and $R \cos \varphi\, d\lambda$ in the east-west direction. The surface area of a sphere in terms of the radius is obtained by performing a double integration on variables of longitude, 0 to 2π, and latitude, $-\pi$ to π (radian units). Surface area on a sphere bounded by latitude/longitude limits is

$$dA = R\, d\phi\, R\cos\phi\ \ d\lambda$$

$$\text{Surface area on a sphere} = R^2 \int_{\lambda_1}^{\lambda_2} \int_{\phi_1}^{\phi_2} \cos\phi\ \ d\phi\ \ d\lambda \tag{4.53}$$

To obtain total surface area, choose limits (in radians) of $\lambda = 0$ to 2π and $\varphi = -\pi$ to π. Then total surface area is computed as

$$A_{Sphere} = R^2 \int_0^{2\pi} \int_{-\pi}^{\pi} \cos\phi\ d\phi\ d\lambda = 2\pi\ R^2 \int_{-\pi}^{\pi} \cos\phi\ d\phi = 2\pi\ R^2 - (-2\pi\ R^2) = 4\pi\ R^2$$

Building on that, the volume of a sphere as illustrated in Figure 4.16 is computed as

$$dV = dA * dR = R^2 \cos\phi\ d\lambda\ d\phi\ dR$$

$$\text{Volume of a Sphere} = \int_0^R \int_{-\pi}^{\pi} \int_0^{2\pi} R^2 \cos\phi\ d\lambda\ dR = 4\pi \int_0^R R^2 dR = \frac{4\pi\ R^3}{3} \tag{4.54}$$

Volume of a Cone

The (truncated) cone is another fundamental shape for which volume needs to be computed. For surveyors and engineers, it often comes disguised as the prismoidal formula for computing earthwork volumes. Referring to Figure 4.17a, calculus is

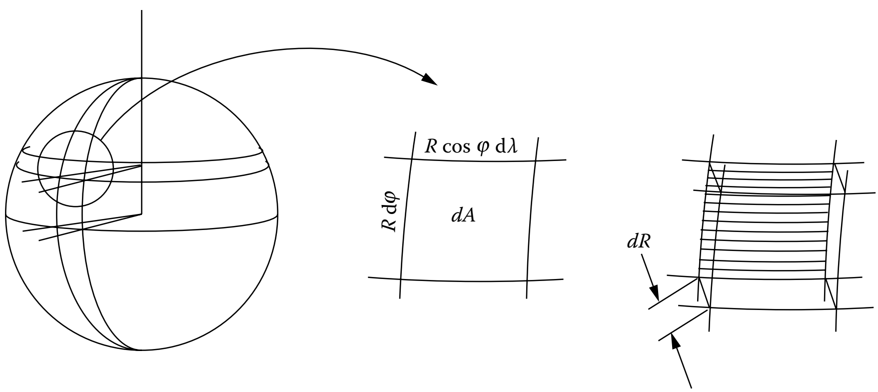


FIGURE 4.16 Volume of a Sphere

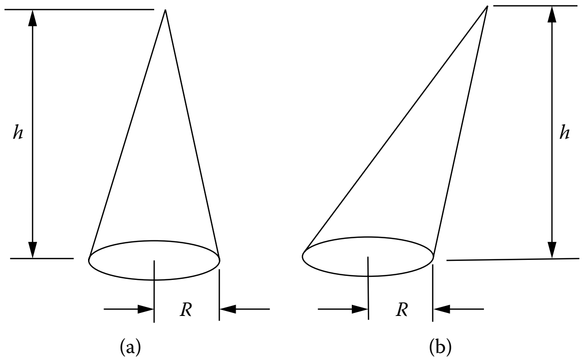


FIGURE 4.17 Volume of a Cone

also used to compute the volume of the cone shown as a cross-section area (a circle) times a differential height. For the cone shown, there is a linear relationship between the height and the radius at that height. Note: at $h = 0$, $R = 0$; and at $h = h$, $R = kh$. The volume of the cone is computed as

$$dV = area * dh = \pi R^2 dh = \pi k^2 h^2 dh$$

$$V = \int_0^b \pi k^2 h^2 dh = \frac{\pi k^2 h^3}{3} = \frac{\pi R^2 h}{3} \tag{4.55}$$

Note that there is no requirement that the cone axis be at a right angle to the base. In Figure 4.17b, the volume of the cone is still computed as the area of the base times the height measured perpendicular to the base.

Prismoidal Formula

Often earthwork volumes on road or highway construction are computed using the average-end-area method, where volume of a given section is computed as the section length times the mean of the cross-sectional areas at the two ends. It is a good approximation, but based on a false assumption. The average-end-area method assumes that area varies linearly from one cross section to another. As will be shown using the example of the cone, the radius may vary linearly, but the cross-sectional area varies exponentially. Consider the two-station length, as illustrated in Figure 4.18. The shape is that of a cone for which volume can be computed without approximation. The volume contained between sections C and E in Figure 4.18 is the difference of the cone AE minus the cone AC. Given that the radius varies linearly from A to E, the area at each cross section is computed accordingly.

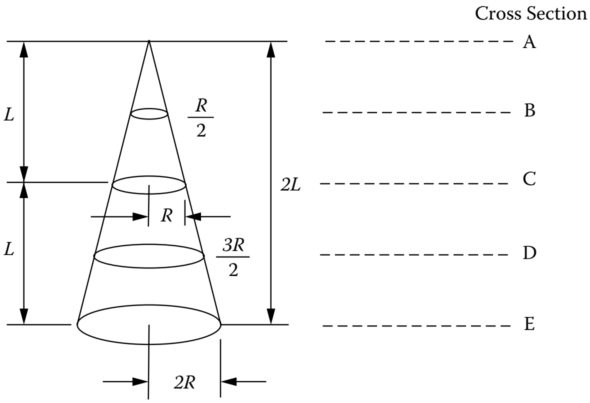


FIGURE 4.18 Derivation of Prismoidal Formula

The total AE cone volume is

$$V = \frac{\pi\,(2\,R)^2\,2\,L}{3} = \frac{8\,R^2 L}{3} \tag{4.56}$$

The volume of the small AC cone is

$$V = \frac{\pi\,R^2\,L}{3}$$

The volume of the remaining CE interval is the difference and is computed as

$$V = \frac{8\,\pi\,R^2 L}{3} - \frac{\pi\,R^2\,L}{3} = \frac{L}{6}\left(14\,\pi\,R^2\right) \tag{4.57}$$

But, note that the cross-sectional areas at sections C, D, and E are as follows:

Cross-sectional area at C = πR^2
Cross-sectional area at D = $\pi (3R/2)^2 = 9 \pi R^2 / 4$
Cross-sectional area at E = $\pi (2R)^2 = 4 \pi R^2$

For the volume CE, consider section C to be the beginning, section D to be the middle, and section E to be the end. The prismoidal formula as derived by others computes prism volume as *L*/6 * (beginning cross-sectional area + 4 * cross-sectional area at midpoint + ending cross-sectional area) or, using the labeling on our example:

$$V = \frac{L}{6}\left(A_{beginning} + 4\,A_{middle} + A_{end}\right) = \frac{L}{6}\left(\pi\,R^2 + 4 * \frac{9\,\pi R^2}{4} + 4\,\pi R^2\right) = \frac{L}{6}\left(14\,\pi R^2\right) \tag{4.58}$$

Note that the results of equations 4.57 and 4.58 are identical. In a backhanded way, we have just derived the prismoidal formula.

The average-end-area method remains an excellent approximation in many cases. Most of the time, "fluffing" will increase the volume of excavated material, or loss of volume by subsequent compaction will far exceed the error caused by not using the prismoidal method. Lack of good cross-sectional data may also be a contributing factor to inaccuracies in computed earthwork volumes. But, the fact remains that the average-end-area method is based upon a false assumption, and it is left up to the user to judge if or when the prismoidal formula should be used instead of the average-end-area method.

Traditional 3-D Spatial Data Models

The volume computations just considered are 3-D, but they really don't address the larger (global) issues facing spatial data users. Expanding attention from the local 2-D and 3-D considerations, traditional spatial data models include a variety of 2-D and 3-D options. Some are more complex than others. Examples along with various measurement units include the following:

- Flat-Earth model used for local mapping, site development, and plane surveying
  1. Units as selected by the user.
  2. Two-dimensional *X*/*Y* (or northing/easting) tangent plane coordinates.
  3. In the third dimension:
     A. Surveyors and engineers use profiles to show grades in terms of centerline stationing and elevation.
     B. Architects show elevation perspectives.
     C. Mappers use hachures and contour lines.
- Spherical Earth model used in geography and navigation
  1. Mixed-mode units: sexagesimal for horizontal, length for vertical
  2. Two-dimensional curvilinear spherical Earth latitude/longitude positions
  3. Third dimension is elevation or altitude in terms of length units
- Ellipsoidal Earth model used in geodesy, cartography, and geophysics
  1. Mixed-mode units: sexagesimal for horizontal, meters for vertical
  2. Two-dimensional curvilinear ellipsoidal Earth latitude/longitude positions
  3. Third dimension based upon
     A. distance from geoid, orthometric height
     B. distance from ellipsoid, ellipsoid height
- Map projection (state plane) spatial data model

1. Units are 2-D grid distances; true distances are distorted.
2. North American Datum—1927: *X*/*Y* coordinates are U.S. Survey Feet.
3. North American Datum—1983: *N*/*E* coordinates, units vary by state.
   A. NGS publishes meters in all state plane coordinate zones.
   B. Plus, some state statutes specify use of U.S. Survey Feet.
   C. Also, some state statutes specify use of International Feet.
4. Elevation, although referenced to a curved surface (approximately mean sea level), is typically given by the *Z* coordinate value.

The 3-D GSDM

Of the models listed, the state plane coordinate system model is called pseudo 3-D because the third dimension is referenced to a nonregular curved surface. In spite of that deficiency, state plane coordinates combined with elevation data have been beneficially used in many applications. But, the fact remains that when performing 3-D computations using state plane coordinates and elevation, equations of solid geometry and vector algebra are valid only to the extent one can assume a flat Earth.

The most appropriate model is often taken to be a rectangular Cartesian coordinate system that defines location of a point as being the perpendicular distances from mutually orthogonal axes. Such a model is simple and, to the extent the Earth is flat, appropriate. This chapter includes extensive discussions of 2-D COGO in the contexts encountered by many spatial data users. But spatial data are 3-D, and spatial data users need simple, standard, and reliable tools and methodology for handling 3-D spatial data. This last section identifies the standard historical methods and coordinate systems that have been used for 3-D data.

Many will choose to use spherical geographic coordinates of latitude/longitude. Others will choose to use the more precise geodetic version of latitude/longitude based upon an ellipsoidal Earth. That choice is related to whether or not one anticipates that a spherical model is sufficiently accurate or whether one needs to use the ellipsoidal Earth model to preserve the geometrical integrity of the data being used. In either case, a study of geodesy supports an understanding of the underlying foundation of each model.

The goal of this book is to highlight features and characteristics of the 3-D GSDM that accommodate existing models, modern measurement technology, digital data storage, and spatial data manipulation practices common to various disciplines. But, before the 2-D concepts as discussed here are extended to 3-D, it is appropriate to consider geodesy and other 3-D relationships of spatial data.

REFERENCES

Beckman, P. 1971. *A history of Pi*. New York: St. Martin's.

Burkholder, E. F. 1984. Coordinates, calculators, and intersections. *ACSM Journal of Surveying & Mapping* 46 (1): 29–39.

———. 1997. Three-dimensional azimuth of GPS vector. *ASCE Journal of Surveying Engineering* 123 (4): 139–46.

Clarke, K. C. 2000. Criteria and measures for the comparison of global geocoding systems. Paper presented at the International Conference on Discrete Global Grids, Santa Barbara, CA, 26–28 March.

Libby, H. W. and E. M. Booth. 1973. *Standard Highway Spiral*, Oregon State Highway Division Technical Bulletin No. 20–revised. Oregon Transportation Commission.

Thompson, W. C. 1974. The TH function. *ACSM Journal of Surveying & Mapping* 34 (2): 151–53 (and discussion in subsequent issues).

5 Overview of Geodesy

INTRODUCTION

Geodesy is geomatics and geomatics is geodesy. Well, not really. Geomatics is a fairly new umbrella term being used to describe both a body of knowledge and the scope of professional activities having to do with the generation, manipulation, storage, and use of spatial data. In a nonexclusive way, geomatics includes traditional disciplines such as surveying, mapping, geodesy, and photogrammetry. It also overlaps with other newer disciplines such as remote sensing, imaging, and information sciences. Of all such disciplines, geodesy provides the geometrical foundation for the rest. A literal meaning of the word geodesy is "dividing the Earth," and geodesy, as a discipline of inquiry, has been concerned with learning more about the size and shape of the Earth since the dawn of civilization.

The scope of geodesy includes both science and art and involves much more content than is summarized here. Without being restricted to either category, the science of geodesy includes issues such as (1) determining the size and shape of the Earth, (2) defining and quantifying the Earth's gravity field, and (3) defining reference systems and coordinate frames. The art and practice of geodesy include using scientific data and various measuring systems to determine the location of points with respect to a defined geodetic framework. The point here is less about classifying an activity as being science or art but more about recognizing a mutual interdependence between the two.

FIELDS OF GEODESY

Distinctions between the fields of geodesy are largely artificial and should not limit one's consideration. But historically the following categories have been used:

- Geometrical geodesy encompasses the 3-D geometrical elements of the ellipsoidal model of the Earth and the location of points relative to that model. Traditional geometrical geodesy includes separate horizontal and vertical datums, while modern practice combines them into a single 3-D database.
- Physical geodesy involves the study of gravity and considers issues such as the distribution of mass within the Earth, the Earth's external gravity field, equipotential surfaces, and the cause-and-effect relationships between them (as evidenced for example by the behavior of a pendulum, the path of artificial satellites, and the shape of the geoid).
- Satellite geodesy deals with the trajectory of missiles in flight and satellites orbiting the Earth. It also includes processing signals received from various orbiting satellites for the purpose of determining the location and/or movement of the receiver.

- Geodetic astronomy has been used extensively to compute positions and directions on the Earth based upon optical (and radio) observations of stars in the sky. With the advent of GPS surveying, geodetic astronomy has lost much of the relevance it once enjoyed.

Given the development of computers and the convergence of other electronic technologies, this book considers how geometrical geodesy supports GIS and spatial data referencing, how physical geodesy relates to geoid modeling and computation of elevations, and how satellite geodesy contributes to an understanding of GPS surveying and the use of GPS data. The GSDM incorporates concepts from various fields of geodesy as needed to support development of a comprehensive model for location and spatial referencing.

GOALS OF GEODESY

Traditional goals of geodesy include determining the size and shape of the Earth, describing the gravity field associated with the Earth, and providing a means of locating points on or near the Earth's surface. Geodetic scientists have defined reference frames and coordinate systems used for referencing spatial data and have determined both the size and shape of the Earth within impressive tolerances. They have also developed sophisticated geoid modeling procedures that can be used to compute both the shape of the geoid and, to a lesser extent, its precise location. Given those definitions and tools, geodetic surveyors and others, using GPS and other positioning technologies, are busy collecting, processing, and using spatial data to determine the location of points and to document the movement of objects on or near the Earth. Based upon what is being accomplished, it could be argued that the traditional goals of geodesy have been met.

The goals of modern geodesy are certainly more comprehensive than the traditional ones summarized here. Two publications by the National Research Council (NRC; 1978, 1990) contain an informative overview of the field of geodesy and a description of future challenges and applications. With the advent of computers, satellites, and other modern technology, the fundamental goals of geodesy are being extended to the oceans, the moon, and the planets. With respect to geodetic networks, the 1978 NRC report states, "The ultimate goal is a global geodetic system providing horizontal and vertical, or three dimensional, coordinates for national and international mapping and charting programs with the confidence that there will be no inconsistencies between the networks produced by individual countries." The ECEF system implemented by the DOD for the GPS lays the foundation for meeting that goal. The GSDM described in this book builds on that foundation.

Although the overall perspective of the 1978 NRC report is excellent, its "crystal-ball" view didn't really anticipate the enormous impact of the digital revolution and makes little mention of the challenges associated with systematically describing spatial data accuracy. The final chapter of the 1978 report discusses future committee activities and includes seven specific important questions. Even though today's view enjoys the luxury of twenty-plus years of hindsight, those questions remain pertinent and should be considered carefully by anyone attempting to articulate the goals of modern geodesy.

The National Research Council also published a follow-up study (NRC 1990) entitled *Geodesy in the Year 2000*. Published only twelve years after the previous study, the 1990 report is much more upbeat about the role of geodesy and begins by noting, "We stand on the threshold of a technological revolution in geodesy. With the introduction of space-based observational techniques over the past two decades, geodesy has undergone and continues to undergo profound changes." The preface of the 1990 report declares, "Geodesy is becoming a truly global science.… And, since the earth's topography and gravity field are continuous across all political and geographic boundaries, mapping them requires careful integration of data collected by various survey techniques in different countries and physiographic regions."

The 1990 NRC report also discusses, among others, how geodesy will use gravity and precise measurements to provide a broader understanding of earthquakes with the idea of better predicting their occurrence and helping to mitigate their impact. And, it discusses the importance of oceanographic research with particular focus on the need for accurate measurements of the geoid and sea-level surface. But, the report also details the need for measurements to show space-time departures from average plate motions predicted by the global rigid-body motion models. The importance of spatial data accuracy is not ignored, but the report fails to anticipate the need for and benefits of using a universal, concise, rigorous stochastic model for spatial data. The report does include five specific recommendations for priorities to "be established in support of scientific and technological opportunities in geodesy for the year 2000." Of the five recommendations, number 3 states, "A global topographic data set should be acquired with a vertical accuracy of about 1 m, at a horizontal resolution of about 100 m." The "Overview and Recommendations" section closes with the statement, "These data [measurements] would also support establishment and maintenance of a conventional terrestrial coordinate system, which is necessary for comparison of results obtained by different space geodetic technologies."

Since 1990, various agencies and organizations have collected an enormous amount of geospatial data, and former Vice President Al Gore (1998) noted in a speech given in January 1998 at the California Science Center in Los Angeles, "The hard part of taking advantage of this flood of geospatial information will be making sense of it—turning raw data into understandable information." The National Spatial Data Infrastructure (NSDI) concept was developed to meet that challenge, and conceptualizers of the NSDI deserve credit for the help it provides. However, at a more fundamental level, the GSDM provides an underlying definition of spatial data and its accuracy. While the NSDI provides technical, organizational, and operational guidelines for using spatial data, the GSDM provides a specific definition of 3-D geometrical relationships and spatial data accuracy. In the rectangular ECEF environment, the rules of solid geometry and vector algebra are universal throughout, and the rules of error propagation can be applied without ambiguity. Issues of data compatibility and interoperability are handled efficiently in the context of the GSDM. From there, derived uses of spatial data are the prerogative of each user. Using the NSDI and GSDM in concert will be more efficient than using them separately. Implementation of an integrated model will not occur immediately, but, in the long run, their combined use will help address the challenge of making sense of the enormous quantities of spatial data being collected.

In the meantime, various initiatives (federal and otherwise) are addressing those challenges in, shall we say, a fragmented manner. Of the many that could be cited, a partial list of organizations, efforts, and relevant web addresses follows.

1. Former President Bill Clinton's 1994 Executive Order establishing the National Spatial Data Infrastructure (NSDI). See http://www.presidency.ucsb.edu/ws/index.php?pid=49945.
2. Former Vice President Al Gore's 1998 speech "The Digital Earth: Understanding Our Planet in the 21st Century." See http://www.isde5.org/al_gore_speech.htm.
3. Federal Geographic Data Committee: Coordinates the development of NSDI and develops standards for meta data (data about data). See http://www.fgdc.gov/metadata/—follow links to "standards" and "publications."
4. The Bureau of Land Management (BLM) and U.S. Forest Service, "National Integrated Land System (NILS)," is an effort to establish a common data model and software tools for collecting, processing, managing, and using spatial data. See http://www.blm.gov/wo/st/en/prog/more/nils.html
5. The Global Spatial Data Infrastructure (GSDI) "is an advocacy group promoting standards and policies for globally-compatible geo-spatial information standards and policies." See http://www.gsdi.org.
6. The Open GIS Consortium is a not-for-profit membership organization founded in 1994 to address the lack of interoperability among systems that process georeferenced data. Their motto is "Spatial connectivity for a changing world." See http://www.opengeospatial.org.

Note that the links listed were valid prior to publication of this book. If a link fails, use a search engine and relevant key words to find the web page containing the information.

Modern geodesy will certainly continue making contributions to the challenges of making sense of spatial data. But, the scope of geodesy includes much more than that. A statement of the goals for modern geodesy is left to the geodesists and other scientists. For example, the International Association of Geodesy (http://www.iag-aig.org) meets periodically and publishes proceedings of their meetings. The current series of publications includes volumes 101–128, the latter entitled, "A Window on the Future of Geodesy," the proceedings of the IAG General Assembly meeting in Sapporo, Japan, 30 June–11 July 2003.

Recent research and development activities have produced efficient reliable positioning tools that can be used to solve almost any location problem. And, it is a self-feeding cycle in that more tools and better tools are used by people who dream up even more applications. In turn, they develop even more elaborate tools. Modern society owes much to geodesy, including both the geodetic scientists and the practitioners. But, the collective appetite of modern society for technical miracles continues. For example, spatial data users ask for GPS that works under heavy canopy, and everyone wants a geoid model that can be used to convert ellipsoid height to reliable orthometric heights on a worldwide basis. Although progress and developments are both apparent, it is beyond the vision of this author to articulate goals for such a diverse discipline.

Prior to the twenty-first century, geometrical geodesy activities were generally separated into considerations of horizontal and vertical. The primary reference for horizontal location was a geodetic datum of latitude/longitude positions monumented at points on the Earth's surface. Elevation, the third dimension, was referenced to the geoid (or to an arbitrary level surface) according to a named vertical datum. Although horizontal and vertical datums both enjoy concise physical definition, they are incompatible in that there is no unambiguous geometrical relationship between mean sea level and the center of mass of the Earth. So, before goals of defining the size and shape of the Earth and locating points on or near the Earth's surface can be mutually fulfilled, the incompatibility of horizontal and vertical datums must be resolved. In place of two datums having separate reference points, a combined 3-D datum having a single origin is required. The implication of a single origin is that elevation will be a derived quantity instead of an observed quantity. If spatial data are attached to a well-defined reference frame such as the North American Datum of 1983 (NAD83) or a named epoch of the ITRF, the GSDM provides a computational environment that has geometrical consistency for any and all points within the birdcage of orbiting GPS satellites. More details are given in chapter 7, "Geodetic Datums."

The shift to using a 3-D datum and a single origin for spatial data may be analogous to the way time is measured and the need for the equation-of-time. John Flamsteed was the first royal astronomer of the Greenwich Observatory, which was built in 1675. The Greeks recognized the existence of the equation-of-time nearly 2,000 years earlier, but it was not until reliable clocks were available and Flamsteed made the necessary observations that the equation-of-time was quantified. Even with the equation-of-time known, most people still reckoned time from the sun's meridian transit at noon each day, and each railroad station had its own version of the correct time. In the United States the problem was solved in 1883 by adopting a system of standard time zones as devised by Charles F. Dowd of Saratoga Springs, New York (Howse 1980). The system was adopted for use worldwide at the International Meridian Conference in Washington, D.C., in 1884, and Greenwich Mean Time became the world standard. Now, most peoples of the world take standard time for granted, but, for scientific purposes, the equation-of-time and other time-scale differences are known, documented, and used by those for whom the difference matters.

Sea level is an intuitive physical reference for elevation and locally serves very well. But, on a global basis, locating the geoid precisely remains a challenge and will be the subject of research for years to come. Even in the United States, where geoid modeling has progressed dramatically in the past twenty years, spatial data users remain dissatisfied when absolute geoid heights are not reliable at the millimeter or centimeter level. Are elevations being done backward? Using time as an analogy, the way elevations are currently determined is like observing the instant the sun crosses the local meridian and re-setting the watch to 12:00:00 noon. Standard time is then found by applying an equation-of-time model to the recently observed solar time. No, current practice is to use standard time routinely, and the fact that the sun crosses the meridian before or after noon is of little consequence to most persons.

Formal arguments for using ellipsoid height for elevation are given in Burkholder (2002, 2006).

Reversing the process with regard to elevations means that ellipsoid height will be used routinely and that geoid height will be used to find orthometric height when needed (this part is already being done—see Meyer, Roman, and Zilkoski 2004). Making that change will involve some or all of the following:

1. The word "elevation" connotes a meaning for most people that should not be taken away. World Vertical Datum of XX (WVDXX) is a designation that could satisfy most spatial data users. In the United States it could be the vertical component of a 3-D datum to be published in 20xx.
2. Most people will remain oblivious to a change in definition that WVDXX elevation is the distance from the reference ellipsoid instead of mean sea level. In fact, a change in definition similar to that occurred in 1973, when mean sea level datum was renamed as the National Geodetic Vertical Datum (NGVD). Admittedly, switching from mean sea level to NGVD was a change of name only and did not involve changing any numbers published for a benchmark (Zilkoski 1992).
3. But, to make a clean break, a specific word or phrase is needed. It will be similar to switching from the National Geodetic Vertical Datum of 1929 (NGVD29) to the North American Vertical Datum of 1988 (NAVD88; see chapter 7). In that case, the change in datums did involve different numbers on each benchmark—due to readjustment. This time, however, the change will be due to redefinition of the reference surface to be compatible with using a single origin for 3-D spatial data. It will also provide an opportunity to update elevations based upon data obtained from recent surveying activities.
4. With adoption of the WVDXX, elevation will be the distance in meters from the reference ellipsoid. Elevation differences obtained from GPS will be used without the need for geoid modeling. Elevation differences from differential leveling (gravity based) will be used as WVDXX elevation differences *except* in those cases where deflection-of-the-vertical is large enough to make a difference at the level of accuracy required on the project. Examples include those persons using dynamic heights to compute hydraulic head in the Great Lakes System and physicists who deal with beam alignments in particle accelerators. Trigonometric height computations will no longer involve curvature of the Earth because that is handled implicitly by the GSDM. Vertical refraction of line of sight is still critical and will need to be considered where the required accuracy warrants it.

HISTORICAL PERSPECTIVE

Many authors generalize about our collective sense of wonder and our asking questions about things observed in nature. When did people first question the shape of the Earth? Hasn't it always been flat? Did Columbus (or was it just his crew) really believe that they would fall off the edge of the world if they sailed too far west? Civilizations develop as subsequent generations build on the body of knowledge developed by their ancestors. Books, libraries, the Internet, and interaction with living professionals are all revered as storehouses of knowledge, and education is the process of drawing from that storehouse to gain a greater understanding of our physical world and the actions of humankind.

Religion, Science, and Geodesy

Scholars and philosophers certainly have much more to say about the interaction of science and religion, but, for the purposes here, consider the following.

Religion

- Truth is something that can't be refuted.
- Faith is accepting as true something that can't be proved.
- Perception is an understanding based upon experience and evidence available. A person with normal eyesight perceives an elephant differently than those blind persons who base their perception only on touching various parts.
- Reality is based upon a collective evaluation of all the evidence, at least all that shared by others. An unsettling realization is that one can never be sure one has all the evidence. None of us really knows what it is we don't know.
- Belief is a conviction one holds based upon experience, training, education, inquiry, and insight. One of the fundamental tenets of democracy is that each person is individually responsible for what he or she believes as well as subsequent decisions and actions. It is impossible to force anyone to believe something. He or she can only be invited to consider the evidence.
- Dogma is characterized by a refusal to consider new or additional evidence: "My mind is made up. Do not confuse me with the facts."

Science

- Science can be defined as the systematic arrangement of knowledge in a logical order in which conclusions are consistent with beginning assumptions and subsequent observations.
    1. Physical science is concerned with physical matter, forces, and objects.
    2. Social science deals with the reasons for and consequences of decisions and actions made by humans (and other life forms).
- Whether physical or social, science is also categorized according to method of inquiry: theoretical or applied.
    1. Theoretical (also called pure) scientific research is conducted for the expressed purpose of gaining a better understanding of the matter, objects, or process of inquiry.
    2. Applied science (also called engineering) is conducted for the purpose of finding or documenting that arrangement of elements or sequence of events that will produce a desired outcome.

The historical development of geodesy could be written as a story of the development of science. Over the centuries many learned persons have contributed to a better understanding of the size and shape of the world, and people living today have the luxury of sharing innumerable triumphs of the human spirit. Measurements were made in the physical world and dutifully recorded. Observations were checked against prevailing thought and accepted, challenged, or ignored. Where warranted, old models were cast aside and new models were proposed and tested. In hindsight we can see where progress was made and where progress was thwarted for various reasons (often having to do with prevailing religious attitudes). Russell (1959) states that one of the most difficult accomplishments for a person is to hold an idea with conviction and detachment at the same time. With regard to the development of science, that story undoubtedly has a thousand variations. Alder (2002), Bell (1937), Berthon and Robinson (1991), Carta (1962), Sobel (1999), Smith (1986, 1999), and Wilford (1981) relate some of them. Many such stories are related to geodesy, and only a few of them are summarized here.

Learned persons have known since before the time of Christ that the Earth is spherical. Pythagoras (born 582 b.c.) declared the Earth to be a globe. Aristotle (384–322 b.c.), upon viewing the Earth's shadow cast upon the moon during an eclipse, concluded the Earth must be spherical. Others pointed out that a curved Earth is apparent from watching the arrival of a ship at port because the top of the sail is readily visible before the ship itself. Admittedly, the local perspective is that the Earth is flat, and that perspective is appropriate for many activities. But, when warranted, the Earth's curvature must be accommodated.

Degree Measurement

At a fundamental level, the process of determining the Earth's size requires two pieces of information from which the third is inferred using the well-known equation $L = R\theta$. Although not always the same, the process has been called degree measurement and consists of measuring an arc on the surface of the Earth along with the corresponding subtended angle. From that length-per-degree, the circumference or radius is readily computed. For example, if $L = 500$ miles and the subtended angle is 1/50 of a circle, the circumference is immediately computed as 25,000 miles and the Earth's radius is found to be about 4,000 miles.

Eratosthenes

Eratosthenes lived approximately 276–195 b.c. in the Mediterranean port city of Alexandria near the mouth of the Nile River in Egypt. He is given credit for first determining the size of the world using the degree measurement method just described. He measured the length of a shadow cast by an obelisk at noon on the longest day of the year in Alexandria. He also knew that on the longest day of the year, the sun shown directly to the bottom of a well located on the Tropic of Cancer near the city of Syene (not far from the present site of the Aswan Dam). From that, he deduced the angle subtended at the center of the Earth to be 1/50 of a circle (see Figure 5.1). Researchers do not agree on how the distance between Alexandria and Syene was measured, but the consensus is that the distance was 5,000 stadia and that there are

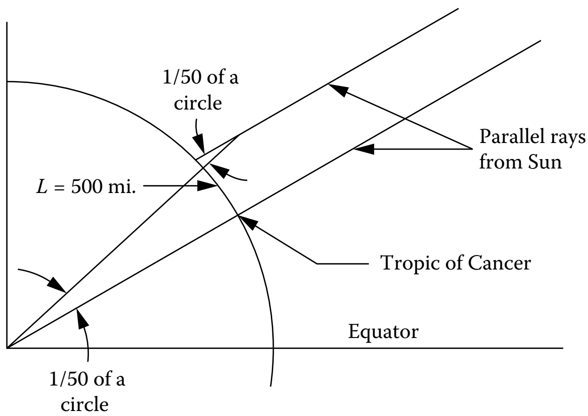


FIGURE 5.1 Erathosthenes' Measurement of the Earth

about 185 meters per stadia. That being the case, the computed circumference is 46,250,000 meters—only about 16 percent larger than the currently known size of the Earth. It is conceded there may have been some compensating errors because Alexandria is not directly north of Syene, as assumed by Eratosthenes; the quality of his arc distance is questionable; and the well was probably not located directly on the Tropic of Cancer (Tompkins 1971). But, his results are a good approximation and show that humans have known for several millennia that the Earth is not flat.

Poseidonius

Poseidonius (135–150 b.c.) was a Greek astronomer who also determined the Earth's size using degree measurements. Alexandria, on the south shore of the Mediterranean, was the south end of his line. The north end was at Rhodes, on an island on the north side of the Mediterranean Sea about twenty kilometers from Turkey. The arc distance was based upon the sailing time of a ship, and the angle subtended at the center of the Earth was deduced from the difference in vertical angle at the two ends of the line to the constellation Canopus as it crossed the meridian. His results were about 10 percent too big.

Caliph Abdullah al Mamun

Arabian efforts included a measurement on the plains near Baghdad by the Caliph Abdullah al Mamun about 827 a.d. that yielded an answer only about 3.6 percent too big. Subsequent reevaluation (Carta 1962) of the unit conversions from ells to barleycorns and Rhineland feet gave another value that was about 10 percent too big. Lack of length standardization still plagues modern interpretation of early work.

Gerardus Mercator

Navigation from the European continent flourished during the thirteenth and fourteenth centuries, and mapmaking became a valuable occupational talent. The person, shipping company, or sovereign possessing the better map enjoyed an enormous advantage. Stories of intrigue abound. Gerardus Mercator (1512–1594) was one of the most famous mapmakers because he published a map of the world in 1569 having parallels of latitude on the map spaced so that one could sail from one port to another on a constant bearing. Later, Mercator's map was shown to be what is now called a conformal projection—the scale distortion at a given point is the same in all directions.

Willebrord Snellius

In 1615 a.d. a Dutchman, Willebrord Snellius (1580–1626), measured an arc more than eighty miles long using a series of thirty-three triangles. His computation of the size of the Earth was too small by about 3.4 percent—not bad for using a telescope with no crosshairs in it and measuring baseline distances with an odometer connected to his carriage wheel. Galileo Galilei (1564–1642) had invented the telescope in 1611, but the Gunter's chain (used in land surveying for several hundred years) was not invented until about 1620.

Jean Picard

The French Academy of Science was founded in 1666 and sponsored, among others, geodetic surveying activities designed to answer questions about the size and shape of the Earth. During 1669–1670, Jean Picard (1620–1682) measured an arc of triangulation from Paris to Amiens using a telescope containing crosshairs and "well seasoned varnished wooden rods" for measuring baseline lengths. Conventional wisdom at the time presumed the Earth to be spherical.

Isaac Newton

Following up on Galileo's work on the pendulum, Isaac Newton (1642–1727) formulated his theory of universal gravitation during the mid-1660s. Using the commonly accepted values for the size of the Earth, Newton was frustrated that the evidence was not consistent with his theory and laid the work aside. It wasn't until after he incorporated Picard's results that Newton's theories were validated and published in his *Principia Mathematica* in 1687.

While his universal gravitational theory was laid aside, Newton became convinced the Earth is flattened at the poles due to the vector sum of two forces, as shown in Figure 5.2. Gravitational attraction pulls an object toward the Earth's center of mass, and centrifugal force acts on the same object perpendicular to the Earth's spin axis. An object can't differentiate between the two forces but responds only to the vector sum of the two forces. Newton also noted that a level surface (sea level) is always perpendicular to the plumb line and that the plumb line will point to the Earth's center only if one is at the equator or at one of the poles. At the equator, gravitational attraction and centrifugal force are colinear and centrifugal force counteracts (reduces) gravitational attraction. At the pole, centrifugal force is zero. Newton argued that an Earth flattened at the poles is the only shape that is consistent with those conditions. The irony is that Newton needed Picard's work to verify

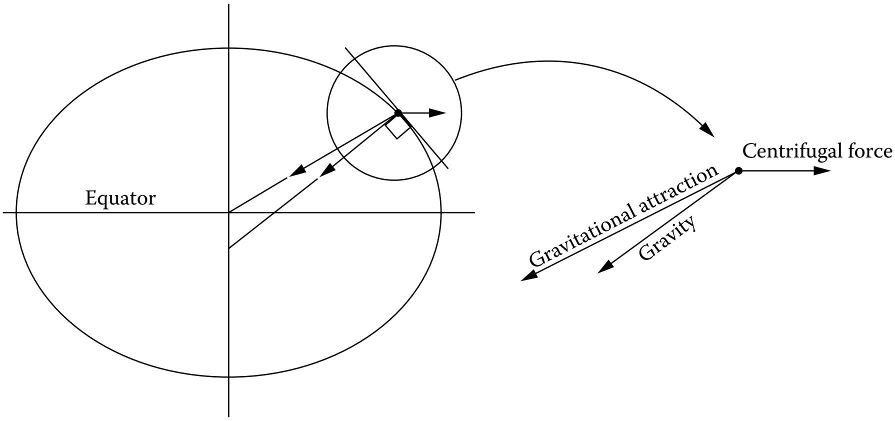


FIGURE 5.2 Newton's Logic for a Flattened Earth

the correctness of his own theories, but Picard and the French Academy refused to accept Newton's theory of a flattened Earth.

Jean-Dominique and Jacques Cassini

Jean-Dominique Cassini (1625–1712) was a brilliant scientist who was lured away from Italy and service to the pope to become the first director of the Paris Observatory, built between 1667 and 1672. In the early 1680s King Louis XIV of France announced Picard's work would be extended from Dunkirk on the north to Collioure in the south. Since Picard had died, Jean-Dominique Cassini directed the survey work. As the work progressed, it became apparent, based upon computations performed separately on that part north of Paris and that part south of Paris, that the radius of the Earth for the northern portion was shorter than the radius for the southern portion. The implication was that the Earth is elongated at the poles, not spherical as had been believed, and not oblate (flattened at the poles) as postulated by Newton.

It became an international debate, with Newton (British) and Christiaan Huygens (1629–1695; Dutch) arguing the Earth is oblate or flattened at the poles and the Cassinis (French) arguing the Earth is prolate or elongated at the poles. After Jean's death in 1718, his son, Jacques Cassini (1677–1756), carried on the work and insisted that the only way to settle the argument was to conduct decisive tests near the equator and near the Arctic Circle. Incidentally, the post of director of the Paris Observatory was held by four generations of Cassinis (Smith 1986).

French Academy of Science

In 1734 King Louis XV directed more geodetic measurements be taken in France. When those measurements failed to settle the argument, the French Academy of Science sponsored two geodetic surveying expeditions as recommended by Jacques Cassini. In 1735 one expedition left for Peru, and another left for Lapland in 1736.

Details of those two expeditions are included in Smith (1986). When the leader of the Lapland expedition returned in 1637 with conclusive proof that Newton was right, Voltaire, a French philosopher and social critic, wrote, "You have flattened the earth and the Cassinis." When the more strenuous Peru expedition returned in 1741, Voltaire commented, "You have found by prolonged toil what Newton learned without leaving home." Whitaker (2004) writes about peripheral challenges encountered during the Peru expedition.

Meter

An idea promoted by Picard was the need for a decimally divided standard of length based upon some natural quantity or simple physical observation. He proposed using the length of a pendulum having a period of 1 second as being the standard of length. His idea was not accepted, in part because the period of a pendulum changes with both latitude and altitude. But, by 1790 the idea of a decimal length standard had caught on to the point that the French Academy of Science sponsored yet another geodetic survey in France, the purpose of which was to determine as precisely as possible the arc distance from the equator to the pole. The French Revolution notwithstanding, the work was completed during the 1790s by Jean Baptiste Joseph, Chevalier Delambre (1749–1822), and Pierre Méchain (1744–1804), and the distance from the equator to the pole, 5,130,740 toises, was set to be exactly 10,000,000 meters. Given further that 1 toise = 864 Paris lines, one can compute 1 meter = 443.296 Paris lines. "The Measure of All Things" (Alder 2002) is a postscript to this section and includes a fascinating account of the measurement of the meter by Delambre and Méchain.

DEVELOPMENTS DURING THE NINETEENTH AND TWENTIETH CENTURIES

By the end of the eighteenth century, the United States had won its independence, Thomas Jefferson had served as minister to France, the French revolution was over, and the decimal meter was defined. A brief summary of geodetic developments during the next 200 years includes the following:

- Least squares: Given various determinations for the size and shape of the Earth, the theory of least squares, published in 1806, was developed by Adrien-Marie Legendre (1752–1833) to give proper weight to the various determinations for the size and shape of the Earth. Credit for inventing least squares is also shared with Carl Friedrich Gauss (1777–1855), who, at the age of eighteen, independently invented the technique of least squares in 1795.
- The United States purchased the Louisiana Territory from France in 1803, and during most of the nineteenth century the rectangular U.S. Public Land Survey System was systematically and permanently etched on the curved surface of the North American continent.

- In 1807 President Thomas Jefferson established the Survey of the Coast, and Ferdinand Hassler was hired as the first director. Ferdinand Hassler (1770–1843) was a Swiss scientist who was well versed in making geodetic surveys and brought an impressive technical library and scientific instruments with him to the United States (including a standard meter bar and three standard toise bars). One of his first tasks as director of the Survey of the Coast was to make a trip to Europe to acquire (additional) appropriate instruments for the task. Over the next 150 years, the Survey of the Coast and its successor, the U.S. Coast and Geodetic Survey, observed arcs of high-order triangulation over the entire United States.
- In 1799 Captain William Lambton (1756–1823) "drew up a project for a mathematical and geographical survey that would extend … from the southern tip to the northern extreme" of India. Formal orders for the survey were issued in 1800, and work commenced with measurement of a base in 1802. George Everest (1790–1866) joined the Great Trigonometric Survey (GTS) in 1819 and quickly established himself as a capable assistant to Superintendent Lambton, who died in 1823. Everest carried on the work until he returned to England in November 1825 on sick leave. In 1829 Everest was named surveyor general of India. He returned to India in October 1830, where he supervised the work until he retired in December 1843 on the pension of a full colonel. Everest then returned to England, where, at the age of fifty-six, he married a twenty-three-year-old woman whose father was six years younger than himself. In all his years in India, George Everest never laid eyes on the highest mountain in the world that was named in his honor in August 1856 (Smith 1999).
- The U.S. Coast and Geodetic Survey (now known as the National Geodetic Survey) completed triangulation arcs and computed massive network adjustments in 1927 and again in 1986 that provide local users reliable geodetic positions for monumented points on the surface of the Earth. These network adjustments were known respectively as the North American Datum of 1927 (NAD27) and the North American Datum of 1983 (NAD83). Leveling networks were also observed and adjusted, and vertical datum elevations were published.
- In October 1957, the Russians launched the first artificial satellite to orbit the Earth, and by 1964 the United States had developed the Transit Doppler satellite positioning system that was used by the U.S. military for positioning Polaris submarines anywhere on the oceans, twenty-four hours a day, rain or shine. The first navigation satellite timing and ranging (NAVSTAR) GPS satellites were launched in 1978, and by 1985 GPS was recognized worldwide as the premier satellite-positioning system. The full constellation of satellites was subsequently completed, and the GPS was declared fully operational in July 1995.
- The transistor and electronic computer were invented in the middle of the twentieth century, and, with regard to surveying and mapping, the electronic distance meter, photogrammetry, GPS, and remote sensing revolutionized the use of spatial data. These developments are all considered part

of the digital revolution. The ubiquitous personal computer (PC), the Internet, and cellular telephone technologies are currently riding the wave of the digital revolution.
- The global spatial data model (GSDM) was first proposed by Earl F. Burkholder in 1997 and is described in a report prepared for the Southeast Wisconsin Regional Planning Commission (Burkholder 1997). The GSDM is intended to be compatible with the use of digital spatial data; the technologies employed for the generation, storage, manipulation, and use of spatial data; and the way humans perceive the world and use spatial relationships.

FORECAST FOR THE TWENTY-FIRST CENTURY

With the dawn of the new millennium, the planet Earth and civilization stand at the threshold of innumerable opportunities. With regard to spatial data and the evolution of geomatics as an umbrella discipline, a few forecasts include the following:

- A 3-D global datum will emerge as a global standard for data exchange.
- Compatibility issues related to horizontal and vertical datums will be addressed in terms of how each is related to the standard global model.
- Spatial data accuracy with respect to a datum will be assessed two ways:
  - A. A collection of points or features (a spatial data set) will have one accuracy ranking based upon the meta data describing the modes of collection and the models used in processing.
  - B. Each point location will have a covariance matrix associated with it that provides statistically reliable standard deviations in three dimensions. Some data sets will also include time as the fourth dimension.
- Statistical correlation between points will be obtained from the stored covariance values. Network accuracy and local accuracy will be computed from the same stored covariance values.
- At the conceptual level, solid geometry, matrices, and vector algebra will be emphasized as appropriate tools for handling spatial data. In the applications arena, spatial data users will be able to work with simple local rectangular coordinate differences that reflect a local perception of a flat Earth. True geometrical integrity over the curved Earth will be preserved as a consequence of using the 3-D GSDM.
- WVDXX elevation (or some other name) will be a derived quantity computed as the distance above or below a named reference ellipsoid. Geodetic scientists will still be concerned about the difference between the ellipsoid and geoid and provide that information to researchers who need to find the geoid. But in local practice, except for specialized applications, the ellipsoid height difference will be used as if it were elevation difference. Before adopting such a practice, the impact of making the change will need to be studied carefully, especially in areas where the deflection-of-the-vertical is large or changes rapidly. Otherwise, the need for elaborate geoid modeling will be drastically reduced.

- Persons using cellular telephones and the Internet for communication along with GPS (and other technology) for position will provide a wide range of location-based services for an increasing segment of the population on a global basis. Data compatibility and interoperability between disciplines are critical.
- A global 3-D spatial database provides the framework within which spatial data from any source can be freely exchanged. Decentralized use of spatial data is supported in that each user is free to manipulate data according to local practice and specifications. Rules for decentralized data manipulation are the responsibility of each user. At the foundation level, the functional and stochastic components of the GSDM stipulate the requirements for meeting any level of accuracy specified by the user.

REFERENCES

Alder, Ken. 2002. *The measure of all things: The seven-year odyssey and hidden error that transformed the world*. New York: Free Press.

Bell, E. T. 1937. *Men of mathematics*. New York: Simon & Schuster.

Berthon, Simon, and A. Robinson. 1991. *The shape of the world*. Chicago: Rand McNally.

Burkholder, E. F. 1997. *Definition of a three-dimensional spatial data model for Southeastern Wisconsin*. Waukesha: Southeastern Wisconsin Regional Planning Commission.

———. 2002. Elevations and the global spatial data model (GSDM). Paper presented at the 58th annual meeting of the Institute of Navigation, Albuquerque, NM, 24–26 June.

———. 2006. The digital revolution: Whither now? *GIM International* 20 (9): 25–27.

Carta, Martha, trans. 1962. *Jordan's handbook of geodesy*. Washington, DC: U.S. Army Corps of Engineers, Army Map Service.

Gore, A. 1998. The digital earth: Understanding our planet in the 21st century. Speech presented at the California Science Center, Los Angeles, January.

Howse, D. 1980. *Greenwich Time and the discovery of longitude*. Oxford: Oxford University Press.

Meyer, T., D. Roman, and D. Zilkoski. 2004. What does height really mean? Part I: Introduction. *Surveying & Land Information Science* 64 (4): 223–33.

National Research Council (NRC). 1978. *Geodesy: Trends and prospects*. Washington, DC: National Research Council, National Academy of Sciences.

———. 1990. *Geodesy in the year 2000*. Washington, DC: National Research Council, National Academies Press.

Russell, Bertrand. 1959. *Wisdom of the West*. London: Rathbone.

Sobel, Dava. 1999. *Galileo's daughter*. New York: Walker.

Smith, J. R. 1986. *From plane to spheroid: Determining the figure of the Earth from 3000 B.C. to the 18th Century Lapland and Peruvian survey expeditions*. Rancho Cordova, CA: Landmark Enterprises.

———. 1999. *Everest: The man and the mountain*. Caithness, Scotland: Whittles.

Tompkins, Peter. 1971. *Secrets of the great pyramids*. New York: Harper & Row.

Wilford, John Noble. 1981. *The mapmakers*. New York: Vintage.

Zilkoski, David. 1992. North American Vertical Datum and International Great Lakes Datum: They are one and the same. In *Proceedings of the U.S. Hydrographic Conference '92*, Baltimore, MD.

6 Geometrical Geodesy

INTRODUCTION

Geometrical geodesy is the branch of science that deals with the Earth's size and shape. Viewed from a large distance, the Earth appears to be nearly spherical, and, by comparison, the Earth is smoother than an orange—even when counting the highest mountains. If our planet Earth were reduced to a globe having a diameter of 1.0000 meter at the equator, the length of the spin axis would be 0.99665 meters, only 3.35 millimeters less. As postulated by Newton, this flattening at the poles is due primarily to the Earth spinning on its axis. The sea-level shape of the Earth is a continuous surface called the geoid. As such, it could be viewed as the surface of the ocean at rest (no tides) and extending coast to coast in a large transcontinental canal. An ellipse rotated about its minor axis—giving a three-dimensional mathematical surface called the ellipsoid—approximates this geoid shape of the world. The distance between the ellipsoid and geoid is called geoid height and varies, plus or minus, up to about 100 meters worldwide. Figure 6.1 is a meridian section of the Earth showing the poles, the equator, the spin axis, the mathematical ellipsoid, the geoid, the normal, and the vertical. Note that the ellipsoid-geoid separation and the Earth's flattening are both exaggerated in Figure 6.1.

An important concept in physical geodesy is that a level surface is perpendicular to the plumb line at every point. The plumb line in Figure 6.1 defines the vertical. In geometrical geodesy, the ellipsoid normal is always perpendicular to the tangent to the ellipse. The angular difference between the normal and the vertical at a point is called the deflection-of-the-vertical and is discussed more in chapter 8, "Physical Geodesy." In this chapter, the deflection-of-the-vertical is taken to be zero.

As described by Newton and discussed in chapter 5, the physical shape of the geoid is determined primarily by gravitational attraction and centrifugal force. The centrifugal force at a point is constant and can be computed with a high degree of certainty. But, due to the irregular distribution of mass and variations of density within the Earth, the force of gravity is not so uniform. Variations in the shape of the geoid caused by gravity anomalies are discussed in chapter 8.

An ellipsoid is the mathematical basis of geometrical geodesy and is an approximation of the sea-level shape of the Earth. From a local perspective, and for many uses, it can be said that the Earth is flat. In that context, horizontal dimensions are flat and vertical dimensions are up. In many ways, such a flat-Earth model is easier to use than is the ellipsoid model, and many spatial data users prefer working with plane rectangular coordinates. But, as has been known for centuries, the Earth is not flat, and numerous applications arise in which the flat-Earth model is not adequate. A more complex model for specifying location on the Earth is needed. One could say that the spatial data (location) model has evolved from saying the Earth is flat, to

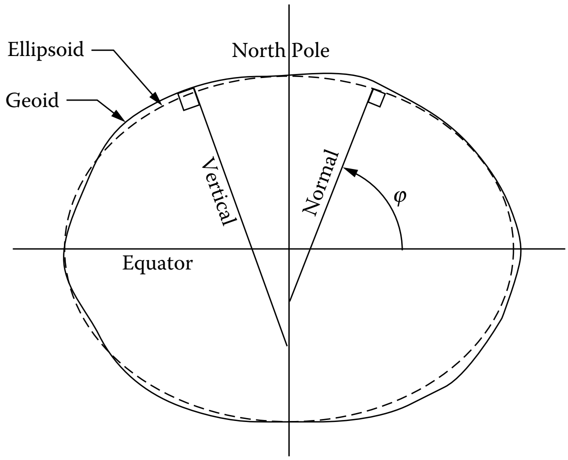


FIGURE 6.1 Ellipsoid and Geoid

using a sphere to approximate the Earth's size and shape, to using an ellipsoid chosen for a regional best fit, to using an ellipsoid selected for a global best fit.

So, even though the spatial data community may prefer using a flat-Earth model that accommodates a local perspective, a better understanding of basic spatial data concepts comes from learning the geometry of a rotational ellipsoid that

- has its origin at (or near) the Earth's center of mass.
- shares its minor axis with the spin axis of the Earth.
- has its major axis in the plane of the Earth's equator.
- has its zero meridian coincident with the Greenwich meridian.
- uses sexagesimal latitude and longitude coordinates on the ellipsoid surface.
- uses elevation (and/or height) in meters for the third dimension.

Esoteric issues (such as polar wandering and the aforementioned deflection-of-the-vertical) need more explanation than is offered in this chapter. But the point made here is that the GSDM is a mechanism that permits spatial data users to work with local rectangular (flat-Earth) differences while the underlying model appropriately accommodates the ellipsoidal shape of the Earth. Not only that, but with careful planning, those esoteric issues, which are still a legitimate concern to geodesists, will have little or no impact on the local spatial data user. Additionally, the GSDM offers well-defined mathematical procedures by which the 3-D accuracy of spatial data can be established, tracked, and utilized—even from a local flat-Earth perspective.

THE TWO-DIMENSIONAL ELLIPSE

The 3-D geometry of the ellipsoid begins with the meridian section of a 2-D ellipse. Figure 6.2 shows a meridian section in which the ellipse major axis is coincident with the Earth's equator and the minor axis is coincident with the Earth's spin axis.

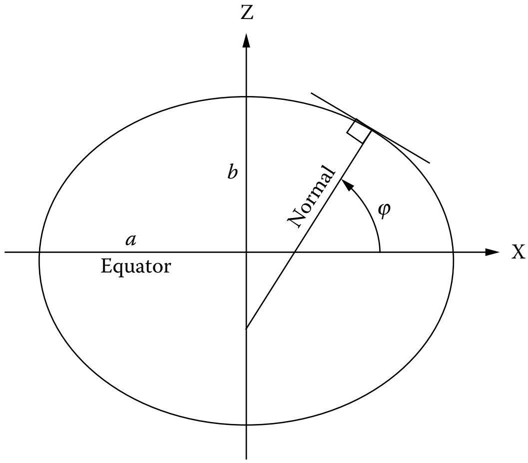


FIGURE 6.2 Two-Dimensional Ellipse

The length of the semimajor axis is denoted as *a*, and the length of the semiminor axis is denoted as *b*. Equation 6.1 is the equation of a 2-D ellipse in the *X/Z* plane whose size and shape are defined by parameters *a* and *b*.

$$\frac{X^2}{a^2} + \frac{Z^2}{b^2} = 1 \tag{6.1}$$

Note that, for the Earth, *a* is greater than *b* and the flattening of the ellipse is expressed several ways.

Flattening is defined as

$$f \equiv \frac{a-b}{a} = 1 - \frac{b}{a} \tag{6.2}$$

Eccentricity squared is

$$e^2 \equiv \frac{a^2 - b^2}{a^2} = 1 - \frac{b^2}{a^2} \tag{6.3}$$

Second eccentricity squared is

$$e'^2 \equiv \frac{a^2 - b^2}{b^2} = \frac{a^2}{b^2} - 1 \tag{6.4}$$

The polar radius of curvature is useful when making comparisons between ellipsoids. The polar radius of curvature is

$$c = \frac{a^2}{b} = \frac{a}{\sqrt{1-e^2}} \tag{6.5}$$

Additional geometrical relationships derived from the definitions above include

$$\lambda e^2 = 2f - f^2; \qquad b = a(1-f) \qquad (6.6 \text{ and } 6.7)$$

$$\frac{b^2}{a^2} = 1 - e^2 ; \qquad e'^2 = \frac{e^2}{1-e^2} \qquad (6.8 \text{ and } 6.9)$$

Several observations with regard to the Earth ellipsoid, a, b, f, e^2, and e'^2 are as follows:

1. If $b = a$, the meridian ellipse is really a circle and the ellipsoid is a sphere. The flattening, the eccentricity, and the second eccentricity are all zero.
2. As b approaches 0, the ellipse degenerates into a line and the ellipsoid becomes a flat plane. In that case, the values of f, e^2, and e'^2 are all exactly 1.
3. For the Earth, values of f, e^2, and e'^2 are all near zero and the ellipsoid is nearly spherical.
4. Some derivations in geodesy are accomplished more efficiently using the second eccentricity instead of the eccentricity. Although equally legitimate, the second eccentricity is used in this book only as required to be consistent with referenced material.

It takes two geometrical elements to define an ellipse. The most obvious pair of elements is the semimajor and semiminor axes. However, over the years the following conventions have been adopted and are used as noted. In each case, the ellipsoid is obtained by rotating the 2-D ellipse about its minor axis.

- a and b: The semimajor axis and the semiminor axis. The Clarke Spheroid of 1866 is defined by a and b and used in North America for the North American Datum of 1927 (NAD27). See Table 6.1.

TABLE 6.1
Selected Geometrical Geodesy Ellipsoids

Ellipsoid	Defined By	Derived Values
Clarke Spheroid of 1866 (used for NAD27)	a = 6,378,206.4 m b = 6,356,583.8 m	$1/f$ = 294.978698214 e^2 = 0.006865799729 c = 6,399,902.5516 m
Geodetic Reference System of 1980 (used for NAD83)	a = 6,378,137.000 m $1/f$ = 298.257222101	b = 6,356,752.3141 m e^2 = 0.0066943800229 c = 6,399,593.6259 m
World Geodetic System of 1984 (used in GPS positioning)	a = 6,378,137.000 m $1/f$ = 298.257223563	b = 6,356,752.3142 m e^2 = 0.0066943799902 c = 3,399,593.6258 m

- a and $1/f$: The semimajor axis and the reciprocal flattening. The Geodetic Reference System of 1980 (GRS80) and the World Geodetic System of 1984 (WGS84) as used in geometrical geodesy are both defined by a and $1/f$.
- a and e^2: The semimajor axis and the eccentricity squared. Most geodetic computations are arranged with the goal of preserving computational strength and efficiency. Although eccentricity squared is not a defining parameter, the quantity $(1 - e^2)$ is used in many geodesy equations and contains two more significant digits than does e^2 by itself.

Note that values of c, the radii of curvature at the poles, for the GRS80 and the WGS84 are within 0.1 mm of being identical even though the $1/f$ (and e^2) values are significantly different. That agreement supports the statement that the GRS80 and WGS84 ellipsoids can be used interchangeably. An important distinction is that, while they may be interchangeable as ellipsoids, WGS84 datum coordinates are not interchangeable with the North American Datum of 1983 (NAD83) datum coordinates if agreement between datums is expected at about the 1 meter level. See chapter 7 on geodetic datums for more details.

An ellipsoid was defined previously as that figure generated by rotating an ellipse about its minor axis. That same definition was used in the past to describe a spheroid, and the Clarke Spheroid of 1866 still carries the word "spheroid" in its name. Although "ellipsoid" and "spheroid" have essentially the same definition in geometrical geodesy and are often used interchangeably, a distinction can be made. A spheroid is generically defined as a body that is nearly spherical, but not quite. That definition permits, but does not require, that the surface be rigorously defined mathematically. The ellipsoid enjoys rigorous mathematical definition in all cases. The convention in this book is to use the word "ellipsoid" when referring to a mathematical approximation of the size and shape of the Earth.

The ellipse shown in Figure 6.2 is taken through the Greenwich meridian and shows the ellipse in terms of the X/Z plane. Figure 6.3 shows the rectangular $X/Y/Z$ geocentric ECEF coordinate system superimposed upon the ellipsoid. The X and Y distances are in the plane of the equator; the Greenwich meridian is in the X/Z plane; the Z-axis is coincident with Earth's spin axis; and the X-axis pierces the equator at the Greenwich meridian. Figure 6.2 also shows a line tangent to the meridian ellipse. The line perpendicular to the tangent is the ellipsoid normal (N) and goes from the ellipse tangent to the spin axis. Geodetic latitude is the angle the normal makes with the equator and is denoted by the Greek letter phi (φ). The geodetic latitude ranges from –90° at the South Pole to 0° on the equator to +90° at the North Pole.

With respect to equation 6.1 and Figure 6.2, expressions for X and Z (in terms of a, e^2, and ϕ) are found by equating the slope of the tangent to the ellipse with the first derivative of equation 6.1 at the same point. The slope of the tangent to the ellipse is given by the trigonometric tangent of the angle (90° + ϕ). Using trigonometric identities,

$$\tan\left(90+\phi\right)=\frac{\sin(90+\phi)}{\cos(90+\phi)}=\frac{\cos\phi}{-\sin\phi}=\frac{-1}{\tan\phi} \tag{6.10}$$

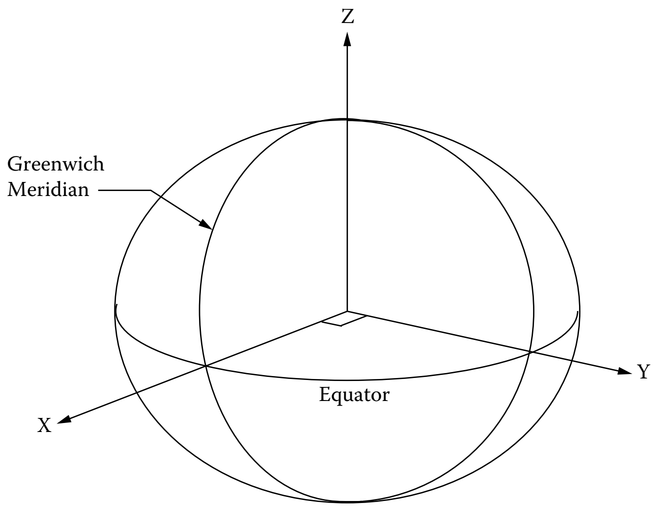


FIGURE 6.3 Three-Dimensional Ellipsoid

By algebraic manipulation, the derivative of equation 6.1 is

$$\frac{2X}{a^2}dX + \frac{2Z}{b^2}dZ = 0 \qquad \text{or} \qquad \frac{dZ}{dX} = -\frac{2b^2Z}{2a^2X} \tag{6.11}$$

Equating equations 6.10 and 6.11, substituting for b^2/a^2, and solving for Z, we get

$$\frac{-1}{\tan\phi} = \frac{-b^2}{a^2}\frac{X}{Z} \rightarrow Z = \left(1-e^2\right)\frac{\sin\phi}{\cos\phi}X \text{ and } Z^2 = (1-e^2)^2\frac{\sin^2\phi}{\cos^{2\phi}}X^2 \tag{6.12}$$

Next, solve equation 6.1 for Z^2,

$$\frac{X^2}{a^2} + \frac{Z^2}{b^2} = 1 \rightarrow Z^2 = b^2\left|1 - \frac{x^2}{a^2}\right| = \left(1-e^2\right)\left(a^2 - X^2\right) \tag{6.13}$$

Now, equate Z^2 in equation 6.12 to Z^2 in equation 6.13, and solve for X.

$$\left(1-e^2\right)\left(a^2 - X^2\right) = \left(1-e^2\right)^2\frac{\sin^2\phi}{\cos^2\phi}X^2$$

$$\cos^2\phi\left(a^2 - X^2\right) = X^2\left(1-e^2\right)\sin^2\phi$$

Use algebraic manipulation on the last two terms in equation 6.13 to solve for X as

$$\cos^2\phi\left(a^2 - X^2\right) = X^2\left(1-e^2\right)\sin^2\phi$$

$$a^2\cos^2\phi - X^2\cos^2\phi = X^2\sin^2\phi - X^2e^2\sin^2\phi$$

$$a^2\cos^2\phi = X^2\left(\cos^2\phi + \sin^2\phi\right) - X^2e^2\sin^2\phi = X^2\left(1-e^2\sin^2\phi\right)$$

$$X^2 = \frac{a^2\cos^2\phi}{1-e^2\sin^2\phi} \text{ and } X = \frac{a\cos\phi}{\sqrt{1-e^2\sin^2\phi}} \tag{6.14}$$

Now substitute equation 6.14 into equation 6.12 to solve for Z:

$$Z = \left(1-e^2\right)\frac{\sin\phi}{\cos\phi}\left(\frac{a\cos\phi}{\sqrt{1-e^2\sin^2\phi}}\right) = \frac{a\left(1-e^2\right)\sin\phi}{\sqrt{1-e^2\sin^2\phi}} \tag{6.15}$$

Note further in Figure 6.2 that the **normal** is the hypotenuse of a right triangle whose base is X. However, the usual statement, also apparent from Figure 6.2, is that

$$N\cos\phi = X \text{ from which } N = \frac{a}{\sqrt{1-e^2\sin^2\phi}} \tag{6.16}$$

Because the Earth is flattened at the poles, the instantaneous radius of curvature in the north-south direction increases as one moves from the equator toward either pole. This is consistent with Newton's oblate Earth theory that was proven conclusively in the mid-eighteenth century by geodetic surveying expeditions to Lapland and Peru (modern-day Finland and Ecuador). The instantaneous radius of curvature in the north-south direction on the ellipsoid at a specified latitude is given by the general mathematical equation for curvature of a function as

$$M = \frac{-\left[1+\left(\dfrac{dZ}{dX}\right)^2\right]^{3/2}}{\dfrac{d^2Z}{dX^2}} \tag{6.17}$$

Starting with equation 6.1, taking the first and second derivatives, and substituting them into equation 6.17 is tedious and involves much algebraic manipulation. But the result is stated concisely (the negative sign is dropped by convention) as

$$M = \frac{a(1-e^2)}{\left(1-e^2 \sin^2 \phi\right)^{3/2}} \tag{6.18}$$

THE THREE-DIMENSIONAL ELLIPSOID

Equations 6.1 to 6.18 are presented in terms of the Greenwich meridian section. Now, consider that the ellipse is rotated about the minor (Z) axis and that the foregoing equations are applicable in any meridian. And, since we are no longer restricted to the meridian ellipse, it is appropriate to speak in terms of the 3-D ellipsoid as depicted in Figure 6.3.

Ellipsoid Radii of Curvature

Parallels of latitude are circles on the ellipsoid in planes parallel to the equator. The location of any parallel on the ellipsoid is measured in degrees, minutes, and seconds north or south from the equator. At each parallel of latitude, the underlying ellipsoid has a radius of curvature in the north-south direction (that is, M, as discussed previously). At the same location, the underlying ellipsoid also has a radius of curvature in the east-west direction (perpendicular to the meridian). The plane perpendicular to the meridian is called the prime vertical and contains the ellipsoid normal. At any point, the radius of curvature in the plane of the prime vertical is N, the length of the ellipsoid normal. Remember, if the Earth were perfectly spherical, the radius would be the same at all points. And, at any given point, the radius of curvature of the underlying sphere would be the same in any direction. But, the Earth is flattened at the poles, and the ellipsoid radius of curvature changes with latitude and with the direction of a line. At any point on a given parallel of latitude, M is the radius of curvature of the underlying ellipsoid in the north-south direction and N is the radius of curvature at the same point in the east-west direction. M and N are collinear, but are not the same length (except at the poles).

Table 6.2 shows the results of tabulating values for both M and N at the equator and at the poles. Note that the value of N at the equator is a, the semimajor axis. The value of M at the equator is somewhat shorter. Except at the poles, the ellipsoid radius of curvature at any point is shorter in the N-S direction than in the E-W direction. At the poles, values for M and N are identical and called c, the polar radius of curvature (see equation 6.5). The reason is that the prime vertical at the pole (90° to a meridian) is itself another meridian.

TABLE 6.2
Comparisons of Radii of Curvature: *M* and *N*

Equation	$M = \dfrac{a(1-e^2)}{(1-e^2\sin^2\phi)^{3/2}}$	$N = \dfrac{a}{\sqrt{1-e^2\sin^2\phi^{3/2}}}$
	Results at 2 Latitudes	
Equator (ϕ - 0°)	$M = a(1-e^2)$	$N = a$
Pole (ϕ = 90°)	$M = \dfrac{a}{\sqrt{1-e^2}} = c$	$N = \dfrac{a}{\sqrt{1-e^2}} = c$

Normal Section Radius of Curvature

The normal section is defined as the intersection of a plane containing the ellipsoid normal and the ellipsoid surface. Given that the deflection-of-the-vertical is zero, the normal section lies in the plane shown by the vertical crosshair of a carefully leveled transit, theodolite, or total station. There are an infinite number of normal sections radiating from a point, and each one has a unique azimuth with respect to the meridian through that point.

Previously the values of *M* and *N* were given as radii of curvature in the north-south and east-west directions. The radius of curvature, R_α, in the plane of a normal section at any azimuth, α, is given by Euler's equation as

$$R_\alpha = \frac{MN}{M\sin^2\alpha + N\cos^2\alpha} \quad \text{or} \quad \frac{1}{R_\alpha} = \frac{\cos^2\alpha}{M} + \frac{\sin^2\alpha}{N} \tag{6.19}$$

Geometrical Mean Radius

Equation 6.19 is as good as it gets and can be used to compute the ellipsoid radius of curvature at a given latitude in a particular direction. But, in some cases, such as reducing horizontal distance to sea level, it is common to use an approximate spherical Earth radius of 6,372,000 m (or 20,906,000 feet). If, for whatever reason, a spherical radius is not good enough, then equation 6.19 could be used. But there are also times when a spherical radius is not good enough, and equation 6.19 represents overkill. A radius of intermediate accuracy is the geometrical mean radius, R_{mean}, which is computed for a given ellipsoid at a given latitude. The azimuth of the normal section at the point is not needed or used. Rapp (1991) calls R_{mean} the Gauss mean radius and derives it as

$$R_{mean} = \sqrt{MN} = \frac{a\sqrt{1-e^2}}{(1-e^2 \sin^2 \phi)} \tag{6.20}$$

ROTATIONAL ELLIPSOID

Equation of Ellipsoid

Given that the Greenwich meridian ellipse is rotated about its minor axis, the resulting figure is an ellipsoid used to approximate the size and shape of the Earth. In terms of the geocentric ECEF rectangular coordinate system, the equation of the ellipsoid is given by

$$\frac{X^2}{a^2} + \frac{Y^2}{a^2} + \frac{Z^2}{b^2} = 1 \tag{6.21}$$

Note that if $Z = 0$, equation 6.21 reduces to a circle in the plane of the equator, and, separately, if $Y = 0$, equation 6.21 reduces to an ellipse as given in equation 6.1. Geodetic latitude and longitude are 2-D coordinates used to represent location on the 3-D ellipsoid surface. Ellipsoid height, the third dimension, is used to specify the distance of a point above or below the ellipsoid.

Geocentric and Geodetic Coordinates

Geodetic coordinates of latitude, longitude, and height have been the computational standard for many years. However, with the advent of GPS technology, the use of the ECEF rectangular geocentric coordinates has much to offer with respect to modern technology and working with digital spatial data. This section first looks at the equations used to convert latitude, longitude, and height to ECEF geocentric *X/Y/Z* coordinates. Then, the inverse computation of starting with *X/Y/Z* and computing latitude, longitude, and height coordinates is considered. With reference to the 3-D GSDM diagram shown in Figure 1.4 and strictly as a matter of convenience, the transformation of latitude, longitude, and height to *X/Y/Z* coordinates is called a BK1 transformation, and transforming *X/Y/Z* coordinates to latitude, longitude, and height is called a BK2 transformation.

Figure 6.4 shows three views of the Earth. Figure 6.4a shows the meridian section quadrant, Figure 6.4b shows the plane of the equator, and Figure 6.4c shows a 3-D view of the ellipsoid.

The following symbols are used as appropriate in the three views.

X/Y/Z = geocentric ECEF rectangular coordinates—meters
φ and λ = geodetic latitude and longitude—sexagesimal units
h = ellipsoid height—meters
N = length of the ellipsoid normal—meters
P = projected distance of point from spin axis in equatorial plane

r = spatial distance from coordinate origin to point
a and b = parameters of underlying ellipsoid

Symbols used in the Vincenty BK2 transformation, but not appearing in Figure 6.4, include the following:

a' & b' = parameters of an auxiliary ellipsoid
h' = approximate ellipsoid height
φ' = approximate intermediate geodetic latitude
T & U = intermediate values used for computational convenience

BK1 TRANSFORMATION

A BK1 transformation uses equations 6.22, 6.23, and 6.24. The relationships for those equations are illustrated in Figure 6.4a and Figure 6.4b. The distance Z in Figure 6.4a is given by equation 6.25, but it is not obvious from the diagram that the factor $(1 - e^2)$ is responsible for removing that part of the normal lying below

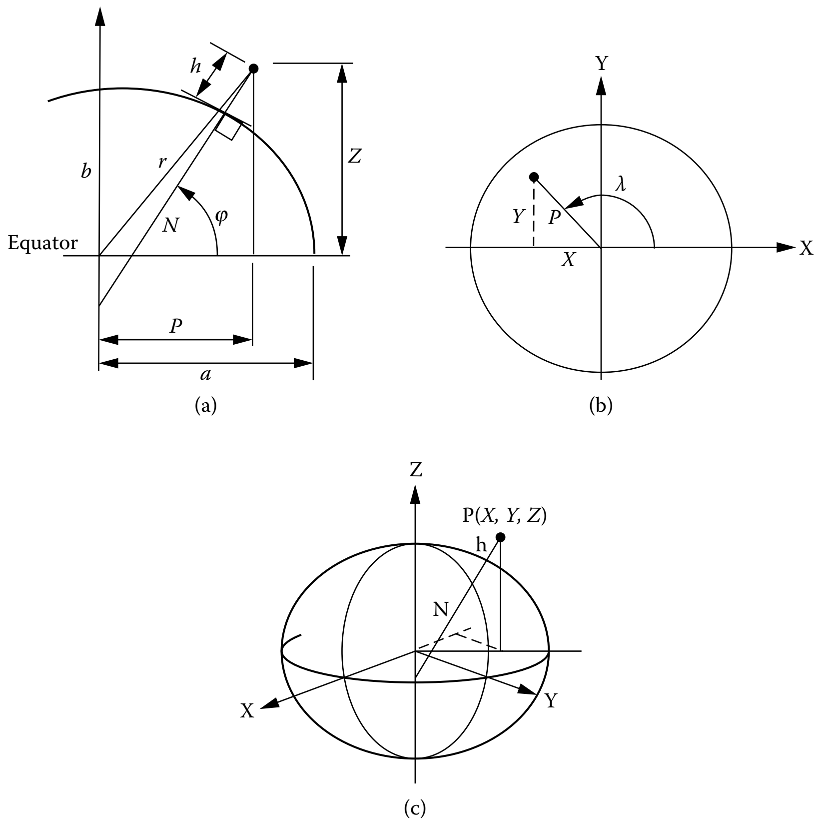


FIGURE 6.4 Geometry and Symbols Used in BK1 and BK2 Transformations

the equator. The equation for Z was derived previously and given as equation 6.15. Note too that equations 6.23, 6.24, and 6.25 give X, Y, and Z values for any point at ellipsoid height, h, but that equations 6.14 and 6.15 are specifically for X and Z on the ellipsoid surface in the Greenwich meridian.

$$P = (N + h)\cos\phi \tag{6.22}$$

$$X = P\cos\lambda = (N + h)\cos\phi\cos\lambda \tag{6.23}$$

$$Y = P\sin\lambda = (N + h)\cos\phi\sin\lambda \tag{6.24}$$

$$Z = [N(1 - e^2) + h]\sin\phi \tag{6.25}$$

These BK1 equations are used to compute the ECEF geocentric coordinates for any point whose position is defined by latitude, longitude, and height on a specified datum or ellipsoid—typically, NAD83 geodetic coordinates or some other datum such as the WGS84 or ITRF reference frame.

BK2 TRANSFORMATION

The BK2 transformation uses geocentric *X/Y/Z* coordinates to compute latitude, longitude, and height referenced to a named Earth-centered ellipsoid. Although a BK2 transformation is not as easy as a BK1 transformation, there are several ways it can be done. The longitude part of a BK2 computation is straightforward and, with reference to Figure 6.4b, uses equation 6.26 with due regard to the quadrant as described in equations 4.11, 4.12, and 4.13. Note, however, the sign convention for longitude in the plane of the equator is consistent with the math/science convention. As such, Y is in the numerator while X is in the denominator when using the inverse tangent function.

$$\tan\lambda = \frac{Y}{X} \tag{6.26}$$

Iteration

Possibly the best approach to a BK2 transformation is to iterate on the geodetic latitude. Once latitude is computed, the ellipsoid height computation is routine. Leick (2004) recommends using equation 6.27 for the iteration. An initial approximation for geodetic latitude is given by equation 6.28. Equation 6.29 is used to compute the ellipsoid normal for the first iteration. Equation 6.30 is used to compute the next latitude approximation based upon previous values of latitude and ellipsoid normal. Values of geodetic latitude and ellipsoid normal are computed repeatedly until the change in successive values is negligible. After the geodetic latitude is found, equation 6.32 can be used to compute the ellipsoid height. Given that the iteration is programmed into a computer; the effort expended in the solution is minimal. A long-hand iteration solution can become rather tedious.

$$\tan\phi = \frac{Z}{P}\left(1 + \frac{e^2 N \sin\phi}{Z}\right) \quad \text{where } P = \sqrt{X^2 + Y^2} \tag{6.27}$$

$$\phi_0 = \arctan\left(\frac{Z}{P(1-e^2)}\right) \tag{6.28}$$

$$N_0 = \frac{a}{\sqrt{1 - e^2 \sin^2 \phi_0}} \tag{6.29}$$

$$\phi_i = \arctan\left[\frac{Z}{P}\left(1 + \frac{e^2 N_{i-1} \sin\phi_{i-1}}{Z}\right)\right] \tag{6.30}$$

$$N_i = \frac{a}{\sqrt{1 - e^2 \sin^2 \phi_i}} \tag{6.31}$$

$$h = \frac{P}{\cos\phi} - N \tag{6.32}$$

Once-Through Vincenty Method

Another approach to the BK2 transformation is to use a "once-through" procedure devised by Vincenty (1980). In his method, the iteration is built in and once through the computation will give an answer tested to be within 0.2 mm for any point within the birdcage of orbiting GPS satellites. No attempt is made to explain all the terms, but it can be inferred from equations 6.37 and 6.38 that the solution is developed with respect to an auxiliary ellipsoid passing through the point at elevation *h*'.

$$b = a(1 - f) \tag{6.33}$$

$$P^2 = X^2 + Y^2 \text{ and } P = \sqrt{X^2 + Y^2} \tag{6.34}$$

$$r^2 = P^2 + Z^2 \text{ and } r = \sqrt{P^2 + Z^2} \tag{6.35}$$

$$h' = r - a + \frac{(a-b)Z^2}{r^2} \tag{6.36}$$

$$a' = a + h' \text{ and } b' = b + h' \quad \text{(6.37 and 6.38)}$$

$$\phi' = \arctan\left[\left(\frac{a'}{b'}\right)^2 \left(\frac{Z}{P}\right)\left(1 + \frac{e^4 h' a(Z^2 - P^2)}{4a'^4}\right)\right] \tag{6.39}$$

$$T = \frac{(P - h'\cos\phi')^2}{a^2} \text{ and } U = \frac{(Z - h'\sin\phi')^2}{b^2} \quad \text{(6.40 and 6.41)}$$

$$h = h' + \frac{1}{2}\left[\frac{T + U - 1}{T/a + U/b}\right] \tag{6.42}$$

$$\phi = \arctan\left[\left(\frac{a}{b}\right)^2 \left(\frac{Z - e^2 h \sin\phi'}{P}\right)\right] \tag{6.43}$$

$$\lambda = \arctan\left(\frac{Y}{X}\right) \tag{6.44}$$

Notes about the BK1 and BK2 transformations:

1. The BK1 and BK2 transformations are critical, and BK2 is probably the most difficult part of the GSDM. But, the rotation matrix in equation 1.29 requires the latitude and the longitude of the standpoint. They are determined from the geocentric *X/Y/Z* coordinates of the standpoint using the BK2 transformation.
2. The BK2 transformation is also described by others. See, for example, Soler and Hothem (1988); Hoffman-Wellenhof, Lichtenegger, and Collins (1992); Wolf and Ghilani (1997); You (2000); and Hooijberg (1997).
3. The integrity of any BK2 solution can be checked by using the computed latitude, longitude, and height in the BK1 equations. It should be possible to duplicate the *X/Y/Z* values used in the BK2 transformation. There is no approximation in the BK1 equations.
4. The accuracy of results may also depend upon the significant digit capacity of the computer being used or by the way the software is written.

EXAMPLE OF BK1 TRANSFORMATION

Ellipsoid: GRS 1980, a = 6,378,137.000 m, and e^2 = 0.006694380023.

Given: Station "Reilly," an "A" order high-accuracy reference network (HARN) station located on the campus of New Mexico State University in Las Cruces (NMSU). Note that equations presume positive latitude north of the equator and negative latitude south of the equator. Longitude east of Greenwich is used as a positive value. If west longitude is used as a negative number, the results will be the same.

Latitude: φ = 32° 16' 55."92906 N
Longitude: λ = 106° 45' 15."16070 W = 253° 14' 44."83930 E
Ellipsoid height: h = 1,166.57 meters

Compute:

$$N = \frac{a}{\sqrt{1 - e^2 \sin^2 \phi}} = 6{,}384{,}235.5313 \text{ m}$$

$$P = (N + h)\cos\phi = 5{,}398{,}396.2940 \text{ m}$$

$$X = P\cos\lambda = -1{,}556{,}177.6148 \text{ m}$$

$$Y = P\sin\lambda = -5{,}169{,}235.3185 \text{ m}$$

$$Z = [N(1 - e^2) + h]\sin\phi = 3{,}387{,}551.7093 \text{ m}$$

EXAMPLE OF BK2 TRANSFORMATION—ITERATION

Ellipsoid: GRS 1980, a = 6,378,137.000 m, and e^2 = 0.006694380023.

Given: Station "K 785," a first-order GPS station located on the campus of the Oregon Institute of Technology in Klamath Falls.

X = –2,490,977.042 m
Y = –4,019,738.192 m
Z = 4,267,460.404 m

Compute:

$$\lambda = \arctan\left(\frac{Y}{X}\right); \lambda = 238° \ 12' \ 50.''646096 \text{ E}$$

$$= 121° \ 47' \ 09.''353904 \text{ W}$$

$$P = \sqrt{X^2 + Y^2}; P = 4{,}728{,}981.0484 \text{ m}$$

TABLE 6.3
Summary of BK2 Iterations

Iteration	Latitude	Difference	Normal	Difference
0	42° 15' 17."133952	—	6,387,812.0556 m	
1	42° 15' 16."993923	–0."140029	6,387,812.0412 m	–0.0144 m
2	42° 15' 16."993408	–0."000515	6,387,812.0411 m	–0.0001 m
3	42° 15' 16."993405	–0."000003	6,387,812.0411 m	–0.0000 m

$$\phi_0 = \arctan\left(\frac{Z}{P(1-e^2)}\right);\ \varphi_0 = 42°\ 15'\ 17.''133952\ \text{N}$$

$$N_0 = \frac{a}{\sqrt{1-e^2\sin^2\phi_0}};\ N_0 = 6{,}387{,}812.0556\ \text{m}$$

$$\phi_1 = \arctan\left[\frac{Z}{P}\left(1+\frac{e^2 N_0 \sin\phi_0}{Z}\right)\right];\ \phi_1 = 42°\ 15'\ 16.''993923\ \text{N}$$

Values for subsequent iterations are shown in Table 6.3.

Once the geodetic latitude is found, the ellipsoid height is computed as

$$h = \frac{P}{\cos\phi} - N;\ h = 1{,}297.8797\ \text{m}$$

EXAMPLE OF BK2 TRANSFORMATION—VINCENTY'S METHOD (SAME POINT)

Ellipsoid: GRS 1980, a = 6,378,137.000 m, and e^2 = 0.006694380023.

Given: Station "K 785," a first-order GPS station located on the campus of the Oregon Institute of Technology in Klamath Falls.

X = –2,490,977.042 m
Y = –4,019,738.192 m
Z = 4,267,460.404 m

Compute:

$$\lambda = \arctan\left(\frac{Y}{X}\right);\ \lambda = 238°\ 12'\ 50.''646096\ \text{E}$$

$$= 121°\ 47'\ 09.''353904\ \text{W}$$

$$P^2 = X^2 + Y^2 \text{ and } P = \sqrt{X^2 + Y^2} \text{ ; } P = 4{,}728.981.0484 \text{ m}$$

$$r^2 = P^2 + Z^2 \text{ and } r = \sqrt{P^2 + Z^2} \text{ ; } r = 6{,}369{,}810.0486 \text{ m}$$

$$h' = r - a + \frac{(a-b)Z^2}{r^2} \text{ ; } h' = 1{,}271.2291 \text{ m}$$

$$a' = a + h' \text{ ; } a' = 6{,}379{,}408.2291 \text{ m}$$

$$b' = b + h' \text{ ; } b' = 6{,}358{,}023.5433 \text{ m}$$

$$\phi' = \arctan\left[\left(\frac{a'}{b'}\right)^2 \left(\frac{Z}{P}\right)\left(1 + \frac{e^4 h' a(Z^2 - P^2)}{4a^4}\right)\right] \text{ ; } \phi' = 42° \ 15' \ 16.''996289 \text{ N}$$

$$T = \frac{(P - h'\cos\phi')^2}{a^2} \text{ ; } T = 0.54950876168$$

$$U = \frac{(Z - h'\sin\phi')^2}{b^2} \text{ ; } U = 0.45049960785$$

$$h = h' + \frac{1}{2}\left(\frac{T + U - 1}{\frac{T}{a} + \frac{U}{b}}\right) \text{ ; } h = 1{,}297.8795 \text{ m}$$

$$\phi = \arctan\left[\left(\frac{a}{b}\right)^2 \left(\frac{Z - e^2 h \sin\phi'}{P}\right)\right] \text{ ; } \varphi = 42° \ 15' \ 16.''993406$$

The advantage of Vincenty's method is that one need go through it only once. Although his method is an approximation, representative tests (even for very large values of *h*) confirm agreement within 0.2 mm. If that is not precise enough, an iteration solution is recommended.

Meridian Arc Length

Throughout history, the size of the Earth has been determined by measuring a portion of a meridian arc and comparing that length to the corresponding angle subtended at the Earth's center (this procedure was described as "degree measurement" in chapter 5). If one knows any two of the three—arc length, central angle, or radius—the third is readily computed. Different values of arc-length-per-degree-of-latitude at various

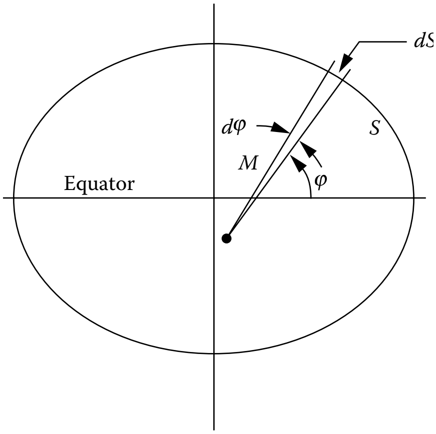


FIGURE 6.5 Meridian Arc Length

latitudes implied that the Earth is flattened at the poles. But, having selected an ellipsoid model for the Earth, differential geometry elements are integrated to compute meridian arc length or portions thereof; see Figure 6.5. Arc length of a differential element, *ds*, equals the instantaneous radius of curvature, *M*, times the differential increment in geodetic latitude, $d\varphi$.

$$dS = M\, d\varphi \tag{6.45}$$

S, the meridian arc distance from one latitude to another, is obtained by integrating equation 6.46 from one latitude to another. The meridian quadrant limits are 0° and 90° (in radian measure, 0 and π/2).

$$S_{\phi 1 \to \phi 2} = \int_{\phi 1}^{\phi 2} M d\phi \tag{6.46}$$

The expression for *M*, equation 6.18, is substituted into equation 6.46, and the constant portion moved outside the integral to get

$$S_{\phi 1 \to \phi 2} = a(1 - e^2) \int_{\phi 1}^{\phi 2} (1 - e^2 \sin^2 \phi)^{-3/2} d\phi \tag{6.47}$$

Equation 6.47 contains an elliptical integral and cannot be integrated in closed form. That is, the expression inside the integral must be expressed in a series expansion containing ever smaller terms that can be integrated individually. A solution is obtained by including all those terms that make a difference in the answer to the accuracy desired. Either a binomial series expansion or the MacLaurin series can be used with identical results (Rapp 1991). The result, equation 6.48, can be used within any specified limits. Computational convention is to start at the southerly of two latitudes and end at the northerly limit. If the convention is switched, the result will be a negative arc distance. Latitude south of the equator should be used as a negative value.

The meridian arc distance (in meters, given that *a* is in meters) between latitude limits selected by the user is

$$S_{\phi_1 \to \phi_2} = a(1-e^2)[A(\phi_2 - \phi_1) - \frac{B}{2}(\sin 2\phi_2 - \sin 2\phi_1) +$$

$$\frac{C}{4}(\sin 4\phi_2 - \sin 4\phi_1) - \frac{D}{6}(\sin 6\phi_2 - \sin 6\phi_1) +$$

$$\frac{E}{8}(\sin 8\phi_2 - \sin 8\phi_1) - \frac{F}{10}(\sin 10\phi_2 - \sin 10\phi_1)] \quad (6.48)$$

where the coefficients A through F are as given in equations 6.49 to 6.54.

$$A = 1 + \frac{3}{4}e^2 + \frac{45}{64}e^4 + \frac{175}{256}e^6 + \frac{11,025}{16,384}e^8 + \frac{43,659}{65,536}e^{10} \quad (6.49)$$

$$B = \frac{3}{4}e^2 + \frac{15}{16}e^4 + \frac{525}{512}e^6 + \frac{2,205}{2,048}e^8 + \frac{72,765}{65,536}e^{10} \quad (6.50)$$

$$C = \frac{15}{64}e^4 + \frac{105}{256}e^6 + \frac{2,205}{4,096}e^8 + \frac{10,395}{16,384}e^{10} \quad (6.51)$$

$$D = \frac{35}{512}e^6 + \frac{315}{2,048}e^8 + \frac{31,185}{131,072}e^{10} \quad (6.52)$$

$$E = \frac{315}{16,384}e^8 + \frac{3,465}{65,536}e^{10} \quad (6.53)$$

$$F = \frac{693}{131,072}e^{10} \quad (6.54)$$

Notes for equation 6.48:

1. The terms within the brackets are all unitless. That means, for example, that the latitude difference given by $\varphi_2 - \varphi_1$ must be in radians. The trigonometric ratios in the other terms are already unitless.
2. The coefficients need to be computed only once for each ellipsoid if they are stored (see Table 6.4). Depending upon the computational environment, it may be more efficient to compute and store them or to compute them as needed.
3. The meridian quadrant is computed by choosing limits of 0° and 90° (0 and 2π in radians). In that case, equation 6.48 reduces to equation 6.55 because the trigonometric sines of the multiple angles are all zero. The meridian

TABLE 6.4
Meridian Coefficients and Quadrant Arc Length for Selected Ellipsoids

	Clarke 1866	GRS 1980	WGS 1984
f	0.00339007530393	0.0033528106812	0.0033528106647
e^2	0.00676865799729	0.0066943800229	0.0066943799901
e^4	0.00004581473108	0.0000448147239	0.0000448147234
e^6	0.00000031010425	0.0000003000068	0.0000003000068
e^8	0.00000000209899	0.0000000020084	0.0000000020084
e^{10}	0.00000000001421	0.0000000000134	0.0000000000134
A	1.00510892038799	1.0050525018131	1.0050525017882
B	0.00511976506202	0.0050631086222	0.0050631085972
C	0.00001086615776	0.0000106275903	0.0000106275902
D	2.1524755323E-08	2.082037857E-08	2.082037826E-08
E	4.1106495949E-11	3.932371371E-11	3.932371294E-11
F	7.5116641593E-14	7.108453403E-14	7.108453229E-14
Quadrant arc	10,001,888.0430 m	10,001,965.7292 m	10,001,965.7293 m

quadrant arc length for any Earth-model ellipsoid should be approximately 10,000,000 meters.

$$S_{0^\circ \to 90^\circ} = a(1-e^2)[A(\pi/2) \tag{6.55}$$

Length of a Parallel

A parallel of latitude on the ellipsoid describes a circle whose plane is parallel to the equatorial plane and whose radius is $N \cos \varphi$. If the Earth were spherical with radius r, the radius of a parallel of latitude would be $r \cos \varphi$. Since each parallel is a circle, its circumference is simply $2\pi N \cos \varphi$. Partial length of a parallel is computed as the proportionate part of the total circumference, or, as illustrated in Figure 6.6, arc length is computed directly using $L = R\theta$ with the longitude difference expressed in radians:

$$L_p = (\lambda_2 - \lambda_1) N \cos\phi = \frac{\Delta\lambda \; a\cos\phi}{\sqrt{1-e^2 \sin^2 \phi}} \tag{6.56}$$

Surface Area of a Sphere

Surface area in a plane is length times width. Area on a curved surface can also be computed, but care must be taken because the distances are no longer "flat." Area of the uniformly curved surface of a sphere is computed using tools of differential

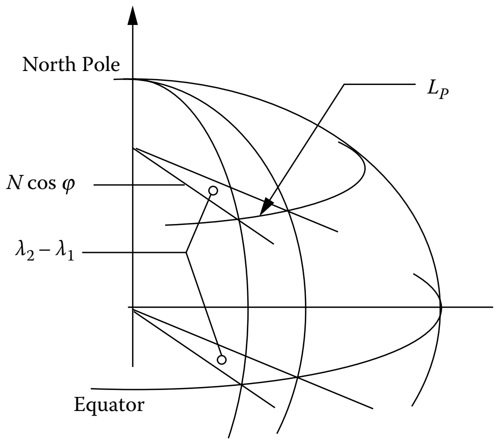


FIGURE 6.6 Length of a Parallel

geometry and integral calculus, as shown in Figure 6.7. For illustration purposes, the sphere is cut into two equal pieces nominally called northern and southern hemispheres. The equator is the dividing plane and, using equation 6.56, is a circle having a circumference of $2\pi R$ ($\Delta\lambda = 2\pi$, $N = R$, and $\cos 0° = 1$). The circumference of the sphere is taken to be the length of a plane rectangle. The width of the rectangle is an infinitesimally small differential element in the north-south direction. Referring to Figure 6.7, the differential north-south distance is $R\, d\varphi$. If the small ring around the sphere is then cut and rolled out flat, the differential area of that ring is computed as

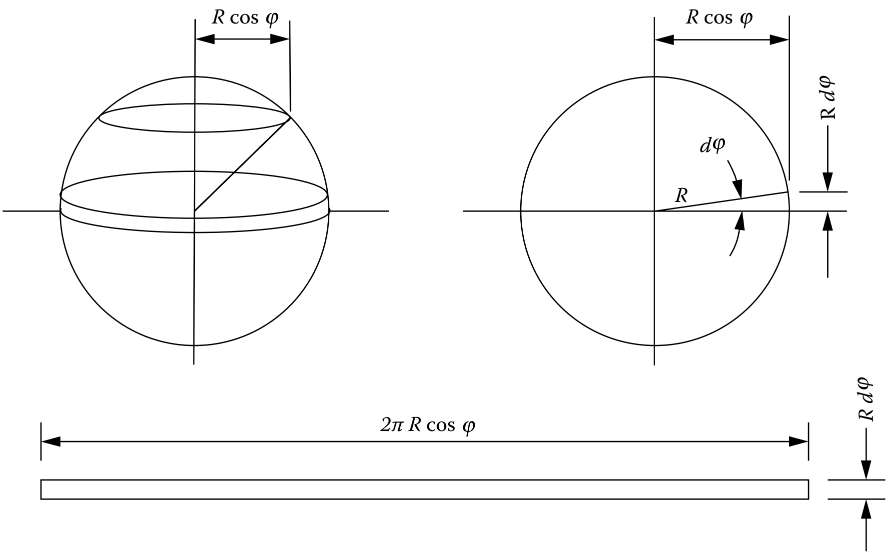


FIGURE 6.7 Surface Area of a Sphere

length ($2\pi R \cos \varphi$) times width ($R\, d\varphi$). The surface area of the entire sphere is found by adding up an infinite number of infinitely thin adjacent rings (integration). The surface area of the entire sphere is found by integrating equation 6.58 from the South Pole to the North Pole, ($\varphi_1 = -90°$ and $\varphi_2 = +90°$).

$$dA = (2\pi \ R\cos\phi) * (R\ d\phi) \quad \text{; surface area of thin ring} \tag{6.57}$$

$$A = \int_{\phi_1}^{\phi_2} 2\pi\ R^2 \cos\phi\ \ d\phi \ = \ 2\pi\ R^2[\sin\phi_2 - \sin\phi_1] \tag{6.58}$$

$$\text{Surface area of entire sphere} = 4\pi R^2 \tag{6.59}$$

Ellipsoid Surface Area

Surface area on the ellipsoid is computed much the same way as surface area on a sphere. A differential surface area element on the ellipsoid is written (see equation 6.60) as the product of an elemental parallel distance and an elemental meridian distance, as shown in Figure 6.8. A double integration is used to compute first the area of a ring using longitude limits of 0 and 2π radians, then the ring areas between latitude limits selected by the user are computed and accumulated using equation 6.62. Equation 6.63 can be used to compute the area of any rectangular block on the ellipsoid surface bounded by parallels and meridians as selected by the user. If limits of 0 to 2π for longitude and limits of –90° to + 90° for latitude are used in equation 6.63, the result (omitting much algebraic manipulation) is equation 6.64, which is used to compute total ellipsoid surface area.

$$dA = (N\cos\phi\ \ d\lambda)\ \ (M\ \ d\phi)\ = \frac{a^2(1-e^2)\cos\phi}{(1-e^2\sin^2\phi)}\ \ d\phi\ \ d\lambda \tag{6.60}$$

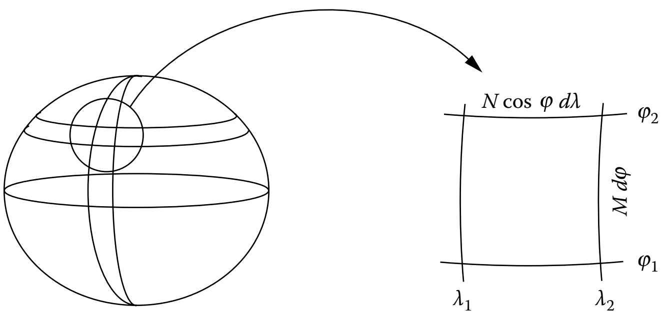


FIGURE 6.8 Ellipsoid Surface Area

$$A = a(1-e^2)\int_{\lambda_1}^{\lambda_2}\int_{\phi_1}^{\phi_2} \frac{\cos\phi}{(1-e^2\sin^2\phi)}\, d\phi\, d\lambda \tag{6.61}$$

$$A = (\lambda_2-\lambda_1)a^2(1-e^2)\int_{\phi_1}^{\phi_2} \frac{\cos\phi}{(1-e^2\sin^2\phi)}\, d\phi \tag{6.62}$$

$$A = \frac{(\lambda_2-\lambda_1)a^2(1-e^2)}{2} \quad *$$

$$\left[\frac{\sin\phi_2}{1-e^2\sin^2\phi_2} - \frac{\sin\phi_1}{1-e^2\sin^2\phi_1} + \frac{1}{2e}\ln\left(\frac{1+e\sin\phi_2}{1-e\sin\phi_2}\right) - \frac{1}{2e}\ln\left(\frac{1+e\sin\phi_1}{1-e\sin\phi_1}\right)\right] \tag{6.63}$$

$$A_{ellipsoid} = 2\pi\, a^2(1-e^2)\left[\frac{1}{1-e^2} + \frac{1}{2e}\ln\left(\frac{1+e}{1-e)}\right)\right] \tag{6.64}$$

An interesting side note is that the equation for the surface area of a sphere should be identical to that of the surface area of an ellipsoid whose eccentricity is zero ($a = R$). If one attempts to insert $e = 0$ into equation 6.64, an impasse is reached when working with the second term within the brackets. First, it is never permissible to divide by zero. As e goes to 0, $1/2e$ becomes infinitely large. The next term involves taking the natural log of a term that goes to 1 as e goes to 0. The second part of the second term goes to zero if one takes the natural log of 1. The interesting part is that l'Hopital's rule can be used to compare the rates of each part of the second term as e goes to zero. The ratio reduces to 1 over 1. Since the first term in the brackets goes to one, one plus one is two (for terms within the brackets). Two times the first part of equation 6.64 gives an identical expression as equation 6.59 for a sphere for $e = 0$.

THE GEODETIC LINE

Description

A geodetic line, also known the geodesic, is defined as the shortest distance between two points on the surface of the ellipsoid. The geodetic line on the ellipsoid is analogous to a great circle on a sphere. When drawn on a Mercator projection, as shown in Figure 6.9, a geodetic line appears to be curved even though on the ellipsoid it bends neither to the right nor to the left.

If one were to start at any point on the equator and travel a geodetic line to a point on the opposite side of the world, the path would cross either the north pole or the south pole before ending up at the antipole (also on the equator). If one were to leave a point on the equator with a beginning azimuth of, say, 00° 01', and travel a geodetic line, the path would miss the North Pole and the final destination on the opposite side of the world would be between the antipole and the "liftoff point." Several comments about the geodetic line:

1. Beginning with an initial azimuth of 00° 01' on the equator, the geodetic line will miss the North Pole by less than 2 kilometers. Increase the azimuth at the equator to 1°, and the geodetic line misses the pole by something over 111 kilometers.
2. In all cases, there is some point along a geodetic line at which the distance to the pole is a minimum. At that point, the latitude is a maximum and the azimuth of the geodetic line is 90° (see Figure 6.9).
3. The geodetic line has no lateral curvature, but the underlying meridians are not parallel. Therefore, it should be apparent that the azimuth of the geodetic line changes continuously as it traverses the globe. One disadvantage of using a Mercator projection of the world to illustrate the behavior of the geodetic line is that meridians on a Mercator projection are parallel, meaning the geodetic line (which does not curve) must be shown as a curved line.
4. If one starts on the equator and travels east or west along the equator, the path of travel is a geodetic line only until reaching the liftoff point. Beyond that, the equator is not a geodetic line. No other parallel of latitude is a geodetic line.
5. Every point on the ellipsoid has an infinite number of geodetic lines going through it. The azimuth of a geodetic line through a point can have any value between 0° and 360°.

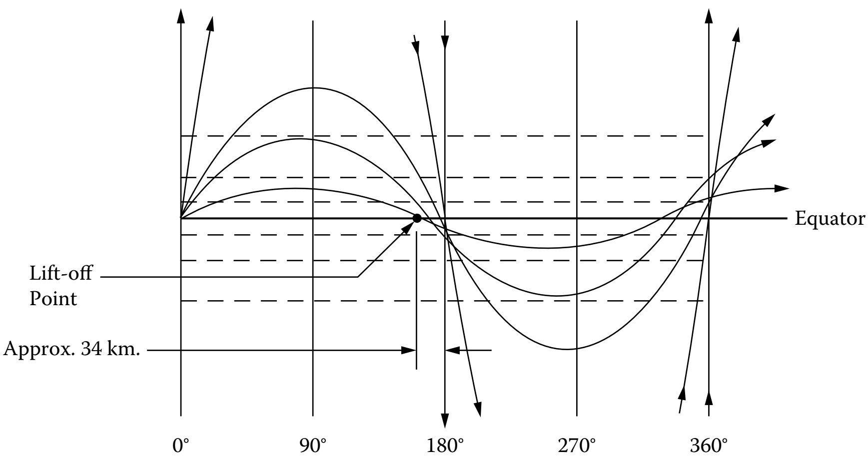


FIGURE 6.9 Geodetic Lines around the Earth

Clairaut's Constant

An important feature of a geodetic line is that each different line has its own unique number. The number, known as Clairaut's constant, is defined as

$$K = N_1 \cos\phi_1 \sin\alpha_1 = N_2 \cos\phi_2 \sin\alpha_2 \tag{6.65}$$

where

N is the ellipsoid normal,
φ is geodetic latitude of the point, and
α is the azimuth of the geodetic line at the point.

Note that the value of Clairaut's constant will be negative for geodetic line azimuths between 180° and 360°. Also note that while each geodetic line has its own unique constant, the value of Clairaut's constant does not change as one travels a parallel of latitude. Hence, an unchanging value of Clairaut's constant is not an exclusive property of a geodetic line.

But, a useful feature of Clairaut's constant is that one can use it to determine the azimuth of a geodetic line at any latitude and, subsequently, the convergence between points. For example, if the azimuth of a line on the GRS80 ellipsoid is 91° 16' 28" at point C in Figure 6.10 (latitude = 38° 48' 05."3342 N), what is the azimuth of the same line at point D (latitude = 38° 12' 22."5464 N)? Relating the problem to Figure 6.10 is important because the inverse sine function has two answers. The first answer (typically given by a calculator or computer) will be less than 90°. That would be a correct answer at point A (at the same latitude as point D), as shown in Figure 6.10, but the second correct answer is the supplement (180° – x) of the first.

Solution: first compute Clairaut's constant using equation 6.65. Then rewrite equation 6.65 to solve for sin α_2, as shown in equation 6.66. The inverse sin function will provide two legitimate answers—each is correct at some point on the geodetic line. The user is expected to choose the correct answer of the two.

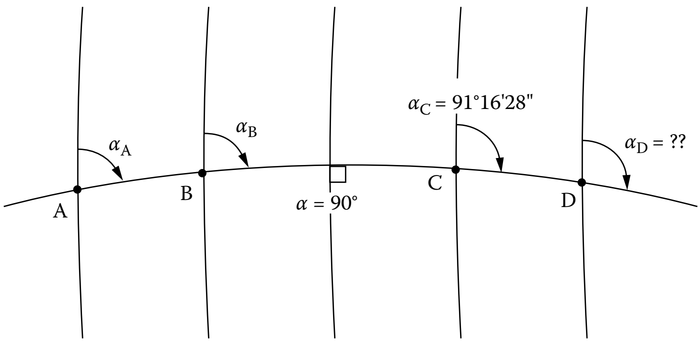


FIGURE 6.10 Geodetic Line Azimuth Using Clairaut's Constant

$$K = \frac{6,378,137.0\cos\ (38° \quad 48'\ 05."3342)\sin\ (91° \quad 16'\ 28")}{\sqrt{1 - 0.006694380023\sin^2(38° \quad 48'\ 05."33423)}} = 4,975,935.539$$

$$\sin\alpha_2 = \frac{K}{N_2\cos\phi_2} = \frac{K\sqrt{1-e^2\sin^2\phi_2}}{a\cos\phi_2} \tag{6.66}$$

$$\sin\alpha_2 = \sin\alpha_A = \sin\alpha_D = \frac{4,969,560.178}{5,011,871.754} = 0.9915577296491$$

$$\alpha_A = 82° \quad 32'\,59" \quad \text{and} \quad \alpha_D = \quad 97° \quad 27'\,01"$$

Convergence of the meridians is defined as the difference in azimuth between two points on the same geodetic line.

$$Conv_{1\to2} \equiv \alpha_2 - \alpha_1 \tag{6.67}$$

In the current example, the convergence from *C* to *D* is

$$Conv_{C\to D} = \alpha_D - \alpha_C = 97° \quad 27'\ 01" \ - \ 91° \quad 16'\ 28" \ = \ 006° \quad 10'\ 33"$$

Knowing that a geodetic line azimuth is 90° (sin 90° = 1) at the maximum latitude, it is also possible to use equation 6.65 to determine the maximum latitude reached by a geodetic line. The solution involves considerable algebraic manipulation, but equation 6.68 is derived by using 1.0 for sin α_2.

$$N_{\max}\cos\alpha_{max}\sin 90° \quad = K$$

$$\cos^2\phi_{\max} = \frac{K^2\left(1-e^2\right)}{a^2 - K^2e^2} \tag{6.68}$$

Finally, the azimuth at which a geodetic line crosses the equator is found from equation 6.65 by using $\varphi_2 = 0°$, which gives

$$\sin\alpha_{eq} = \frac{K}{a} \tag{6.69}$$

Geodetic Azimuths

An azimuth is the angle a line on the surface of the Earth makes with the meridian through the same point. The geodetic azimuth on the ellipsoid is the azimuth of the geodetic line, but, when working with 3-D spatial data, it is often more convenient to work with the 3-D azimuth. The two are very nearly identical and, except for very precise applications, can be used interchangeably. Various azimuths are summarized here, but a detailed analysis of the differences is given in Burkholder (1997):

1. If the angle is measured in the horizontal plane defined as perpendicular to the local plumb line, a Laplace correction is required to obtain an equivalent angle in the tangent plane to the ellipsoid at that point. As discussed in chapter 8, if the deflection-of-the-vertical is zero or insignificant, the Laplace correction need not be applied.
2. If the angle is referenced to the physical spin axis of the Earth instead of the adopted mean geodetic position of the North Pole, a correction for polar motion is required. The polar motion correction is quite small and beyond the scope of this text. Additional information on polar motion can be found in texts such as Bomford (1971), Vaníček and Krakiwsky (1986), or Leick (2004), or on an appropriate web site.
3. When looking through the telescope of a carefully leveled surveying instrument to a target at the other end of the line, there is a common line in 3-D space between point A and point B. A normal section is a line on the ellipsoid from point to point formed by the intersection of a plane containing the normal at the standpoint and the target at the forepoint. Interestingly enough, the normal section from point A to point B on the ellipsoid is not the same as the normal section from point B to point A because the directions of the ellipsoid normals at different latitudes are not parallel. The spatial vector between telescope and target (each way) is common to both planes. But, when the vertical plane at each end of the line is projected to the ellipsoid, the line (normal section) on the ellipsoid from point A to point B will be slightly different than the line from point B to point A. The difference between normal sections is exaggerated and shown in Figure 6.11.

   But, a geodetic line is defined as the shortest distance between two points on the ellipsoid surface. Several observations are: first, the geodetic line has no lateral curvature but curves only in the plane containing the

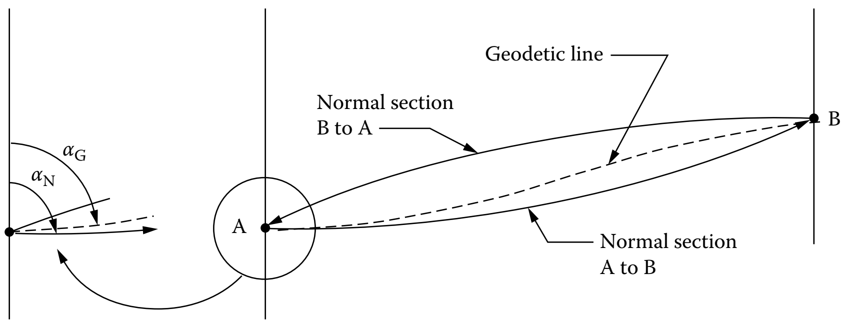


FIGURE 6.11 Normal Sections and the Geodetic Line

instantaneous normal section. Figure 6.11 gives the mistaken impression that a geodetic line has a double curvature. Second, the underlying meridians are not parallel on a globe, but, on a Mercator map, the meridians appear as parallel lines. This explanation for the wrong impression is that a three-dimensional phenomenon is portrayed on a two-dimensional diagram. The important point is that the geodetic line is one line between points on the ellipsoid surface and that the normal sections between points are slightly different depending on whether they are going from point A to point B or from point B to point A.

4. If the angle is measured to a target some distance above or below the ellipsoid, a target height correction may be required. As shown in Figure 6.12, the target height correction is required because the normal through the target (forepoint) is not parallel with the normal through the standpoint (instrument station). The elevation of the theodolite or total station above or below the ellipsoid is immaterial because the vertical axis of instrument is coincident with the vertex of the dihedral angle being measured.
5. The 3-D azimuth (Burkholder 1997) lies in the tangent plane at the standpoint and is computed as the inverse tangent of ($\Delta e/\Delta n$), the local geodetic horizon components of a 3-D vector defined by $\Delta X/\Delta Y/\Delta Z$.

In summary, the following three azimuths are all very close to being identical and are often used interchangeably as a geodetic azimuth. The size of each correction can be used to decide whether the difference is significant or not.

- The geodetic line azimuth is the traditional standard as it is used in geodetic computations on the ellipsoid surface.
- The normal section azimuth is the azimuth of the line on the ellipsoid from standpoint to forepoint (on the ellipsoid) as observed from the standpoint.

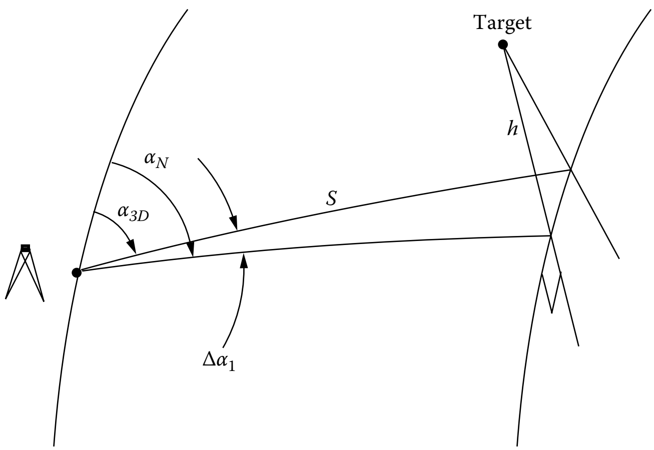


FIGURE 6.12 Target Height Correction

- The 3-D azimuth is the spatial direction from the standpoint to the forepoint projected into the tangent plane at the standpoint. The 3-D azimuth is readily obtained from GPS data and is the azimuth utilized in the GSDM.

There are two corrections that relate these three azimuths to each other. One is the target height correction. The other is called the geodesic correction from the normal section to the geodetic line. In many cases, the 3-D azimuth is the one routinely used. If and when a true geodetic line azimuth is required, the logical sequence would be to make the target height correction to obtain the normal section azimuth from the 3-D azimuth, then the geodesic correction is applied to get the geodetic line azimuth from the normal section azimuth.

Target Height Correction

The azimuth of the normal section is computed from the 3-D azimuth by adding the target height correction, as shown in Figure 6.12.

$$\alpha_N = \alpha_{3D} + \Delta\alpha_1 \tag{6.70}$$

$$\Delta\alpha_1 = \frac{\rho\, h\, e^2 \cos^2 \phi_1}{2N_1(1-e^2)}\left(\sin 2\alpha_{3D} - \frac{S}{N_1}\sin\alpha_{3D}\tan\phi_1\right) \tag{6.71}$$

where

ρ = 206,264.806247096355156 seconds per radian,
h = height of target in meters above or below the ellipsoid,
e^2 = eccentricity squared of the ellipsoid,
φ_1 = geodetic latitude of the standpoint,
N_1 = length of the ellipsoid normal in meters at the standpoint,
S = distance from the standpoint to the forepoint in meters, and
α_{3D} = 3-D azimuth from the standpoint to the forepoint.

Notes related to using the target height correction:

1. Rarely will the target height correction be greater than 0.5 arc seconds.
2. The correction is always added (subtraction is adding a negative number). Whether the correction is positive or negative is determined by sin ($2\alpha_{3D}$).
3. If using equation 6.70 backwards to compute the 3-D azimuth from the azimuth of the normal section, the normal section azimuth can be used in computing the correction instead of the 3-D azimuth.

Geodesic Correction

Given the azimuth of a normal section, the azimuth of a geodetic line is found by adding a correction, as shown in equation 6.72:

$$\alpha_g = \alpha_N + \Delta\alpha_2 \quad (6.72)$$

$$\Delta\alpha_2 = -\frac{\rho\, e^2 S^2}{12 N_1^2} \cos^2 \phi_m \sin 2\alpha_N \quad (6.73)$$

where

ρ = 206,264.806247096355 seconds per radian,
e^2 = eccentricity squared of the ellipsoid,
S = distance from standpoint to forepoint,
N_1 = ellipsoid normal at standpoint,
φ_m = mean latitude between standpoint and forepoint, and
α_N = azimuth of normal section from standpoint to forepoint.

Notes regarding use of the geodesic correction:

1. The magnitude of the geodesic correction is quite small and can be ignored in most cases. For example, regardless of standpoint location or azimuth of line, the geodesic correction will never exceed 0.0003 seconds of arc on a 10 kilometer line or 0.03 seconds of arc on a 100 kilometer line.
2. As written, the geodesic correction is a negative value. Depending upon the azimuth of the normal section, the correction may be positive or negative.
3. If equation 6.72 is rewritten to find the azimuth of the normal section given the azimuth of the geodetic line, it is permissible to use the azimuth of the geodetic line in equation 6.73 instead of the normal section azimuth.

GEODETIC POSITION COMPUTATION: FORWARD AND INVERSE

Geometrical geodesy relationships are used extensively when computing geodetic traverses and inverses on the ellipsoid. A two-dimensional (2-D) geodetic traverse is typically called the geodetic forward (or direct) computation, and computing the 2-D direction and distance between points is called the geodetic inverse. In order to avoid confusion with other similarly named computations (such as state plane coordinate forward and inverse transformations), the 2-D geodetic forward is henceforth referred to as a BK18 computation, and the 2-D geodetic inverse is called a BK19 computation. Other BK-designated computations are listed in Table 1.1 in the description of the GSDM.

PUISSANT FORWARD (BK18)

The geodetic forward (or BK18) computation computes the latitude and longitude of an unknown point based upon the known latitude and longitude of the beginning point and measured (or given) direction and distance from the known point to the unknown point. The direction is the geodetic line azimuth at the beginning point, and the distance is the geodetic line distance along the ellipsoid surface. Due to ellipsoid flattening, there is no closed-form equation that can be used without some approximation. Over the years, geodesists have devised numerous methods whereby the effect of required approximations is minimized and BK18 computations are performed according to specific procedures. One of the most popular procedures is the Puissant method, which is quite accurate for lines up to about 60 miles in length. The Puissant method is illustrated in Figure 6.13 and summarized as

$$\phi_2 = \phi_1 + \Delta\phi \tag{6.74}$$

$$\Delta\phi'' = SB\cos\alpha_1 - S^2C\sin^2\alpha_1 - D(\Delta\phi'')^2 - hS^2E\sin^2\alpha_1 \tag{6.75}$$

where

$B = \dfrac{\rho}{M_1}$ seconds per meter,

$h = SB\cos\alpha_1$ seconds,

$C = \dfrac{\rho\tan\phi_1}{2M_1N_1}$ seconds per meter^2,

$D = \dfrac{3e^2\sin\phi_1\cos\phi_1}{2\rho(1-e^2\sin^2\phi_1)}$ per second, and

$E = \dfrac{(1+3\tan^2\phi_1)(1-e^2\sin^2\phi_1)}{6a^2}$ per meter.

The constants h, B, C, D, and E are computed using ρ = 206,264.8062470964 seconds per radian, S = the ellipsoidal distance, α_1 = the geodetic line azimuth, M = the radius of curvature in the meridian (see equation 6.18), N = the ellipsoid normal (see equation 6.16), a = the ellipsoid semimajor axis, and e^2 = the eccentricity squared. Note that Δφ" appears on both sides of equation 6.75, requiring an iterative solution. Use Δφ" = 0 for the first iteration, and pay close attention to the units of each term.

The longitude of point 2 is computed as

$$\lambda_2 = \lambda_1 + \Delta\lambda \quad \text{(east longitude)} \tag{6.76}$$

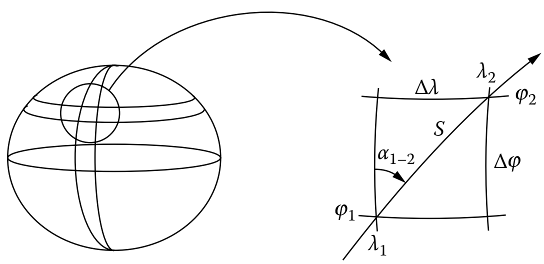


FIGURE 6.13 Geodetic Forward (BK18) and Inverse (BK19) Computations

$$\Delta\lambda'' = \frac{\rho\, S \sin\alpha_1}{N_2 \cos\phi_2} = \frac{\rho\, S \sin\alpha_1 \sqrt{1 - e^2 \sin^2\phi_2}}{a \cos\phi_2} \tag{6.77}$$

Note that the latitude of point 2 must be computed before the longitude at point 2 because φ_2 is used in equation 6.77. The azimuth from point 2 to point 1 (the back azimuth) can be computed using Clairaut's constant, but the Puissant method uses the following solution based upon convergence, $\Delta\alpha$:

$$\alpha_{2\to1} = \alpha_1 + \Delta\alpha + 180^\circ \tag{6.78}$$

$$\Delta\alpha'' = \frac{\Delta\lambda'' \sin\phi_m}{\cos\left(\dfrac{\Delta\phi}{2}\right)} + \frac{(\Delta\lambda'')^3 \sin\phi_m \cos^2\phi_m}{12\rho^2}; \text{ where } \phi_m = \frac{\phi_1 + \phi_2}{2} \tag{6.79}$$

Puissant Inverse (BK19)

The geodetic inverse, BK19, begins with the latitude and longitude of two points and computes the geodetic direction and distance between them. The Puissant method for BK19 uses the same *B*, *C*, *D*, and *E* coefficients as defined in BK18 and computes intermediate *x* and *y* values from which direction and distance are computed.

$$\Delta\phi = \phi_2 - \phi_1 \tag{6.80}$$

$$\Delta\lambda = \lambda_2 - \lambda_1 \quad \text{(east longitude)} \tag{6.81}$$

$$x = \frac{\Delta\lambda'' N_2 \cos\phi_2}{\rho} = S \sin\alpha_1 \tag{6.82}$$

$$y = \frac{1}{B}\left[\Delta\phi'' + C\, x^2 + D\, (\Delta\phi'')^2 + E\, (\Delta\phi'')\, x^2\right] = S \cos\alpha_1 \tag{6.83}$$

$$\alpha_1 \text{ (azimuth from north)} = \arctan\left(\frac{S \sin\alpha_1}{S \cos\alpha_1}\right) = \arctan\frac{y}{x} \tag{6.84}$$

$$S \text{ (distance)} = \sqrt{x^2 + y^2}\ ; \text{ in meters} \tag{6.85}$$

Advantages of the Puissant method include (1) it is quite accurate for distances up to about 100 kilometers (62 miles), and (2) it is a well-defined procedure that can be done by hand or by computer. Disadvantages of the Puissant method are (1) the uncertainty of really knowing how accurate it is, (2) it is not obvious how various pieces fit in the solution, and (3) distance wise, it is limited in scope. Other traditional methods have similar advantages and disadvantages.

Numerical Integration

Jank and Kivioja (1980) published a numerical integration procedure for BK18 and BK19, which is summarized here and recommended for use as appropriate. The method is quite tedious if used for longhand computations, but, by user choice, the results can be as accurate as desired over any length of line. If the procedure is programmed, even on a small handheld computer, the tediousness objection becomes moot. The Jank-Kivioja method utilizes differential geometry relationships for ellipsoidal triangles that are small enough to be treated as plane triangles. The key to preserving computational accuracy is maintaining the correct azimuth for each geodetic line element. That is done by using Clairaut's constant.

BK18 by Integration

Figure 6.14 shows a diagram of the BK18 computation in which the computation begins at point 1 and ends at point 2. The overall geodetic distance is broken into as many increments as needed to preserve computational accuracy at the endpoint. Clairaut's constant is used to update the geodetic line azimuth at the midpoint of each computational element. Conceptually, the process begins by getting on the geodetic line at point 1, moving up to the approximate midpoint of the element, using Clairaut's constant to find an average azimuth for the entire element, computing latitude and longitude increments based upon midpoint values, adding the latitude and longitude increments to the beginning element values to get latitude and longitude at the end of the element, using Clairaut's constant to update the azimuth at the element endpoint, then using the ending values of one element as beginning values for the next element, and repeating for the number of elements chosen by the user. Special considerations are required to compute across the meridian through the point of maximum latitude.

The following equations are written from the diagram in Figure 6.14, using the assumption that each element is small enough to be treated as a plane triangle.

$$M_m(\Delta\phi) = \Delta S \cos\alpha_m \Rightarrow \Delta\phi = \frac{\Delta S \cos\alpha_m}{M_m} \tag{6.86}$$

$$N_m \cos\phi_m \Delta\lambda = \Delta S \sin\phi_m \Rightarrow \Delta\lambda = \frac{\Delta S \sin\alpha_m}{N_m \cos\phi_m} \tag{6.87}$$

Steps for performing a geodetic line BK18 numerical integration are as follows:

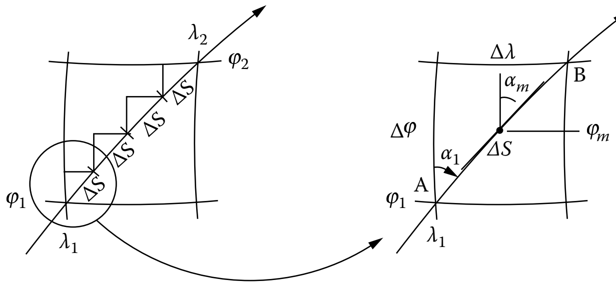


FIGURE 6.14 Geodetic Line Numerical Integration

1. $\Delta S = S$ divided by the number of increments chosen by the user.
2. Compute values of M and N at point 1 (point A for first element; see equations 6.18 and 6.16).
3. Compute Clairaut's constant for the line (see equation 6.65).
4. Compute the approximate change in latitude for the element using equation 6.86 and the azimuth, radius of curvature M, and latitude at the *beginning* of the element.
5. Find the latitude of the element midpoint by adding half of approximate $\Delta\varphi$ to the latitude of the beginning of the element.
6. Compute M and N at the element midpoint.
7. Use Clairaut's constant to update the geodetic line azimuth at midpoint.
8. Use equations 6.86 and 6.87 to compute latitude and longitude increments.
9. Add latitude and longitude elements to the latitude and longitude of the beginning of the element.
10. Compute M_m and N_n at the element midpoint.
11. Use Clairaut's constant to update the azimuth at the endpoint.
12. Use endpoint values as beginning values for the next element.
13. Repeat for the number of elements chosen by the user.

Jank and Kivioja (1980) claim that millimeter accuracy can be realized for 20,000 km lines (halfway around the world) if elements are limited to a length of 200 meters. For shorter lines, longer elements can be used while maintaining millimeter accuracy. Practically, increasing the number of increments and noting any change in the final computed position can be used to check the accuracy of a computed position for any line length. For example, geodetic latitude and longitude to five decimal places of seconds correspond to about 0.0003 meters on the ellipsoid and are often used as a computational standard. If increasing (say, doubling) the number of increments gives the same answer to five decimal places of seconds, enough elements were used the first time. The BK18 integration printout in Figure 6.15 shows a 2,000.000 meter line broken into four elements (ΔS = 500.000 meters) with individual values printed out for each element. At the bottom of the printout, the results obtained using ten elements (sections) show the final results without the intermediate printouts. No change to six decimal places of seconds is noted. Conclusion: results shown at the end of

```
BK18 Geodetic Forward Computation                User: Earl F. Burkholder
GFORWARD (Ver B 11/91)                           Date: August 3, 2000

Reference ellipsoid: GRS 1980
   Semi-major axis of ellipsoid =           6,378,137.000 meters,
   Reciprocal flattening of ellipsoid =     298.2572221008800

Beginning Point:
   Latitude        =   32 16 55.929060        CC    = 3,816,545.2425 m
   Longitude       =  253 14 44.839300        Dist  =      2000.0000 m
   Azi.(North)     =   45 00 00.000000        DS    =       500.0000 m

Section Number 1
  Latitude(1)      =   32 16 55.929060        M1    = 6,353,629.8253 m
  Longitude(1)     =  253 14 44.839300        N1    = 6,384,235.5313 m
  Half Dphi        =          5.738894 sec    MM    = 6,353,631.4314 m
  ϕ (Mid pt)       =   32 17 01.667954        NM    = 6,384,236.0693 m
  Azimuth(m)       =   45 00 03.608278
  Latitude(2)      =   32 17 07.406645        Dphi  =      11.477585 sec
  Longitude(2)     =  253 14 58.350998        Dlam  =      13.511698 sec
  Azimuth(2)       =   45 00 07.216777

Section Number 2
  Latitude(2)      =   32 17 07.406645        M2    = 6,353,633.0376 m
  Longitude(2)     =  253 14 58.350998        N2    = 6,384,236.6072 m
  Half Dphi        =          5.738691 sec    MM    = 6,353,634.6437 m
  ϕ (Mid pt)       =   32 17 13.145336        NM    = 6,384,237.1452 m
  Azimuth(m)       =   45 00 10.825626
  Latitude(3)      =   32 17 18.883823        Dphi  =      11.477178 sec
  Longitude(3)     =  253 15 11.863641        Dlam  =      13.512643 sec
  Azimuth(3)       =   45 00 14.434695

Section Number 3
  Latitude(3)      =   32 17 18.883823        M3    = 6,353,636.2499 m
  Longitude(3)     =  253 15 11.863641        N3    = 6,384,237.6831 m
  Half Dphi        =          5.738487 sec    MM    = 6,353,637.8561 m
  ϕ (Mid pt)       =   32 17 24.622310        NM    = 6,384,238.2211 m
  Azimuth(m)       =   45 00 18.044114
  Latitude(4)      =   32 17 30.360593        Dphi  =      11.476770 sec
  Longitude(4)     =  253 15 25.377230        Dlam  =      13.513589 sec
  Azimuth(2)       =   45 00 21.653754

Section Number 4
  Latitude(4)      =   32 17 30.360593        M4    = 6,353,639.4623 m
  Longitude(4)     =  253 15 25.377230        N4    = 6,384,238.7591 m
  Half Dphi        =          5.738283 sec    MM    = 6,353,641.0685 m
  ϕ (Mid pt)       =  32 17 36.098876         NM    = 6,384,239.2971 m
  Azimuth(m)       =  45 00 25.263743
  Latitude(5)      =  32 17 41.836956         Dphi  =      11.476363 sec
  Longitude(5)     = 253 15 38.891765         Dlam  =      13.514535 sec
  Azimuth(5)       =  45 00 28.873954

Increments used for a better position = 10   (Only the final values are printed.)
  Latitude(10)     =  32 17 41.836956
  Longitude(10)    = 253 15 38.891765
  Azimuth(10)      =  45 00 28.873954
```

FIGURE 6.15 BK18 Numerical Integration Printout section 4 in Figure 6.15 are accurate at least within 0.00003 meters (one magnitude better than the "standard").

BK19: Numerical Integration

The BK19 numerical integration is really a misnomer in that numerical integration is only part of the procedure. Essentially the BK19 computation utilizes a conventional geodetic inverse computation to obtain approximate answers (direction and distance). Those answers are then used in the BK18 computation to see how close the computed endpoint position is to the given values of latitude and longitude. Corrections to the approximate direction and distance are computed based upon the misclosure. With corrections applied, a second BK18 computation is performed and another misclosure is computed. The procedure (iteration) terminates when the user is satisfied that the computed position is acceptably close to the given latitude/longitude position. The direction and distance used in such a final BK18 computation are taken to be the BK19 solution.

Given two geometrical parameters for an ellipsoid and the latitude and longitude of two points, a listing of the steps given by Jank and Kivioja (1980) to find the direction and distance from one point to another (a BK19 solution) is as follows:

1. Find the azimuth of the normal section from point 1 to point 2. This is an approximation of the final geodetic line azimuth.

$$\cot A_n = \left(\frac{\tan\phi_2}{(1+e'^2)\tan\phi_1} + \frac{e^2 V_2 \cos\phi_1}{V_1 \cos\phi_2} - \cos\Delta\lambda \right) \frac{\sin\phi_1}{\sin\Delta\lambda} \tag{6.88}$$

where

$$e'^2 = \frac{a^2-b^2}{b^2} = \frac{a^2}{b^2} - 1 = \frac{e^2}{1-e^2};$$

$$V_1 = \sqrt{1+e'^2\cos^2\phi_1} \quad \text{and} \quad V_2 = \sqrt{1+e'^2\cos^2\phi_2} \,.$$

2. Compute Clairaut's constant for the normal section azimuth.

$$K = \frac{a\cos\phi_1 \sin A_n}{\sqrt{1-e^2\sin^2\phi_1}} \tag{6.89}$$

3. Use Clairaut's constant to compute the azimuth at the midlatitude between points 1 and 2.

$$A_m = \arcsin\frac{K\sqrt{1-e^2\sin^2\phi_m}}{a\cos\phi_m} \quad \text{where} \quad \phi_m = \frac{\phi_1+\phi_2}{2} \tag{6.90}$$

4. Compute radii of curvature for prime vertical and meridian sections at midlatitude.

$$M_m = \frac{a(1-e^2)}{(1-e^2\sin^2\phi_m)^{3/2}} \quad \text{and} \quad N_m = \frac{a}{\sqrt{1-e^2\sin^2\phi_m}} \qquad \text{(6.18) and (6.16)}$$

5. Use Euler's formula (equation 6.19) to compute the ellipsoid radius of curvature for the line (specific azimuth at given latitude).

$$R_{A_m} = \frac{M_m N_m}{M_m \sin^2 A_m + N_m \cos^2 A_m} \tag{6.91}$$

6. Use the spherical law of cosines to compute the angle subtended at the center of the Earth by points 1 and 2 on the ellipsoid surface. Use this angle and radius from step 5 to compute the preliminary distance between points 1 and 2.

$$P_1P_2 \text{ (in radians)} = \arccos\ [\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos(\Delta\lambda)] \tag{6.92}$$

$$S_c = Dist(P_1P_2) = R_{A_m}(P_1P_2)_{rad} \tag{6.93}$$

7. With the approximate distance known, compute the difference in azimuths of the geodetic line and the normal section. Apply the difference to the previously known normal section azimuth to get the azimuth of the geodetic line from point 1 to point 2. Update the value of Clairaut's constant to reflect traversing the geodetic line instead of the ellipsoid normal. Units in equation 6.94 are seconds of arc.

$$\Delta A'' = A_n - A_g = \frac{\rho(V_1^4 - V_1^2)}{12} * \frac{b^2}{a^4} * S_c^2 \sin(2A_m) \tag{6.94}$$

$$A_g = A_n - \Delta A'' \tag{6.95}$$

$$K = \frac{a\cos\phi_1 \sin A_g}{\sqrt{1-e^2\sin^2\phi_1}} \tag{6.96}$$

Although the difference is quite small, the value of K in equation 6.96 represents an improvement over the value in equation 6.89 because the BK18 check computation follows the geodetic line, not the normal section.

8. Start from point 1 and use the preliminary distance along with the azimuth of the geodetic line in a BK18 computation to compute the latitude and longitude of point 2. The computed position of point 2 should be very close to given values, as illustrated in Figure 6.16. The azimuth at the computed point is computed using Clairaut's constant as

$$A_c = \arcsin\left(\frac{K}{N_2 \cos\phi_2}\right);$$ note approximation of N_2 and φ_2 for N_c and φ_c.

9. If the approximation of azimuth from step 7 and the approximate distance from step 6 are good enough, the latitude/longitude position computed in step 8 will be close to the given position. In that case, the computation is done. Otherwise, corrections to the preliminary direction and distance need to be computed, and the BK18 computation needs to be used again.
10. Specific steps for computing the corrections to direction and distance are given in Jank and Kivioja (1980) and only summarized here. With reference to Figure 6.16, the corrections are

$$\Delta\phi_c = \phi_2 - \phi_c; \quad \text{misclosure in north-south direction} \tag{6.97}$$

$$\Delta\lambda_c = \lambda_2 - \lambda_1; \quad \text{misclosure in east-west direction, positive east } \lambda \tag{6.98}$$

$$P_cP_2 = \sqrt{(N_2 \cos\phi_2\ \Delta\lambda_c)^2 + (M_2\ \Delta\phi_c)^2}; \quad \text{distance from } P_c \text{ to } P_2 \tag{6.99}$$

$$\tan\beta = \frac{M_2 \Delta\phi_c}{N_2 \Delta\lambda_c \cos\phi_2}; \quad \text{note that } \Delta\varphi_c \text{ and } \Delta\lambda_c \text{ are in radians.} \tag{6.100}$$

$$\gamma = A_c - \beta; \quad \text{this equation is correct for the case shown. Others exist.} \tag{6.101}$$

From Figure 6.17, it is readily seen that

$$\Delta S = P_cP_2 \sin\gamma; \quad \text{this is a correction to the distance.} \tag{6.102}$$

$$\Delta A'' = \frac{\rho\ P_2P_c \cos\gamma}{S_c}; \quad \text{this is a correction to the geodetic line azimuth.} \tag{6.103}$$

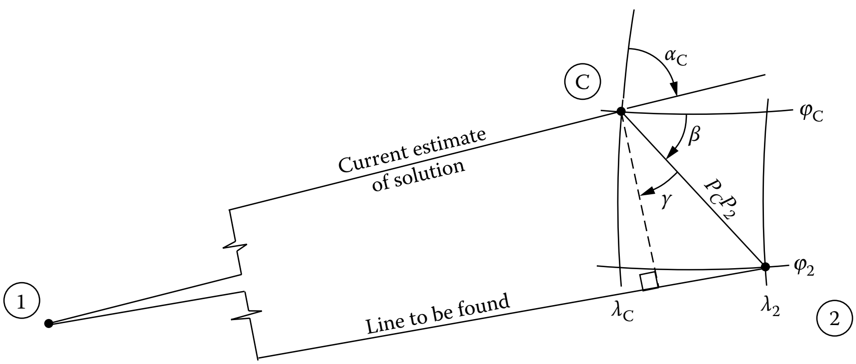


FIGURE 6.16 Misclosures in the Trial BK18 Computation

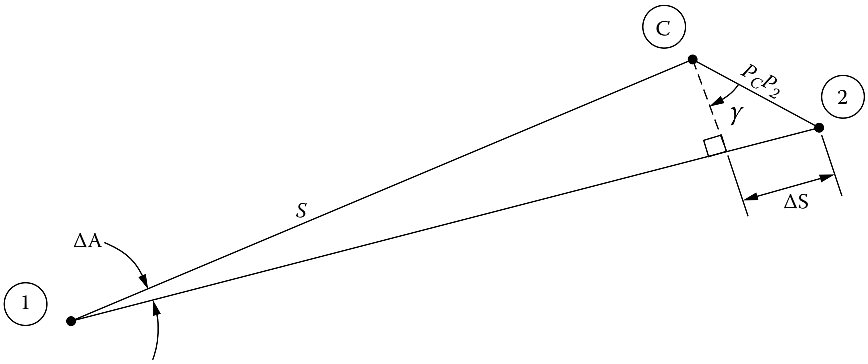


FIGURE 6.17 Corrections to Previous Direction and Distance

11. With the corrections determined above, the BK18 computation is used again with new (better) values for distance and azimuth. Figure 6.17 shows only one of many possible combinations for applying the computed corrections. The user will need to assure that correct signs are used in applying the corrections. The misclosure should be much smaller at the end of the second BK18 computation. If the agreement of the computed position with the given position is acceptable, the following direction and distance are the inverse (BK19) solution. If not, steps 7, 8, 9, 10, and 11 are repeated.

$$A_2 = A_g + / - \Delta A; \tag{6.104}$$

$$S_2 = S_c + / - \Delta s; \tag{6.105}$$

Comments on the Jank-Kivioja BK19 procedure:

1. The Jank-Kivioja numerical integration procedure is tedious to perform, but it does a good job of "choking the elephant" (see the preface) and puts computational control in the hands of the user.
2. Ultimate accuracy of the Jank-Kivioja method is limited only by the significant digit capacity of the computer being used. But computational procedures and equations are arranged such that only modest significant digit capacity is required to achieve impressive results for both BK18 and BK19 computations.
3. The Jank-Kivioja procedure is included to show how "classical" geodesy approaches the BK19 computation. The 2-D computation is performed strictly on the ellipsoid surface.
4. The GSDM can be used in place of equations 6.88 through 6.96. If needed, equations 6.97 through 6.105 can be used to improve the GSDM solution. See the subsequent section in this chapter on GSDM 3-D geodetic computations.

Geodetic Position Computations Using State Plane Coordinates

State plane coordinates can also be used to perform geodetic direct and geodetic inverse computations. The use of state plane coordinates is described in chapter 10, "Map Projections." In the context of a map projection, the geodetic forward and inverse computations are performed using simple 2-D COGO relationships. The complex part is converting latitude and longitude to plane coordinates and plane coordinates to latitude and longitude (BK10 and BK11 computations). Computers have been programmed to perform those transformations, and users are largely oblivious to the complexity involved. The advantage of using state plane coordinates for geodetic computations is the ease with which traverse and inverse computations are performed with plane coordinates. The disadvantage to using state plane coordinates to perform a geodetic direct computation is that local directions and distances must first be converted to grid azimuths and grid distances before performing the simple 2-D COGO computations. Similarly, the disadvantage to using state plane coordinates for the geodetic inverse computation is that the answer comes out in grid azimuth and grid distance. Many users desire local tangent plane distances and true bearings. That means grid azimuth and grid distance obtained from a state plane coordinate inverse must be converted to local tangent plane direction and distance or to ellipsoid direction and distance.

Procedures for using state plane coordinates and for computing a state plane traverse are included in chapter 10. Steps for performing a single-course geodetic traverse using state plane coordinates include the following:

1. Start with latitude/longitude of point 1 (state and zone must be named).
2. Convert latitude/longitude to state plane coordinates (use reliable software).
3. Reduce measured slope distance to grid distance. Steps include the following:
   - A. Slope to horizontal (include curvature and refraction if needed).
   - B. Horizontal to sea level or ellipsoid (ellipsoid is more elegant).
   - C. Ellipsoid distance to grid distance (use the Simpson 1/6 rule for long lines).
   - D. (Steps B and C are often done together using a combined factor.)
4. Compute the grid azimuth of line from point 1 to point 2:
   - A. Use field-measured angle from known grid azimuth of reference line.
   - B. Use solar/Polaris observation at point 1 and, as appropriate, apply
     1. Laplace correction to convert astronomic azimuth to geodetic azimuth.
     2. convergence at point 1 to convert geodetic azimuth to grid azimuth.
5. Compute state plane grid coordinates of point 2 using grid distance and grid azimuth (BK16).

$$E_2 = E_1 + (\text{grid distance}) \sin(\text{grid azimuth}) \tag{6.106}$$

$$N_2 = N_1 + (\text{grid distance}) \cos(\text{grid azimuth}) \tag{6.107}$$

6. Convert *E/N* state plane coordinates of point 2 to latitude/longitude (BK11).

Steps for performing a geodetic inverse using state plane coordinates include the following:

1. Start with latitude/longitude of point 1 and point 2 (state and zone must be named).
2. Convert latitude/longitude of each point to state plane coordinates.
3. Compute grid distance and grid azimuth between points (BK16).

   A. $Dist = \sqrt{(E_2 - E_1)^2 + (N_2 - N_1)^2}$ (grid distance)

   B. $\tan\alpha = \frac{E_2 - E_1}{N_2 - N_1} = \frac{\Delta e}{\Delta n}$ (grid azimuth—see equations 4.11 to 4.13)

7. Convert (see chapter 10):
   A. grid distance to sea level (or ellipsoid) distance.
   B. sea level (or ellipsoid) distance to horizontal ground distance.
8. Convert grid azimuth to geodetic azimuth using convergence:
   A. At point 1 to find geodetic azimuth point 1 to point 2
   B. At point 2 to find geodetic azimuth point 2 to point 1

Note: the geodetic azimuth point 1 to point 2 will not be the same (+/– 180°) as the geodetic azimuth point 2 to point 1 because meridians on the Earth are not parallel.

GSDM 3-D Geodetic Position Computations

With the exception of state plane coordinate computations, the geodetic position computations described so far are part of classical geometrical geodesy computations and are conducted only on the ellipsoid surface (e.g., 2-D). And, the argument could be made that since a map projection is strictly a 2-D model, the state plane coordinate exception is moot. The GSDM provides an alternative to classical geodesy methods and can be used to obtain results to any accuracy desired. The first solution is generally sufficient, but, if needed, refinement of the solution uses the same misclosure and correction methods as the Jank-Kivioja method—see equations 6.97 to 6.105.

Geodetic position computations using the GSDM require extensive use of BK1 and BK2 computations. BK1 is very straightforward and uses equations 6.22 through 6.25. BK2 is more complex and uses either an iteration procedure (equations 6.26 through 6.32) or a lengthier "recipe" procedure given by equations 6.33 through 6.44. Given the procedures and equations described, the GSDM geodetic forward and inverse computations are given below.

Forward (BK3)

The steps for performing a GSDM geodetic forward computation are as follows:

1. Given: latitude and longitude for point 1. Ellipsoid height may also be given, but for a strict comparison with other examples, point 1 should be on the ellipsoid, $h = 0.00$ m.
2. Use BK1 to convert *latitude/longitude/height* coordinates of point 1 to geocentric *X/Y/Z* ECEF coordinates.
3. The position of point 2 is computed as

$$X_2 = X_1 + \Delta X \tag{6.108}$$

$$Y_2 = Y_1 + \Delta Y \tag{6.109}$$

$$Z_2 = Z_1 + \Delta Z \tag{6.110}$$

   where $\Delta X/\Delta Y/\Delta Z$ are components from a GPS vector or $\Delta e/\Delta n/\Delta u$ components from total station observations rotated from the local geodetic horizon perspective to the geocentric perspective using the BK9 rotation matrix (equation 1.22).
4. The geodetic latitude and longitude of point 2 are computed by transforming the *X/Y/Z* coordinates of point 2 to *latitude/longitude/height* using a BK2 transformation. Of course, the ellipsoid height, whether zero or not, is also a product of the BK2 computation.

Inverse (BK4)

Given the latitude and longitude (ellipsoid height must be zero) of points A and B, the objective is to find the geodetic azimuth and ellipsoid distance between them using the GSDM. Steps in the GSDM geodetic inverse computation are as follows:

1. First, the latitude, longitude, and height of each point are used to compute the ECEF geocentric coordinates for point A and point B.
2. Then, the geocentric inverse is computed as

$$\Delta X = X_2 - X_1 \tag{6.111}$$

$$\Delta Y = Y_2 - Y_1 \tag{6.112}$$

$$\Delta Z = Z_2 - Z_1; \tag{6.113}$$

3. The local $\Delta e/\Delta n/\Delta u$ components are computed using the BK8 rotation matrix as described in equation 1.21.
4. Various azimuths and distances are computed depending upon the choice of the user.
   A. The 3-D azimuth is computed as

$$\alpha_{3D} = \arctan(\Delta e/\Delta n). \tag{6.114}$$

If warranted, the geodetic azimuth is computed from the 3-D azimuth by applying the correction given in equation 6.73. The target height correction is zero because $h = 0.0$ at the forepoint.

B. Given $h = 0.0$ m at both ends of the line, the mark-to-mark chord distance is either

$$D_{M-M} = \sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2} \text{ or } D_{M-M} = \sqrt{\Delta e^2 + \Delta n^2 + \Delta u^2} \tag{6.115}$$

C. The local tangent plane distance [$HD(1)$], from Point A to Point B is not really part of the geodetic inverse, but can be easily computed as:

$$HD(1) = \sqrt{\Delta e^2 + \Delta n^2} \tag{6.116}$$

D. Using the concept of $L = R\theta$ (θ in radians), the geodetic arc distance is obtained from the chord distance between point A and point B and the "best" radius of the ellipsoid between the two points. Use equation 6.115 for the chord distance and equation 6.19 for the radius. If the line is not very long, equation 6.20 may suffice for the radius.

$$D_{geod.} = R_\alpha * 2\left(\arcsin \frac{D_{M-M}}{2R_\alpha}\right) \tag{6.117}$$

where

R_α = radius from equation 6.19 or 6.20, and
D_{M-M} = chord distance from equation 6.115.

Note that for long lines, the midpoint latitude geodetic line azimuth should be computed using Clairaut's constant. This midpoint azimuth will give a better value of R_α.

GSDM Inverse Example: New Orleans to Chicago

The great circle arc distance between New Orleans and Chicago was given as in chapter 3 with promise of this similar ellipsoid computation in chapter 6. Using the GRS80 ellipsoid and the latitude/longitude positions (remember $h = 0.0$ m) listed in chapter 3, the geocentric $X/Y/Z$ coordinates of the two points are computed using BK1:

GRS80: $a = 6{,}378{,}137.000$ m and $e^2 = 0.006694380023$

New Orleans (1)	Chicago (2)
φ = 30° 02’ 17” N	φ = 42° 07’ 39” N
λ = 90° 09’ 56” W	λ = 87° 55’ 12” W
269° 50’ 04” E	272° 04’ 48” E
h = 0.0 m	*h = 0.0 m*
N = 6,383,493.2334 m	N = 6,387,764.6511 m
X = –15,967.7193 m	X = 171,947.3712 m
Y = –5,526,123.0752 m	Y = –4,734,389.6050 m
Z = 3,174,026.4177 m	Z = 4,256,117.7017 m

Using equations 6.111 through 6.113, the geocentric components are

$\Delta X = X_2 - X_1 = 187{,}915.0905$ m
$\Delta Y = Y_2 - Y_1 = 791{,}733.4702$ m
$\Delta Z = Z_2 - Z_1 = 1{,}082{,}091.2840$ m

Using the rotation matrix in equation 1.21 at New Orleans, the local perspective components are as follows:

$\Delta e = 185{,}626.6036$ m
$\Delta n = 1{,}333{,}351.1820$ m
$\Delta u = -144{,}197.4535$ m

The mark-to-mark chord distance (both points are on the ellipsoid) is computed using either geocentric components or local (New Orleans) perspective components as

$$D_{M-M} = \sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2} = \sqrt{\Delta e^2 + \Delta n^2 + \Delta u^2} = 1{,}353{,}911.192\ m \quad (6.115)$$

$$\alpha_{3D} = \tan^{-1}\left(\frac{\Delta e}{\Delta n}\right) = 7^\circ\ 55'\ 32.''4001 \quad (6.114)$$

Equations 6.72 and 6.73 are used to compute the azimuth of the geodetic line at the standpoint (New Orleans).

$$\alpha_g = \alpha_{3D} - \frac{\rho\, e^2 D_{M-M}^2}{12 N_1^2} \cos^2 \phi_m \sin(2\alpha_{3D}) \qquad (\varphi_m = 36^\circ\ 04'\ 58'')$$

$$\alpha_g = 7^\circ\ 55'\ 32.''40016 - 0.''92344 = 7^\circ\ 55'\ 31.''47666 \quad (6.118)$$

In order to convert the chord distance to the equivalent ellipsoid arc distance using equation 6.117, we need the azimuth of the line and the R_α radius of curvature at the midlatitude. Use Clairaut's constant (computed at New Orleans) to get the midpoint azimuth and equations 6.16, 6.18, and 6.19 to get the ellipsoid distance.

$K = N_1 \cos\phi_1 \sin\alpha_1$ = 6,383,493.232 m cos (30° 02' 17") sin (7° 55' 31."4767) = 761,966.1120 m

$$M_m = \frac{a(1-e^2)}{(1-e^2 \sin^2 \phi_m)^{3/2}} = 6{,}357{,}570.4002 \text{ m}$$

$$N_m = \frac{a}{\sqrt{1-e^2 \sin^2 \phi_m}} = 6{,}385{,}555.1108 \text{ m}$$

The geodetic line azimuth at the midpoint latitude is found using equation 6.66 as

$$\sin\alpha_m = \frac{K}{N_m \cos\phi_m} = \frac{761{,}966.113 \text{ m}}{5{,}160{,}594.603 \text{ m}} = 0.1476508371 \quad ; \alpha_m = 8° \, 29' \, 26."92885$$

The "best" radius of curvature at the midpoint of the line is found using equation 6.19 as

$$R_\alpha = \frac{M_m N_m}{M_m \sin^2 \alpha_m + N_m \cos^2 \alpha_m} = 6{,}358{,}177.8727 \text{ m}$$

$$D_{ellipsoid} = R_\alpha * 2 \left(\sin^{-1} \frac{D_{M-M}}{2R_\alpha} \right) \frac{\pi}{180} = 1{,}356{,}482.2905 \text{ m} \qquad (6.117)$$

As geodetic inverse computations go, New Orleans to Chicago is a long line. For shorter lines, the direction in equation 6.114 and the distance in equation 6.117 may be good enough, but whether the line is long or short, the results should be checked using the BK18 forward computation. Remember that the Earth is ellipsoidal, not spherical as was assumed in equation 6.117. To check how good the answers are, we need to perform a numerical integration forward (BK18) computation using equations 6.86 and 6.87. The BK18 printout shown in Figure 6.18 uses the latitude/longitude at New Orleans, the geodetic azimuth at New Orleans, and the ellipsoid distance from New Orleans to Chicago. Note that the line was broken into twenty pieces, each 67,824.1145 meters long. Those elements are not short enough for a good answer—but it puts us in the ballpark, as shown at the end of element 20. In the same computation the total distance was broken into 10,000 pieces (each 135.64829 meters long), but the intermediate results were not printed. A comparison of the results at the end

```
BK18 Geodetic Forward Computation                    User: Earl F. Burkholder
By: E.Burkholder 11/1991 (Ver B)                     Date: 4 August 2007

Reference ellipsoid: GRS1980
   Semimajor axis of ellipsoid =              6378137.000 meters,
   Reciprocal flattening of ellipsoid =       298.2572221008827

Beginning Point:
   Latitude    =   30  2 17.000000            CC     =     761,966.1105 m
   Longitude   =  269 50  4.000000            Dist   =   1,356,482.2905 m
   Azimuth(N)  =    7 55 31.476600            DS     =      67,824.1145 m

Section Number 1
  Latitude     =   30  2 17.000000            M1     =   6,351,413.8649 m
  Longitude    =  269 50  4.000000            N1     =   6,383,493.2334 m
  Half Dphi    =       1,090.788917 sec       MX     =   6,351,707.5680 m
  Mid point    =   30 20 27.788917            NX     =   6,383,591.6274 m
  Azimuth      =    7 56 59.514816
  Latitude     =   30 38 38.347137            Dphi   =   2,181.347137 sec
  Longitude    =  269 55 55.203701            Dlam   =     351.203701 sec
  Azimuth      =    7 58 28.891436

Section Number 2
  Latitude     =   30 38 38.347137            M1     =   6,352,002.9810 m
  Longitude    =  269 55 55.203701            N1     =   6,383,690.5911 m
  Half Dphi    =       1,090.556747 sec       MX     =   6,352,300.1327 m
  Mid point    =   30 56 48.903885            NX     =   6,383,790.1343 m
  Azimuth      =    7 59 59.642822
  Latitude     =   31 14 59.223962            Dphi   =   2,180.876825 sec
  Longitude    =  270  1 50.813333            Dlam   =     355.609632 sec
  Azimuth      =    8  1 31.767361
.
.
     Sections 3 through 19 were deleted to save space.
.
.
Section Number 20
  Latitude     =   41 31 29.042976            M1     =   6,363,502.1192 m
  Longitude    =  271 56 57.351362            N1     =   6,387,540.4366 m
  Half Dphi    =       1,085.173096 sec       MX     =   6,363,837.0364 m
  Mid point    =   41 49 34.216072            NX     =   6,387,652.4954 m
  Azimuth      =    9 12 41.626309
  Latitude     =   42  7 39.009923            Dphi   =   2,169.966947 sec
  Longitude    =  272  4 47.840849            Dlam   =     470.489486 sec
  Azimuth      =    9 15 19.694725

Increments used for following computed position = 10,000  (Only the final values are printed.)

  Latitude     =   42  7 38.997194
  Longitude    =  272  4 48.034168
  Azimuth      =    9 15 19.692856
```

FIGURE 6.18 Trial Geodetic Line: New Orleans to Chicago—BK18 of element 20 and the "final" solution shows very good agreement. But, the computed position in Chicago falls slightly south of the given latitude of 42° 07' 39."000 N. Also note that the computed longitude is slightly east of the given value of 272° 04' 48."000 E. Yes, the solution is close, but a better inverse solution is available if we use that information to compute corrections to both the approximate azimuth and distance. With the corrected values, the numerical integration BK18 computation will be performed again with better results.

At this point, the 3-D inverse has done as well as it can do. From here onward, the method of computing misclosures and corrections is taken from the Jank and Kivioja (1980) inverse algorithm.

The misclosures and corrections are computed using equations 6.97 through 6.105, as follows:

$$\Delta\phi_c = \phi_2 - \phi_c = \text{ 42° 07' 39."00000 – 42° 07' 38."997194 = 0."002806} \quad (6.97)$$

$$\Delta\lambda_c = \lambda_2 - \lambda_c = \text{ 272° 04' 48."00000 – 272° 04' 48."034168 = –0."034168} \quad (6.98)$$

$$M_2 = \frac{a(1-e^2)}{(1-e^2\sin^2\phi_2)^{3/2}} = \text{ 6,364,172.255 m} \quad (6.18)$$

$$N_2 = \frac{a}{\sqrt{1-e^2\sin^2\phi_2}} = \text{ 6,387,764.651 m} \quad (6.16)$$

$$P_cP_2 = \sqrt{(N_2\cos\phi_2\Delta\lambda_c)^2 + (M_2\Delta\phi_c)^2} = \text{ 0.7895 m } (\Delta\varphi \text{ and } \Delta\lambda \text{ in radians}) \quad (6.99)$$

$$\tan\beta = \frac{M_2\Delta\phi_c}{N_2\Delta\lambda_c\cos\phi_2} = \frac{0.08658\text{ m}}{0.7848\text{ m}} \quad \Rightarrow \quad \beta = 6°\ 17'\ 44'' \quad (6.100)$$

From the printout at the end of Figure 6.18, we get α_c = 9° 15' 20".

$$\gamma = A_c - \beta \text{ = 9° 15' 20" – 6° 17' 44" = 002° 57' 36"} \quad (6.101)$$

$$\text{Correction to distance: } \Delta S = P_cP_2\sin\gamma \text{ = 0.0408 m} \quad (6.102)$$

$$\text{Correction to azimuth: } \Delta\alpha = \frac{P_cP_2\cos\gamma}{S_c} \text{ = 0."120} \quad (6.103)$$

Using these corrections, the corrected direction and distance are

Trial 2: azimuth = 7° 55’ 31.4766 – 0.”120 = 7° 55’ 31.”3567 (6.104)

distance = 1,356,482.2905 – 0.0408 m = 1,356,482.2497 m (6.105)

These corrected values were used in the (BK18) iteration software again, and the final answer came within 1 cm of the latitude/longitude position given in chapter 3. Another iteration could have been used to make the agreement even better. This method is tedious, but it puts the user in control and provides tools that can be used to obtain the best answer possible.

Notes about the GSDM geodetic inverse:

1. The 3-D azimuth as computed by equation 6.114 contains no approximation. But, especially for long lines, the geodesic correction should be used to find the azimuth of the geodetic line. The BK18 computation provides a way to check the solution.
2. Since the Earth is ellipsoidal and not spherical, the geodetic distance computed with equation 6.117 is an approximation—albeit a very good one. The integrity of the GSDM geodetic inverse can be checked the same way as the Jank-Kivioja iteration method. Once an approximate direction and distance (equations 6.114 and 6.117) are obtained, the geodetic forward (BK18) is used to compute the latitude and longitude of point 2. If the misclosure is significant, corrections are computed and applied as outlined in equations 6.97 through 6.105. As in the iteration inverse, subsequent corrections are computed and applied as often as required to obtain a sufficiently precise answer.
3. The GSDM geodetic inverse is valid in true 3-D space and not limited to the 2-D example given above. Although the GSDM inverse can be used to duplicate a 2-D inverse on the ellipsoid, the GSDM inverse is more versatile and can be used to find any of the several horizontal distances described by Burkholder (1991).

REFERENCES

Bomford, G. 1971. *Geodesy*, 3rd ed. Oxford: Oxford University Press.

Burkholder, E. F. 1991. Computation of horizontal/level distances. *Journal of Surveying Engineering* 117 (3): 104–16.

———. 1997. The 3-D azimuth of a GPS vector. *Journal of Surveying Engineering* 123 (4): 139–46.

Hoffman-Wellenhof, B., H. Lichtenegger, and J. Collins. 1992. *GPS: Theory and practice*, 2nd ed. Vienna: Springer-Verlag.

Hooijberg, M. 1997. *Practical geodesy using computers*. Berlin: Springer-Verlag.

Jank, W., and L. A. Kivioja. 1980. Solution of the direct and inverse problems on reference ellipsoids by point-by-point integration using programmable pocket calculators. *Surveying & Mapping* 40 (3): 325–37.

Leick, A. 2004. *GPS satellite surveying*, 3rd ed. New York: John Wiley.

Rapp, R. 1991. *Geometrical geodesy: Part I*. Columbus: Ohio State University, Department of Civil and Environmental Engineering and Geodetic Science.

Soler, T., and L. Hothem. 1988. Coordinate systems used in geodesy: Basic definitions and concepts. *Journal of Surveying Engineering* 114 (2): 84–97.

Vaníček, P., and E. Krakiwsky. 1986. *Geodesy: The concepts*. New York: North-Holland.

Vincenty, T. 1980. On the transformation from three-dimensional to ellipsoidal coordinates (in German). *Zeitschrift für Vermessungswexen*, 11 (1980): 519–521. Stuttgart, Germany.

Wolf, P. R., and C. D. Ghilani. 1997. *Adjustment computations: Statistics and least squares in surveying and GIS*. New York: John Wiley.

You, R. 2000. Transformation of Cartesian to geodetic coordinates without iteration. *Journal of Surveying Engineering* 126 (1): 1–7.

7 Geodetic Datums

INTRODUCTION

A datum is a reference to which other values are related. In surveying, a vertical datum could be as simple as an arbitrary benchmark assigned an orthometric height of 100.000 meters (or feet, etc.). A horizontal datum could be defined by a stake pounded in the ground for a point of beginning (P.O.B.) and assigned arbitrary coordinates such as east = 5,000.000 meters, north = 10,000.000 meters. If both horizontal and vertical values are assigned to the same point, the result could be called a 3-D datum. The definition of units of measurement, orientation of azimuth, and coordinate system are all implicit in such a definition. Another important implicit assumption is that horizontal is perpendicular to the plumb line, or, as is the case with the GSDM, horizontal is taken to be perpendicular to the ellipsoid normal through the standpoint or through the P.O.B. as defined in chapter 1. Presumably rectangular Cartesian coordinates, either 2-D or 3-D, are used to describe the location of all points with respect to the datum origin and with respect to each other. The permanence and value of such an assumed datum depend upon the stability of monumented points, the quality of coordinates on those points, and the extent to which assumptions associated with establishment of the origin are documented, followed, and made available to others. When using the GSDM, the user selects the P.O.B. and works with local flat-Earth components in a well-documented system.

In the larger view, a datum must accommodate more than a flat-Earth perspective. It has been known since before the birth of Christ that the Earth is not flat. Eratosthenes determined the size of the Earth several hundred years b.c., and the value he obtained (Carta 1962) was within about 16 percent of today's accepted value. Locally, it still makes sense to reference the horizontal location of an object with plane rectangular coordinates, but, when describing the location of a point on the globe, it is more convenient to use the latitude/longitude graticule. Latitude is reckoned in angular (sexagesimal) units north or south from the equator, and longitude is similarly reckoned east and west from the arbitrary meridian through Greenwich, England. And, mean sea level has been used as a reference for vertical datums for many years. An observation here is that the use of mixed units—angular sexagesimal units for horizontal and length units for vertical—introduces a level of computational complexity for 3-D data that many users would like to avoid.

On one hand, the goal is to keep the datum definition as simple as possible. A simple 2-D or 3-D flat-Earth coordinate system is used successfully in many local applications to describe the location of points and/or objects. On the other hand, the definition of a datum should be sufficiently comprehensive to accommodate the entire world with geometrical integrity. According to Taylor (2004), Eratosthenes devised a rudimentary grid of latitude/longitude about 250 b.c., and several hundred years later Claudius Ptolemy's (c. 100–170) *Geographica* "included a catalog of some eight thousand place-names, rivers, mountains, and peninsulas, each of them with its position defined by degrees of latitude and longitude." Presumably the relative location of those positions was more valid than an absolute location due, in part, to uncertainties associated with a formal definition of a datum. Modern geodetic practice includes very detailed datum definitions, and the location of points anywhere within the birdcage of orbiting GPS satellites can be determined within very small tolerances—either relative or absolute.

A datum and the GSDM are similar in that each is used to define a computational environment for handling geospatial data. But, there is also an important distinction—the GSDM provides a set of rules, equations, and relationships that can be used with various 3-D datums. It is also specifically noted that the GSDM should only be used on one datum at a time. The GSDM is not a tool for combining data from two separate datums. It is the users' responsibility to know at all times what the underlying datum is for the data being manipulated. If a user is faced with combining data from several different 3-D datums, a separate transformation (such as a seven-parameter transformation, described later in this chapter) should be employed.

Sometimes the exception can also be instructive. While it is true that the GSDM should not be used as a tool for combining datums, the stochastic model portion of the GSDM allows the user great latitude with regard to working with spatial data. If the datum differences are at the 1 meter level and the data being used have a larger standard deviation (say, 5 meters), then the datum difference becomes insignificant by comparison and both data sets can be used along with their standard deviations. But the user must take responsibility for how the tool is used—in this case, lumping systematic error with larger random errors.

HORIZONTAL DATUMS

Brief History

Reviewing some of the history associated with horizontal and vertical datums provides additional perspective and promotes a better understanding of the GSDM. The following quote is found in the section "U.S. Horizontal Datums" written by Joseph Dracup, a former geodesist for the U.S. Coast & Geodetic Survey, now the National Geodetic Survey (NGS), http://www.ngs.noaa.gov/PUBS_LIB/geodetic_survey_1807.html:

> In 1879 the first national datum was established and identified as the New England Datum. Station PRINCIPIO in Maryland, about midway between Maine and Georgia, the extent of the contiguous triangulation[,] was selected as the initial point with its position and azimuth to TURKEY POINT determined from all available astronomical data, i.e. 56 determinations of latitude, 7 of longitude, and 72 for azimuth.
>
> Later its position was transferred to station MEADES RANCH in Kansas and the azimuth to WALDO by computation through the triangulation. The Clarke Spheroid of 1866 was selected as the computational surface for the datum in 1880, replacing the Bessel spheroid of 1841 used after 1843. Prior to 1843, there is some evidence that the Walbeck 1819 spheroid was employed.

> The datum was renamed the U.S. Standard Datum in 1901 and in 1913 the North American Datum (NAD) as Canada and Mexico adopted the system. In 1927 an adjustment of the first-order triangulation of the U.S., Canada and Mexico was began and completed about 1931. The end result was the North American Datum of 1927 (NAD27).

More recent information is provided by Schwarz (1989) in connection with the readjustment of the horizontal network in the United States. He describes the development of the North American Datum of 1983 (NAD83) as being based upon the Geodetic Reference System of 1980 (GRS80) ellipsoid and includes other significant historical details as well.

The *Geodetic Glossary* (National Geodetic Survey [NGS] 1986) and the *Glossary of the Mapping Sciences* (American Society of Civil Engineers, American Congress on Surveying & Mapping, and American Society of Photogrammetry & Remote Sensing 1994) each contain descriptions and definitions for various datums used in the United States. For example, a number of new datums were established in various parts of Alaska when it was not convenient or possible to tie a new project to previously existing survey control. In each case, the goal was to define a mathematical model appropriate for that portion of the Earth being surveyed.

With regards to development of horizontal datums, a conflicting goal was to avoid making "unnecessary" changes. Consider that the Clarke Spheroid of 1866 was adopted by the U.S. Coast and Geodetic Survey (USC&GS) in 1880 and used for projects throughout the United States. The International Ellipsoid was derived in 1909 by John Hayford of the USC&GS and adopted in 1924 by the International Association of Geodesy, which recommended it for use by all member countries. But, the USC&GS continued using the Clarke Spheroid of 1866 for the NAD27 adjustment because the coordinates of many stations were already based upon the Clarke spheroid, because computational tables for the Clarke spheroid were already published, and because the newer ellipsoid differed only slightly from the older one (Schwarz 1989, ch. 4).

North American Datum of 1927 (NAD27)

Prior to the satellite era, it was impossible to establish accurate intercontinental ties. Consequently, effective application and extension of a datum were limited to a specific region or to one of the continental land masses. Datums lacking global extent are called regional geodetic datums. With the advent of the space age, the tools of satellite geodesy made it possible to survey the world as a whole, and regional datums were no longer able to accommodate "big-picture" observations adequately. The solution was to develop a best-fitting mathematical model for the whole Earth with the origin located at the Earth's center of mass. Such a model is called a global geodetic datum. By contrast, a regional geodetic datum has its origin located at some point on or near the surface of the Earth. The NAD83 is a global geodetic datum, and the NAD27 is a regional geodetic datum with the origin located at triangulation station "Meades Ranch" in Kansas. Parameters that define the NAD27 are as follows (NGS 1986):

a = semimajor axis, Clarke 1866 spheroid = 6,378,206.4 m
b = semiminor axis, Clarke 1866 spheroid = 6,356,583.8 m
φ = geodetic latitude of station = 39° 13' 26."686 N
λ = geodetic latitude of station = 98° 32' 30."506 W
α = geodetic south azimuth to Station Waldo = 75° 28' 09."64
N = geoid height = 0.00 m

An implied condition is that the minor ellipsoid axis is parallel with the spin axis of the Earth. It was also intended that deflection-of-the-vertical be zero at Station Meades Ranch, but subsequent refinements in the geoid model indicate residual deflection components at the initial point. The NAD27 is a 2-D datum.

North American Datum of 1983 (NAD83)

Adjustment of the national horizontal network and publication of the NAD27 were enormous accomplishments, and that network served the control needs of a growing nation for more than half a century. But, the NAD27 was not without its problems. Some problems were associated with the sparseness of the data, some problems were associated with the accumulation of newer high-quality data that were expected to "fit" the 1927 adjustment, and some problems were associated with the lack of computing "horsepower." Localized readjustments were used during the intervening decades to fix problems in various parts of the network. Such stopgap measures accumulated to a point where the decision was made to readjust the entire North American network and, in the process, to define a new datum. The *North American Datum* (Committee on the North American Datum 1971) is a report prepared for the U.S. Congress by the National Academy of Sciences that justifies allocating resources for performing the readjustment. The NAD83 project began on July 1, 1974, and was completed on July 31, 1986, at a cost of approximately $37 million.

The NAD83 is a global geodetic datum defined as follows:

1. The datum origin was located at the Earth's center of mass as best determined at the time. As discussed later in this chapter, subsequent observations have yielded slightly different results.
2. The Z-axis is in the direction of the Conventional Terrestrial Pole as defined by the International Earth Rotation Service (IERS).
3. The X-axis coincides with the Greenwich meridian.
4. A reference ellipsoid is defined by four physical geodesy parameters:
   A. a = semimajor axis of Geodetic Reference System of 1980
   B. GM = the Geocentric Gravitational Constant
   C. J_2 = zonal spherical harmonic coefficient of second degree
   D. ω = rotational velocity of the Earth

A mathematical ellipse is defined by two geometrical parameters. In the case of an ellipsoid for the Earth, the two parameters are typically a and b, a and $1/f$, or a and e^2 as noted in chapter 6. The geometrical parameters for the Geodetic

Reference System of 1980 (GRS80) are derived from four physical geodesy parameters as adopted by the International Union of Geodesy and Geophysicists meeting in Canberra, Australia, in December 1979 and reported by Moritz (1980). The four parameters are used in an iterative algorithm to compute a value of e^2, which, in turn, is used to compute the value of $1/f$ for the GRS 1980. The following derived value of $1/f$ was computed to sixteen significant digits (Burkholder 1984).

a = semimajor axis (exact) = 6,378,137.000 meters
$1/f$ = reciprocal flattening (derived) = 298.2572221008827

Although angle and distance observations were reduced to the geoid for the NAD27 adjustment, angle and distance observations used in the NAD83 adjustment were reduced to the GRS80 ellipsoid. Scale and orientation for the overall NAD83 network were provided by a combination of a precise transcontinental traverse (TCT), Doppler data, lunar laser ranging (LLR), very long baseline interferometry (VLBI), satellite laser ranging (SLR), and astronomical azimuths. Gravity data were used to compute geoid heights and deflections-of-the-vertical at all occupied control points. The end product of the NAD83 adjustment was the two-dimensional latitude and longitude coordinates for each station. But, the NAD83 is called a 3-D datum because it is ultimately based upon the underlying ECEF coordinate system defined by the DOD (see chapter 1). Ellipsoid height is the third dimension.

World Geodetic System 1984 (WGS84)

The DOD has been engaged in navigation and mapping activities all over the world for many years. They were among the first to develop geodetic models for the entire world and have continued to refine early efforts. The initial World Geodetic System was dated 1960 (WGS60). Since then, the DOD has variously used WGS66, WGS72, and WGS84, and practice has evolved to the point that WGS84 is used both as an ellipsoid and as a datum. Given the high level of DOD implementation, the spatial data user should also understand that there are differences between a datum definition, a reference frame, and the realization of the datum in the form of coordinates. The WGS84 as a datum also incorporates gravity, equipotential surfaces, and other physical geodesy concepts. However, the summary here is limited to describing how the WGS84 relates to the collection and manipulation of spatial data primarily via GPS equipment and observations. Materials for additional study include a comprehensive report by the National Imagery and Mapping Agency (NIMA 1997) and other standard geodesy texts. A web search can also be very productive.

The geometrical parameters for the WGS84 ellipsoid were originally intended to be the same as for the GRS80 ellipsoid, but at one point in the computational process the DOD truncated an intermediate value prematurely and the resulting value of $1/f$ is noticeably different. Although the numbers in the $1/f$ value for WGS84 are different, the impact of that difference is very small. For example, in Table 6.1, the computed value of c, the polar radius of curvature, is a large number and differs by only 0.0001 meters between the GRS80 and WGS84 ellipsoids. The geometrical parameters for the WGS84 ellipsoid are as follows:

a = semimajor axis (exact) = 6,378,137.000 meters
$1/f$ = reciprocal flattening (derived) = 298.257223563

There is no practical difference between GRS80 and WSG84 so far as ellipsoids are concerned. But, NAD83 coordinates and WSG84 coordinates may differ up to a meter or more—not because the ellipsoids are different but because the datum origins are at different locations and/or the coordinates were derived from disparate survey operations (Schwarz 1989, ch. 22).

International Terrestrial Reference Frame (ITRF)

The IERS provides the International Terrestrial Reference Service (ITRS) as one of several services and defines a reference frame for scientific uses the world over. A quote from the IERS web site (http://www.iers.org/MainDisp.csl?pid=97-108) is as follows: "The ITRS is realized by estimates of the coordinates and velocities of a set of stations observed by VLBI, GPS, SRL, and DORIS [Doppler orbitography and radio positioning integrated by satellite]. Its name is the International Terrestrial Reference Frame (ITRF)." The ITRF uses the meter as the unit of length and shares the center of mass of the Earth as its origin with other 3-D datums. Initially the orientation of the ITRF was the same as that of WGS84, but the "time evolution of orientation" for the ITRF is chosen such that there is a net zero rotation with regard to horizontal tectonic plate movements over the entire Earth. The ITRF uses the same ECEF rectangular coordinate system as other 3-D datums, and, although ECEF coordinates are the defining values, the GRS80 ellipsoid should be used when expressing ITRF positions in latitude/longitude/height.

The point is that the NAD83, WGS84, and ITRF are all 3-D datums having their origin at the Earth's center of mass. They all use the meter as the unit of length and all employ the ECEF rectangular coordinate system. With those characteristics, the GSDM can be used equally well with any of them—one at a time. The difference between the datums lies in the location of the center of mass, the orientation of the coordinate system, and the realization of coordinates within the respective systems.

Miscellaneous comments are:

1. The GPS satellites orbit the Earth's physical center of mass. The NAD83 origin was positioned quite well with publication of the NAD83 adjustment, but, since then, better data for the position of the Earth's center of mass are consistent at the centimeter level and different from the NAD83 origin by about two meters. The NAD83 is realized by the coordinates of "fixed" points on the North American tectonic plate. Earthquake zones being an exception, that means the NAD83 coordinates for monumented points throughout North America change very little, if at all, over time. New NAD83 coordinates are best determined by adding precise *relative* $\Delta X/\Delta Y/\Delta Z$ baseline components to higher-accuracy control points.
2. The WGS84 datum is updated periodically so that GPS satellite orbits are computed with respect to the current approximation of the Earth's center of mass. A WGS84 datum update consists of computing and using revised

coordinates for the DOD GPS tracking network. The updates are identified by WGS84(G730), WGS84(G873), and WGS84(G1150), where "G" means the update is based upon GPS observations and the number following is the GPS week since January 5, 1980. In general, *absolute* WGS84 coordinates are obtained from code-phase GPS observations and referenced to the broadcast ephemeris of the satellites. See chapter 9 for more details.

3. The ITRF is developed by collaborators in the international community who compute yearly updates to the coordinates of a larger network of GPS tracking stations. The ITRF solution is quite close to the WGS84 realization, and comparisons are made between the ephemerides of WGS84 and ITRF on a daily basis. Although the differences between WGS84 and ITRF do exist, they are small and statistically insignificant. A general statement is that ITRF utilizes both absolute and relative positioning data, whereas NAD83 is based primarily on relative GPS data and WGS84 primarily uses absolute GPS data.
4. NAD83, WGS84, and ITRF are similar (they are all 3-D datums), but they are also quite different. Each exists for specific reasons, and applications vary accordingly. Points to be remembered:
   A. The NAD83 is "fixed" to the North American plate and is very stable. Typically, if newer coordinates for a control monument as published by the NGS are different than previous ones, the changes really reflect greater consistency in the network. The changes are most likely due to improved observations and readjustments by the NGS. The possibility of tectonic movement exists—especially in earthquake-prone areas.
   B. The WGS84 is "native" to the NAVSTAR satellite system and is maintained and updated by the DOD. The quality is very high, but modifications are strictly the prerogative of the DOD. Absolute positioning (as opposed to relative positioning) is the primary goal.
   C. The ITRF is a collaborative product of the international community and serves a greater user base than either the NAD83 or the WGS84. The ITRF uses the "best" technology available irrespective of the source and is updated on a yearly basis to reflect the best current solution for coordinates of tracking stations and computation of precise orbits. Like the Internet, it lacks the commitment of a sovereign to guarantee permanence. It could be argued that commitment of the user community is a better guarantee.
5. Relative geocentric differences are determined by carrier-phase GPS observations and define a vector from one point to another. For short lines (say, less than 20 km), these $\Delta X/\Delta Y/\Delta Z$ components are very nearly identical in each of the three datums, ITRF, WGS84, and NAD83. A careful evaluation would need to be made for very precise applications and/or long lines.
6. If the measurement system is operational at the 1 ppm level (e.g., code-phase GPS) and determines absolute ECEF coordinates for a point on or near the Earth's surface, there is still an allowable uncertainty of about 6.38 meters at the antenna due to the distance from the Earth's center of mass. A 0.1 ppm system would still have an error budget of 0.64 meters. But, that

is not the whole story. Such autonomous positions are often improved or augmented by the application of "corrections" determined from data collected at a nearby base station or by wide area differential GPS (WADGPS) procedures—see chapter 9.

7. In the past, standard geodetic surveying practice included building a network of precise vectors and attaching such a network to fixed points of greater accuracy. With the development of real-time-kinematic (RTK) GPS surveying procedures, radial configurations are becoming more popular. Issues of redundancy notwithstanding, the GSDM accommodates either well-formulated network adjustments or "unchecked" data collected in the radial mode and affords many options for the spatial data analyst. Admittedly, the GSDM permits reckless use (such as mixing datums or using unchecked data). But the GSDM supports many beneficial uses, and the intent is that the GSDM will be used responsibly and competently. If a project is based upon appropriate control information and if observed spatial data components are entered with their appropriate standard deviations (and covariances), subsequently derived answers will reflect legitimate statistical properties of those data. Issues of local accuracy and network accuracy also need to be examined carefully (Burkholder 1999; Pearson 2004).

High Accuracy Reference Network (HARN)

During the twelve years that the NAD83 was being computed, GPS became a viable positioning tool embraced by the user community. One justification for readjusting the NAD27 was that, over a period of decades, people in the user community gained access to better equipment and were able to survey more precisely than the network to which they were expected to attach their results. Ironically, GPS positioning technology was not used in the NAD83 readjustment, and, by the late 1980s, that justification repeated itself. Immediately following publication of the NAD83, the NGS came under pressure to support the control needs of GPS-equipped persons and organizations in the user community. To forestall creation of local or proprietary networks, the NGS adopted a policy of upgrading portions of the network—state by state—based upon GPS control surveys. These upgrades were known as High Precision Geodetic Networks (HPGN). Those GPS surveys were conducted primarily to improve the horizontal latitude/longitude position of the GPS control points and to establish new control points in places more accessible to the user public (Bodnar 1990). But, this time there was a big difference: the datum did not change; only the coordinates for the reobserved control points were improved. Although many parts of the national network have been readjusted and HPGN values have been published since 1990, it is still the NAD83. It is not a new datum. The counterargument is "If the coordinates for the same point are different, it is a new datum." Education helps spatial data users understand those differences.

Subsequently, the HPGN acronym gave way to the High Accuracy Reference Network (HARN), which is essentially the same as a HPGN. D'Onofrio (1991), Strange and Love (1991), and Doyle (1992) use the HARN acronym and describe the technical, logistical, and political considerations for upgrading the NAD83 to support modern GPS positioning capabilities.

Much could be said (see, e.g., Doyle 1994) about the evolving character and quality of the National Geodetic Reference System (NGRS)—also called the National Spatial Reference System (NSRS). An oversimplification is that the NAD83 datum evolved from a 2-D datum to a 3-D datum (hence the name NGRS) and that the name was revised to NSRS in an attempt to avoid the more intimidating use of the word "geodetic." The original NAD83 coordinates were simply called NAD83 coordinates (as opposed to NAD27 coordinates). The newly adjusted GPS-derived HARN coordinates in a state are referred to as NAD83 (19xx), where "xx" is the year in which the GPS-derived values were published. In time, the original NAD83 coordinates came to be known as NAD83(86) values. Two considerations for using NAD83 coordinates are (1) to be clear as to the (epoch) version of the coordinates being used and (2) to avoid mixing coordinates from two different adjustments. Of course, there are times when coordinates from different adjustments can be used without detrimental consequences, but a user should not attempt to start a survey on a NAD83(19xx) control point, close on a NAD83(86) point, and expect reliable results—it is like mixing apples and oranges or using different datums.

Another issue to be addressed was the boundary of adjustments between states. Although the newer adjustment results were typically within 0.5 meters of the previously published values, that level of difference needed to be "feathered" into the positions of adjoining states so that users could continue to enjoy consistent results even if starting a control survey in one state and closing it on control points in an adjacent state. With publication and use of the NAD83(2007), the boundaries between previous HARN, adjustments have disappeared.

Continuously Operating Reference Station (CORS)

A GPS receiver connected to a permanently mounted antenna and collecting data continuously is called a continuously operating reference station, or CORS. Depending upon the capabilities of the receiver, the design of the antenna, and associated software, a CORS station may be a sophisticated high-end automated installation that collects meteorological data as well as satellite signals and that posts collected data to the Internet in real time. Some CORS stations (in the United States) are owned and operated by the NGS, some are owned and operated by the U.S. military, and many others are owned and operated by federal, state, and local agencies, corporations, businesses, or individuals. The NGS maintains a clearinghouse of CORS data that meet standards established for operation and maintenance of a national CORS network. CORS data collected by the NGS are available via the Internet from the NGS archives to users worldwide. The precise positions of the fundamental CORS network stations are computed daily with respect to the ITRF, and modifications are made to the published positions as warranted. The point here is that CORS positions published by NGS are very high quality and are the "best" available to users in the United States. Using appropriate datum transformation software, the same ITRF CORS positions, standard deviations, and velocities may also be published on other datums such as the NAD83 or WGS84.

At the other end of the spectrum, a GPS receiver running continuously and broadcasting raw data to remote receivers operating within local radio range of the "base" station is also referred to as a CORS station. Such an installation typically serves a small area and maybe even only one user. The reference position of the base station may have been established with respect to the NAD83, WGS84, or ITRF at the prerogative of the owner, and the position of the remote receiver(s) is determined during a survey with respect to the CORS within the guidelines (proven or unproven) chosen by the user/owner. Proving the quality or integrity of such results is the responsibility of the owner/user. It is not uncommon for local RTK surveying activities to be served by a local GPS CORS or, more recently, a network of CORS operating in concert and known as a real-time GPS CORS network (RTN). The value of such RTN's is enhanced to the extent that the integrity and quality of the results are proven to be consistent and compatible with the NSRS, maintained by the NGS.

In the late 1990s the NGS embarked upon a program known as "height modernization" in which state-by-state projects were implemented with the idea of improving the height component of the NSRS. Height modernization observations are typically referenced to a network of CORS whose positions are determined rigorously by the NGS with respect to the ITRF datum. Those ITRF coordinates are converted to NAD83(CORS) coordinates using a fourteen-parameter transformation as documented in HTDP—see the "Datum Transformations" section, below. CORS data are available to anyone via the Internet.

More recently, the NGS has completed a project to readjust the entire collection of HARN networks to the CORS stations. The project was similar to the massive readjustment performed for the NAD83 but was accomplished with many fewer people over a much shorter time frame. The datum did not change, but the readjusted values are referred to as NAD83(2007). Completion of the readjustment was announced on February 10, 2007, to coincide with the 200th anniversary of the establishment of the Survey of the Coast, the predecessor of the NGS.

VERTICAL DATUMS

As stated at the beginning of this chapter, a simple vertical datum can be defined by assigning an arbitrary elevation to a specified benchmark and referencing other elevations to that assumed value. That practice may be legitimate for local relative elevation differences, but it is not pursued here. More formally, a vertical datum is an equipotential surface used as a reference for elevation. In this sense, elevation is an absolute term associated with the third dimension. Other terms associated with elevation include "altitude" and "height." The goal here is to retain the rigor of the formal definitions while acknowledging and building on the intuitive understanding of the reader.

Mean sea level (MSL) is widely understood as being a reference for elevation. Most humans stand erect and have some concept of oceans and sea level. "Up" goes higher, and sea level seems to be a good place from which to start. But sea level moves up and down according to the tides, changing barometric pressure, ocean currents, and other factors. As discussed more in chapter 8, "Physical Geodesy," the geoid is the formal reference surface for elevation and is approximated by the idealized mean sea level at rest. Although finding the precise geoid is quite difficult, the geoid (or MSL) has been used worldwide as a reference for elevation.

Mean Sea Level Datum of 1929 (now NGVD29)

The Mean Sea Level Datum of 1929 was established in the United States on the basis of extensive readings of twenty-six tide gauges scattered along the North American coast—twenty-one of them in the United States and five of them in Canada. Zero elevation as determined by the mean readings at each tide gauge was held in the adjustment, and elevations were computed for thousands of benchmarks on the differential level loops that had been run throughout the United States.

In the decades following publication of the 1929 mean sea level elevations, it became apparent, based upon precise level loops throughout North America, that mean sea level as determined by the tide gauge readings was less accurate than that carried through the leveling network. That means that a zero elevation is not exactly the same as mean sea level. As a consequence, the name of the datum was changed from the Mean Sea Level Datum of 1929 to the National Geodetic Vertical Datum of 1929 (NGVD29) in May 1973. No published elevations were changed; only the name of the datum was changed (Berry 1976).

International Great Lakes Datum

When attempting to compute accurate hydraulic head for water stored in the Great Lakes system, scientists in the United States and Canada realized that a system of geopotential numbers and dynamic heights would provide a better model for their computations than does the concept of elevation. As described in more detail in chapter 8, equipotential surfaces are separated by units of work (force × distance), where gravity is the force variable and elevation difference is the distance variable. The geopotential number for an undisturbed water surface is a constant and is computed as the infinite summation of the product of a force times distance accumulated along the path of the maximum gradient. A dynamic height is computed as the geopotential number divided by normal gravity. A dynamic height consists of numbers that look like an elevation, but dynamic height is not the distance from the geoid. Dynamic heights are everywhere the same for the surface of a body of water at rest (e.g., one of the Great Lakes) and can be used for accurate hydraulic head computations.

Understanding the need for dynamic heights and recognizing that the Hudson Bay region is still rebounding from the glacial burden of 10,000 years ago, the International Great Lakes Datum (IGLD55) was established jointly by U.S. and Canadian scientists in 1955 for the Great Lakes region of North America. It was also anticipated that the IGLD would need to be readjusted every thirty years or so due to continuing crustal rebound. The current version is IGLD85 (see, e.g., IGLD85 2006).

North American Vertical Datum of 1988

Following readjustment of the horizontal network, the NGS also performed a readjustment of the vertical network (Zilkoski, Richards, and Young 1992) and published the results as the North American Vertical Datum of 1988 (NAVD88). Several significant differences between the NGVD29 and the NAVD88 include (1) the internal consistency of the loops covering the nation is better and (2) while the NGVD29 adjustment was constrained to the tide gauge readings around the coast of North America, the NAVD88 elevations are computed with respect to a single tide gauge elevation at Father's Point/Rimouski, Quebec, Canada.

Several notes:

1. The single benchmark at Father's Point had already been selected as the initial benchmark for the IGLD85 by the Coordinating Committee on Great Lakes Basic Hydraulic and Hydrologic Data.
2. Using the Father's Point benchmark also met the requirement that the datum shift from NGVD29 to NAVD88 would, to the extent possible, minimize recompilation of the national mapping products by the U.S. Geological Survey.
3. Both the IGLD85 and the NAVD88 datums have the same geopotential numbers on each published benchmark. From those geopotential numbers, the IGLD85 dynamic heights are obtained by dividing the geopotential number by normal gravity at latitude 45° (980.6199 gals), and the NAVD88 Helmert orthometric heights are obtained from the same geopotential numbers by dividing by the value of gravity at the station.

3-D DATUMS

A formal definition of a 3-D datum will include physical geodesy concepts and is left to the scientists. But, for the purposes of spatial data and using the GSDM, a 3-D datum consists of the following:

- An Earth-centered Earth-fixed (ECEF) right-handed rectangular coordinate system.
- An origin located at the Earth's center of mass.
- The Z-axis is directed to the Conventional Terrestrial Pole as defined by the IERS.
- The *X*/*Y* plane is the plane of the equator, and the X-axis is directed to zero longitude. The Y-axis completes the right-handed system.
- The meter (or metre) is the unit of length.

Subject to choices of the user with respect to datums, derived quantities include the following:

- Latitude, longitude, and ellipsoid height based upon some chosen ellipsoid
- Orthometric height based upon ellipsoid height and geoid height
- State plane, UTM, or map projection coordinates based upon the user's choice

- Direction, distance, and rectangular components between points in
  1. local tangent plane through standpoint, or
  2. tangent plane through P.O.B. selected by the user

DATUM TRANSFORMATIONS

The topic of datum conversions is very important and deserves more discussion than is provided here. The GSDM provides an efficient environment for working on a given datum, but the GSDM is not a tool for transforming data from one datum to another. However, as a convenience to the reader, the following is a summary of specific bidirectional datum transformations that the spatial data user may encounter:

1. NAD27 to NAD83: 2-D problem, use the CORPSCON Windows-based program developed by the U.S. Army Corps of Engineers.
2. NAD83 to HPGN: 2-D/3-D problem, use CORPSCON.
3. NGVD29 to NAVD88: 1-D problem, use CORPSCON.
4. NAD83(xx) to ITRF 3-D: 3-D problem, use HTDP program from NGS.
5. NAD(83) and numerous other datums to WGS84: 3-D problem, see NIMA (1997, appendices B, C).
6. From one epoch to another: Use HTDP from NGS for 3-D data.
7. Write your own 3-D datum conversion: Use seven- (or fourteen-) parameter transformation.
8. NAD83 or NAD83(xx) to NAD83(2007): 3-D problem. Tools are still being developed. For example, see number 7, above.

The following transformation procedures are all bidirectional, meaning there is a unique set of range and domain values in each case. Any transformation procedure should be checked by performing the reverse computation to confirm the integrity of the software and the process. The user should also be aware that most datum transformation procedures are approximations (although most are very good) and should not be used for critical or very precise applications unless done so under the responsible supervision of a knowledgeable professional.

NAD27 TO NAD83(86)

This 2-D datum transformation can be performed using the NADCON program available from the NGS. The NADCON program was incorporated into the overall CORPSCON Windows-based program developed by the U.S Army Corps of Engineers and is available free of charge.

NAD83(86) TO HPGN

This is a separate process. To complete the transformation of NAD27 values to the NAD83(xx) in each state, the NGS provided an additional set of transformation parameters for each state and region as required to transform between the original NAD83(86) adjustment and the GPS statewide upgrades. The CORPSCON combines the two-step process into a single operation. However, the user should confirm that the CORPSCON version being used contains the current HPGN files for the specific region needed.

NGVD29 to NAVD88

This 1-D datum transformation can be performed using the VERTCON program available from the NGS. The VERTCON program was also incorporated into the overall CORPSCON Windows-based program developed by the U.S. Army Corps of Engineers and is available free of charge.

HTDP

The horizontal time dependent positioning (HTDP) program is available gratis from the NGS web site either as a download or to use interactively. Based upon observed tectonic movements, the program permits the user to interpolate positions for any recent epoch anywhere within the United States and its territories. The HTDP software can also be used to convert between NAD83 and ITRF.

Software Sources

The following web sites can be used to find free software.

CORPSCON: http://crunch.tec.army.mil/software/corpscon/corpscon.html
NADCON: http://www.ngs.noaa.gov (follow link to software)
VERTCON: http://www.ngs.noaa.gov (follow link to software)
HTDP: http://www.ngs.noaa.gov (follow link to software)

Seven- (or Fourteen-) Parameter Transformation

A seven-parameter transformation is often used when converting 3-D coordinates from one reference frame to another. There are three translation parameters, three rotation parameters, and one scale parameter. A generalized form of the seven-parameter transformation in matrix form is:

$$\mathbf{X}_2 = \mathbf{K} + S\,\mathbf{R}\,\mathbf{X}_1 \tag{7.1}$$

where

$\mathbf{K}$ = translation vector,
S = scalar—often close to 1.0,
$\mathbf{R}$ = rotation matrix frame 1 to frame 2,
$\mathbf{X}_1$ = vector of frame 1 coordinates, and
$\mathbf{X}_2$ = vector of frame 2 coordinates.

The seven-parameter transformation algorithm as stated contains no approximation. In a perfect world, the equations are exact, and data in one reference frame can be converted perfectly to another. But in the real world, coordinates are obtained from measurements and measurements contain errors. Therefore, the stochastic properties of the data must be evaluated to determine what is good enough and what isn't in a transformation. The stochastic features of the GSDM accommodate those issues as well—see chapter 11. Note that a fourteen-parameter transformation accommodates velocities on the seven parameters.

REFERENCES

American Society of Civil Engineers, American Congress on Surveying & Mapping, and American Society of Photogrammetry & Remote Sensing. 1994. *Glossary of the mapping sciences*. New York and Bethesda, MD: American Society of Civil Engineers, American Congress on Surveying & Mapping, and American Society of Photogrammetry & Remote Sensing.

Berry, R. M. 1976. History of geodetic leveling in the United States. *Surveying and Mapping* 36 (2): 137–53.

Bodnar, A. Nicholas. 1990. National Geodetic Reference System statewide upgrade policy. In *Proceedings of the ACSM GIS/LIS 1990 Fall Convention*, Anaheim, CA, 7–10 November.

Burkholder, Earl F. 1984. Geometrical parameters of the Geodetic Reference System 1980. *Survey and Mapping* 44 (4): 339–40.

———. 1999. Spatial data accuracy as defined by the GSDM. *Surveying and Land Information Systems* 59 (1): 26–30.

Carta, Martha, trans. 1962. *Jordan's handbook of geodesy*, vol. 3. Washington, DC: U.S. Army Corps of Engineers, Army Map Service.

Committee on the North American Datum. 1971. *North American Datum*. Washington, DC: National Academy of Sciences.

D'Onofrio, Joseph D. 1991. High precision geodetic network for California. Paper presented at ASCE Specialty Conference, Sacramento, CA, 18–21 September.

Doyle, David R. 1992. High accuracy reference networks: Development, adjustment, and coordinate transformation. Paper presented at ACSM Annual Convention, New Orleans, LA, 15–18 February.

———. 1994. Development of the National Spatial Reference System. http://www.ngs.noaa.gov/PUBS_LIB/develop_NSRS.html.

Dracup, Joseph F. 2006. Geodetic surveys in the United States: The beginning and the next one hundred years; U.S. horizontal datums. http://www.ngs.noaa.gov/PUBS_LIB/geodetic_survey_1807.html.

IGLD85. 2006. International Great Lakes Datum. http://www.ngs.noaa.gov/TOOLS/IGLD85/igld85.html.

Moritz, Helmut. 1980. Geodetic Reference System 1980. *Bulletin Geodesique* 54 (3). http://www.gfy.ku.dk/~iag/handbook/geodeti.htm.

National Geodetic Survey (NGS). 1986. *Geodetic glossary*. Rockville, MD: National Geodetic Survey, National Ocean Service, National Oceanic and Atmospheric Administration.

National Imagery and Mapping Agency (NIMA). 1997. Department of Defense World Geodetic System 1984: Its definition and relationships with local geodetic systems. Technical Report 8350.2, 3rd ed. Bethesda, MD: National Imagery and Mapping Agency.

Pearson, Chris. 2004. National spatial reference systems readjustment of NAD83. http://www.ngs.noaa.gov/NationalReadjustment/Items/CFP.The%20national%20readjustment-1.html.

Schwarz, Charles R., ed. 1989. *North American Datum of 1983*. Rockville, MD: National Geodetic Survey, National Ocean Service, National Oceanic and Atmospheric Administration, U.S. Department of Commerce.

Strange, William E., and John D. Love. 1991. High accuracy reference networks: A national perspective. Paper presented at ASCE Specialty Conference, Sacramento, CA, 18–21 September.

Taylor, Andrew. 2004. *The world of Gerard Mercator: The mapmaker who revolutionized geography*. New York: Walker.

Zilkoski, D., Richards, J., and Young, G. 1992. Results of the general adjustment of the North American Vertical Datum of 1988. *Surveying & Land Information Systems* 52 (3): 133–49.

8 Physical Geodesy

INTRODUCTION

Physical geodesy is the branch of science that relates the internal distribution of mass within the Earth to its corresponding gravity field. Under ideal circumstances, there would be no hills, valleys, mountains, or oceans on the Earth, and the distribution of mass within the Earth would be uniform. Given those conditions, the plumb line (the vertical) would be coincident with the ellipsoid normal and there would be no difference between the ellipsoid and the geoid. But, the geoid is an equipotential surface that is always perpendicular to the local plumb line and the direction of the plumb line is dictated by the vector sum of forces acting on the plumb bob; for example, gravity is the sum of gravitational attraction and centrifugal force, as shown in Figure 8.1. The centrifugal force component is very predictable and can be computed. But, due to the Earth's topography and due to variations of density within the Earth, gravitational attraction varies from point to point and, although the difference is generally quite small, the resulting vertical is rarely coincident with the ellipsoid normal. That being the case, the geoid is not parallel to the ellipsoid, and separation between the two surfaces varies with location. Ongoing geodetic research continues to improve knowledge of the relationship between the ellipsoid and the geoid. Evidence of continuing progress and improvement in geoid modeling in the United States is seen in the publication of various geoid models (dated 1990, 1993, 1996, 1999, 2003, and so on).

Physical geodesy is also of concern to the geophysicist who studies gravity anomalies in search of patterns that indicate the presence of substrata oil deposits and to the geodesist who computes the trajectory of missiles and satellites flying above the Earth. Inferring the characteristics of underground strata and being able to predict how objects move are both important concepts, but they are beyond the scope of this book. The focus of this book is documenting where things are in terms of geometrical geodesy and the GSDM. Historically, physical geodesy and geometrical geodesy have shared the challenge of finding the elusive geoid and using it as a reference for elevation. The approach in this book is different in that elevation is viewed as a derived quantity computed from GPS ellipsoid height and the ever improving knowledge of geoid height.* Therefore, while a fundamental description of physical geodesy is viewed as useful in understanding the broader context of geospatial data and how they are used, the reader is encouraged to consult other sources for additional information on physical geodesy.

* Of course, the time-honored practice of starting on a benchmark of known elevation and taking simple backsight and foresight readings is still a legitimate method for determining the elevation of an unknown point.

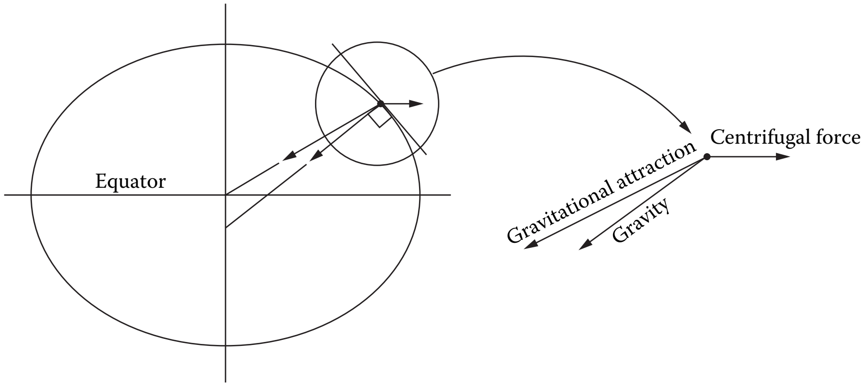


FIGURE 8.1 Vector Components of Gravity

GRAVITY

Gravity is a vector quantity composed of the sum of gravitational attraction and centrifugal force due to the Earth's rotation. Centrifugal force is always parallel to, and is greatest at, the equator. Centrifugal force is zero at the poles, and gravity at the poles is the same as gravitational attraction. But, at the equator, centrifugal force is colinear with gravitational attraction, and, because it acts in the opposite direction, the force of gravity is smaller at the equator than is gravitational attraction. Therefore, on the same equipotential surface, the force of gravity at the pole is greater than the force of gravity at the equator. The global implication is that level surfaces are not parallel. See Figure 8.2.

As described by Newton, gravitational attraction is the mutual attractive force between each and every particle in the universe.

$$F = \frac{kM_1M_2}{D^2} \tag{8.1}$$

where

k = universal gravitational constant,
M_1 and M_2 = masses of the particles, and
D = the distance between particles.

The magnitude of attraction between respective paired centers of mass decreases by the square of the increasing distance between them. Particle-pairs with a large separation react minimally, and attractions at very large distances tend to be ignored. But, taken as a large collection of particles (a large body such as the Earth or sun can be treated as a point mass located at its center), the gravitational attractions interact to keep the planets in orbit about the sun in addition to keeping the moon and satellites in orbit about the Earth. Conventional practice on Earth is to express gravity as the force per unit mass with respect to the mass of the Earth concentrated at its center.

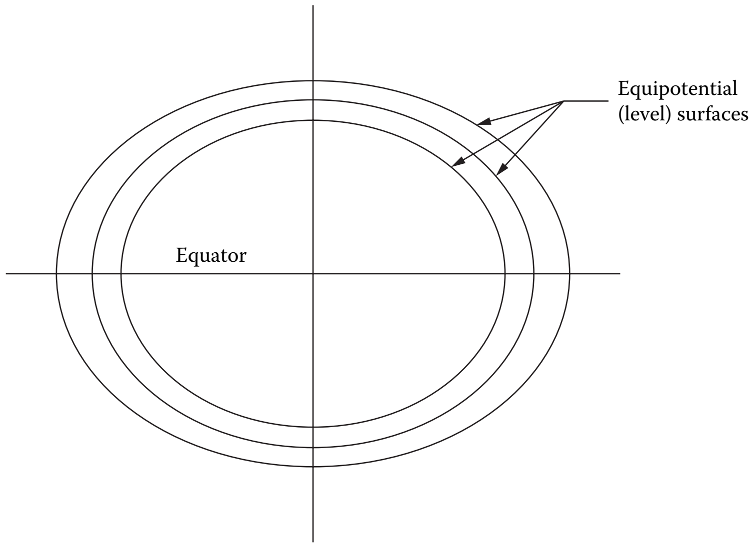


FIGURE 8.2 Level Surfaces Are Not Parallel

Two important consequences of the gravity vector are:

1. As postulated by Newton and shown in Figure 8.1, the Earth is flattened at the poles because the gravity vector does not point directly to the Earth's center of mass and because the geoid is always perpendicular to the plumb line.
2. On a global scale, equipotential surfaces are not parallel, and, as illustrated in Figure 8.2, the distance between level surfaces is not constant. That means the definitions of elevation, level surfaces, orthometric heights, and other terms need to be very specific.

The intuitive equipotential surface most commonly understood is sea level. Sea level is physical, readily visible, and observed worldwide. Due to the influence of tides and other factors, the term "mean sea level" has long been associated with the geoid and served as the reference surface for the Mean Sea Level Datum of 1929 in the United States. As described in chapter 7, the name of the datum was changed to NGVD29 in 1973.

DEFINITIONS

The topic of units deserves particular attention when discussing concepts of physical geodesy. Spatial data users are primarily concerned with distance units (meters, feet, etc.) as related to location and elevation. However, when building on fundamental physical concepts, the physical separation between two equipotential surfaces is defined in terms of work (i.e., force × distance). The following definitions presume an understanding that gravity is the force part of "work" and that elevation difference is the distance part. The following definitions are intended to be consistent with common usage, recent publications, and standard references such as the National Geodetic Survey (NGS; 1986); American Society of Civil Engineers, American Congress on Surveying & Mapping, and American Society of Photogrammetry & Remote Sensing (1994); and Meyer, Roman, and Zilkoski (2004).

Elevation (Generic)

Elevation is the distance above or below a reference surface. The geoid is an equipotential reference surface that has been widely used and is closely approximated by mean sea level. Unless specifically stated otherwise, a mean sea level elevation should probably be viewed as a generic elevation.

Equipotential Surface

An equipotential surface is a continuous surface defined in terms of work units with regard to its physical environment. Although not perfect, mean sea level is often given as an example. Two objects at rest, having the same mass, and located on the same equipotential surface store the same amount of potential energy. Work is required to move any objects to a higher elevation. If the strength of gravity at point A is greater than at point B, then, for identical objects and for expenditure of the same work, the distance moved by the object at point B will be greater than at point A. The implication is that equipotential surfaces are parallel if and only if gravity is the same at both points on each respective surface. Although defined differently, a level surface and an equipotential surface are very nearly identical and, for most purposes, can be used interchangeably.

Level Surface

A level surface is a continuous surface that is always perpendicular to the local plumb line. A level surface can be at any elevation. Due to the Earth's curvature and variations of density within the Earth, the direction of the plumb line changes as one moves from point to point on or near the surface of the Earth. Consequently, a level surface (which is always perpendicular to the plumb line) is said to be "lumpy" due to these random changes in direction of the plumb line. But, in most cases, changes in the direction of the plumb line are gradual and "lumps" in the geoid are gradual as well.

Geoid

The geoid is an equipotential surface most closely represented by mean sea level in equilibrium all over the world (i.e., constant barometric pressure at the surface, no winds, no currents, uniform density layers of water, etc.). The "ideal" conditions do not exist, and locating the geoid precisely on a global scale is an enormous challenge.

Geopotential Number

A geopotential number is a relative value computed as the infinite summation of the product of force times distance accumulated along a path of maximum gradient. A geopotential number has units of work and is rarely used in surveying and mapping applications. Dynamic heights are often used instead—see the next subsection.

Dynamic Height

Dynamic height is the geopotential number at a point divided by a constant reference gravity. Often, normal gravity at latitude 45° is used. Dynamic heights and geopotential numbers are useful when working with precise hydraulic grade lines over a large area. Standard geodesy texts contain additional information on these topics and their applications.

Orthometric Height

Orthometric height is the curved distance along the plumb line from the geoid to a point or surface in question. Few users make the distinction between the curved-line distance and the straight-line distance between the plumb line endpoints. In the past, orthometric height has been computed as the geopotential number of the equipotential surface divided by gravity at the point (Zilkoski, Richards, and Young 1992). More recently (Meyer, Roman, and Zilkoski 2004), orthometric height is more specifically called a Helmert orthometric height and is computed using ellipsoid heights (from GPS) and geoid-modeling procedures. The accuracy of such a derived height is dependent upon both the quality of the GPS data and the integrity of the geoid modeling. Although elevation and orthometric height can often be used interchangeably, elevation is considered generic while orthometric height is specific.

Ellipsoid Height

Ellipsoid height is the distance as measured along the ellipsoid normal above or below the mathematical ellipsoid.

Geoid Height

Discounting curvature of the plumb line, geoid height is taken to be the distance along the ellipsoid normal between the ellipsoid and the geoid. Geoid height is computed as the ellipsoid height minus orthometric height. Within the conterminous United States, the orthometric height is always greater than the ellipsoid height, which means the geoid height is a negative number. But, on a worldwide basis, the simple relationship is

$$h = H + N \tag{8.2}$$

where

h = ellipsoid height,
H = orthometric height, and
N = geoid height.

GRAVITY AND THE SHAPE OF THE GEOID

Everyone knows that the Earth is flat. At least, that is our experience until we become convinced otherwise. That happens as we look at the bigger picture and watch ships sailing beyond the horizon or view the curved shadow of the Earth on the moon during an eclipse. Learning that the Earth is round and rotating on its axis also helps us to understand the sunrise and sunset each day and the traverse of stars across the night sky. Furthermore, coming to understand that the Earth also revolves yearly around the sun helps explain why the sun and the stars appear to traverse the heavens at different rates. In considering observable physical phenomena and studying the definitions in the previous section, everyone should agree that a flat-Earth view of the world is misleading.

On a global scale, Newton rationalized that the Earth must be flattened at the poles because a level surface is always perpendicular to the plumb line and, due to the addition of force vectors, the plumb line does not point directly to the center of the Earth—unless one is on the equator or at one of the poles. But, that level of analysis does not differentiate between the various land masses of the continents or include the fact that some parts of the Earth are denser than others. For example, the land mass and ice loading at the south pole give rise to an identifiable bulge in the geoid in the southern hemisphere. Therefore, the geoid is referred to as pear shaped.

Understanding that gravity is a vector having both direction and magnitude and realizing that the gravity vector is the sum of all external forces acting on a given mass at a given location, it should be understandable that the scope of physical geodesy encompasses much more material than is presented here. The magnitude of gravity affects the spacing of equipotential surfaces, and variations in the direction of gravity affect the shape of an equipotential surface. That means that the geoid is not a regularly curved surface. Since the resultant geoid is related to both the magnitude and direction of gravity, investigation of cause-effect relationships must include considerations of both. Stated differently, it is said that the magnitude of gravity affects where the geoid is located and that the direction of gravity determines the shape of the geoid. On a global scale, a stronger value of gravity will tend to pull the geoid in closer to the Earth's center, but the shape of the geoid, both globally and locally, is always perpendicular to the direction of gravity.

Figure 8.3 illustrates an apparent paradox. While it is true that a stronger value of gravity tends to pull the geoid in closer to the Earth's center, a local mass concentration lying next to a mass deficiency (mountains and oceans) will deflect the plumb bob toward the mass excess, and, contrary to statements in the previous paragraph, the geoid will rise under the mountains while dipping over the ocean. It remains a challenge to determine the location of the geoid precisely because the actual location of the geoid is the physical realization of many factors.

LAPLACE CORRECTION

The Laplace correction is used to relate a geodetic azimuth to an astronomical azimuth. Such a correction is made at a point called a Laplace Station, where the deflection-of-the-vertical components are known. In a broader sense, the Laplace

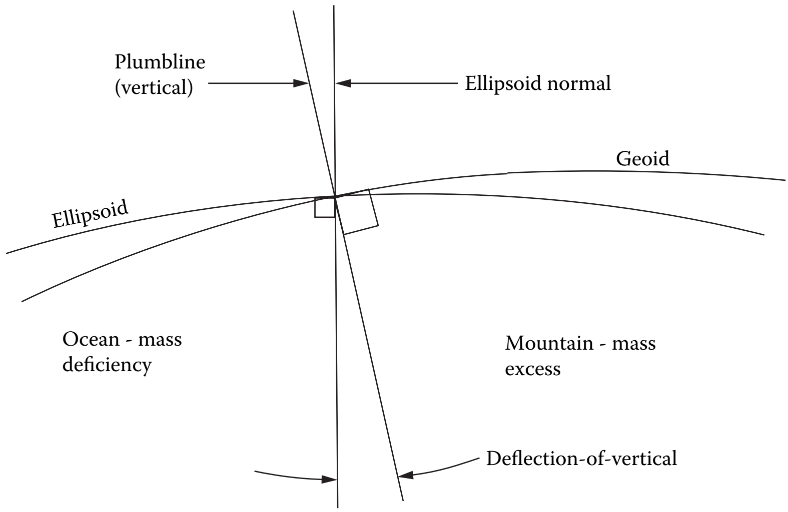


FIGURE 8.3 Deflection-of-the-Vertical

equations are used to connect the physical world with a mathematical representation. The mathematical ellipsoid normal is a computational standard and is perpendicular to the tangent to the ellipsoid at that point. In the physical world, the direction of the plumb line is the result of physical forces acting on an object at a point. Although the difference is not significant for many applications, the two are different and must be considered in geodetic applications. The direction of the normal at a point is well defined and computable in the ECEF environment, but, given the nonuniform distribution of mass within the Earth, the precise direction of the plumb line is less predictable. The irony of human experience is that we, standing erect, view the plumb line as being vertical. That is, we reference our view of the world to a changing feature that appears constant. In our view, up is always up. But, when we speak of the plumb bob being deflected by the mass of a mountain range, we are, in fact, referencing our perspective to the ellipsoid normal.

The difference between the ellipsoid normal and the vertical plumb line is called deflection-of-the-vertical and is realized in several ways. First, the ellipsoid surface is not necessarily parallel with the level surface—the geoid. Given the two surfaces are not coincident, except for where they cross, there must be a physical distance between them. That difference is called geoid height and is studied under the name of geoid modeling. Second, the angular amount by which the two surfaces are not parallel is given by the deflection-of-the-vertical and expressed in terms of a north/south component and an east/west component. Any measurement or observation made with an instrument having a level bubble on it (whether a carpenter's level, a surveyor's total station, a differential level, a telescope in an observatory, or an inertial measuring unit) is physically referenced to the local vertical. On the other hand, vectors obtained from GPS observations are referenced to the ellipsoid normal.

Specifically, astronomical observations of directions to stars are gravity based and yield astronomical positions, while GPS-derived positions are normal based and provide geodetic positions. The two systems are related by the Laplace equations based upon the deflection-of-the-vertical components at a point that, using Greek letters, are called

north/south: xi = ξ, and
east/west: eta = η.

The relationships between astronomical latitude, longitude, and azimuth and geodetic latitude, longitude, and azimuth are

$$\phi = \Phi - \xi \quad \text{latitude} \tag{8.3}$$

$$\lambda = \Lambda - \frac{\eta}{\cos\phi} \quad \text{longitude} \tag{8.4}$$

$$\alpha = A - \frac{\eta}{\tan\phi} \quad \text{azimuth} \tag{8.5}$$

where

φ = geodetic latitude,
λ = geodetic longitude,
α = geodetic azimuth,
Φ = astronomic latitude,
Λ = astronomic longitude, and
A = astronomic azimuth.

Equations 8.3–8.5 can be used two ways—if the geodetic latitude and longitude and the astronomic latitude and longitude are known, then deflections-of-the-vertical can be computed at that point. If the deflection-of-the-vertical components are known, then geodetic latitude, longitude, and azimuth can be computed from astronomic latitude, longitude, and azimuth and vice versa. Probably equation 8.5 is the most commonly used, where the astronomic azimuth (observed from sun shot or Polaris observation) is converted to the equivalent geodetic azimuth.

In the United States, deflection-of-the-vertical components are readily obtained from the NGS web site (http://www.ngs.noaa.gov) using the "deflect" program.

MEASUREMENTS AND COMPUTATIONS

One way to understand physical geodesy and the geoid better is to look at some of the physical quantities that can be measured and to describe the quantities that might be derived from those observations. Given recent advances in technology, there are changes in the combinations of what is measured and what can be reliably computed from those measurements. The goal is that, as technology and procedures evolve, the geoid can be located with greater certainty than in the past and the computational burden for spatial data users will be reduced.

Interpolation and Extrapolation

The following discussion should also be viewed in terms of estimating and the difference between interpolation and extrapolation. Ideally, a reliable observation is available at every point where such a measurement is needed. That is, however, rarely the case. Representative measurements are made, and other values are estimated—either by interpolation (often acceptable) or by extrapolation (often used as a last resort). The validity of any estimation hinges on factors such as:

1. What is the accuracy (uncertainty) of the available measurements?
2. Are the number and spacing (density) of the measurements consistent with observed rates of change? Would a greater density of measurements merely confirm what can be obtained by interpolation, or are more measurements needed to provide a better picture of the quantity being estimated?
3. What is the rate of change of the quantity being measured? Is the observed rate of change reasonable and/or consistent with anticipated changes based upon current theories and the models being applied?
4. How sensitive are the indirect values being computed to changes in the quantity being measured?

Geoid height cannot be measured directly but must be inferred or computed from other quantities. With regard to locating the geoid or determining geoid heights, the list of measurable physical quantities includes:

- Gravity:
    1. Magnitude
        A. Absolute
        B. Relative
    2. Direction (deflection-of-the-vertical)
- Tide levels (location of mean sea level)
- Differential levels (changes in orthometric height)
- Ellipsoid heights (distance from ellipsoid)
- Time—related to gravity and location of geoid (Kleppner 2006)

There is a direct correlation between deflection-of-the-vertical and changes in geoid height. Deflection-of-the-vertical values express the "slope" of the geoid at a given point with respect to the ellipsoid normal. In the past, geoid heights were computed from an accumulation of deflection-of-the-vertical determinations at stations in the national triangulation network where the observed astronomical position was compared to the computed geodetic position. Such dual-value stations are called Laplace stations.

Gravity

If a precise value of the magnitude of gravity were known at all points, the location and shape of the geoid could be computed without ambiguity. But gravity values are not known everywhere, and, because the value of gravity changes from place to place, obtaining sufficient high-quality gravity measurements for precise geoid computations is not feasible. Absolute gravity measurements at the one-part-per-billion level (of gravitational acceleration at the Earth's surface) are possible (NGS 2005), but they require expensive equipment and rigorous observing procedures. Relative gravity measurements are more popular because they are easier to make and less expensive to obtain. A network of relative gravity measurements needs to be attached to an absolute gravity network in order to obtain information that is useful for geodetic computations.

In addition to ground-based measurements, gravity measurements are made at sea (both on the surface and underwater) and in the air (in balloons, airplanes, and spacecraft). High-quality, land-based gravity measurements are typically more accurate because, while the measurement is being made, the gravity meter is motionless with respect to the Earth. Gravity measurements in a moving environment must be corrected for the location and motion of the platform—ship, airplane, and so on. There is a trade-off between more costly high-quality land-based gravity measurements (access to private property may also be an issue) and mobile gravity measurements obtained from a moving vehicle or platform.

Measuring the direction of gravity is distinctly different than measuring the strength (magnitude) of gravity. A simple plumb bob is probably the best example of determining the direction of gravity—no matter where you go, the plumb bob always points down! Other devices for measuring the direction of gravity include the level vial as used on a carpenter's level, the bull's-eye bubble as used on many surveying instruments, striding levels as used on precision theodolites, and, for high-accuracy applications, the undisturbed surface of a pool of mercury in a vacuum (precisely perpendicular to the local plumb line). Sensors in modern inertial measuring systems also detect the direction of gravity.

With regard to the direction of gravity, a challenge is to answer the question "With respect to what?" Ironically, the direction of gravity is always vertical! After all, that is the definition. But, due to land masses and the nonuniform distribution of density within the Earth, the direction of gravity changes from point to point, and ultimately the question is "What is the direction of gravity at a point with respect to the ellipsoid normal at the same point?" That difference is called the deflection-of-the-vertical and is not measured directly but is inferred from other measurements and computations. During the era of triangulation, it was commonplace to compute the difference between the astronomic position observed at a point and the computed geodetic position for the same point based upon some datum and initial point. Deflection-of-the-vertical components were derived from those differences and defined the slope of the geoid at that point. With a sufficient number of dual-value Laplace stations and associated deflections, the shape of the geoid in that area could be inferred. Given that the shape of the geoid is attached to some initial point where the geoid height is zero, then absolute geoid heights can be determined in other locations throughout the country.

The U.S. Army Map Service (1967) published a three-sheet map, "Geoid Contours in North American," based upon astrogeodetic deflections. The map shows 1 m contours throughout North America on the North American Datum of 1927. The geoid height was held to be zero ($N = 0.0$ meters) at the Initial Point, Meades Ranch, Kansas. Although the map is based upon astrogeodetic deflections, gravity data can be used to verify and/or augment the deflection-of-the-vertical computations.

Tide Readings

Tide gauge data have been collected at numerous locations by various organizations for many years. The First General Adjustment of geodetic leveling in the United States was published in 1900 and based upon mean sea level as determined by tide gauges in five locations. Subsequent leveling network adjustments (in 1903, 1907, and 1912) connected to more tide gauges, and, as discussed in chapter 7, the 1929 General Adjustment (which became known as the Mean Sea Level Datum of 1929) was based upon mean sea level as determined at twenty-six tide gauge locations—twenty-one in the United States and five in Canada (Berry 1976).

Differential Levels

As described by Zilkoski, Richards, and Young (1992), loops of precise levels (orthometric height differences) within the United States were observed in support of the readjustment and definition of the North American Vertical Datum of 1988 (NAVD88). The loop results are very good, but, because the tide gauges do not define the same geoid surface, the relative loop differences could not be attached to absolute tide gauge elevations without distorting the observed relative differences. Therefore, one station—Rimouski at Fathers Point, Quebec—was selected as the datum point, and all NAVD88 orthometric heights are stated with respect to that one arbitrary point. That being the case, it can be said that all NAVD88 orthometric heights are relative and that there is no basis for absolute vertical accuracy statements with respect to the geoid. Such a statement is consistent with conventional leveling standards (Bossler 1984) that are given in terms of relative differences. Admittedly, NAVD88 orthometric heights, being consistent with the definition of "absolute" in chapter 2 and being on the same datum, are often used as absolute values.

Ellipsoid Heights

With the advent of GPS and through the combined efforts of many people, the location* of the center of mass of the Earth is known within 1 cm (Schwarz 2005). Holding the center of mass as the origin and using the rectangular ECEF coordinate system

* By definition, the location of the center of mass of the Earth is fixed and does not move—land masses and points on the surface of the Earth move with respect to the center of mass. The computed relative location of points on the surface of the Earth and orbit of satellites are consistent within 1 cm.

as defined by the DOD, all spatial data computations within the birdcage of orbiting GPS satellites follow the time-honored rules and equations of solid geometry. In that computational environment, ellipsoid heights are a derived quantity based upon the geocentric *X*/*Y*/*Z* position of the point and the ellipsoid parameters chosen by the user. Several relevant issues are that spatial data components (coordinate differences) are derived from a variety of sensors and those relativistic effects should be modeled in the process of determining those spatial data components. Of course, the user must take responsibility for choosing and working on an appropriate datum such as NAD83, WGS84, or ITRF.

A related issue here is that GPS-derived ellipsoid heights are typically quoted as less precise than the horizontal position. One reason is that those results are based only on signals from satellites visible from one side of the world. More data are available. Taken as a whole, GPS is, among other things, a huge interpolation system for spatial data. Given that signals from orbiting satellites are transmitted to the Earth from all sides, it is conceivable, if data from all satellites are used simultaneously and if the worldwide GPS network is treated as a deformable solid, that the observed vertical (radial) dimension could turn out to be the strongest component of a GPS-derived position. That being the case, ellipsoid height can be determined very precisely for many points around the world (Burkholder 2003). The hypothesis is that absolute ellipsoid height can be determined routinely within a centimeter or less with respect to the Earth's center of mass. Of course, with the Earth's tides, a monumented point (or CORS position) on the Earth's surface will rise and fall twice a day as our moon goes around the Earth. A mean value of ellipsoid height for a point will be adopted and used for most purposes, but, much like polar motion components, the daily differences should be available to those needing them. Many local spatial data applications will continue using relative ellipsoid height differences attached to a network of precisely determined absolute ellipsoid height points. Tectonic uplift and subsidence are nonregular movements monitored by those having responsibility for maintaining and updating the National Spatial Reference System.

Comparing ellipsoid height with orthometric height at a point is one of the best ways to find the geoid height. The problem is that, although obtaining a reliable ellipsoid height at a point is fairly routine, rarely are there enough reliable high-order benchmarks available for making the comparison. Consider two extremes: if high-quality gravity values were known at a sufficient number of points in an area, it would be possible to compute geoid heights using gravity data. On the other hand, if sufficient high-quality orthometric heights were available in an area, it would be possible to compute geoid heights using equation 8.2 rewritten as $N = h - H$. The challenge is to find and use the best (most efficient, reasonable, and practical) combination of observations and procedures to determine acceptable geoid height values. Once primary reliable values for the geoid height are determined in a given area, other geoid heights in the same area can be interpolated using standard modeling techniques.

Time

Although this author will leave it to others, a recent hypothesis is that the absolute location of the geoid can be inferred from precise time measurement. In discussing "the great geoid search," Kleppner (2006, 11) states,

> In the not-too-distant future, our ability to compare atomic frequency standards and clocks at different laboratories will be limited by our knowledge of the geoid.
>
> The obvious way to deal with the geoid problem is to reverse the argument and employ the gravitational redshift to explore the geoid. If, for instance, one had a portable atomic frequency standard accurate to 1 part in 10^{18} and if it could be compared to a primary standard with the same accuracy, the position of the geoid could be independently and relatively quickly determined to 1 cm.

The point of discussing the measurements and computations related to the geoid height is to focus on the quantities that can be determined with the greatest certainty. From a geometrical geodesy perspective, it appears that the combination of precise ellipsoid heights and high-order orthometric heights is the best way to determine geoid heights. Gravity measurements are used as collaborating data to verify results. But, from a physical geodesy perspective, a worldwide gravimetrically derived geoid is used as the basis for developing geoid models such as Geoid03. Many more details are available from the NGS (2006a) web site. And, if using atomic clocks to locate the geoid independently ever comes to fruition, current geoid-modeling procedures may be significantly modified.

USE OF ELLIPSOID HEIGHTS IN PLACE OF ORTHOMETRIC HEIGHTS

Given the elusive nature of the geoid (Kleppner 2006) and given that ellipsoid heights are readily available in the user community, it has been proposed (Burkholder 2002, 2006; Kumar 2005) that ellipsoid heights be used in place of orthometric height for routine 3-D spatial data applications. Selling that change will be difficult because most people are comfortable with using sea level as a vertical reference. Sea level is a physical reference surface that people can see and understand. But, starting in 1973 with the change in the name of the vertical datum, the association of mean sea level with the zero elevation was broken, and with publication of the NAVD88 the zero elevation is no longer "connected" to the tide gauges. Technically, the change to using ellipsoid heights for orthometric heights can be managed effectively in the same way as any other datum upgrade. But, although selling that change may be difficult, there are several precedent-setting examples to be considered—including the definition of the North Pole and the definition of time.

The difference between the physical position of the North Pole (it moves) and the mathematically adopted Conventional Terrestrial Pole (CTP) is an important concept but used only by a relatively small number of people. However, polar motion is routinely factored into those computations needed to reconcile the physical position of the spin axis with the CTP, and GPS users the world over confidently use GPS-derived positions without needing to model polar wandering (Leick 2004). The fact that few people actually use polar motion coordinates does not diminish the scientific importance of determining those quantities with exactitude.

Time is another analogy. In years past, noon was defined as the instant the sun crossed one's local meridian and the beginning of each day was defined as midnight, the instant the sun crosses one's meridian on the other side of the world. This definition is simple, physical, and easily understood, and relies on the apparent motion of the sun across the sky. Two problems—the sun-meridian definition of noon is location dependent, and, throughout the year, the time interval from noon to noon varies. Using the noon definition of time, each railroad train station in the 1800s had its own version of correct time, and developing reliable train schedules was a real problem. The problem was solved (Howse 1980) by adopting a system of standard time zones in the United States in 1883. The worldwide system of time zones was adopted by the International Meridian Conference in Washington, D.C., in 1884. Yes, selling standard time zones to the public was an enormous task but, in hindsight, apparently well worth the effort.

The second time problem is measuring the uniformity with which time progresses. The equation-of-time is the difference between time as determined by the motion of the Earth and the uniform progression of atomic time. The ancient Greeks recognized the existence of the equation-of-time, but it was not until the late 1600s that John Flamsteed, the first royal astronomer of the Greenwich Observatory, quantified the equation-of-time. Since then, measurement of the uniform progression of time has evolved from the pendulum clock to the quartz crystal clock to atomic clocks capable of accuracies in the range of one part in 10^{14} (Jones 2000). Now most people take time zones for granted, and the fact that the sun crosses the local meridian before or after 12:00:00 noon is of little consequence. But, for scientific and other purposes, the equation-of-time and other time scale differences are known, documented, and used by those for whom the difference matters.

Similarly, the location of the center of mass of the Earth is easier to locate, is more stable than the geoid, and offers obvious advantages when used as a single reference for terrestrial 3-D spatial data. Making such a change would in no way diminish the importance of geoid-modeling research, but it would relieve the spatial data user community of a significant computational burden imposed by the continued use of separate origins for horizontal and vertical data. Issues associated with such a change include:

1. Water flows downhill. In most cases, ellipsoid height differences are sufficiently accurate to establish grades for highway alignments and gravity sewer flows. As an example, typical storm sewer manholes are spaced about 100 meters apart. If one assumes an acceptable as-built tolerance of 0.005 m for the invert orthometric height at each manhole (0.005' * $\sqrt{2}$ = .007 m), the tolerance for slope is arctan (.007 m / 100 m), or about 15 seconds of arc (deflection-of-the-vertical is less than 15 seconds of arc in most places). Given that engineering involves finding, documenting, and using acceptable approximations, a study needs to be conducted and published to identify the severity of slope for other conditions and to identify acceptable tolerances for other assumptions. Of course, more critical applications need to model

and accommodate differences between gravity-based measurements (e.g., differential leveling) and normal-based measurements (e.g., GPS). Examples include hydraulic grade lines (dynamic heights) over large areas, tunneling through a mountain from two ends, and establishing a true plane for atomic particle colliders. In such cases, geoid modeling should still be used by those for whom the difference matters.

2. In the continental United States, the geoid lies below the ellipsoid. If ellipsoid heights are used in place of orthometric heights, there are many places along the coast where one is clearly standing on dry ground but the ellipsoid height for the point is a negative number. We've learned to accept negative orthometric heights in places like Death Valley, California, but negative elevations on dry land along the coast are rather dramatic reminder that zero elevation is not the same as mean sea level.

THE NEED FOR GEOID MODELING

Before getting into the details of geoid modeling, it may be beneficial to look at several assumptions associated with using or choosing not to use geoid height.

1. For the first approximation, the difference between the ellipsoid and the geoid can be assumed to be zero—that is, ignoring geoid heights altogether. On a global basis, the error of such an assumption could be as much as 100 meters. In the United States the error of ignoring the geoid height is limited to about 50 meters. If the orthometric heights being used have a standard deviation of 100 meters or more, the noise of the data is greater than the feature being modeled, and it makes little or no difference whether one uses ellipsoid height or orthometric height.
2. Two other assumptions are that the world is flat (for a small area) and that the ellipsoid and the geoid are two parallel planes. Under this assumption, if the geoid height at one place is known, the geoid height at nearby locations is the same. For some applications this assumption may be appropriate, but there is a real danger that limits of the assumption may not be recognized. The user should be aware of unintended, possibly even severe, consequences.
3. A better assumption (but still assuming a flat Earth) is that the ellipsoid and geoid are nonparallel plane surfaces—that is, one is tilted with respect to the other. If geoid heights are known at three points (required to define a plane) in a specified area, other geoid heights in the same general area can be obtained by linear interpolation techniques. Extrapolation may provide reasonable answers in the same general area, but uncontrolled extrapolation is to be avoided. If geoid heights are known for more than three points scattered throughout the project area, a local "best-fit" geoid model may be the appropriate choice.
4. The best assumption underlying geoid modeling is that the two surfaces are both curved and not concentric. Curvature of the ellipsoid is mathematically well defined and can be computed. But, even though the geoid is a continuous surface, the geoid curves in an irregular manner, and writing a

mathematical function for the separation between the two curved surfaces involves complex modeling. The reader is referred to the NGS (2006a) web site for additional information on geoid modeling and the latest geoid model available to the user community.

In spite of the advantages of using ellipsoid heights for orthometric heights, that change may never occur. We need to recognize that current practices evolved to be what they are for specific reasons (we are where we are because of where we came from), and the prudent user needs to be familiar with current geoid-modeling practices.

In spatial data applications, geoid modeling is used primarily to relate ellipsoid height to orthometric height. As noted in the previous section, that application may be significantly reduced if ellipsoid heights are adopted and used in place of orthometric heights. A secondary application using geoid heights is reducing a horizontal distance at some elevation to either the geoid or the ellipsoid. Whether using the ellipsoid or the geoid as the computational surface, demand for geoid modeling is driven by the desire for physical measurements to be modeled at a level that preserves their geometrical integrity. The next section considers the impact of geoid height on reducing horizontal distance to the ellipsoid.

Typically, a slope distance is measured between a standpoint and a forepoint, reduced to an equivalent horizontal distance, and further reduced to the ellipsoid for use in geodetic computations (Burkholder 1991). One could say that geoid modeling is not required if it makes no significant difference whether orthometric height or ellipsoid height is used for the elevation reduction computation. But, we also need to acknowledge practices of the past as we make decisions about current standards, specifications, and procedures. For example, triangles and quadrilaterals in the national horizontal network defining the North American Datum of 1927 (NAD27) were computed on the geoid as if it were the Clarke Spheroid of 1866. Of course, both the shape of the ellipsoid and the location of the datum origin at Meades Ranch, Kansas, were selected so that, in the United States, the geoid and the ellipsoid were close to each other. That proximity mitigated the consequences of performing the computations on the geoid instead of the ellipsoid.

However, the North American Datum of 1983 (NAD83) was computed on the ellipsoid, not on the geoid. The shape and location of the GRS 1980 ellipsoid were chosen for a global best fit rather than a continental best fit as was done for the NAD27. The origin of the NAD83 was located at the Earth's center of mass (as determined at the time), and the shape of the GRS 1980 ellipsoid approximates the global shape of the geoid. An unintended consequence of a global best fit for the geoid is that differences between the ellipsoid and the geoid in North America are greater on NAD83 than they are on NAD27. Separately, the intended internal consistency of the NAD83 was to be at least one magnitude better than for NAD27. These factors, both singularly and collectively, mean that geoid height is an important consideration in distance reduction and should not be ignored in the geodetic computation of positions on the NAD83.

In an attempt to quantify the severity of using an orthometric height instead of ellipsoid height in the distance reduction, Burkholder (2004) examines the accuracy of the elevation reduction factor. Equation 8.6 (from that article) can be used to compute the elevation reduction factor standard deviation for any combination of values and is given as

$$\sigma_{ElevationFactor} = \sqrt{\left[\frac{h}{(r+h)^2}\right]^2 \sigma_r^2 + \left[\frac{-r}{(r+h)^2}\right]^2 \sigma_h^2} \quad (8.6)$$

where

h = ellipsoid height used in reduction,
r = radius of the Earth,
σ_h = uncertainty of ellipsoid height, and
σ_r = uncertainty of the Earth's radius.

The four examples in Table 8.1 were computed using equation 8.6 and are intended to show the following:

1. The uncertainty in ellipsoid height is a prominent factor in the reduction of a horizontal distance to the ellipsoid.
2. Ellipsoid height and the uncertainty in the Earth's radius both contribute to the uncertainty in the elevation reduction factor, but not much.
3. Using ellipsoid height instead of orthometric height or vice versa (ignoring a geoid height of fifty meters) may be acceptable at the 1:100,000 level.
4. Ignoring a geoid height of ten meters will affect results at the 1 ppm level.

But, the real question to be addressed is whether geoid modeling is needed at all. Because the presumed answer is "yes," the question deserves careful consideration. When looking at computation and use of spatial data components, the GSDM stores geocentric *X*/*Y*/*Z* coordinates and covariance values as the primary record of a point. A new point is established from an existing point using the 3-D forward (BK3) computation—either on the basis of GPS-derived $\Delta X/\Delta Y/\Delta Z$ or on the basis of local normal-based $\Delta e/\Delta n/\Delta u$ components rotated to $\Delta X/\Delta Y/\Delta Z$. The computation takes place in 3-D space, not on the geoid and not on the ellipsoid. No geoid modeling is required. The exception is if the slope of the geoid with respect to the ellipsoid normal is needed to correct vertical-based measurements to normal-based measurements. In

TABLE 8.1
Four Examples of Elevation Reduction Factor Uncertainty

Uncertainty of Height	Height of Standpoint	Uncertainty in Earth's Radius	Resulting Uncertainty Elevation Factor
1. 50 meters	500 meters	1,000 meters	1:127,460
2. 50 meters	2,000 meters	5,000 meters	1:127,457
3. 10 meters	500 meters	1,000 meters	1:637,267
4. 10 meters	2,000 meters	5,000 meters	1:629,882

that case, deflection-of-the-vertical data are needed, not the precise geoid height. If the geodetic line distance on the ellipsoid between the standpoint and the forepoint is desired, it can be computed from a geodetic inverse (BK19)—geoid height is not an issue.

As described in chapter 2, rectangular spatial data components are computed as coordinate differences in the geocentric rectangular coordinate system and rotated (without distortion or loss of integrity) to the local perspective for use as 3-D rectangular flat-Earth components. In that environment, horizontal distance from the standpoint to the forepoint lies in the local tangent plane through the standpoint. The 3-D azimuth (Burkholder 1997) from standpoint to forepoint is computed simply as arctan ($\Delta e/\Delta n$), and ellipsoid height is derived directly from the *X*/*Y*/*Z* coordinates (see equation 1.7). Geoid modeling is not an issue unless one needs to relate the third dimension to orthometrics heights on existing benchmarks. Two points:

1. Yes, orthometric heights are still used, and geoid modeling is important as a way to find orthometric height from ellipsoid height. But for many flat-Earth applications (e.g., RTK construction layout staking over limited distances), the Δu component of a local vector can be used as an orthometric height difference. If a curved-Earth height difference is needed, the ellipsoid height difference (Δh) of the vector will suffice unless the slope of the geoid in that area is quite severe and/or very high precision is required for vertical. In that case, geoid modeling will be needed.
2. Reducing a horizontal distance to the ellipsoid for geodetic computation is no longer needed. Instead, a geodetic 3-D forward computation (BK3) based upon slope distance, zenith direction, and azimuth of the line is used to compute *X*/*Y*/*Z* coordinates of the forepoint. *Latitude*/*longitude*/*height* of the forepoint are computed from the forepoint *X*/*Y*/*Z* values.

GEOID MODELING AND THE GSDM

An important concept is that the simplest model that supports the integrity of the data is the most appropriate model to use. The 1-D flat-Earth model for leveling is probably the simplest spatial data model and is used extensively. The 2-D flat-Earth model of rectangular plane coordinates is used all over the world for simple local applications. A local 3-D rectangular model is also used for applications such as describing condominium space, local area topographic maps, and construction stakeout. But, geographers quickly find themselves beyond the range of an acceptable flat-Earth model, and they utilize spherical latitude/longitude coordinates to express horizontal location. Orthometric height (or altitude) is used to describe vertical location. At an even higher level of complexity, geodesists, geophysicists, and other scientists need to use the flattened ellipsoid model in order to preserve geometrical integrity of measurements made all over the Earth. As noted earlier, the spherical and ellipsoidal Earth models involve the use of mixed units, that is, angular units for horizontal and length units for vertical. Map projections (see chapter 10) were invented as a way to represent a portion of the curved Earth on a flat map. That essentially solves the mixed-unit problem but map projections are severely limited by, among other reasons, the fact that a map projection is strictly 2-D. Spatial data are 3-D, and modern practice needs to combine horizontal and vertical into the same database.

Whether using the NAD83, WGS84, or ITRF datum, the GSDM defines any point within the birdcage of orbiting GPS satellites by a triplet of geocentric rectangular *X*/*Y*/*Z* coordinates. Furthermore, each stored point has a covariance matrix associated with it and interdependencies between point-pairs are obtained from the appropriate covariance submatrix. From these values stored in a BURKORD™ database, the user can select and compute derived quantities. In some cases, the assumptions are contained within the definition of the GSDM, while in other cases the user must specify additional assumptions about the derived quantities. And, in all cases, the computations are bidirectional, meaning data can be transformed either way without loss of geometrical integrity.

Examples include:

1. Quantities computed directly from the *X*/*Y*/*Z* coordinates and covariance matrix for a point are latitude, longitude, ellipsoid height, and standard deviations for the point in either the geocentric reference frame or the local reference frame. The local "up" standard deviation is the standard deviation of the ellipsoid height. The implicit assumptions are that the user has selected the datum and the ellipsoid parameters.
2. Other quantities that can be computed from the latitude and longitude of a point include UTM coordinates, state plane coordinates, or user-defined projection coordinates. The implicit assumptions are that the user is responsible for using appropriate transformation equations, using the correct units, and staying on the same datum. The local reference frame (*e*/*n*/*u*) standard deviations are not changed by such a transformation.
3. Orthometric height can be computed from the ellipsoid height using geoid modeling and a rearrangement of equation 8.2 as $H = h - N$. The geoid modeling program Geoid03 (and other versions) is described in the next section. The standard deviation of the computed orthometric height is based upon the standard deviation of the ellipsoid height and the standard deviation of the geoid height. It can be done either way, but this is one place where the user must exercise caution—the standard deviation of an absolute geoid height at a point is typically inferior to the standard deviation of the geoid height **difference** between points, as described later.
4. Other derived quantities are listed here for the sake of completeness but are discussed in more detail in chapter 10. Given that a user selects a pair of points, the derived quantities include the mark-to-mark distance, the 3-D azimuth from standpoint to forepoint, the zenith (or vertical) direction from standpoint to forepoint, the local tangent plane horizontal distance from standpoint to forepoint, and the standard deviation of each quantity. Furthermore, if appropriate covariance submatrices have been stored, both the network and local accuracies can be computed for these derived quantities. The explicit condition here is that all quantities are with respect to the ellipsoid normal at the standpoint.

5. Finally, if the user selects a P.O.B. datum point (see chapter 1), then any and all other points in the database can be viewed from that perspective, and those local coordinate **differences** can be used as local flat-Earth coordinates with respect to the origin selected by the user. Standard deviations of all derived quantities (direction, distance, area, volume, etc.) are available and routinely reported. The implicit assumptions are that all horizontal distances are within the tangent plane through the P.O.B. and that all directions are grid directions with respect to the true meridian through the P.O.B. A further assumption is that, unless modified by the user (see chapter 10), the "up" component is the perpendicular distance from the forepoint to the tangent plane through the P.O.B. As will be described in chapter 10, these derived values will see extensive use throughout the spatial data user community in a wide variety of applications.

USING A GEOID MODEL

The following procedures for determining orthometric heights via GPS and geoid modeling are similar, but not identical, to those procedures given by NGS in the draft publication "Guidelines for Establishing GPS-Derived Orthometric Heights Version 1.4" (NGS 2006b). Given that one is to fulfill the requirements of the NGS specification, those criteria supercede the suggestions offered here and should be followed. For purposes of local practical application, the following procedures have been proven to provide excellent results.

As implemented in Geoid03 and other geoid-modeling programs, the printout typically gives the geoid height to three decimal places of meters. The geoid height at that point is not necessarily within 1 mm, as implied by the printout, but the **difference** in geoid heights between two neighboring points does need that many decimal places. Stated differently, the relative accuracy is better than the absolute accuracy. That means the shape of the geoid is known better than the location of the geoid. The following procedure is recommended to take advantage of that characteristic of geoid modeling.

Given:

Known orthometric height at point A = H_A
GPS-based ellipsoid heights at points A and B = h_A and h_B
Geoid03 geoid heights at points A and $B = N_A$ and N_B
Fundamental relationships: $h = H + N$, $H = h - N$, and $\Delta H = \Delta h - \Delta N$

Problem: As illustrated in Figure 8.4, determine the orthometric height of point B using the orthometric height at point A, Geoid03 (or another geoid model), and the GPS vector from point A to point B.

Solution:

1. Determine the latitude, longitude, and ellipsoid height of point A and point B from the geocentric $X/Y/Z$ coordinates of the points. If the covariance

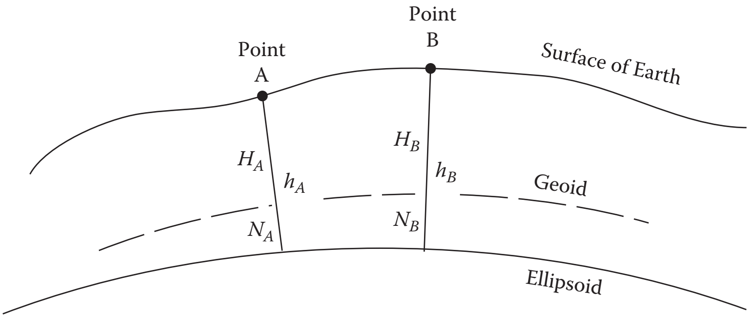


FIGURE 8.4 Determining Orthometric Height Using GPS and Geoid Modeling

matrix of each point is also stored, a program such as BURKORD™ will also provide, among others, the standard deviation of the height component.

2. Use Geoid03 and the latitude/longitude of each point to compute the geoid height for each point, N_A and N_B. Note that in the continental United States, the geoid lies below the ellipsoid and values of N, geoid height, are negative. The sign conventions for ellipsoid height, geoid height, and orthometric height are pretty much standard the world over, and the equations used herein are consistent with Figure 8.4, but the negative sign of N must be handled properly in the algebraic sense.
3. Combine the various components as follows:

$$\Delta h = h_B - h_A \text{ (from GPS and/or geocentric } X/Y/Z\text{)}$$

$$\Delta N = N_B - N_A \text{ (from Geoid03 or similar model)}$$

$$\Delta H = \Delta h - \Delta N$$

$$H_B = H_A + \Delta H = H_A + (h_B - h_A) - (N_B - N_A) \tag{8.7}$$

Computing the standard deviation of the orthometric height at point B is a natural extension of this discussion. A detailed example of using GPS to establish a reliable NADV88 elevation on a HARN station, along with standard deviation, is posted at http://www.globalcogo.com/ReilElev.pdf.

Geoid03 is available gratis from the National Geodetic Survey (NGS) web site, http://www.ngs.noaa.gov.

REFERENCES

American Society of Civil Engineers, American Congress on Surveying & Mapping, and American Society of Photogrammetry & Remote Sensing. 1994. *Glossary of the mapping sciences*. New York and Bethesda, MD: American Society of Civil Engineers, American Congress on Surveying & Mapping, and American Society of Photogrammetry & Remote Sensing.

Berry, Ralph Moore. 1976. History of geodetic leveling in the United States. *Surveying & Mapping* 36 (2): 137–53.

Bossler, John (chairman). 1984. *Standards and Specifications for Geodetic Control Networks*. Federal Geodetic Control Committeee, Rockville, Maryland.

Burkholder, Earl F. 1991. Computation of level/horizontal distance. *Journal of Surveying Engineering* 117 (3): 104–116.

———. 1997. Three-dimensional azimuth of GPS vector. *Journal of Surveying Engineering* 123 (4): 139–46.

———. 2002. Elevations and the global spatial data model (GSDM). In *Proceedings of the 58th Annual Meeting of the Institute of Navigation*, Albuquerque, NM, 24–26 June.

———. 2003. The digital revolution begets the global spatial data model (GSDM). *EOS, Transactions, American Geophysical Union* 84 (15): 140–41.

———. 2004. Accuracy of elevation reduction factor. *Journal of Surveying Engineering* 130 (3): 134–37.

———. 2006. The digital revolution: Whither now? *GIM International* 20 (9): 25–27.

Federal Register. 1973. 38 (94): 12840.

Howse, Derek. 1980. *Greenwich time and the discovery of longitude*. Oxford: Oxford University Press.

Jones, Tony. 2000. *Splitting the second*. Boston: Institute of Physics Publishing.

Kleppner, Daniel. 2006. Time too good to be true. *Physics Today*, March, 10–11.

Kumar, Muneendra. 2005. When ellipsoidal heights will do the job, why look elsewhere! Paper presented at the American Congress on Surveying & Mapping annual meeting, March, Las Vegas, NV, http://www.acsm.net/kumar05.pdf.

Leick, Alfred. 2004. *GPS satellite surveying*. New York: John Wiley.

Meyer, Thomas H., Daniel R. Roman, and David B. Zilkoski. 2004. What does height really mean? Part I: Introduction. *Surveying and Land Information Science* 64 (4): 223–33.

National Geodetic Survey (NGS). 1986. *Geodetic glossary*. Rockville, MD: National Geodetic Survey, U.S. Department of Commerce. http://www.ngs.noaa.gov/PUBS_LIB/DraftGuidelinesForEstablishingGPSDerivedOrthometricHeights.pdf.

———. 2005. Gravimetry: Absolute gravity. http://www.ngs.noaa.gov/GRD/GRAVITY/ABSG.html.

———. 2006a. Geoid03. http://www.ngs.noaa.gov/PC_PROD/Geoid03/.

———. 2006b. Guidelines for establishing GPS-derived orthometric heights (standards: 2 cm and 5 cm) version 1.4. http://www.ngs.noaa.gov.

Schwarz, K-P. 2005. Personal communication. University of Calgary, Alberta, Canada.

U.S. Army Map Service. 1967. Geoid contours in North America [map]. Washington, DC: U.S. Army Corps of Engineers.

Zilkoski, David B., John H. Richards, and Gary M. Young. 1992. Results of the general adjustment of the North American Vertical Datum of 1988. *Surveying & Land Information Systems* 52 (3): 133–49.

9 Satellite Geodesy and Global Navigation Satellite Systems (GNSS)

INTRODUCTION

Goals of satellite geodesy include using Earth-orbiting satellites to determine the size and shape of the Earth, to obtain a greater understanding of the Earth's gravity field, and to define the position or movement of points on the Earth or in near space. The first two goals are primarily scientific in nature, but the third goal is more applications oriented and of interest to many more people. The utility of position fostered by the pervasive use of Global Positioning System (GPS) technology in various applications means that many spatial data users are interested in learning more about fundamental concepts of satellite positioning. Many principles and systems are involved in meeting those goals, and, because there are so many details, this chapter should be considered as an overview.

Satellite-positioning technology evolved during the Cold War—primarily following the launch of Sputnik I by the Russians in 1957. Since then, various satellite systems have been used for positioning—the TRANSIT Satellite System, the NAVSTAR system of GPS satellites, the Russian Global Navigation Satellite System (GLONASS) system, and, finally, the GALILEO satellite-positioning system being developed by the Europeans. The purpose of including a chapter on global navigation satellite systems (GNSS) in this book is to make the point that the GSDM provides an appropriate bridge between the exacting rigor of the methods used by the "rocket scientists" and the much simpler flat-Earth methods typically employed in the spatial data user community.

Spatial data were defined in chapter 2 as distances between the endpoints of a line in 3-D space. As such, the definition is rather meaningless until such data are expressed with respect to a coordinate system—local, geodetic, or geocentric. The GSDM includes each of those coordinate systems and defines a common geometrical computational environment for all geospatial data users, including those who develop satellite-positioning systems, those who collect and process physical measurements to obtain coordinates, those who use spatial data for an endless array of applications, and those who manage the collection, storage, manipulation, and use of geospatial data.

"The Global Positioning System: Charting the Future" (National Academy of Public Administration [NAPA] 1995) is a report prepared by the National Academy of Public Administration for the U.S. Congress and the U.S. Department of Defense that documents the development, applications, and future of GPS. In the executive summary, the report predicts worldwide revenue for GPS-related products and services to exceed $30 billion by 2005. Current estimates of global economic impact for GPS and spatial data are difficult to find, but Navteq, a firm devoted to building and marketing a global geospatial database was purchased in October 2007 by Nokia for $8.1 billion and a GPS World (2007, 34) information item states that GPS purchases are expected to generate $4.1 billion in sales in 2007. The geospatial data arena includes GIS, GPS, and various other technologies. The same arena includes many talented persons who are experts at building measurement systems and generating quality 3-D spatial data.

"Geographic Information for the 21st Century" (NAPA 1998) is a report also prepared by the National Academy of Public Administration—in this case, for the U.S. Bureau of Land Management, U.S. Forest Service, U.S. Geological Survey, and National Ocean Service—that is a formal study of the civilian federal surveying and mapping activities. This arena includes many talented persons who use the measurement systems and related technology to generate and manipulate 3-D spatial data. But, perhaps more significantly, this arena also includes countless applications of spatial data used within the context of the modern Geographic Information System (GIS).

Even though one could say the focus of those two publications is very different, there is an enormous overlap in that both rely on a shared foundation of 3-D spatial data. Given the impact of the digital revolution (Burkholder 2003), spatial data are now characterized as digital and three-dimensional (3-D) and modeled by the GSDM. Built on the premise of a single origin for 3-D data, the GSDM simultaneously accommodates the rigorous methods of the GPS arena and the flat-Earth practicality of GIS applications. Establishing and tracking spatial data accuracy efficiently comprise an added bonus for all users.

Entire books are written on satellite geodesy and the processes used to convert physical measurements into spatial data components. The interested reader is referred to comprehensive texts such as Leick (2004), Seeber (1993), or Hoffman-Wellenhof, Lichtenegger, and Collins (1994) for more details. Current magazines such as *GPS World*, *Inside GNSS*, and *Geospatial Solutions* provide additional information, and web searches on GNSS will reveal other sources—especially for information on the newer satellite-positioning systems. While not exhaustive, material in this chapter is intended to provide context for spatial data users who wish to gain a better understanding of GNSS satellite data and how they are collected and used. When physical observations are reliably processed to the point of being included in the GSDM as geocentric $X/Y/Z$ coordinates on a specific datum, then others should be able to use those spatial data with confidence.

Although the Russian GLONASS system and the European GALILEO system both deserve consideration, GPS is the primary satellite system discussed here. Coverage of GLONASS and GALILEO is included by extension of GPS concepts. That is done because the focus of this book is on the fundamental geometrical model for spatial data, and the objective of this chapter is to describe how spatial elements and components are obtained from satellite data. As related to satellite geodesy and GNSS positioning, the following points are important:

- As stated in chapter 7 on datums, spatial data users encounter both absolute positions as represented by *X*/*Y*/*Z* coordinates and networks of relative positions built from observed GPS baselines. Absolute and relative concepts are both considered in this chapter.
- Unlike traditional ground-based measurements, spatial data components derived from satellite data are computed indirectly from measurements of physical quantities such as voltage, current, time, and phase shifts. Direct measurements of angles and distances are not made by satellite geodesy or GNSS surveying. Instead, by modeling the physical and geometrical circumstance of each measurement, coordinates and spatial data components are computed from the raw data. Those components are analyzed for quality, adjusted as needed for geometrical consistency, and attached to previously "fixed" points, and the new points are either used immediately or stored as coordinates in a database. Information from the database is then used to derive other quantities such as the distance and direction between points, areas, volumes, velocities, and so on.
- When discussing layout for construction projects, the traditional procedure is to lay out angles and distances relative to control points previously established on-site. GPS may have been used to establish those control points, and construction drawings may show new features located with respect to such points. Those traditional layout procedures may be used for years to come, but with the newer technology of GPS base station networks and real-time-kinematic (RTK) surveying procedures, the control monuments are effectively taken to be the satellites in the sky, and, using satellite orbits as the control points, stakes are established according to construction drawings. Carried further, the construction stake is eliminated entirely because the design feature is defined in an electronic file, the file is loaded into a computer carried on board the construction scrapper or bulldozer, a GPS unit receives signals from the satellites, and the current position of the cutting blade is relayed to the onboard computer, which computes the difference between "actual" and "design" locations. The equipment operator, guided by an electronic display, cuts or fills according to the instructions shown on the display—all in the comfort of the heated or air-conditioned cab.
- Each spatial data component, whether measured directly or indirectly, has a standard deviation associated with it. And, if correlation exists between components, the appropriate covariance matrix should be part of the record for that point. Given that covariance information for each point and between points is stored in a 3-D database, the standard deviation of any derived quantity is readily available via the stochastic model component of the GSDM.
- The quality of spatial data is not established just because it came from satellites, GPS, or other measuring systems. Meta data, statistical verification, quality control, or personal testimony may be needed to "prove" the quality of spatial data. However, once the quality of spatial data is established, reliable information management procedures are essential for preserving the value of that information. The GSDM provides for efficient numerical storage of stochastic information for each point in the database.

- The GSDM provides a standard interface between two categories of activities: those efforts devoted to generating reliable spatial data and those efforts associated with using spatial data. Although some people undoubtedly operate in both categories, the goal in developing the GSDM was for the designers and builders of measurement systems to know specifically how far the processing needs to go until the spatial data can be stored and/or turned over to the user. By the same token, spatial data users deserve to know with assurance that data from a given measurement system and/or database conform to an underlying standard and that reliable information on spatial data accuracy is readily available.

BRIEF HISTORY OF SATELLITE POSITIONING

It is generally agreed that the space age dawned in October 1957 with the launch of the Russian Sputnik satellite—the first to achieve Earth orbit. Scientists at Johns Hopkins University were able to track the signal broadcast by the satellite and, noting the Doppler effect, were able to reconstruct the orbit of the satellite. The process was then inverted, making it possible to compute the position of a receiver on the ground by knowing the satellite orbit and observing the Doppler shift of the signals received from the satellites passing overhead. That was the basis for development of the U.S. TRANSIT satellite system, which was first funded in December 1958, operational in 1964, and released for commercial use in July 1967—all within a 10-year time frame.

Since the launch of Sputnik I in 1957, many spacecraft have been launched for various purposes: research, experimentation, communication, navigation, mapping, exploration, and, yes, even sending humans to the moon. An early tabulation of objects in orbit and decayed objects gives dates and other details of launch and decay for several hundred satellites between 1957 and 1963 (Mueller 1964, sec. 2.64).

With regard to geodesy, passive satellites and active satellites have both been used beneficially. The Echo 1 and Echo 2 satellites were large metallic balloons that could be seen from the Earth on clear nights in the early 1960s. Multiple precisely timed photographic exposures generated images (small dots) of the satellite moving across a background of stars. By synchronizing exposures recorded on glass plates at widely separated locations, it was possible to work out the geometrical separation of the observing stations. This ability to compute a geometrical connection from one continent to another represented significant progress for geodesy. Such use of satellites for positioning was sufficiently successful to justify the launch of a dedicated geodetic balloon satellite in 1966 called PAGEOS (*PA*ssive *GEO*detic *S*atellite) that was used to establish a geometrical worldwide network of forty-five stations observed with BC4 cameras over a period of about 6 years.

Geodetic research has both contributed to and benefited from various satellite programs. But geometrical geodesy has probably benefited most from two satellite systems that were designed and built for navigation purposes—the TRANSIT system and the NAVSTAR GPS. In the early 1960s, the U.S. Polaris submarine fleet relied heavily upon inertial navigation instruments that needed periodic position updates, typically based upon astronomical observations. Such an update could be done only during clear weather and not if the vessel was submerged. After the TRANSIT navigational system became operational in 1964, a position update could be obtained 24 hours a day, rain or shine, and, in the case of a submarine, only required exposure of a radio antenna instead of the vessel.

Decommissioned in 1996, the TRANSIT system was released for commercial use in July 1967 and, in addition to military applications, was used extensively worldwide for both sea and land navigation until replaced by GPS. Recognizing that high-quality geodetic positions could be obtained from Doppler measurements, a number of companies developed instruments designed for land-based surveying applications. In addition to private sector use of Doppler positioning, the BLM made extensive use of the Magnavox MX 1502 Doppler receiver during the 1970s and 1980s; and, from 1973 to 1978, the U.S. National Geodetic Survey (Schwarz 1989) collected Doppler data at approximately 600 stations in all fifty states in support of the readjustment of the North American Datum of 1983 (NAD83). Additional information on development and use of the TRANSIT system can be found in Hoar (1982) and Stansell (1978).

The DOD first started development on the *NAV*igation *S*atellite *T*iming *A*nd *R*anging (NAVSTAR) satellite system in 1973 and called it the Global Positioning System, or GPS. GPS was developed for the purpose of providing timely, reliable positioning and navigation support for military activities all over the world. The first GPS satellite was launched in 1978, and, after completing a constellation of satellites that provided global coverage 24 hours per day, initial operational capability (IOC) for GPS was declared on December 8, 1993. Full operational capability (FOC) was declared on July 17, 1995.

GPS consists of three segments—the space segment, the control segment, and the user segment. The space segment consists minimally of twenty-four satellites in six different planes, with each plane inclined 55° to the equator. Each satellite orbits the Earth twice a day at an altitude of 20,183 km and broadcasts information on two frequencies, L1 and L2. The L1 carrier frequency is 1575.42 MHz and has a wavelength of 19 cm, while the L2 carrier frequency is 1227.60 MHz and has a wavelength of 24 cm. Information about the satellite clock performance, data on the health and status of each satellite, and a prediction of (ephemeris for) each satellite orbit are modulated onto the carrier frequencies broadcast by each satellite. Additional information and operational details on the GPS can be accessed from http://www.navcen.uscg.gov/gps/ and other sites.

The control segment is wholly owned, controlled, and operated by the DOD, and includes six tracking stations located around the world—Cape Canaveral, Florida; Hawaii in the Pacific Ocean; Ascension Island in the South Atlantic Ocean; Diego Garcia in the Indian Ocean; Kwajalein in the North Pacific Ocean; and Colorado Springs, Colorado. Data from all tracking stations are transmitted to the Master Control Station in Colorado Springs, where the data are processed and the predicted orbit (ephemeris) for each satellite is computed. Ephemerides, clock corrections, and other messages are then transmitted back to each GPS satellite once a day (optionally more often) from one of three upload stations located at Ascension, Diego Garcia, and Kwajalein.

TABLE 9.1
ECEF Coordinates of DOD Global Control Stations

Station	X	Y	Z (meters)
Colorado Springs	–1,248,597.295	–4,819,433.239	3,976,500.175
Ascensión	6,118,524.122	–1,572,350.853	–876,463.990
Diego García	1,916,197.142	6,029,999.007	–801,737.366
Kwajalein	–6,160,884.370	1,339,851.965	960,843.071
Hawaii	–5,511,980.484	–2,200,247.093	2,329,480.952
Cape Canaveral	918,988.120	–5,534,552.966	3,023,721.377

The WGS84 ECEF coordinates of the electrical center of each antenna for the six tracking stations as determined for the epoch G1150 are as shown in Table 9.1 (National Imagery and Mapping Agency 1997).

The user segment consists of all those persons and organizations that collect the GPS signal from the satellites and use it for an increasing variety of applications. From the user perspective, GPS is a passive system in that the end user only receives signals from the satellite. However, with the evolution of technology, it could be said that some GPS receivers are active because the signal received at a given location is immediately rebroadcast to other GPS units operating in the same local area for purposes of performing RTK surveys. At some continuously operating stations (CORS), the GPS signal is collected and retransmitted via radio, cell phone, or the Internet. Even so, GPS is still considered a passive navigation, timing, and positioning system. The end user does not transmit signals to the GPS satellites.

During the 1980s, while the GPS constellation was still being developed, manufacturers began building GPS receivers for civilian use. Two modes of operation were developed: one based upon the coarse acquisition (C/A) code modulated onto the L1 frequency and the other based upon observing the phase shift of the fundamental L1 carrier frequency as received at two separate locations. Using information broadcast by the GPS satellites, the autonomous location of a C/A code GPS unit could be determined within about 100 meters worldwide. In the early years of GPS, the DOD purposefully degraded the C/A code signal to deny ultimate accuracy to nonmilitary users. That policy of selective availability (SA) was discontinued May 1, 2000, and since then, the autonomous accuracy of a C/A code receiver worldwide is within about 10 meters. With enhancements such as augmentation or using differential corrections, submeter accuracy can be achieved using C/A code receivers.

On the other hand, it might seem that the carrier phase mode of GPS operation was developed specifically for surveying and mapping because, using GPS, it is now possible to determine the location of an unknown point with respect to a known point very precisely (within mm or cm) even though the stations may be separated by 20, 50, or even 100 or more kilometers. Since 1985, the use of GPS has completely revolutionized the surveying and mapping professions. Using GPS, the surveyor no longer needs intervisibility between ground points and no longer needs to travel to a remote mountaintop to gain access to the NSRS. In fact, if a modern surveyor is working in an area covered by a GPS CORS network (see chapter 11), he or she takes a portable pole-mounted unit to any point having sky visibility and determines the position of a point within centimeters (or less) in real time. The overall point of this discussion is that by the time GPS was declared operational in 1993, civilian uses for the GPS signal had effectively outstripped military applications, and now the utility of GPS is taken for granted by many users. Depending upon the application, GPS receivers (code phase, carrier phase, or both) are now relatively inexpensive and are being used beneficially worldwide by novice and expert alike.

The Russians (formerly, the Soviet Union) have also built a satellite-positioning system called the GLONASS that is very similar to GPS but with specific differences. For example, the GLONASS satellites are closer to the Earth, and their orbit period is slightly shorter than the GPS orbit period. GPS satellites all broadcast on the same two frequencies, and each satellite is identified by a unique code. The GLONASS satellites broadcast on different frequencies. The first GLONASS satellite was launched n October 1982, and the system was declared operational on September 24, 1993. Various manufacturers either build or plan to build equipment capable of using signals from GPS, GLONASS, and/or GALILEO satellites.

More recently, the European Space Agency has committed to building, launching, and operating a satellite-positioning system called GALILEO, which will both complement and provide competition for the two existing systems. The first GALILEO satellite was launched in December 2005. Following testing and additional development, initial operational capability is optimistically speculated to occur about 2010. Many details are yet to be determined, but the GALILEO system is being developed as a subscription-based system, whereas the GPS and GLONASS systems are supported respectively by the U.S. and Russian governments with no charges made for using the signals.

MODES OF POSITIONING

Many variables need to be considered in using satellites for positioning. Some of the most obvious variables are the instantaneous location of each satellite, the time each signal is broadcast or received, the number of satellites broadcasting signals, the frequencies of signals being broadcast, the location of each receiver, the number of receivers collecting data, and the atmosphere through which the signals travel. Even with all those variables, GNSS positioning is ultimately based upon some combination of three physical concepts: elapsed time, Doppler shift, and interferometry.

ELAPSED TIME

Distance is the product of rate and time. The time interval for the transit of a signal from a satellite to the receiver is measured very precisely. The speed of light (radio signal) is modeled for the intervening atmosphere and the distance from satellite to receiver is computed as the product of time interval and rate. Distances from a minimum of three satellites are needed to compute an intersection in three-dimensional space on or near the Earth. But, since a small correction to the receiver clock is also required, a minimum of four satellites must be observed—three for position and one for a clock correction. If signals from additional satellites are available, a position may be determined quicker and with greater accuracy.

Doppler Shift

As described by Christian Doppler (1803–1853) in the 1840s, the frequency received at a given place depends upon the frequency transmitted and whether the source and receiver are stationary with respect to each other, moving closer together, or moving farther apart. Standing at a train station and listening to the whistle of a train going by is often used as an example. To a person riding on the train while the whistle sounds, the pitch heard is uniform and continuous (no movement between the source and the receiver). That is the trivial case. The sound heard by a person standing on the platform is of more interest. If the train is approaching the station, the frequency heard is higher than the frequency broadcast because the distance between the source and receiver is decreasing while the sound is traveling. On the other hand, if the train is going away from the station, the frequency heard is lower than the frequency broadcast. That is because the distance is growing larger while the sound is traveling from the source to the receiver.

GPS satellites broadcast a steady precise frequency. As illustrated in Figure 9.1, the frequency received on the ground is higher or lower than the transmitted frequency depending upon the location of the observer and whether the distance between the satellite and observer is decreasing or increasing. Not surprisingly, the instant at which the two frequencies match—the minimum distance between satellite and receiver—can be identified very precisely. That instant figures prominently in the calculations of the observer's position using Doppler data. But, it takes many passes of a TRANSIT Doppler satellite to determine an unknown position precisely. It is far easier to use Doppler data to verify the integrity of a computed solution than it is to compute an unknown position solely using Doppler data. Elapsed time and Doppler data are often used in combination to determine an autonomous position.

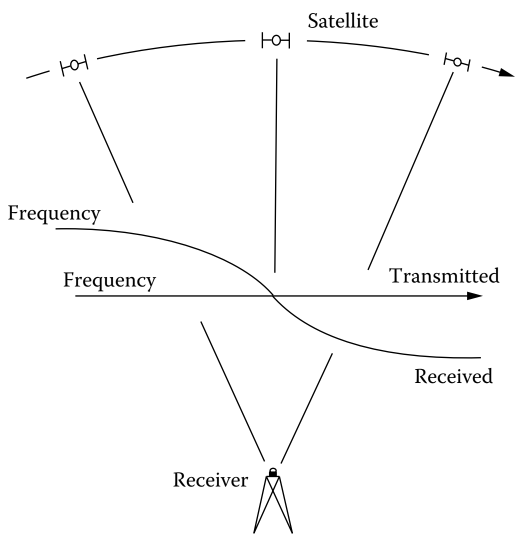


FIGURE 9.1 Example of Doppler Shift for Signal Received

INTERFEROMETRY

Using the concept that light (or radio signals) can be represented by a sine wave, interferometry is the term used to describe both the constructive and destructive interaction of two waves arriving simultaneously at a single location. When the signals are in phase, a double "high" is recorded. On the other extreme, when the phase of one signal is shifted 180° with respect to the other, the high of one signal cancels out the low of the other. The pattern of the changing phase shift, double-high-to-zero-and-back, is driven by a *difference* in the distance traveled by the signals and is itself a sine wave. Figure 9.2 illustrates the interference of light waves after passing through two slots in a barrier.

Interferometry concepts are used in processing GPS carrier frequency measurements. In general, the L1 signal is broadcast from a satellite and collected by a GPS receiver. The distance from the satellite to the receiver is a huge number of 19 centimeter wavelengths plus a fractional part of a wavelength. The large number of integer wavelength is not known when the observations start, and it is called the integer ambiguity. However, as time progresses, the receiver keeps track of (counts) the wavelengths received and, at specified intervals (1 second, 5 seconds, 15 seconds, etc.), records the phase shift (fractional wavelength) of the incoming signal. If the signal is interrupted, it is said the receiver "loses lock" and the count of full wavelengths must start over. Such an interruption is called a cycle slip in the data. Uninterrupted signals collected simultaneously at two GPS carrier phase receivers over a period of time (e.g., 5 to 60 minutes) from four or more satellites are used to compute a three-dimensional space vector between receivers in terms of $\Delta X/\Delta Y/\Delta Z$ components in the ECEF geocentric coordinate system. And, because each 19 centimeter wavelength can be resolved into one hundred fractional parts, it is said that the ultimate resolution of a carrier phase GPS measurement is about 1.9 mm.

As described in the next section, a C/A code (elapsed time) observation typically provides an economical autonomous *absolute* position with respect to a specific datum using one receiver. On the other hand, interferometric carrier phase processing requires data from two receivers, and the resulting vector between points is a *relative* measurement of one point with respect to another. Carrier phase GPS instruments are often referred to as survey grade instruments and typically cost more than C/A code receivers.

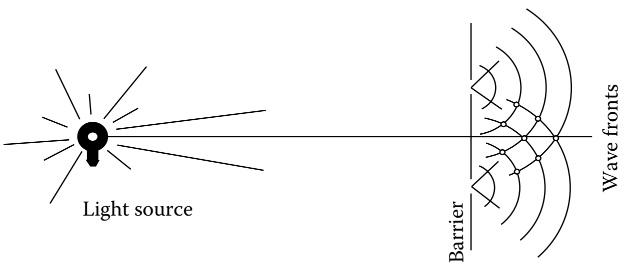


FIGURE 9.2

SATELLITE SIGNALS

Development of the NAVSTAR GPS grew out of a concept known as Very Long Baseline Interferometry (VLBI), which consists of collecting random pattern radio signals from distant quasars. Quasars are located extremely far from our solar system and emit radio waves that are received here on Earth. Two characteristics of interest are the pattern of signals received and the direction to the quasar. In order to determine the precise direction to a quasar, the same pattern must be received at widely separated locations and correlated according to their respective arrival times (as determined by very precise atomic clocks). Once the direction to a quasar is known, the process can be inverted to determine the 3-D space vector between receivers (Earth-based radio telescopes). Seeber (1993) reports a standard deviation of 1.7 cm on a 5,998 km baseline between the United States and Germany on data collected between 1984 and 1991.

The GPS is different from VLBI in that

- the source of the GPS signal is not from a fixed direction in deep space but from one of many satellites orbiting the Earth, and
- the GPS signal is not a random radio noise pattern but consists of very stable frequencies that are coded with important data related to the orbit, performance, and health of each satellite.

Accurate timing is the heart of GPS positioning. Jones (2000) describes the development of atomic clocks: hydrogen maser (stable to $1{:}10^{15}$), cesium (stable to $1{:}10^{14}$), and rubidium (stable to $1{:}10^{13}$). Although hydrogen maser clocks demonstrate superior performance, cesium clocks are preferred over hydrogen maser clocks for their longer term stability, and rubidium atomic clocks are preferred for GPS satellites due to their adequate performance, compact size, and comparatively lower cost.

The NAVSTAR satellite system also provides an enormous benefit to everyone by providing continuous access to accurate atomic time all over the world. GPS time began at 0^h Coordinated Universal Time (UTC) on January 6, 1980, and progresses at the same rate as the International Atomic Time (TAI) scale. UTC is the time scale used all over the world by the general population and runs at the same TAI rate. Earth's rotation on its axis is quite regular and was used as a time standard before the era of atomic clocks. Now, however, because the rate of atomic clocks is more uniform than the rotation of the Earth, UTC is modified by a leap second from time to time to keep midnight at midnight. Therefore, GPS time differs from UTC by an integer number of seconds—14 seconds as of January 1, 2006 (http://tycho.usno.navy.mil/gpstt.html). Additional information on atomic time and time scales is also available from Leick (2004), Kleppner (2006), or a web search.

Details of the electromagnetic spectrum are listed in Table 9.2 and include gamma rays, X-rays, ultraviolet light, visible light, infrared rays, microwaves, and radio waves. The GPS signal is located in the radio wave portion of the spectrum. During World War II, the radar band portion of the spectrum was assigned capital-letter designators for various wavelengths. The range of frequencies from 1,000 MHz to 2,000 MHz was called the L-band and includes both the GPS L1 and L2 frequencies. The GPS signal structure is based upon a fundamental oscillator

TABLE 9.2
Electromagnetic Spectrum

	Gamma Rays	X-Rays	Ultraviolet Rays	Visible Light	Infrared Rays	Radio Waves		
						(micro)	GPS L1 / L2	
Hertz	10^{21}	10^{18}	10^{15}		10^{12}	10^{9}	10^{6}	10^{3}
Wavelength	10^{-14} m	10^{-11} m	10^{-8} m		10^{-5} m	10^{-2} m	10^{1} m	10^{4} m

frequency of 10.23 MHz. The L1 frequency of 1,575.42 MHz (wavelength = 19.0 cm) is obtained as 154 times the fundamental frequency, and the L2 frequency of 1,227.60 MHz (wavelength = 24.0 cm) is 120 times the fundamental frequency. Given the underlying carrier frequencies, the C/A code is modulated onto the L1 frequency at 1.023 MHz, and the precision (P) code is modulated onto both the L1 and L2 carrier frequencies without alteration (i.e., at the original 10.23 MHz). The navigation message containing the broadcast ephemeris for each satellite, GPS time, and other system parameters (health, etc.) is modulated onto both L1 and L2 frequencies at a rate of 50 bits per second (bps).

The original intent was for all users to have access to the standard positioning service (SPS) based upon the C/A code on L1. The P-code is modulated onto both L1 and L2 frequencies and supports what is known as the precise positioning service (PPS). But, access to the P-code is reserved for military users. In addition to civilian users not having access to the P-code, the L1 signal was purposefully degraded so that C/A code users could not count on autonomous stand-alone positioning accuracy any better than about 100 meters. This policy of SA was intended to provide the U.S. military access to the full capability of the GPS system while denying ultimate functionality to others—especially users hostile to interests of the United States. As it turns out, that intent of SA was thwarted by the invention and adoption of differential positioning techniques that obviated the impact of SA. Therefore, SA was formally discontinued on May 1, 2000, at the direction of the president of the United States.

A second security measure is known as anti-spoofing (A-S), which "guards against fake transmissions of satellite data by encrypting the P-code to form the Y-code." A-S was exercised intermittently through 1993 and implemented on January 31, 1994 (http://tycho.usno.navy.mil/gpsinfo.html, May 2006).

Enhancements have been made to both the hardware and the signal structure since the first GPS satellites were built and launched. For example, Leick (2004) describes A-S and the P(Y)-code as related to military uses and notes that manufacturers have developed proprietary methods that make the P(Y)-code a nonissue for civilian uses. Some changes have to do with policy issues (e.g., S/A), and others are related to system performance. Changes of note for civilian users also include the planned addition of a new civil code on L2 (to be known as L2C) and a third civil frequency known as L5, which will enhance robustness by mitigating effects of interference and supporting increased availability of precision navigation. Not surprisingly, some improvements have also come about in response to competition from the two newer systems, GLONASS and GALILEO (Lavrakas 2007). A wealth of current information and the status of all three GNSS systems are available via a web search.

C/A Code

The observable for a C/A code phase receiver is the time delay of the signal to travel from the satellite to the receiver. The C/A code is time-tagged as it leaves the satellite and matched to a duplicate of the same code in the receiver by shifting the unique pattern on the receiver clock time scale. As illustrated in Figure 9.3, the shift required for the best match is a measure of the time interval of the signal from satellite to receiver. The distance from the satellite to the receiver is the product of the speed of light and the observed time interval. Theoretically, three such distances can be used to solve for the 3-D location of the unknown receiver. Regretfully, the signal does not travel in a vacuum and the clock in the receiver is not as precise as the atomic clocks in the satellites. Relativity is also a consideration. The time delay of the signal through the atmosphere is modeled to a close approximation (such details are beyond the scope of this book), but the receiver clock error must be observed and treated as an unknown. Therefore, four satellites must be observed to solve for four unknowns—the geocentric *X*/*Y*/*Z* coordinates of the antenna and a correction to the receiver clock. Note in equations 9.2 to 9.5 that the receiver clock correction shows up as a correction to the distance to each satellite and is the same regardless of the satellite from which the signal originated.

The C/A code phase solution can be described as a three-dimensional application of the Pythagorean theorem, which states,

$$Dist = \sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2} \tag{9.1}$$

Using that form and writing an equation for the distance from the single receiver to each of four satellites gives the following:

$$\text{To satellite A: } D_A + \Delta t * c = \sqrt{(X_A - X_R)^2 + (Y_A - Y_R)^2 + (Z_A - Z_R)^2} \tag{9.2}$$

$$\text{To satellite B: } D_B + \Delta t * c = \sqrt{(X_B - X_R)^2 + (Y_B - Y_R)^2 + (Z_B - Z_R)^2} \tag{9.3}$$

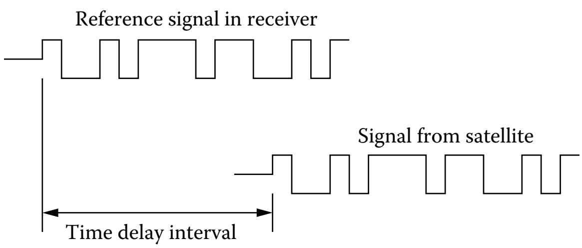


FIGURE 9.3 Matching Signals from Receiver and Satellite

To satellite C: $$D_C + \Delta t * c = \sqrt{(X_C - X_R)^2 + (Y_C - Y_R)^2 + (Z_C - Z_R)^2} \quad (9.4)$$

To satellite D: $$D_D + \Delta t * c = \sqrt{(X_D - X_R)^2 + (Y_D - Y_R)^2 + (Z_D - Z_R)^2} \quad (9.5)$$

where

D_A, D_B, D_C, D_D = observed distance to satellites A, B, C, and D, respectively;
X_A, Y_A, Z_A = known coordinates of satellite A;
X_B, Y_B, Z_B = known coordinates of satellite B;
X_C, Y_C, Z_C = known coordinates of satellite C;
X_D, Y_D, Z_D = known coordinates of satellite D;
X_R, Y_R, Z_R = unknown geocentric coordinates of receiver;
Δt = correction to receiver clock;
c = speed of light in a vacuum; and
$\Delta t * c$ = distance correction to each satellite (small).

With the observed distance to each satellite and the known *X*/*Y*/*Z* coordinates of each satellite, there are only four unknowns in the four equations—Δt, X_R, Y_R, and Z_R. Equations 9.2 to 9.5 look innocent enough, but using them to solve for the position as displayed by a GPS receiver is no trivial task. The GPS receiver initially solves for geocentric *X*/*Y*/*Z* coordinates, then converts those values to *latitude*/*longitude*/*height* using the BK2 transformation equations.

Finding the autonomous position of a point with a small handheld receiver (or a chip in your cell phone) to the nearest 10 meters with respect to the equator and the Greenwich meridian is an incredible feat. But scientists and manufacturers are continually working to find ways to improve that performance. Differential corrections and augmentation are discussed in a subsequent section.

Carrier Phase

The observable for a survey grade carrier phase GPS receiver is the fractional part of the 19 cm wavelength received at a specific time. The fractional parts (or phase shifts) recorded at two separate receivers from four or more satellites over time are used to compute a precise vector from one antenna to the other. The material discussed here is conceptual, and the actual algorithms used by the various vendors involve many more details. Also, be aware that GPS equipment designers do not restrict themselves to using just one principle or concept, but draw upon various techniques to build equipment that customers will purchase. For example, a vector (as used in surveying) from one point to another is determined using carrier phase observations, but that is not to say that interferometry is the only concept used in the processing algorithm. Likewise, neither is it correct to say that carrier phase observations are never used in determining the autonomous position of a handheld receiver.

The fundamental concept exploited in carrier frequency processing is that the distance between two points is represented by a sine wave. No matter the distance or the wavelength, there is always an integer number of waves and a fractional portion of the same wave as shown in Figure 9.4. When a GPS measurement starts (a receiver "locks" onto a satellite), the observation is the fractional part of the wavelength. The integer number of wavelengths is unknown. But the receiver keeps track of (counts) the number of wavelengths as long as it maintains lock on the satellite and records the phase shift at specified time instants.

So long as a receiver maintains lock on a given satellite, the **change** in distance between the satellite and receiver is directly observed. But, if the signal is interrupted, a cycle slip is said to have occurred and the integer count for that satellite must begin again. When observing numerous satellites, a cycle slip on a given satellite may be inconsequential and may have little impact on the solution. But, in the early years of baseline processing, especially when attempting to get a good solution using a limited number of satellites, the process of fixing cycle slips typically involved user intervention and could be tedious and time-consuming. In some cases, fixing cycle slips was not possible, and no solution could be obtained. In that case, the baseline had to be reobserved or eliminated from the solution. With the current full constellation of satellites and more robust automated baseline processing software, cycle slips are not the nuisance they once were. In fact, the modern user is often unaware of their occurrence.

DIFFERENCING

The comment has been made that electrical engineers rule the world because nearly every aspect of modern life, in one way or another, relies upon electronic signal processing. That appears to be especially true with GPS data. A common procedure when processing GPS signals is to subtract one (sine wave) signal from the other. Various advantages are realized depending upon how the differences are formed and used. While details can be found in references such as Leick (2004); Seeber (1993); Hoffman-Wellenhoff, Lichtenegger, and Collins (2001); and U.S. Army Corps of Engineers (1990), here is a summary of differencing options.

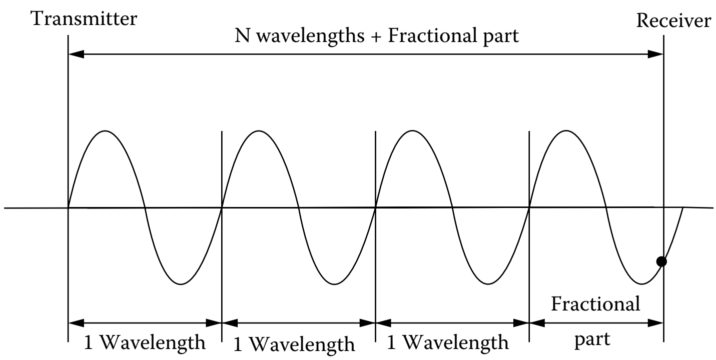


FIGURE 9.4 Fundamental Wave Measurement

Single Differencing

Three kinds of single differences are possible. First, and probably the most commonly used, the signal from one satellite is collected at a common time (epoch) by two different receivers. When these data are merged and differenced, the satellite clock error, orbit errors, and atmospheric delays cancel out because they are common to the signals received at both locations. Second, a single difference can be formed from the data collected at one receiver from two separate satellites. This difference combination effectively eliminates the impact of the receiver clock error. Third, a single difference can be formed between epochs by a given satellite-receiver pair. The Doppler effect shows up in these data and is modeled accordingly.

Double Differencing

A double difference is obtained as the difference between the single differences of a given type. Other double differences can be used, but when processing GPS carrier phase data, the most common double-differencing procedure involves finding the differences between two receivers observing two satellites at the same epoch. This procedure effectively eliminates both the receiver clock errors and the satellite clock errors from the baseline solution.

A characteristic of such a double difference solution is that the resulting wavelength ambiguity is an integer. That is, a baseline solution may be found quicker if, during the solution, the integer variable is allowed to take on a real number (known as a float solution). However, the unknown number of wavelengths in physical three-dimensional space is an integer, and a stronger solution (known as a fixed solution) is obtained by constraining the solution to integer values for the number of wavelengths.

Triple Differencing

Triple differencing involves a difference of the double differences over time and is useful because the integer ambiguity cancels out of the observation equations. But, because of the time interval involved, the triple difference solution must also accommodate the Doppler effect. In the early days of GPS vector processing, triple difference solutions were sometimes used, especially for longer lines where fixing the integers was problematic. However, with the development of more robust processing algorithms for fixing integers and with the use of data from more than four satellites, triple difference solutions are seldom the final solution. But triple differencing remains a valuable tool for preliminary (or intermediate) solutions.

RINEX

Of many things that could be said about the GNSS signals, this section includes a short description of the *R*eceiver *IN*dependent *EX*change (RINEX) format. Given more than one way to manipulate the signals received from the GPS (and other) satellites, it is not surprising that data from one brand receiver are not necessarily compatible with data collected by other brands. Computing a baseline vector from data collected by Brand X receiver on one point and Brand Y receiver on another point is impossible without having the data in a common format. Therefore, the RINEX format was proposed as a convenience for both the manufacturers and GPS users so that data collected by any of various brands of equipment could be combined in computing either baselines or networks.

At a minimum, the RINEX format includes definitions for (there are more)

- observation file,
- meteorological file, and
- navigation file.

A RINEX observation file uses ASCII characters and includes the basic observables, the phase measurement on both L1 and L2 (in cycles), and the range (in meters) to each satellite referenced to the receiver clock according to GPS (not UTC) time. The observation file header also contains information on the antenna type and the field-measured antenna height as determined by the user.

A Google search on "rinex" provides several hundred thousand hits, but not all sites involve the use of satellite data. http://gps.wva.net/html.common/rinex.html is a site containing excellent information on RINEX details, format, and history.

PROCESSING GPS DATA

Processing GPS data involves a number of "flavors." In the early years of GPS, two primary modes of processing were code phase (autonomous) processing and carrier phase (vector) processing. Given the evolution of technology and practice, each "flavor" begins to look more like the other and exclusive descriptions become meaningless. The purpose here is to describe the two basic procedures and acknowledge the evolution of practices. Current literature (theoretical, technical, and promotional) is rich with detail. The point, as described below, is that the GSDM supports all geospatial data processing, including GNSS-derived data.

In years past, processing GPS data involved a lot more user interaction than it does now. The computational load for the spatial data user has been significantly reduced and, in some cases, eliminated entirely. In the not-too-distant future, if not already being done, the collected data will be processed according to options and tolerances input by the user and the "final" answer will be available and used in real time. That is already happening with aircraft landing approaches, GPS-enabled earth-moving equipment, wide-area augmentation for navigation, and, to some extent, RTK surveying practices and procedures. Irrespective of the "flavor" and with regard to answers, two issues are involved: (1) getting an instant answer in the units and datum expected, and (2) being assured of the quality of the answer, that is, spatial data accuracy. In terms of answers, the user must be knowledgeable of the options, must know specifically what kind of spatial data are expected, and must know what constitutes a finished task, job, or project. In the case of an aircraft landing, the job is complete once the plane has landed and the tolerance for unacceptable results is very small. There are many other applications for real-time positioning—each with its own spatial data accuracy criterion. In each case, the GSDM provides for comparison of "actual" and "desired" differences and provides an efficient procedure for assessing the accuracy of such comparisons (Burkholder 1999).

Spatial Data Types

Spatial data types are listed in chapter 2, and most of them relate to geospatial data as obtained from GNSS. Several important points of which each user should be aware are the following:

1. The data types are independent of whether the observations were collected with GPS, GLONASS, or GALILEO.
2. The underlying spatial data types are also independent of the mode (code phase or carrier phase) by which the data were collected and processed. But, the covariance matrix associated with each point is greatly influenced by the mode of collection and processing. Generally the carrier phase-based data will have a smaller standard deviation than the code phase-derived data.
3. The GSDM handles all spatial data the same—independent of the mode of collection (code phase or carrier phase) or whether the data were collected and processed on the NAD83, WGS84, or ITRF.
4. The GSDM does not move data from one datum to another. But neither does the GSDM discriminate. For example, a user could input point A as defined on the NAD83 and point B as being on the WGS84. Clearly, that is not appropriate. But, if the standard deviation of either (or both) points is significantly larger than the known datum difference, then the 3-D inverse between point A and point B is legitimate *at the level of uncertainty derived from the 3-D inverse*. Used that way, the GSDM can be a dangerous tool.

It is presumed that position coordinates are associated with a named datum and represent the GPS antenna location. Of course, known offsets may also be involved. Such an offset may be from the antenna to the cutting edge of a bulldozer or earthmover, or, in the case of surveying, the offset is to the mark on the ground determined via the antenna height measurement. Offset measurements are the user's responsibility.

The seven spatial data types from chapter 2 are as follows:

1. Absolute geocentric *X*/*Y*/*Z* coordinates: primary values stored in a 3-D database.
2. Absolute geodetic coordinates of *latitude*/*longitude*/*height*: derived from the stored *X*/*Y*/*Z* primary values.
3. Relative geocentric coordinate differences: this is the primary result of GPS baseline vector processing. Relative geocentric coordinates are also obtained from local geodetic horizon coordinate differences that have been rotated into the geocentric reference frame.
4. Relative geodetic coordinate and ellipsoid height differences: obtained as the difference of compatible (common datum) geodetic coordinates.
5. Relative local coordinate differences: these are the local components of a GPS vector rotated to the local *e*/*n*/*u* perspective. These local components

are also the product of surveying total station observations and can often be used as flat-Earth components in the local tangent plane with an origin as selected by the user.

6. Absolute local coordinates: *e/n/u* are distances from some origin whose definition may be mathematically sufficient in three dimensions, two dimensions, or one dimension. Examples (see comments in chapter 2) are as follows:
   - Point-of-beginning (P.O.B.) datum coordinates as defined by the GSDM
   - Map projection (state plane) coordinates
   - Elevations on some named datum
7. Arbitrary local coordinates: may be 1-D (assumed elevations), 2-D (assumed plane coordinates), or 3-D (spatial objects). Although useful in some applications, arbitrary local coordinates are generally not compatible with geospatial data and have limited value in the broader context of georeferencing.

Often, the type of spatial data expected will be dictated by the context. For example, a graphical display of local coordinate differences may be used to navigate to a point. It is one thing to return to camp following a hike in the woods (based upon code phase observations) and something completely different to navigate to a point to be staked for construction or to find a fire hydrant buried under a snow bank (based upon RTK surveying procedures). Of course, the user may select a different type of data to be displayed (e.g., latitude/longitude/height or plane coordinates). In any case, the spatial data user is responsible for knowing specifically what to expect and verifying that the data being obtained are those expected.

Back to those automated aircraft landings—a Google search on "GPS aircraft landings" returns over 1 million hits. One of the more informative web site hits is http://waas.stanford.edu/pubs/phd_pubs.html, which lists approximately fifty Ph.D. dissertations on GPS-related positioning completed since 1993. Innovative applications and research opportunities involving spatial data abound.

Admittedly, the following descriptions of "autonomous" and "vector" processing are oversimplified, but they can be useful for learning more about underlying concepts of how GPS data are processed and used.

Autonomous Processing

This GPS processing option typically includes handheld C/A code receivers used for a variety of applications—recreation, navigation, tracking, and GIS. Once turned "on" and receiving data from a minimum number of satellites, an autonomous position is obtained from a single GPS receiver with little or no input by the user. The results will be displayed in the coordinate system, datum, and units as specified by the user. Internally, the GPS signals are processed to find the geocentric *X/Y/Z* coordinates of the antenna (spatial data type number 1). But, since those rectangular *X/Y/Z* values are difficult to visualize, the results are converted into the coordinate system as selected by the user (typically, spatial data type number 2—latitude/longitude/height). Other options may be available such as specifying a datum, a state plane coordinate zone, UTM coordinates, a national grid designation, or another user-defined system. User-selectable options for handheld C/A receivers typically include the following.

Datum

World Geodetic System 1984 (WGS84)
North American Datum of 1983 (NAD83)
International Terrestrial Reference Frame (ITRF), epoch XX
Other

Units

Meters (standard)
International feet
U.S. Survey feet
Statute miles
Nautical miles
Other

Display

Geocentric Earth-centered Earth-fixed (ECEF) $X/Y/Z$
Degrees, minutes, and seconds (decimal)
Degrees and minutes (decimal)
Decimal degrees

Time

Universal Time Coordinated (UTC)
Time zone offset
24-hour or a.m./p.m. (12-hour) mode

With SA enabled, routine accuracy expected from a C/A code receiver was originally +/– about 100 meters. But, SA was discontinued in May 2000, and the expected autonomous accuracy dropped to about 10 meters. That is impressive, but, using advanced processing techniques, vendors routinely claim submeter accuracy for nonsurvey grade GPS receivers.

Vector Processing

When processing carrier phase GPS data, two separate receivers are needed to collect data from common satellites. The simplest form of vector processing uses GPS signals recorded simultaneously in two receivers located at two different points. The data file and station message file from each receiver and an ephemeris file from one of the receivers are transferred to a computer having the appropriate processing software. Using those data, the baseline between stations is computed and reported in terms of $\Delta X/\Delta Y/\Delta Z$ between stations (spatial data type number 3). The covariance matrix for the baseline is also computed and reported. *The ECEF coordinate differ-*

ences and their covariances are the primary answers obtained from baseline processing. Other pieces of information such as local component differences ($\Delta e/\Delta n/\Delta u$), slope distance, ellipsoid height difference, forward geodetic azimuth, back geodetic azimuth, and standard deviations of the various answers are derived from those primary answers.

Depending upon which of several options are used, various results can be obtained when computing GPS baselines. Typically, a preliminary triple difference solution is computed first, even for shorter baselines, because the integer ambiguity value cancels out of the observation equation. If the baseline is quite long, cycle slips may be more of a problem and a triple difference solution may be an important intermediate step in determining a preliminary solution. However, once cycle slips are identified and fixed, a double difference solution is typically better than a triple difference solution. But, there are two levels of double difference solutions. The first double difference solution solves for the integer ambiguities as a real number. That is called a float solution because the values found for the integer wavelength numbers are allowed to "float" and to be something other than an integer. There are cases where the float solution may be the strongest solution available for a given set of data. But, the preferred solution is obtained when the integer ambiguities are "fixed" and forced to be integers—because, physically, the path of the signal contains an integer number of wavelengths plus a remainder. Modern baseline processing software is quite robust, and, unlike the early days of GPS baseline processing, the user rarely needs to interject decisions as the processing proceeds. However, once baseline processing is complete, networks are formed, adjustments are performed, and statistics of the results are developed, the judgment of the user is still important in deciding upon the ultimate acceptability of the results. If the results are not acceptable, corrective measures may include actions like changing the elevation mask, eliminating signals from (turning off) a particular satellite, or reobserving a given baseline. Vendor manuals and other GPS users offer many suggestions for various corrective actions.

Multiple Vectors

Added complexity in carrier phase processing arises if data are collected simultaneously with three or more receivers. Although the underlying algorithm for computing $\Delta X/\Delta Y/\Delta Z$ components for each baseline may be the same, when processing multiple baselines, two additional considerations are (1) avoiding the use of trivial vectors and (2) handling the correlation between baselines sharing a common station. A nonexclusive definition of a trivial baseline is any vector computed using two data sets that have already been used in computing another baseline. There are different ways of determining a trivial baseline, but perhaps the easiest way is to look specifically at nontrivial vectors. A nontrivial vector is a baseline computed using, at least in part, data not used previously. If data collected at both ends of a line have been used previously, the result may be a trivial vector. If data at one end have been used but data at the other end have not been used, it is a nontrivial vector. Many baseline-processing packages permit the user to choose the nontrivial baselines, but default choices built into most processing software are also popular.

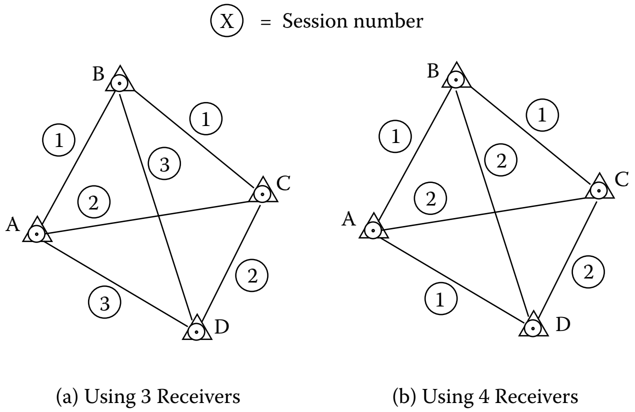


FIGURE 9.5 GPS Receivers and Observing Sessions for Nontrivial Vectors

Figure 9.5 shows two views of a network of four points to be surveyed using GPS. The points are labeled A, B, C, and D, and GPS baselines between all points are represented by the six connecting lines. Scenario 1 is not illustrated, but view (a) illustrates using three GPS receivers and three observing sessions for the six nontrivial vectors, while view (b) illustrates using four receivers and two observing sessions for the same six vectors.

Scenario 1: Two GPS receivers are used to occupy each baseline in six separate sessions. It is a tedious and time-consuming method of GPS data collection, but each observed baseline is a nontrivial vector.

Scenario 2: Three GPS receivers are used simultaneously, and two nontrivial vectors are obtained in each of three sessions. For example, in session 1, receivers occupy stations A, B, and C. Nontrivial vectors are AB (two new data sets) and AC (one new data set at C). For session 2, receivers occupy stations A, C, and D. Nontrivial vectors are CA and CD. In session 3, receivers occupy stations D, B, and A, and the nontrivial vectors are DB and DA. Other combinations could also be used so long as each new vector includes station data not used previously.

Scenario 3: Four GPS receivers are used simultaneously, and three nontrivial vectors are obtained in each of two sessions. For example, in session 1, nontrivial vectors could be AB (two new data sets), BC (one new data set at C), and AD (one new data set at D). For session 2, the nontrivial vectors are CD (two new data sets), CA (one new data set at A), and DB (one new data set at B).

Traditional Networks

The simplest network scenario is to build a GPS network of independent vectors. That would be the case if a network was built using independent baseline data collected from just two receivers. When combining vectors and computing such a multisession network adjustment, it is standard practice to anchor the network to a single fixed 3-D (*X*/*Y*/*Z*) point and to compute a minimally constrained least squares adjustment. The purpose of computing a minimally constrained network is to confirm the absence of blunders. If blunders exist, the offending baseline is reprocessed sans the blunder, reobserved, or eliminated entirely. Once the observed baselines are verified blunder-free, the network is constrained to the *X*/*Y*/*Z* values of two or more existing high- (or higher-) order stations and the network is recomputed. A successful least squares adjustment provides the adjusted *X*/*Y*/*Z* coordinates (and covariance matrices) for each new station established during the survey. Such a network is both practically and statistically legitimate. A well-documented least squares network adjustment of independent vectors is posted at http://www.globalcogo.com/nmsunet1.pdf.

This is not a book on adjustments. However, as illustrated in Figure 9.5, where three or more GPS receivers collect data simultaneously, there is correlation between those nontrivial vectors sharing a common station. When developing the weight matrix for a multistation session, the correlation between baselines should also be included in the covariance matrix of the observations. The network adjustment then becomes a multistation multisession solution, which is described in more detail by Seeber (1993). Examples include the adjustment of a statewide HARN network, the national readjustment of the NAD83 datum network, or, ultimately, even a global network of GPS stations.

Whether building a network of independent vectors or building a multistation multisession network, the GSDM provides a "natural" computational environment for performing those computations. The final coordinates (spatial data type number 1) from a least squares adjustment are absolute values based upon the fixed absolute control selected by the user (also spatial data type number 1) and the observed relative baseline vectors (spatial data type number 3). The least squares adjustment also provides the covariance matrix for each point and the covariance submatrix from which correlation between points can be obtained. A BURKORD™ database stores the *X*/*Y*/*Z* coordinates for each point, the covariance matrix for each point, and (optionally) the covariance submatrix representing correlation between point pairs.

Advanced Processing

The concepts presented here are advanced in that they move beyond the fundamental procedures described earlier. But they are not advanced in that GPS professionals routinely work with many of these "advanced" concepts. First, there are several issues deserving consideration that are omitted. It is well-known and understood that the quality of long baselines can be enhanced by using a precise ephemeris instead of the (default) broadcast ephemeris. Fine-tuning the baseline computational process is left to others. And, in some cases, GPS data are commingled with inertial data. That discussion is also very relevant but beyond the scope of this book. However, a closing thought is that once inertial data are "corrected" for deflection-of-the-vertical (see chapter 8), even gravity-based measurements (total stations, levels, or an inertial measuring unit) can be used along with appropriate standard deviations in the GSDM.

Issues of absolute and relative spatial data need to be considered in moving from traditional network adjustments to procedures such as differential positioning, augmentation, RTK positioning, and using the NGS On-line User Positioning Service

(OPUS). It would be nice if absolute and relative issues could be considered separately, but as technology and operational procedures evolve, the simple categories are no longer applicable. Admittedly, it doesn't matter for many because, in all honesty, many end users are really interested in obtaining a reliable position with appropriate statistics without needing to worry about how it was obtained.

Differential GPS positioning (DGPS) involves observing an autonomous position at two locations—one known, and the other unknown. The difference of "known" minus "observed" at the known location is called a correction. That correction is applied at the unknown location to give a differentially corrected position. The process of applying the correction can take any of various forms. Here is an abridged list of possible scenarios (some give better results than the others):

1. One (simple handheld) receiver is used by one person at multiple locations. One of the locations visited is the known point. The correction is computed and applied manually. The assumption is that the correction does not vary over (short periods of) time or with location.
2. Two GPS receivers are used in the field. One receiver remains on the known location (base), while a second (rover) is taken to various points. The correction as determined at the base is applied (manually) for each point visited. This procedure documents whether the correction is "fixed" or whether the correction varies with time. This procedure is better than the first but not very efficient, especially if security considerations mandate that a person stay with the base receiver.
3. The location of a permanent (secure) base station is surveyed precisely, and the base station receiver records data continuously. A second (remote) receiver is used by one person to collect data at various points. Back at the office, data from the base station are retrieved and the correction is applied to points collected with the roving receiver. Software for automated processing increases efficiency.
4. The correction is computed automatically at the base station and transmitted via radio or cell phone from the base to the remote receiver. The "corrected" position at the rover is determined in real time and used immediately or stored for use elsewhere or by others.
5. The correction at the base is determined, not as a location difference but as a time-delay modification to the signal received from each satellite. The correction sent to the remote is not a location difference but a time-delay correction from each observed satellite. The DGPS position observed at the remote station is better than that obtained with other methods and is achieved with greater computational efficiency.

General statements are that DGPS provides an improved absolute position obtained from a C/A code receiver. The quality of such an observed position is enhanced by observing more than four satellites, by using the strongest geometry afforded by the existing satellite constellation, by observing for a longer time period, by using more than one receiver (for redundancy), and by observing the same point on several different days.

Augmentation can be described as DGPS on a big scale. The Federal Aviation Administration (FAA) has established a wide-area augmentation system (WAAS) for determining differential corrections at ground-based locations and transmitting that information to geosynchronous communications satellites that retransmit the signals over a large area—most of the United States. WAAS provides absolute position in real time and was designed primarily to support civil aviation.

A summary of approximate *absolute* positions based upon code phase receivers is:

1. Original C/A code position with selective availability imposed: 100 m
2. C/A code position without selective availability imposed: 10 m
3. Typical DGPS: 3 to 5 m
4. Typical WAAS positional accuracy: < 3 m

The reader should also be aware that some manufacturers claim submeter accuracy for their autonomous GPS receivers. Such accuracy may be obtained by exploiting information on the carrier frequency in addition to the code phase data. Those claims, equipment, and observational processes should be discussed with the manufacturer. Tests can be conducted to establish the veracity of such claims.

RTK surveying procedures are based upon processing carrier phase data received at two separate locations and provide a *relative* baseline vector between the two points. Data must be collected simultaneously and combined to compute a vector. One scenario is to collect data with two or more receivers (one receiver on a known point) and bring the data back to the office for processing. But, if the raw carrier phase data at the base station are transmitted to the remote unit, then the processing can take place at the remote unit in real time. Given further that the base station occupies a precisely surveyed point, then the position of the remote unit can be computed with mm or cm in real time. Here, too, the answer is an absolute value (coordinates), but those coordinates are based upon a *relative* vector tied to the known base station position. One drawback of RTK surveying is lack of redundancy. That can be overcome by occupying the unknown point a second time or by computing a baseline from a second base station. Many localities either have or are in the process of establishing GPS base station networks to cover specific regional areas such that precise RTK surveying can be conducted with one unit and one person.

Another processing option is the OPUS available via the Internet from the NGS. At least 2 hours of dual frequency data are collected on an unknown point. The data are submitted via the Internet to NGS along with antenna type and measured antenna height. Shortly, often within a matter of minutes, an e-mail is returned to the sender and contains the position of the point computed on the basis of continuously operating reference stations (CORS) in the general region. The answer is reported in both geocentric *X*/*Y*/*Z* and conventional *latitude*/*longitude*/*height* coordinates and in both NAD83 and ITRF coordinates. An estimate of the accuracy is also provided.

The NGS also supports a service called OPUS-RS (rapid static), which will provide a solution with as little as 15 minutes of static GPS data. Details are available at the NGS web site—see http://www.ngs.noaa.gov/OPUS/OPUS-RS.html.

The GSDM fundamentally supports all varieties of GPS processing including both absolute and relative considerations, NAD83 or ITRF (not at the same time), and any GNSS system (GPS, GLONASS, or GALILEO). Furthermore, the GSDM handles spatial data accuracy for all types of spatial data. See chapter 11 for a discussion of spatial data accuracy.

THE FUTURE OF SURVEY CONTROL NETWORKS

This section contains certain speculative predictions about GNSS technology and the use of spatial data. Some parts of the speculation are already possible though not necessarily commonly realized in practice. However, even modest extrapolation exposes exciting possibilities for many spatial data users.

At the risk of rubbing the crystal ball too hard, the future of monumented control points is bleak and limited. Using existing technology, it is possible, and becoming ever more practical, to determine a position anywhere in the world based upon signals received from satellites in the sky. Under that scenario, it is said the satellite orbits become the permanent reference monuments—see items 1 and 2 in the following list. Of course, old habits die hard, and, to some extent, there will always be a demand for a reliable physical P.O.B. Even so, the following issues need to be considered when evaluating access to the NSRS:

1. Satellite visibility (or lack thereof) will obviate some of the following points.
2. Traditional relative positioning methods will serve as a backup if and when GNSS technology is not available or appropriate.
3. However, GNSS technology and equipment are already being used to establish locations all over the world within impressive levels of tolerance. Yes, it takes more sophisticated equipment and more exacting observing procedures to position a point within 1 cm than it does to position a point within 1 meter. Are smaller tolerances feasible? Yes.
4. Realized (surveyed) positions may be absolute or relative. In some cases, they may need to be both.
   A. An absolute position is given by the coordinates of the surveyed location. Many applications are satisfied with absolute data. Examples include answering questions such as “Where is the point source pollution?” “Where did that accident occur?” “Where is the defective transformer?” and other inventory-related questions.
   B. A relative position is important in other applications where the user needs to know a location with respect to other points. How far is it to my destination? How far have I traveled today? How far is the back of the curb from the right-of-way line? What is the direction or distance from one property corner to the next? How far is it from the airplane to the runway (automated landings)?
5. The results of a survey may be consumed instantaneously or archived for future use. The value of an instantaneous position may be the comfort of knowing where I am—no one with a GPS unit ever needs to admit to not knowing where they are. On the other hand, information archived for future

use may involve storing the waypoint of the trailhead so I can return to my vehicle at the end of a hike. I may need to go back to a specific location so I can capture accurate "before" and "after" photographs. Or, I may need to return to the site of a recently discovered petroglyph or an underwater wreck that has been there for many years.

6. In order for a coordinate position to be meaningful, it needs to be compared with a previous value. Whether one is a hiker, photographer, anthropologist, or maritime archeologist (as noted above), the comparison may be casual, time-delayed, or approximate. However, when using GPS to land an airplane or to fly an unmanned aerial vehicle (UAV), the stakes are higher. The location (3-D representation) of the runway must be in the database (computer memory), and the instantaneous position of the aircraft must be compared with values in the database to determine the separation between the aircraft and the runway. Whether one is landing an aircraft or engaging in other intelligent vehicle navigation application, accuracy is critical and spatial separation changes rapidly. Three important considerations are as follows:
   A. The accuracy of data in the database must be of proven quality—see number 7, following.
   B. The accuracy of the observed instantaneous position must be realized within an acceptable tolerance.
   C. The comparison needs to occur instantly, and answers must be available in real time.

   Understandably, all three issues become moot once the aircraft is safely on the ground.

7. In surveying and other related applications, real-time considerations may be less critical than when landing an airplane, but relative/absolute and spatial data accuracy issues still need to be considered. Of course, as technological improvements keep coming, the tolerances will become smaller, the comparisons will become more economical, and there will be many more applications.

8. Especially with regard to high-end applications and comparisons, data in the database must be compatible with the position derived from satellites orbiting the Earth's center of mass. Scientists, engineers, and manufacturers build equipment that determines the location of the receiver and antenna. That is only half of the solution. The quality of a relative location depends heavily upon the quality of information in the database. Three possible kinds of information in the database include the following:
   A. Design locations (virtual) are stored in the database and represent errorless quantities.
   B. Staked positions (intended) are marked on the ground in accordance with design locations using equipment and procedures capable of providing answers within a given tolerance. The error (spatial data accuracy) of such locations is determined by the equipment and the procedures used during the layout process.
   C. Surveyed location (actual) is a measurement of the position of a marked point. In this case, the quality of the information in the database reflects both the integrity with which the point (monument) was established

and the quality with which it was reobserved. Ideally those two error sources should both be controlled and "small." However, it could be that either the staked location of the point or the surveyed location of the point as stored in the database contains a large error. In either case, regardless of how good the current instantaneous position might be, a subsequent comparison with the value stored in the database will be bogus.

A huge caveat to this entire discussion is the understanding that all data, regardless of their quality, must be on the same datum. If database positions are not expressed in the same datum as the currently observed satellite position, then computed relative positions may contain unacceptable error.

The overall point of this chapter is that GNSS computations expressing location, whether conducted on ITRF, WGS84, or NAD83, can and should be accomplished in the ECEF environment of the GSDM. Two compelling justifications are that both high-level scientists and "flat-Earth" end users can use the same rectangular solid geometry equations and that spatial data accuracy is easily established, tracked, and made available in terms of the GSDM.

REFERENCES

Burkholder, Earl F. 1999. Spatial data accuracy as defined by the GSDM. *Surveying and Land Information Systems* 59 (1): 26–30.

———. 2003. The digital revolution begets the global spatial data model (GSDM). *EOS, Transactions, American Geophysical Union* 84 (15): 140–41.

GPS World. (2007) GPS purchases to generate $4.1 billion in sales in 2007. 18(5):34.

Hoar, Gregory J. 1982. *Satellite surveying: Theory, geodesy, map projections—applications, equipment, operations*. Torrance, CA: Magnavox Advanced Products and Systems.

Hoffman-Wellenhof, B, H. Lichtenegger, and J. Collins. 2001. *GPS: Theory and practice*, 5th ed. Vienna: Springer-Verlag.

Jones, Tony. 2000. *Splitting the second*. Boston: Institute of Physics Publishing.

Kleppner, Daniel. 2006. Time too good to be true. *Physics Today* 59 (3): 10–11.

Laurakas, John W. 2007. GNSS in the year 2017, *GPS World,* 18(5): 14–20.

Leick, Alfred. 2004. *GPS satellite surveying*. New York: John Wiley.

Mueller, Ivan I. 1964. *Introduction to satellite geodesy*. New York: Frederick Ungar.

National Academy of Public Administration (NAPA). 1995. The Global Positioning System: Charting the future. Report by a Panel of the National Academy of Public Administration and a Committee of the National Research Council for the U.S. Congress and the Department of Defense. Washington, DC: National Academy of Public Administration.

———. 1998. Geographic information for the 21st century: Building a strategy for the nation. Report by a panel of the National Academy of Public Administration for the Bureau of Land Management, Forest Service, U.S. Geological Survey, and National Ocean Service. Washington, DC: National Academy of Public Administration.

National Imagery and Mapping Agency. 1997. Addendum to NIMA TR 8350.2: Implementation of the World Geodetic System 1984 (WGS 84) Reference Frame G1150. In Department of Defense World Geodetic System 1984: Its definition and relationships with local geodetic systems. Technical Report TR8350.2, 3rd ed. Washington, DC: National Imagery and Mapping Agency. http://earth-info.nga.mil/GandG/publications/tr8350.2/Addendum%20NIMA%20TR8350.2.pdf.

Schwarz, Charles R., ed. 1989. North American Datum of 1983. NOAA Professional Paper NOS 2. Rockville, MD: National Geodetic Survey, National Ocean Service, National Oceanic and Atmospheric Administration.

Seeber, Gunter. 1993, *Satellite geodesy: Foundations, methods, and applications*. Berlin: Walter de Gruyter.

Stansell, Thomas A. 1978. *The TRANSIT Navigation Satellite System*: *Status, theory, performance, applications*. Torrance, CA: Magnavox Government and Industrial Electronics.

U.S. Army Corps of Engineers. 1990. *Engineering and design: NAVSTAR Global Positioning System surveying*. Washington, DC: Department of the Army, U.S. Army Corps of Engineers.

10 Map Projections and State Plane Coordinates

INTRODUCTION: ROUND EARTH—FLAT MAP

A map projection is a 2-D model whereby the curved surface of the Earth is portrayed on a flat map. If one looks only at a small portion of the Earth's surface such as a city map, it appears that local features on the map correctly portray the same features as viewed by a person walking or driving in that area. However, when dealing with larger and larger portions of the Earth's surface, distortion and the challenge of true map representation grow exponentially. An extreme example of distortion is recognized by many elementary schoolchildren who look at a comparison of Alaska and Brazil—first on the globe, then on a Mercator world map. On the globe, Alaska appears noticeably smaller than Brazil, but on a Mercator map projection of the world, Alaska appears much larger than Brazil. The problem is that, on the globe, all meridians converge at the North Pole and South Pole. On the projection at the equator, the spacing of the meridians is identical to the meridian spacing on the globe. But the meridians remain parallel on the projection, and features nearer the poles appear grossly exaggerated in size. In fact, neither the North Pole nor the South Pole can be shown on a Mercator map, which touches the Earth at the equator as shown in Figure 10.1a.

A spherical Earth is shown in Figure 10.1a with rays originating at the center and piercing the globe at each 15° degrees of latitude before striking the cylindrical surface of the Mercator projection. Following such graphical projection, the cylinder is cut down the back and rolled out flat to give the appearance of the graticule shown in Figure 10.1b. Notice that the 15° blocks of latitude and longitude near the equator are nearly square but that the same 15° blocks become elongated further from the equator. An area lying near either pole is grossly exaggerated in size when shown on such a Mercator projection.

Cartography is the science of making maps and includes various graphical portrayals of spatial data. Using cartographic definitions, a graticule is the grid-like appearance of parallels and meridians covering the Earth, and a map projection is defined as a systematic arrangement of the graticule on a flat surface. The challenge is going from a curved surface to a flat map without distorting any geometric element. It can't be done. Most people know that when you peel an orange (even children enjoy making those pieces as big as possible), the curved peel will not lay flat on a table unless one presses it flat. In so doing, the peel is distorted. Either the peel tears or other parts of the peel are artificially compressed in the process of being flattened. However, if one considers only a small portion of the orange peel, it appears to be smooth and flat—even though it originated from a spherical "whole." So it is

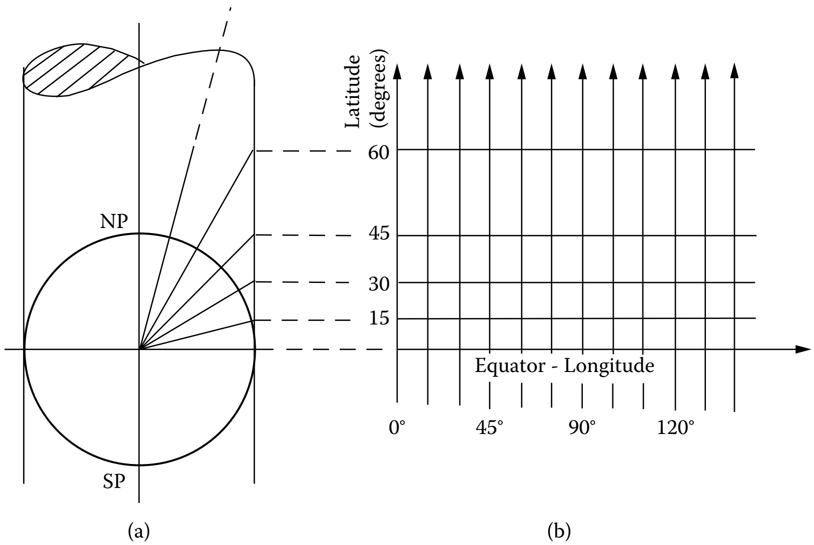


FIGURE 10.1 Globe and Mercator Grid

with the Earth. Small portions of the surface can be represented very well using the assumption that the Earth is flat. But, when dealing with larger and larger areas on the Earth, the inevitable distortions that occur between the curved surface and the flat map must be accommodated.

Maps range from simple to complex and serve many purposes. In some cases, a map communicates best by grossly distorting geometrical detail. In other cases, the geometrical detail of a map is the basis of its value to the user. Given that most spatial data are now digital and given further the proliferation of computerized data visualization tools, the opportunities for cartographers to develop creative and innovative representations of spatial data have become innumerable. The GSDM provides a concise set of rules equally applicable worldwide for generating, storing, manipulating, viewing, and otherwise using digital geospatial data. Although beneficial uses still exist, map projections have lost some of their utility because a map shows only 2-D relationships from a fixed perspective. Modern practice must accommodate 3-D digital spatial data, and many users prefer the option of choosing a perspective. The GSDM allows flat-Earth relationships (including 2-D ones) to be computed, viewed, and used as local coordinate differences while the underlying ECEF coordinates retain their geometrical integrity and global uniqueness.

PROJECTION CRITERIA

Use of the GSDM notwithstanding, concepts of map construction are still important and are summarized herein. It is impossible to generate a flat map that depicts the curved surface of the Earth accurately without distorting two or more of the following geometrical elements: angles, distances, or area. In the past, questions to be answered included "What elements will be distorted and by how much?" Another part of the same question is "What element can be preserved in the projection without being distorted?" Various answers dictate a particular class of projection, and maps are made accordingly. But, the digital revolution has changed much of that because computers, equations, and talented cartographers can now manipulate digital spatial data in creative ways not previously feasible. Therefore, a basic understanding of map projections is still important.

Mathematical set theory includes the concept of range and domain. Bringing that analogy to map projections, the location of a point on the curved Earth is considered to be the domain, while an equivalent expression for the same point on a flat map is the range. A map projection is the function (set of rules) whereby a discrete point in the domain is given equivalent expression in the range. And, the transformation rules must be bidirectional in order to preserve the unique point-to-point match between the range and the domain.

Given the impossibility of portraying a curved Earth on a flat map with true geometrical integrity, the "rules" used in the transformation function will distort some combination of angles, distances, and area. Of course, the reader should realize that if the area being mapped is relatively small, the distortion of a given geometrical element may be small enough to be of no consequence (the magnitude of permissible distortion will be discussed later). But, when considering larger and larger areas, a cartographer has the option of designing a map projection such that one of the elements—angles, distances, or area—can be preserved on the map. That choice provides the mathematical basis for three important classes of map projections:

Conformal: A conformal map projection is one in which a horizontal angle on the Earth is unchanged in its representation on the flat map. Distances and areas are distorted on a conformal projection, but the distortion is controlled by limiting the area of the Earth being projected. Conformal map projections are used extensively in surveying and mapping applications and are discussed later in this chapter.

Equidistant: An equidistant map projection is one in which the distances on the Earth are faithfully represented on the map. Angles and areas are distorted.

Equivalent: An equivalent map projection is one in which the area of a given portion of the Earth's surface is truthfully represented on the map. In geography, the equal-area map is used beneficially in many applications.

Another important issue is that map projections can be developed by graphical construction (often used for illustration purposes) or mathematical equations. Graphical projections are categorized by the origin of the imaginary light ray. A gnomonic projection is one where the light ray originates at the center of the Earth—see Figure 10.2a. A stereographic projection is one in which the light rays originate on the opposite side of the world, as shown in Figure 10.2b. And, an orthographic projection is one in which the rays all arrive perpendicular to the surface—in other words, the rays originate from a point source at an infinite distance (see Figure 10.2c).

When using the P.O.B. option of the GSDM, the relative position of each point is plotted with respect to the P.O.B. according to their local latitudes and departures. The result is an orthographic projection of a point cloud to the tangent plane through

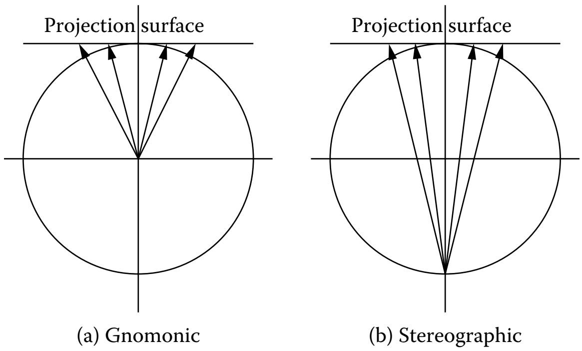


FIGURE 10.2 Location of Light Sources for Projection

the P.O.B. selected by the user. If the point cloud is a scanned ortho photo image stored pixel by pixel in a BURKORD™ database, the resulting image is a rectified map of all the pixels in the cloud. The impact of this feature will be significant for softcopy photogrammetry, use of scanned images, and photogrammetric mapping. Although the orthographic projection is not conformal over large areas, the distortion of angles and distances for small areas is inconsequential, and each user has the option of selecting successive P.O.B.'s such that the spacing between P.O.B.'s will control distortion at an acceptable level. This feature of the GSDM needs additional study and further clarification.

The conformal map projections discussed in the remainder of this chapter are mathematical projections even though the graphical mode is used for illustration purposes.

PROJECTION FIGURES

Map projections are also categorized according to whether the projection surface is a plane, a cone, or a cylinder, as illustrated in Figure 10.3. Conceptually, the basic difference between all three cases is the location of the apex of the cone. In one extreme, the apex of the cone is on the curved surface, resulting in a tangent plane projection (see Figure 10.3a). The cylinder illustrates the other extreme in which the apex of the cone is infinitely distant from the Earth (see Figure 10.3c). Between those extremes, the cone contacts the Earth along a standard parallel of latitude as determined by the distance between the curved surface and the apex of the cone. As the distance to the apex becomes larger and larger, the standard parallel of latitude moves closer and closer to the equator. Gerard Mercator (1512–1594) is credited with devising the cylindrical conformal Mercator projection, while Johann Heinrich Lambert (1728–1777) is credited with developing the transverse Mercator projection and the conic conformal projection as an extension of Mercator's work.

The word "zone" is often used to describe a specific portion of the Earth's surface that can be mapped with a single projection without exceeding some limit of distance distortion. For the state plane coordinate system (SPCS), the distortion limit in a zone is 1 part in 10,000; and for the Universal Transverse Mercator

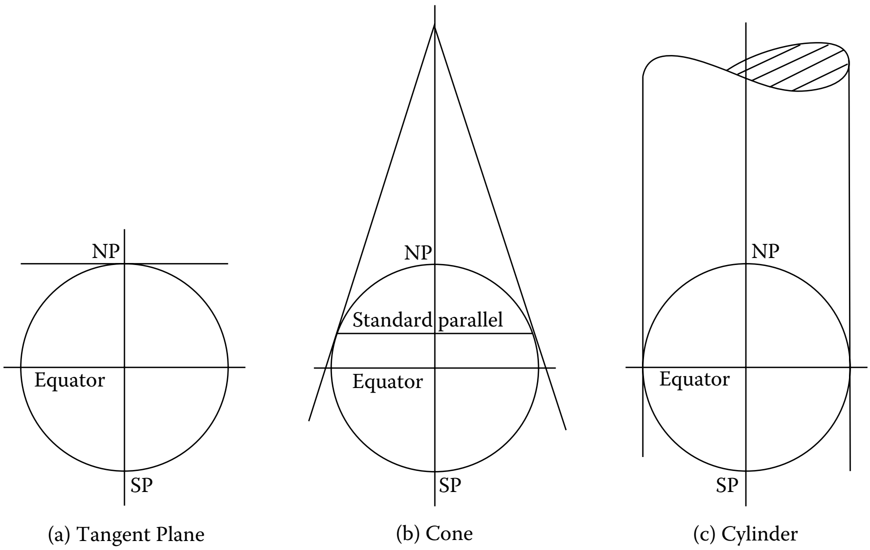
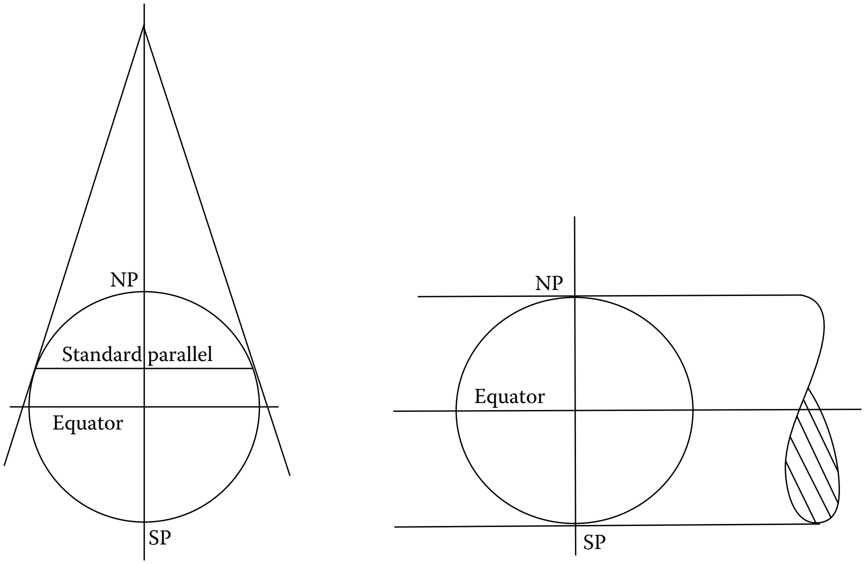


FIGURE 10.3 Map Projection Surfaces: Three Apex Locations

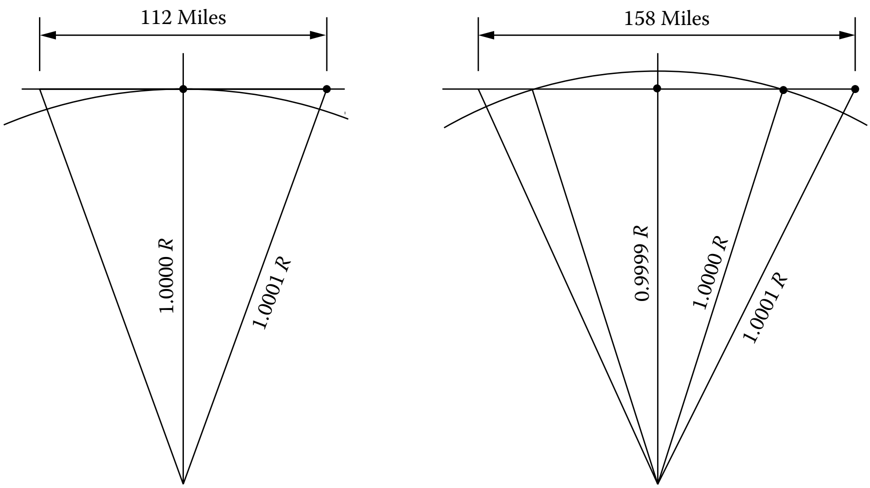
(UTM) projections, the distortion limit in a zone is 1 part in 2,500. Of course, the goal is to include as much area in a zone as may be practical without exceeding the specified distortion limit. That means the zone must be as wide as possible and as long as practical. Two choices are typical when looking at zone length. The cartographic designer can choose either a Lambert conic conformal projection, as shown in Figure 10.4a, or a transverse Mercator projection, as illustrated in Figure 10.4b. If a state to be covered has a long east-west extent (e.g., Tennessee), then a conic projection is appropriate. If the state to be covered has a long north-south extent (e.g., Illinois), then a transverse Mercator projection is better. In the case of the UTM projections, each zone is 6° wide and extends from latitude 80° S to 84° N.

When attempting to maximize the width of a zone without exceeding a given distortion limit, a further consideration is that the projection surface may be tangent to the Earth, as shown in Figure 10.5a, or secant, as shown in Figure 10.5b. Distortion on a tangent projection stretches the distance from the curved surface to the mapping plane—the distortion is one-sided. A wider zone is possible if a secant projection is used and the distortion is two-sided—that is, if the distortion includes both compression and expansion. On a secant projection, the distortion near the center of the projection compresses a distance element. Moving away from the center of the projection, the distortion diminishes, goes to zero where the two surfaces intersect, then increases without limit as one moves further and further away from the center of the zone. That means the nominal width of a zone is determined by a (arbitrary) choice of the designer regarding maximum allowable distortion.



(a) Lambert Conic Conformal (b) Transverse Mercator

FIGURE 10.4 Lambert Conic Conformal and Transverse Mercator Projections



(a) Tangent (b) Secant

FIGURE 10.5 Projection Types

PERMISSIBLE DISTORTION AND AREA COVERED

The grid scale factor is used to describe distance distortion. The grid scale factor is represented by the letter *k* and defined as the distance on the flat map grid divided by the curved distance on the ellipsoid.

$$k \equiv \frac{\text{distance on the grid}}{\text{distance on the ellipsoid}}$$

For example:

$$k = \frac{99.990 \text{ meters}}{100.00 \text{ meters}} = 0.99990 \tag{10.1}$$

Grid scale factor distortion can also be expressed as a ratio or as parts per million (ppm). A distortion limit of 1 part in 10,000 is often used in the state plane coordinate system projections. The following are three equivalent expressions of the same grid scale factor.

$$k = 0.99990$$
$$k = 1 \text{ part in } 10{,}000$$
$$k = 100 \text{ ppm}$$

When the 1 in 10,000 limit is applied to the state plane coordinate projections, the range of grid scale factors is as summarized in Table 10.1.

Using a radius of a spherical Earth as 6,372,000 meters (approximately 20,906,000 feet) and the grid scale factors above, the maximum zone widths in Figure 10.5 are computed as

$$\text{Tangent projection: Zone Width} = \frac{\sqrt{(1.0001R)^2 - R^2}}{1000} * 2 = 180 \text{ km} \approx 112 \text{ miles} \tag{10.2}$$

TABLE 10.1
Comparison of Grid Scale Factors

Tangent Projection	
Center of the zone	$k = 1.0000$
Edge of the zone (imposed by 1/10,000 criterion)	$k = 1.0001$
Secant Projection	
Center of zone	$k = 0.9999$
Intersection of curved surface and mapping plane	$k = 1.0000$
Edge of the zone (imposed by limit of 1 in 10,000)	$k = 1.0001$

Secant projection:

$$\text{Zone Width} = \frac{\sqrt{(1.0001R)^2 - (0.9999R)^2}}{1000} * 2 = \quad 255 \text{ km} \approx 158 \text{ miles} \quad (10.3)$$

When making decisions about what projection to use on the NAD83 datum, mapping professionals in several states chose to relax the 1:10,000 distance distortion criterion so that the entire state could be covered by a single zone. On NAD83, the State of South Carolina uses a single zone with a grid scale factor of 0.999793656965 (1:4,846), the State of Montana uses 0.999392636277 (1:1,646), and the State of Nebraska uses 0.999658595062 (1:2,938). All three are Lambert conic conformal projectsions and the effective zone widths are

South Carolina:

$$\text{Zone Width} = \frac{\sqrt{(1.000206343R)^2 - 0.9997936570R^2}}{1000} * 2 = 366km \approx 227 \text{ miles}$$

Montana:

$$\text{Zone Width} = \frac{\sqrt{(1.00060736R)^2 - 0.999392636R^2}}{1000} * 2 = 628km \approx 390 \text{ miles}$$

Nebraska:

$$\text{Zone Width} = \frac{\sqrt{(1.000341405R)^2 - 0.999658595R^2}}{1000} * 2 = 471km \approx 293 \text{ miles}$$

THE U.S. STATE PLANE COORDINATE SYSTEM (SPCS)

The SPCS zones in the United States were designed in the 1930s for use on the NAD27. Although other projection options were considered for use on the NAD83, the defining SPCS zone parameters were largely unchanged for implementation on the NAD83. The SPCS on the NAD83 consists of fifty-four transverse Mercator projections, sixty-eight Lambert conic conformal projections, and one oblique Mercator projection. Some states are covered by a single zone, but most states require more than one zone due to the limiting width of 158 miles and due to choosing SPCS zone boundaries to follow county boundaries. Other incidental changes were made during the transition from NAD27 SPCS to NAD83 SPCS and can be gleaned from two important publications. Claire (1968) is the "bible" for working with SPC on the NAD27, and Stem (1989) is the "bible" for working with SPC on the NAD83. Each booklet contains a description of the underlying map projections, a listing of the defining parameters for each zone, and a list of equations that can be used to perform bidirectional transformations between latitude/longitude positions and plane coordinates on the respective datum.

History

The following quote is found in a section entitled "SPCS—UTM and Oscar S. Adams" by Joseph Dracup, former geodesist for the U.S. Coast & Geodetic Survey (USC&GS), now the National Geodetic Survey (NGS), http://www.ngs.noaa.gov/PUBS_LIB/geodetic_survey_1807.html (accessed 12 June 2007).

> In 1933–34, Oscar S. Adams ably assisted by Charles N. Claire developed the State Plane Coordinate System (SPCS) at the request of George F. Syme a North Carolina Highway engineer. Syme died shortly after the North Carolina system was developed being succeeded by O.B. Bestor to carry on the cause. Bestor was in charge of the State local control project established in 1933, later identified as the North Carolina Geodetic Survey. Most State and the few county projects involved in this program also were so named. Colonel C. H. Birdseye of the USGS, with a strong interest in Statewide coordinate grids[,] also participated in the several conferences leading to the decision to honor Syme's request.
>
> The first tables for computing Lambert coordinates were developed for North Carolina and the first tables for the transverse Mercator grid were for New Jersey. Tables were prepared for all States early in 1934. For the first time all horizontal control stations previously defined only by latitudes and longitudes would be available in easy to use plane coordinates.

Features

"Special Publication 235" (Mitchell and Simmons [1945] 1977) is a booklet that describes details of the state plane coordinate system. It is of both practical and significant historical value because it documents surveying policies and practices prior to the electronic revolution. Several important features of the SPCS described in "Special Publication 235" include the following:

- The state plane coordinate system provides a method by which the latitude/longitude positions of the national triangulation network can be represented by plane coordinates. That meant local surveyors and/or engineers could continue using plane surveying procedures yet realize the benefits of basing their work on the national network of geodetic control points established by the federal agencies. This item is still valid in the 2-D arena (a subset of the 3-D arena). But, spatial data are 3-D and the GSDM does for 3-D data what the SPCS does for 2-D data.
- Normal land-surveying measurements in the 1930s were made with a transit and steel tape. Expected accuracies were often in the range of 1:5,000 to 1:8,000 or better. Under those circumstances, a routine distance distortion of 1:10,000 could be tolerated without making a scale factor correction and without significant detrimental impact on the quality of the survey. With newer technology, this assumption is no longer valid because measurement accuracies today routinely exceed those of eighty years ago. Better accuracy

is not a problem because high-quality computational results are obtained by applying the grid scale factor correction. With the corrections applied, the SPCS is fundamentally sound for 2-D applications. Elevation is typically used to handle the third dimension.

- There are two distance "corrections" to be made when working with the SPCS (Burkholder 1993a): (1) the grid scale factor is used to correct for the distortion between the ellipsoid and the grid, and (2) the elevation factor is needed to reduce a ground-level horizontal distance to the ellipsoid. These two corrections are often combined into one "combination factor" (the product of the grid scale factor and the elevation factor). The grid distance between the plumb lines through two points is the product of the horizontal ground distance and the combination factor. "Special Publication 235" explains both factors quite well, but, as discussed later, this is the primary disadvantage of using the SPCS. Regretfully, when using the SPCS, a foot on the grid is not necessarily a foot on the ground. In many cases, such as centerline stationing on a highway project, the difference between grid and ground distances becomes intolerable (see Burkholder 1993b, app. 3).
- Although the NGS has always performed and computed its geodetic surveys in meter units, the NAD27 state plane coordinates were published in foot units—see the sidebar discussion of the U.S. Survey Foot on page 259.

It is not true, as some have said, that the state plane coordinate systems distort distances by 1:10,000. It is true to say that, when compared to a distance on the map, the equivalent distance on the ellipsoid *may* be distorted by up to 1:10,000. On a secant projection, the distortion is zero along the lines of exact scale where the two surfaces intersect and the distance on the map is the same as the distance on the ellipsoid. At the center of the zone, the distance is compressed by 1:10,000 or by whatever distortion value was selected by the zone designer. In some cases, a zone width of 158 miles was not quite sufficient to cover the area required, and the distance distortion at the center of the zone is greater than 1:10,000 (i.e., the grid scale factor at the zone center is less than 0.9999)—see constants for California Zone 1, both Oregon zones, Zone 10 in Alaska, North Carolina, South Carolina, four of the five Texas zones, Utah Central Zone, and the offshore zone for Louisiana.

The grid scale factor is only part of the distortion. The elevation factor also contributes to the difference between a horizontal ground distance and the state plane grid distance. Modern practice looks more closely at the grid-ground distance difference (as a result of using the combination factor), and many resort to using surface coordinates or project datums as a way to avoid the mismatch between grid and ground distances. More recently, the use of "low-distortion projections" has been discussed as being a way to minimize the grid-ground distance distortion. The distance distortion issue is largely moot when using the GSDM.

NAD27 AND NAD83

The NAD27 was the only logical datum choice available when the state plane coordinate zones were developed during the 1930s. The zones were selected by matching the projection type with the state's general configuration. Lambert conic projections were selected for states long in the east-west dimension, while transverse Mercator projections were selected for states oriented primarily north-south. Some states have only one projection, other states require more than one zone to cover the needed width, and some states have more than one projection type. For example, the State of Florida utilizes two transverse Mercator projections and one conic projection, New York employs three transverse Mercator projections and one conic projection, and the State of Alaska uses nine transverse Mercator projections, one conic projection, and one oblique Mercator projection.

In the 1930s, the USC&GS developed a "model law," which was promoted by the Council of State Governments for several decades. By 1971 the SPC model law was adopted in one form or another by twenty-six states (Mitchell and Simmons [1945] 1977). However, the Michigan Legislature adopted a different projection than that proposed by the USC&GS. Originally, Michigan was to be covered by three transverse Mercator projections, but when the state plane coordinate law was written, professionals within the state opted instead for three conic conformal projections based upon an elevated reference surface selected to minimize the need for the elevation reduction. The elevated system worked as intended and was deemed very beneficial, but, because it was "nonstandard," there was confusion both in practice and in the published literature about computing the correct combination factor for a line (Burkholder 1980). The Michigan state plane coordinate law for NAD83 returned the reference surface to the ellipsoid.

Relationship between the Meter, the International Foot, and the U.S. Survey Foot

1. The length of the meter was established in the 1790s as 1/10,000,000 of the distance from the equator to the North Pole as determined by a geodetic survey in France.
2. In the early 1800s, prototype meter bars were made and distributed to the nations of the world.
3. Although the meter has been used as the standard of length for geodetic surveys in the United States since the establishment of the Coast Survey (predecessor to the NGS) in 1807, the meter length unit was declared legal for trade in the United States in 1866. The relationship between the foot and meter was stated in 1866 to be 39.37 feet = 12.00 meters exactly.
4. Leading up to and during World War II, Canada, the United States, and Great Britain each used a slightly different relationship between the foot and meter.
   United States: 1.00 meter = 39.37 inches, or 1 inch = 2.540005 cm
   England: 1 inch = 2.539997 cm
   Canada: 1 inch = 2.540000 cm
5. Following World War II, machinists and aircraft mechanics, working under the auspices of NATO, discovered that parts of aircraft engines built according to the same blueprints were not interchangeable due to differences in

unit definitions. A compromise was reached that adopted the Canadian relationship (1 inch = 2.54 centimeters) as the International Foot (1 foot = 0.3048 meters).

6. However, to avoid recomputing and republishing thousands of existing state plane coordinates, the United States retained use of 12 meters = 39.37 feet and gave that long-standing relationship a name—the U.S. Survey Foot. A 1959 *Federal Register* notice ("*Federal Register* Notice" 1959) stated that the U.S. Survey Foot should be used "until such time as it becomes desirable to readjust the basic geodetic networks in the United States, *after which the ratio of a yard, equal to 0.9144 meter, shall apply*" (emphasis added).
7. In 1960 the Eleventh General Conference of Weights and Measures redefined the meter, but not the length. The redefinition made it possible to duplicate the 1.00 meter distance in terms of wavelengths of Krypton 86 gas instead of relying upon the distance between two marks on a prototype bar.
8. The definition of the length of the meter was changed again in 1983—this time in terms of the distance light would travel in a vacuum in 1/299,792,458 seconds. The new definition is the equivalent to saying that light travels 299,792,458 meters in one second.

Although the definition used for duplicating the length of the meter has evolved over the years, the fundamental unit of length has not changed. The relationship of 12.00 meters = 39.37 feet has existed in the United States for over one hundred years. The name "U.S. Survey Foot" was developed in 1959 to describe the relationship already in existence. "International Foot" is the name given to the relationship used before 1959 by Canada (1 foot = 0.3048 meters) and adopted for use around the world (except for surveying and mapping in the United States). Neither the U.S. Survey Foot nor the International Foot is part of the International System of Units (SI) adopted by the Eleventh General Conference on Weights and Measures in 1960.

When the NAD27 datum was readjusted and published as the NAD83, the legislative intent was for the International Foot to be used as an alternate to meters. Recognizing that, a number of states included the International Foot in the state plane coordinate legislation written and adopted to accommodate the NAD83. Other states objected and ultimately won. A notice published in the *Federal Register* on May 16, 1998, closes by saying, "The effect of this notice is to allow the U.S. Survey Foot to be used indefinitely for surveying and mapping in the United States. No other part of the 1959 notice is in any way affected by this notice." The NGS still uses meter units for all geodetic surveying operations.

The upshot is that NAD83 state plane coordinates in the United States may be meters, U.S. Survey Feet, or International Feet. Although the GSDM is based exclusively on metric units, each user has the option of specifying different linear units when displaying or printing P.O.B. results. That is, provision is made for other derived units in the P.O.B. datum option. However, it is intended that the underlying ECEF coordinates will always be metric when using the GSDM.

Current Status: NAD83 State Plane Coordinate Systems

Although developed for use on the NAD27, design of the SPCS was revisited prior to publication of the readjusted NAD83 in North America. Arguments were advanced for taking advantage of the standardization offered by the UTM system, and using 2° UTM zones on the NAD83 was considered. After many discussions and consideration of various alternatives, the decision was to adopt parameters of a different ellipsoid (the GRS 1980 in place of Clarke 1866) and to move the datum origin from "Meade's Ranch," Kansas, to the Earth's center of mass. But, with exceptions, the existing SPCS projections and zone parameters were retained for use on the NAD83. Notable exceptions include the following:

- The reference surface for Michigan was returned to the ellipsoid instead of being computed at an elevation of 800 feet.
- Zone 7 in California was eliminated. Zone 5 now covers that area.
- The states of Montana, Nebraska, and South Carolina elected to relax the arbitrary 1:10,000 criteria and to cover each state respectively with one zone.

Advantages

The advantages of using the SPCS today are largely the same as when the SPCS was first implemented. A map projection flattens a portion of the Earth and allows one to perform 2-D rectangular surveying computations within a defined zone using plane Euclidean geometry. Standardization and wide acceptance are two huge benefits. An incidental benefit of the SPC is that the back azimuth of a line is the same as the forward azimuth + 180°. This feature could also be called a disadvantage because it belies the fact that meridians are not parallel, but converge at the poles.

Disadvantages

A disadvantage of the SPCS for the GIS community is the absence of uniqueness. For inventory, and other purposes, it is highly desirable for the description of any point location to be globally unique. State plane coordinates are unique within a zone but not globally. In addition to knowing the state plane coordinate values for a point, the spatial data user must also know what zone or map projection is associated with the point. Two points having the same (or nearly so) coordinate values may appear to be the same or very close together although they are, in fact, many kilometers apart. A triplet of ECEF rectangular *X/Y/Z* metric coordinates used in the GSDM is unique within the "birdcage" of orbiting GPS satellites.

In the surveying, mapping, and engineering communities, the biggest disadvantage of using map projections and the SPCS is that they are strictly 2-D mathematical models and spatial data users work with 3-D data. The GSDM is a rigorous 3-D model. Specific drawbacks to using the SPCS are listed by Burkholder (1993a) as follows:

- Lack of accessibility: control points are not easy to visit, permission, and so on.
- Lack of proximity: control points are too far away.

- Lack of quality: the published positions are not of sufficiently high quality.
- Lack of understanding: spatial data users need to learn more about the SPCS.
- Mapping distortion: ground distance may differ too much from grid distance.

With the advent of GPS, continued densification of the control network, higher levels of support from NGS, and greater awareness within the spatial data user community, the first four disadvantages have been significantly mitigated. But, the grid-ground difference is more of a problem than ever because more and more people are using equipment and computational processes in which that systematic difference cannot be tolerated. An argument is that more education and better enforcement of minimum standards could overcome those disadvantages. Without discounting the benefits of more education, it is suggested that using the GSDM is another alternative in which spatial data users can fully exploit the three-dimensional characteristics of their data and in which 2-D applications are still supported as a subset of the 3-D model.

PROCEDURES

Recognizing that full implementation of the GSDM will take some time; this section is included to help readers become more comfortable with the transition. Although a competent 3-D least squares network adjustment can more fully utilize the 3-D characteristics of GPS measurements, current practices involving 1-D or 2-D data will not be replaced instantaneously. Therefore, this section provides a summary of procedures commonly used when working with state plane coordinates. The overall point to remember is that computing a state plane coordinate traverse is the same as computing a regular plane surveying traverse with the following exception—one must use grid azimuths and grid distances. There are many sources of information and software available for instructions on how to compute and adjust a traverse. Those points are only summarized here.

Grid Azimuth

Conformal projections are used for all the state plane coordinate systems in the United States. That means that an angle measured on the ground is the same as the angle on the map and that a field-measured angle added to or subtracted from a known grid azimuth will give a grid azimuth. The implication is that one should always start with a grid azimuth. Two common methods for beginning with a grid azimuth are:

1. Backsight another point having known state plane coordinates. Doing that, the grid azimuth from standpoint to the backsight is computed using the plane coordinate inverse, $\tan \alpha = \Delta e/\Delta n$, as used in equation 4.11, 4.12, or 4.13, depending upon the quadrant.
2. Perform an astronomical observation using a star or the sun as the backsight to determine the astronomical azimuth to the foresight. Depending upon the quality of the grid azimuth required, two corrections are needed. A Laplace correction (equation 8.5) is used to convert the observed and/or computed astronomical azimuth to a geodetic azimuth, and the convergence (between

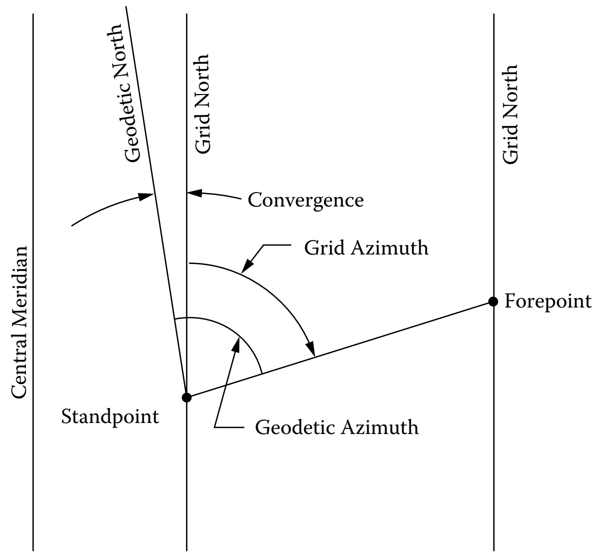


FIGURE 10.6 Convergence of Meridians

geodetic north and grid north at the station) is used in equation 10.4 to convert the geodetic azimuth to a grid azimuth. A generic diagram is shown in Figure 10.6 and shows that geodetic azimuth = grid azimuth + convergence.

$$\text{grid azimuth} = \text{geodetic azimuth} - \text{convergence} \tag{10.4}$$

Note that if the point lies west of the central meridian, the convergence is a negative quantity but equation 10.4 remains valid. Another point is that for lines over about 2 kilometers long, an arc-to-chord (known as the "t-T") correction may be needed to preserve high-quality results. See Stem (1989).

Grid Distance

An important design feature of the state plane coordinate systems is that the grid distance will approximate the ellipsoid distance within 1 part in 10,000. The grid scale factor is used to convert an ellipsoid distance to a grid distance. Regretfully, most surveys are conducted at some elevation and not on the ellipsoid. Therefore, an additional reduction is required to convert a ground-level horizontal distance to an ellipsoid distance. And, going back one step further, there are several options for computing a precise horizontal distance from observed slope distance and vertical (zenith) angles. For surveys of nominal accuracy, horizontal distance, HD(1), is computed as the right triangle component of slope distance and vertical angle. For surveys of higher accuracy and distances over about 2 km, horizontal distance is taken to be HD(2), the tangent plane distance between plumb lines, and involves computing a correction due to the plumb lines not being parallel. Such details are beyond the scope of this book but can be found in Burkholder (1991).

Reliable horizontal distance is critical when using state plane coordinates because it is reduced from horizontal to ellipsoid, then from the ellipsoid to the state plane grid. A high-quality grid distance relies upon the integrity of each part of the computational process. By contrast, the slope distance in 3-D space is used by the GSDM, and the underlying model obviates reduction of slope distance to grid. However, once a geocentric *X*/*Y*/*Z* position at the standpoint is computed, various definitions of horizontal distance between points can still be computed from traditional inverse computations.

When reducing a horizontal distance to the ellipsoid, the user often is faced with the choice of reducing horizontal distance to the ellipsoid or to sea level (the geoid). The "best" choice is to use ellipsoid height at the station and reduce horizontal distance to the ellipsoid rather than the geoid. The difference is whether or not one uses the geoid height portion of the elevation reduction equation. The summary below includes both.

Finally, the ellipsoid distance must be reduced to the state plane grid. For lines less than 1 km long, one can use the grid scale factor at either end of the line or at the middle. For lines longer than 1 km and less than about 4 km, it is acceptable to use the average grid scale factor. For lines longer than 4 km, the grid scale factor should be computed using the Simpson 1/6 Rule—see Stem (1989, 50). Figure 10.7 shows a diagram illustrating the distance reductions.

As a summary, traditional state plane grid distances are computed as follows:

1. Slope distance to horizontal:

$$HD(1) = (\text{slope distance}) * \sin(\text{zenith direction}) \tag{10.5}$$

HD(2) = A more precise option—see Burkholder (1991).

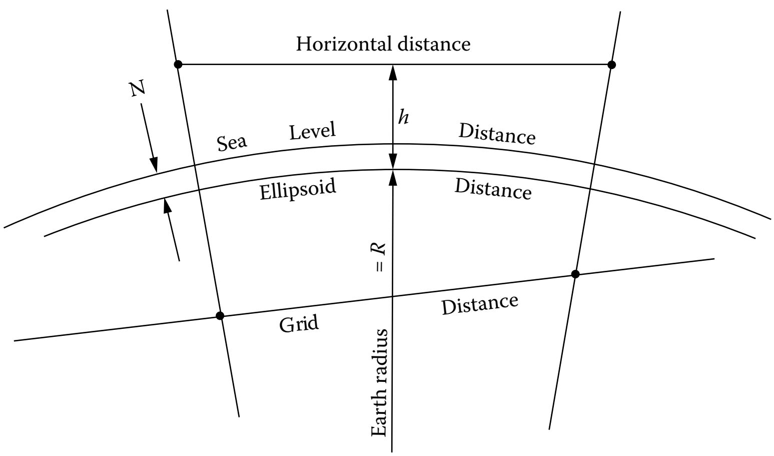


FIGURE 10.7 Horizontal, Sea Level, Ellipsoid, and Grid Distance

2. Horizontal to ellipsoid and/or sea level

a. $$\textit{ellipsoid distance} = \textit{horizontal distance} * \left(\frac{R}{R+h}\right) \quad (10.6)$$

b. $$\textit{sea level distance} = \textit{horizontal distance} * \left(\frac{R}{R+h-N}\right) \quad (10.7)$$

where

R = radius of Earth,
h = ellipsoid height, and
N = geoid height.

Equation 10.6 is recommended and should be used. However, it can be argued that it makes little or no difference which is used—equation 10.6 or 10.7. For an analysis of the difference, see Burkholder (2004).

3. Ellipsoid to grid: equation 10.8 is valid for all except very long lines. The difference is in computation of the appropriate grid scale factor.

$$\text{grid distance} = \text{ellipsoid distance} * \text{grid scale factor} \quad (10.8)$$

A. For "short" lines, the grid scale factor for any part of the line can be used.
B. For lines 2 to 4 km, the average grid scale factor for the line gives good results.
C. For lines over 4 km long, use the Simpson 1/6 Rule to compute the grid scale factor for a long line.

Steps 2 and 3 above are often combined into a single step by using the combined factor for a line. The combined factor is the product of the grid scale factor and the elevation factor **at a point** and makes converting grid distance to ground distance (and vice versa) more efficient. But the temptation is to use a single combined factor for an entire project without investigating how it changes from point to point. It is true the same combined factor can be used for a given elevation over a specified area, but each user should be aware of how the factor changes.

$$\text{combined factor} = \text{grid scale factor} * \text{elevation factor} \quad (10.9)$$
$$\text{grid distance} = \text{horizontal distance} * \text{combined factor} \quad (10.10)$$
$$\text{horizontal distance} = \text{grid distance} / \text{combined factor} \quad (10.11)$$

TRAVERSES

The primary advantage of using a state plane coordinate traverse is that one can use simple plane surveying procedures to establish a "big picture" position on each traverse point—be it latitude/longitude or state plane coordinates. Two other huge benefits of using state plane coordinates are that the procedures are long adopted (standard) and that a state plane traverse can begin on one point and close on another, even distant, point. Otherwise, a traverse must return to the beginning point in order to determine the traverse misclosure as a check on possible blunders.

Loop Traverse

A loop traverse is one that begins and ends at the same point, forming a closed loop. If one begins with state plane coordinates at the point of beginning and uses grid azimuths and grid distances, then it is a state plane coordinate traverse. The sum of the latitudes (north/south components of each course) and the sum of the departures (east/west components of each course) should each add up to zero. Any difference is the traverse misclosure and provides the basis of a traverse adjustment. Typically a loop traverse is adjusted by the Compass Rule. Other methods exist, but the Compass Rule is quite simple to apply and, if used properly, delivers good results. Although a loop traverse may be quite useful, a point-to-point traverse is preferred because it provides better azimuth control and helps prevent possible scaling problems.

Point-to-Point Traverse

A point-to-point traverse is a mathematically closed traverse that starts on one known point and ends on another. The traverse misclosure is determined as the difference of observed (computed) value minus the published (known) value, and the traverse is typically adjusted by the Compass Rule. When using a point-to-point traverse, an angular misclosure and adjustment should be completed before the latitude/departure misclosures are computed. The purist will argue that a least squares adjustment is better than a Compass Rule adjustment, and that may be true. But, by comparison, a Compass Rule adjustment is very easy to perform and achieves most of the benefits of a least squares adjustment.

ALGORITHMS FOR TRADITIONAL MAP PROJECTIONS

Although the focus of this book is the 3-D GSDM, the topic of map projections remains vital for many applications—especially for data visualization. Three excellent sources of information on the broad topic of map projections include Pearson (1990), Richardus and Alder (1972), and Snyder (1987). However, since various geomatics applications make extensive use of the state plane coordinate systems, the state plane coordinate map projection algorithms are included for the benefit of those reading them. The algorithms for the BK10 (forward) and BK11 (inverse) computations, described in chapter 1, are given (as used in the northern hemisphere) for the Lambert conic conformal projection, the transverse Mercator projection, and the oblique Mercator projection. A subsequent section references Burkholder (1993a) and describes how those same equations can be modified to accommodate low-distortion projections, project datums, or the use of surface coordinates as described in chapter 1. Specific rigorous equations for the BK14 and BK15 transformations (there are others) are described at the end of this chapter.

Although not technically prohibited from being used on the WGS84 datum or the ITRF datum, the following algorithms are specifically intended to be used with the NAD83 on the GRS 1980 ellipsoid. Except where noted, the equations and symbols are intended to be consistent with those used in "NOAA Manual NOS NGS 5, State Plane Coordinate System of 1983" (Stem 1989). Within a very small tolerance, the results obtained using these equations should be identical to those obtained using the equations and procedures given in said manual. State plane coordinate zone parameters are listed in appendix A. The following algorithm is included in a paper presented by Burkholder (1985), which contains a description, the algorithm, a flowchart, and a FORTRAN listing of a computer program to compute zone constants and perform transformations (BK10 and BK11) on all three projections.

Lambert Conic Conformal Projection

Any reference ellipsoid could be used, but state plane coordinates are based upon the GRS 1980 ellipsoid as listed here.

$$a = \text{semimajor axis} = 6{,}378{,}137.000 \text{ m} \tag{10.12}$$

$$1/f = \text{reciprocal flattening} = 298.2572221008827 \tag{10.13}$$

Compute ellipsoid constants:

$$e^2 = 2f - f^2 \quad \text{and} \quad e = \sqrt{e^2} \quad \text{eccentricity squared and eccentricity} \tag{10.14}$$

$$c_2 = \frac{e^2}{2} + \frac{5e^4}{24} + \frac{e^6}{12} + \frac{13e^8}{360} + \frac{3e^{10}}{160} \tag{10.15}$$

$$c_4 = \frac{7e^4}{48} + \frac{29e^6}{240} + \frac{811e^8}{11{,}520} + \frac{81e^{10}}{2{,}240} \tag{10.16}$$

$$c_6 = \frac{7e^6}{120} + \frac{81e^8}{1{,}120} + \frac{3{,}029e^{10}}{53{,}760} \tag{10.17}$$

$$c_8 = \frac{4{,}279e^8}{161{,}280} + \frac{883e^{10}}{20{,}160} \tag{10.18}$$

$$c_{10} = \frac{2{,}087e^{10}}{161{,}280} \tag{10.19}$$

Equations 10.15 through 10.19 are used to compute the following F coefficients.

$$F_0 = 2\left(c_2 - 2c_4 + 3c_6 - 4c_8 + 5c_{10}\right) \tag{10.20}$$

$$F_2 = 8\left(c_4 - 4c_6 + 10c_8 - 20c_{10}\right) \tag{10.21}$$

$$F_4 = 32\left(c_6 - 6c_8 + 21c_{10}\right) \tag{10.22}$$

$$F_6 = 128\left(c_8 - 8c_{10}\right) \tag{10.23}$$

$$F_8 = 512c_{10} \tag{10.24}$$

The *F* coefficients are in radian units and are used in the BK11 and BK15 transformations to compute geodetic latitude without iterating.

Input the defining parameters for the Lambert projection zone of the user's choice:

φ_n = latitude of north standard parallel.
φ_s = latitude of south standard parallel.
φ_b = latitude of false origin (usually where northing = 0.0 meters).
λ_0 = longitude of central meridian; east is +, and west is –.
E_0 = false easting on central meridian (meters).
N_b = northing on false origin (usually 0.0 meters).

Compute projection constants:

ln = natural logarithm
exp(x) = ε^x where ε = 2.71828… (base of natural logarithms)
Q_i = isometric latitude for corresponding geodetic latitude
W_i = intermediate computational value at φ_n and φ_s

$$Q_n = \frac{1}{2}\left[\ln\left(\frac{1+\sin\varphi_n}{1-\sin\varphi_n}\right) - e\ln\left(\frac{1+e\sin\varphi_n}{1-e\sin\varphi_n}\right)\right] \tag{10.25}$$

$$Q_s = \frac{1}{2}\left[\ln\left(\frac{1+\sin\varphi_s}{1-\sin\varphi_s}\right) - e\ln\left(\frac{1+e\sin\varphi_s}{1-e\sin\varphi_s}\right)\right] \tag{10.26}$$

$$Q_b = \frac{1}{2}\left[\ln\left(\frac{1+\sin\varphi_b}{1-\sin\varphi_b}\right) - e\ln\left(\frac{1+e\sin\varphi_b}{1-e\sin\varphi_b}\right)\right] \tag{10.27}$$

$$W_n = \sqrt{1 - e^2 \sin^2 \varphi_n} \quad \text{and} \quad W_s = \sqrt{1 - e^2 \sin^2 \varphi_s} \qquad \text{(10.28 and 10.29)}$$

$$\varphi_0 = \sin^{-1}\left[\frac{\ln\left(\mathrm{W}_n \cos \phi_s\right) - \ln\left(W_s \cos \phi_n\right)}{Q_n - Q_s}\right] \quad \text{latitude of central parallel} \qquad (10.30)$$

$$K = \frac{a \cos \varphi_s \exp\left(Q_s \sin \varphi_0\right)}{W_s \sin \varphi_0} = \frac{a \cos \varphi_n \exp\left(Q_n \sin \varphi_0\right)}{W_n \sin \varphi_0}$$

mapping radius of equator (10.31)

$$Q_0 = \frac{1}{2}\left[\ln\left(\frac{1 + \sin \varphi_0}{1 - \sin \varphi_0}\right) - e \ln\left(\frac{1 + e \sin \varphi_0}{1 - e \sin \varphi_0}\right)\right] \quad \text{isometric latitude of } \varphi_0 \qquad (10.32)$$

$$R_b = \frac{K}{\exp\left(Q_b \sin \varphi_0\right)} \quad \text{mapping radius of latitude of origin} \qquad (10.33)$$

$$R_0 = \frac{K}{\exp\left(Q_0 \sin \varphi_0\right)} \quad \text{mapping radius of central parallel} \qquad (10.34)$$

$$k_0 = \frac{R_0 \tan \varphi_0 \sqrt{1 - e^2 \sin^2 \varphi_0}}{a} \quad \text{grid scale factor at center of zone} \qquad (10.35)$$

The preceding zone constants need be computed only once for a given projection, but they are used repeatedly in the following BK10 (forward) and BK11 (inverse) computations.

BK10 (Forward) Transformation on Lambert Conic Conformal Projection

Input:

φ = geodetic latitude (positive north)
λ = geodetic longitude (positive east)

Compute:

$$Q_\varphi = \frac{1}{2}\left[\ln\left(\frac{1+\sin\varphi}{1-\sin\varphi}\right) - e\ln\left(\frac{1+e\sin\varphi}{1-e\sin\varphi}\right)\right] \text{ isometric latitude of point } (\varphi, \lambda) \quad (10.36)$$

$$R_\varphi = \frac{K}{\exp\left(Q_\varphi \sin\varphi_0\right)} \text{ mapping radius of point } (\varphi, \lambda) \quad (10.37)$$

$$\gamma = \left(\lambda - \lambda_0\right)\sin\varphi_0 \text{ convergence at point } (\varphi, \lambda) \quad (10.38)$$

$$k = \frac{R_\varphi \sin\varphi_0 \sqrt{1-e^2\sin^2\varphi}}{a\cos\varphi} \text{ grid scale factor at point } (\varphi, \lambda) \quad (10.39)$$

$$E = E_0 + R_\varphi \sin\gamma \text{ easting for point } (\varphi, \lambda) \quad (10.40)$$

$$N = R_b + N_b - R_\phi \cos\gamma \text{ northing for point } (\varphi, \lambda) \quad (10.41)$$

BK11 (Inverse) Transformation on Lambert Conic Conformal Projection

Input:

E = easting of point within defined map projection
N = northing of point within defined map projection

Compute:

$$R' = R_b - N + N_b \text{ intermediate value} \quad (10.42)$$

$$E' = E - E_0 \text{ intermediate value} \quad (10.43)$$

$$\gamma = \tan^{-1}\left(\frac{E'}{R'}\right) \text{ convergence at point } (E, N) \quad (10.44)$$

$$R_\varphi = \sqrt{R'^2 + E'^2} \quad \text{mapping radius at point } (E, N) \tag{10.45}$$

$$Q_\varphi = \frac{\ln K - \ln R_\varphi}{\sin \varphi_0} \quad \text{isometric latitude at point } (E, N) \tag{10.46}$$

$$\chi = 2\tan^{-1}\left(\frac{\exp(Q_\varphi) - 1}{\exp(Q_\varphi) + 1}\right) \quad \text{conformal latitude at point } (E, N) \tag{10.47}$$

$$\varphi = \chi + \sin\chi\cos\chi\left(F_0 + \cos^2\chi\left(F_2 + \cos^2\chi\left(F_4 + \cos^2\chi\left(F_6 + F_8\cos^2\chi\right)\right)\right)\right)$$
$$\text{geodetic latitude at point } (E, N) \tag{10.48}$$

$$\lambda = \lambda_0 + \frac{\gamma}{\sin\varphi_0} \quad \text{east longitude at point } (E, N) \tag{10.49}$$

$$k = R_b \sin\varphi_0 \frac{\sqrt{1 - e^2 \sin^2\varphi}}{a\cos\varphi} \quad \text{grid scale factor at point } (E, N) \tag{10.50}$$

The State of Oregon uses a Lambert projection for its state plane coordinate system. Figure 10.8 is a computer printout showing example Lambert conic conformal BK10 and BK11 transformations at Station "Median 2" on the campus of the Oregon Institute of Technology, located in Klamath Falls. The transformations were computed using the equations in this section, and numerical values match those shown on the NGS data sheet for the same station.

Transverse Mercator Projection

Any reference ellipsoid could be used, but state plane coordinates are based upon the GRS 1980 ellipsoid as listed here.

$$a = \text{semimajor axis} = 6{,}378{,}137.000 \text{ m} \tag{10.51}$$

$$1/f = \text{reciprocal flattening} = 298.2572221008827 \tag{10.52}$$

Compute ellipsoid constants:

```
                    PROGRAM: LOCALCOR
          COPYRIGHT 2001 BY GLOBAL COGO, INC.
              LAS CRUCES, NEW MEXICO 88003
                  WWW.GLOBALCOGO.COM


USER: Earl F. Burkholder
DATE: 12 June 2007


LAMBERT CONIC CONFORMAL COORDINATE TRANSFORMATIONS
PROJECTION NAME: Oregon South Zone - 3602

REFERENCE ELLIPSOID: GEODETIC REFERENCE SYSTEM 1980
                       A =  6378137.0000 METERS
                     1/F =  298.2572221008827

ZONE PARAMETERS:
     NORTH STANDARD PARALLEL              44  0   .000000
     SOUTH STANDARD PARALLEL              42 20   .000000
     FALSE ORIGIN LATITUDE                41 40   .000000
     CENTRAL MERIDIAN (W)                120 30   .000000
     FALSE EASTING ON CM                     1500000.0000 METERS
     NORTHING AT FALSE ORIGIN                       .0000 METERS

ZONE CONSTANTS:
     CENTRAL PARALLEL                     43 10  6.919559
     SCALE FACTOR ON CENTRAL PARALLEL    .999894607592090
     MAPPING RADIUS OF EQUATOR             12033772.69836 METERS
     MAPPING RADIUS OF FALSE ORIGIN         6976289.23822 METERS
     NORTHING OF CENTRAL PARALLEL ON CM      166836.95660 METERS
     CONFORMAL LATITUDE CONSTANTS:  F(0) =  .006686920927
          F(2) =  .000052014583     F(4) =  .000000554458
          F(6) =  .000000006718     F(8) =  .000000000089


TRANSFORMATIONS:

NAME OF STATION: Median 2 - PID NY0996              FORWARD (BK10)

     LATITUDE:    42 15 15.611960   NORTHING     66102.3042 METERS
     LONGITUDE:  121 47 25.985950   EASTING    1393505.6444 METERS
     CONVERGENCE:      0-52 58.54   SCALE FACTOR:   1.000020826193


NAME OF STATION: Median 2 - PID NY0996              INVERSE (BK11)

     LATITUDE:    42 15 15.611959   NORTHING     66102.3042 METERS
     LONGITUDE:  121 47 25.985950   EASTING    1393505.6444 METERS
     CONVERGENCE:      0-52 58.54   SCALE FACTOR:   1.000020826194
```

FIGURE 10.8 Example BK10 and BK11 Transformations for Lambert Projection

$$e^2 = 2f - f^2 \qquad \text{and} \qquad e = \sqrt{e^2} \quad \text{eccentricity squared and eccentricity} \tag{10.53}$$

$$n = \frac{f}{\left(2 - f\right)} \quad \text{intermediate value} \tag{10.54}$$

$$r = a \quad \left(1 - n\right)\left(1 - n^2\right)\left(1 + \frac{9n^2}{4} + \frac{225n_4}{64}\right) \quad \text{intermediate value} \tag{10.55}$$

$$u_2 = \frac{-3n}{2} + \frac{9n^3}{16} \tag{10.56}$$

$$u_4 = \frac{15n^2}{16} - \frac{15n^4}{32} \tag{10.57}$$

$$u_6 = \frac{-35n^3}{48} \tag{10.58}$$

$$u_8 = \frac{315n^4}{512} \tag{10.59}$$

These intermediate values of u are used only in the equations that follow.

$$U_0 = 2 \ \left(u_2 - 2u_4 + 3u_6 - 4u_8\right) \tag{10.60}$$

$$U_2 = 8 \ \left(u_4 - 4u_6 + 10u_8\right) \tag{10.61}$$

$$U_4 = 32 \ \left(u_6 - 6u_8\right) \tag{10.62}$$

$$U_6 = 128u_8 \tag{10.63}$$

These values of U are used to compute zone constants and in the BK10 transformation.

$$v_2 = \frac{3n}{2} - \frac{27n^3}{32} \tag{10.64}$$

$$v_4 = \frac{21n^2}{16} - \frac{55n^4}{32} \tag{10.65}$$

$$v_6 = \frac{151n^3}{96} \tag{10.66}$$

$$v_8 = \frac{1097n^4}{512} \tag{10.67}$$

These values of v are used only in the equations that follow.

$$V_0 = 2\left(v_2 - 2v_4 + 3v_6 - 4v_8\right) \tag{10.68}$$

$$V_2 = 8\left(v_4 - 4v_6 + 10v_8\right) \tag{10.69}$$

$$V_4 = 32\left(v_6 - 6v_8\right) \tag{10.70}$$

$$V_6 = 128v_8 \tag{10.71}$$

These values of V are used in the BK11 transformation.

Input the defining parameters for a transverse Mercator projection of the user's choice:

λ_0 = longitude of central meridian; east longitude is +, and west is –.
E_0 = false easting on central meridian (meters).
k_0 = grid scale factor on central meridian.
φ_0 = latitude of false origin, usually where northing = 0.0 meters.
N_0 = false northing at false origin (usually 0.0 meters).

Compute coordinate system projection constants:

$$\omega_0 = \varphi_0 + \sin\varphi_0 \cos\varphi_0 \left(U_0 + \cos^2\varphi_0 \left(U_2 + \cos^2\varphi_0 \left(U_4 + U_6 \cos^2\varphi_0\right)\right)\right)$$

rectifying latitude of origin (10.72)

$$S_0 = rk_0\omega_0 \text{ distance on grid from equator to origin} \quad (10.73)$$

These constants are computed only once for each zone. After that, they are used in computing the BK10 and BK11 transformations. The nominal grid scale factor used on the central meridian for the state plane coordinate systems is 0.9999. Specific values are given in appendix A. Other values are chosen when designing a custom system.

BK10 (Forward) Transformation for Transverse Mercator Projection

Input:

φ = geodetic latitude (positive north)
λ = geodetic longitude (positive east)

Compute:

$$L = (\lambda - \lambda_0)\cos\varphi \text{ , } L \text{ in radians and positive east of central meridian} \quad (10.74)$$

$$t = \tan\varphi \quad (10.75)$$

$$\eta^2 = \frac{e^2\cos^2\varphi}{1-e^2} \quad (10.76)$$

$$\omega = \varphi + \sin\varphi\cos\varphi\left(U_0 + \cos^2\varphi\left(U_2 + \cos^2\varphi\left(U_4 + U_6\cos^2\varphi\right)\right)\right)$$

rectifying latitude (10.77)

$$S = rk_0\omega \text{ arc distance on grid to parallel through point } (\varphi, \lambda) \quad (10.78)$$

$$R = \frac{ak_0}{\sqrt{1-e^2\sin^2\varphi}} \quad (10.79)$$

$$A_1 = -R \tag{10.80}$$

$$A_2 = \frac{1}{2}\ R\ t \tag{10.81}$$

$$A_3 = \frac{1}{6}\left(\ 1 - t^2 + \eta^2\ \right) \tag{10.82}$$

$$A_4 = \frac{1}{12}\ \left(5 - t^2 + \eta^2\left(9 + 4\eta^2\right)\right) \tag{10.83}$$

$$A_5 = \frac{1}{120}\ \left(5 - 18t^2 + t^4 + \eta^2\left(14 - 58t^2\right)\right) \tag{10.84}$$

$$A_6 = \frac{1}{360}\ \left(61 - 58t^2 + t^4 + \eta^2\left(270 - 330t^2\right)\right) \tag{10.85}$$

$$A_7 = \frac{1}{5,040}\ \left(61 - 479t^2 + 179t^4 - t^6\right) \tag{10.86}$$

$$E = E_0 + A_1 L\left(1 + L^2\left(A_3 + L^2\left(L_5 + A_7 L^2\right)\right)\right) \text{ easting of point } (\varphi, \lambda) \tag{10.87}$$

$$N = N_0 + S - S_0 + A_2 L^2\left(1 + L^2\left(A_4 + A_6 L^2\right)\right) \text{ northing of point } (\varphi, \lambda) \tag{10.88}$$

$$C_1 = -t \tag{10.89}$$

$$C_2 = \frac{1}{2}\ \left(1 + \eta^2\right) \tag{10.90}$$

$$C_3 = \frac{1}{3}\left(1 + 3\eta^2 + 2\eta^4\right) \tag{10.91}$$

$$C_4 = \frac{1}{12}\left(5 - 4t^2 + \eta^2\left(9 - 24t^2\right)\right) \tag{10.92}$$

$$C_5 = \frac{1}{15}\left(2 - t^2\right) \tag{10.93}$$

$$\gamma = C_1 L\left(1 + L^2\left(C_3 + C_5 L^2\right)\right) \text{ convergence at point } (\varphi, \lambda) \tag{10.94}$$

$$k = k_0\left(1 + C_2 L^2\left(1 + C_4 L^2\right)\right) \text{ grid scale factor at point } (\varphi, \lambda) \tag{10.95}$$

BK11 (Inverse) Transformation for Transverse Mercator

Input:

E = easting of point within defined map projection
N = northing of point within defined map projection

Compute:

$$\omega = \frac{N - N_0 + S_0}{k_0 r} \tag{10.96}$$

$$\phi_f = \omega + \sin\omega\cos\omega\left(V_0 + \cos^2\omega\left(V_2 + \cos^2\omega\left(V_4 + V_6\cos^2\omega\right)\right)\right) \tag{10.97}$$

$$\eta_f^2 = \frac{e^2\cos^2\phi_f}{1 - e^2} \tag{10.98}$$

$$R_f = \frac{ak_0}{\sqrt{1 - e^2\sin^2\phi_f}} \tag{10.99}$$

$$Q = \frac{E - E_0}{R_f} \text{ radian units} \tag{10.100}$$

$$B_2 = \frac{-t_f\left(1+\eta_f^2\right)}{2} \tag{10.101}$$

$$B_3 = \frac{-\left(1+2t_f^2+\eta_f^2\right)}{6} \tag{10.102}$$

$$B_4 = \frac{-\left(5+3t_f^2+\eta_f^2\left(1-9t_f^2\right)-4\eta_f^4\right)}{12} \tag{10.103}$$

$$B_5 = \frac{\left(5+28t_f^2+24t_f^4+\eta_f^2\left(6+8t_f^2\right)\right)}{120} \tag{10.104}$$

$$B_6 = \frac{\left(61+90t_f^2+45t_f^4+\eta_f^2\left(46-252t_f^2-90t_f^4\right)\right)}{360} \tag{10.105}$$

$$B_7 = \frac{-\left(61+662t_f^2+1320t_f^4+720t_f^6\right)}{5040} \tag{10.106}$$

$$L = Q\left(1+Q^2\left(B_3+Q^2\left(B_5+B_7Q^2\right)\right)\right) \tag{10.107}$$

$$\phi = \phi_f + B_2Q^2\left(1+Q^2\left(B_4+B_6Q^2\right)\right) \quad \text{geodetic latitude of point } (E, N) \tag{10.108}$$

$$\lambda = \lambda_0 + \frac{L}{\cos\phi_f} \quad \text{geodetic longitude (east) of point } (E, N) \tag{10.109}$$

$$D_1 = t_f = \tan\phi_f \tag{10.110}$$

$$D_2 = \frac{1+\eta_f^2}{2} \quad \text{G}_2 \text{ in "Manual NOS NGS 5"} \tag{10.111}$$

$$D_3 = \frac{-\left(1+t_f^2-\eta_f^2-2\eta_f^4\right)}{3} \tag{10.112}$$

$$D_4 = \frac{1+5\eta_f^2}{12} \quad \text{G}_4 \text{ in "Manual NOS NGS 5"} \tag{10.113}$$

$$D_5 = \frac{2+5t_f^2+3t_f^4}{15} \tag{10.114}$$

$$\gamma = D_1 Q\left(1+Q^2\left(D_3+D_5 Q^2\right)\right) \text{, convergence at point } (N, E) \tag{10.115}$$

$$k = k_0\left(1+D_2 Q^2\left(1+D_4 Q^2\right)\right) \text{, grid scale factor at point } (N, E) \tag{10.116}$$

The State of New Mexico uses a transverse Mercator projection for its state plane coordinate system. Figure 10.9 is a computer printout showing example transverse Mercator BK10 and BK11 transformations at Station "Reilly" on the New Mexico State University campus. The transformations were computed using the equations in this section, and numerical values match those shown on the NGS data sheet for the same station.

Oblique Mercator Projection

Any reference ellipsoid could be used, but NAD83 state plane coordinates in Alaska Zone 1 are based upon the GRS 1980 ellipsoid as listed here.

$$a = \text{semimajor axis} = 6{,}378{,}137.000 \text{ m} \tag{10.117}$$

$$1/f = \text{reciprocal flattening} = 298.2572221008827 \tag{10.118}$$

```
                  PROGRAM: LOCALCOR
         COPYRIGHT 2001 BY GLOBAL COGO, INC.
             LAS CRUCES, NEW MEXICO 88003
                 WWW.GLOBALCOGO.COM


USER: Earl F. Burkholder
DATE: 12 June 2007


TRANSVERSE MERCATOR PROJECTION TRANSFORMATIONS
PROJECTION NAME: New Mexico Central Zone - 3002

REFERENCE ELLIPSOID: GEODETIC REFERENCE SYSTEM 1980
                       A = 6378137.0000 METERS
                     1/F = 298.2572221008827

ZONE PARAMETERS:

     CENTRAL MERIDIAN (W)                106 15   .000000
     LATITUDE OF FALSE ORIGIN             31  0   .000000
     FALSE NORTHING AT FALSE ORIGIN                 .0000 METERS
     FALSE EASTING ON CENTRAL MERIDIAN        500000.0000 METERS
     SCALE FACTOR ON CENTRAL MERIDIAN       .999900000000

ZONE CONSTANTS:

     RECTIFYING SPHERE RADIUS              6367449.1458 METERS
     RECTIFYING LATITUDE CONSTANTS:
          U(0) =   -.005048250776    V(0) =    .005022893948
          U(2) =    .000021259204    V(2) =    .000029370625
          U(4) =   -.000000111423    V(4) =    .000000235059
          U(6) =    .000000000626    V(6) =    .000000002181

     RECTIFYING LATITUDE OF FALSE ORIGIN     30 52 21.720626
     GRID MERIDIAN ARC TO FALSE ORIGIN   3430631.2260 METERS


TRANSFORMATIONS:

NAME OF STATION: Reilly – PID AI5445                FORWARD (BK10)

     LATITUDE:    32 16 55.929060   NORTHING    142268.7414 METERS
     LONGITUDE:  106 45 15.160700   EASTING     452506.4804 METERS
     CONVERGENCE:      0-16  9.48   SCALE FACTOR:    .999927806946


NAME OF STATION: Reilly – PID AI5445                INVERSE (BK11)

     LATITUDE:    32 16 55.929059   NORTHING    142268.7414 METERS
     LONGITUDE:  106 45 15.160700   EASTING     452506.4804 METERS
     CONVERGENCE:      0-16  9.48   SCALE FACTOR:    .999927806946
```

FIGURE 10.9 Example BK10 and BK11 Transformations for Transverse Mercator Projection

Compute ellipsoid constants:

$$e^2 = 2f - f^2 \quad \text{and} \quad e = \sqrt{e^2} \quad \text{eccentricity squared and eccentricity} \tag{10.119}$$

$$e'^2 = \frac{e^2}{1-e^2} \quad \text{second eccentricity squared} \tag{10.120}$$

$$c_2 = \frac{e^2}{2} + \frac{5e^4}{24} + \frac{e^6}{12} + \frac{13e^8}{360} + \frac{3e^{10}}{160} \tag{10.121}$$

$$c_4 = \frac{7e^4}{48} + \frac{29e^6}{240} + \frac{811e^8}{11{,}520} + \frac{81e^{10}}{2{,}240} \tag{10.122}$$

$$c_6 = \frac{7e^6}{120} + \frac{81e^8}{1{,}120} + \frac{3{,}029e^{10}}{53{,}760} \tag{10.123}$$

$$c_8 = \frac{4{,}279e^8}{161{,}280} + \frac{883e^{10}}{20{,}160} \tag{10.124}$$

$$c_{10} = \frac{2{,}087e^{10}}{161{,}280} \tag{10.125}$$

Equation 10.119 and equations 10.121 through 10.125 are used once in computing the F coefficients. These are the same F coefficients as used in equations 10.20 through 10.24.

$$F_0 = 2\left(c_2 - 2c_4 + 3c_6 - 4c_8 + 5c_{10}\right) \tag{10.126}$$

$$F_2 = 8\left(c_4 - 4c_6 + 10c_8 - 20c_{10}\right) \tag{10.127}$$

$$F_4 = 32\left(c_6 - 6c_8 + 21c_{10}\right) \tag{10.128}$$

$$F_6 = 128\left(c_8 - 8c_{10}\right) \tag{10.129}$$

$$F_8 = 512c_{10} \tag{10.130}$$

The F coefficients are radian units and are used in the BK11 and BK15 transformations to compute geodetic latitude without iterating.

Input the defining parameters for the oblique Mercator projection of the user's choice:

φ_c = latitude of local origin
λ_c = longitude (east) of local origin
k_0 = grid scale factor along projection axis
N_0 = false northing at (u, v) origin
E_0 = false easting at (u, v) origin
α_c = positive skew axis (u axis) azimuth at local origin

Compute zone constants:

$$B = \sqrt{1 + e'^2 \cos^4 \phi_c} \tag{10.131}$$

$$W_c = \sqrt{1 - e^2 \sin^2 \phi_c} \tag{10.132}$$

$$A = \frac{aB\sqrt{1-e^2}}{W_c^2} \tag{10.133}$$

$$D = \frac{Ak_c}{B} \tag{10.134}$$

$$Q_c = \frac{1}{2}\left[\ln\left(\frac{1+\sin\phi_c}{1-\sin\phi_c}\right) - e\ln\left(\frac{1+e\sin\phi_c}{1-e\sin\phi_c}\right)\right] \tag{10.135}$$

$$C = \cosh^{-1}\left(\frac{B\sqrt{1-e^2}}{W_c\cos\phi_c}\right) - BQ_c \tag{10.136}$$

Note: $\cosh^{-1} x = \ln\left(x + \sqrt{x^2 - 1}\right)$.

$$F = \sin\phi_0 = \frac{a \sin\alpha_c \cos\phi_c}{AW_c} \tag{10.137}$$

$$G = \cos\alpha_0 \tag{10.138}$$

$$I = \frac{Ak_c}{a} \tag{10.139}$$

$$\lambda_0 = \lambda_c + \frac{1}{B}\sin^{-1}\left[\frac{\sin\alpha_0 \sinh\left(BQ_c + C\right)}{\cos\alpha_0}\right] \tag{10.140}$$

Note: $\sinh x = \dfrac{\exp(x) - \exp(-x)}{2}$, where exp(x) = ε^x and ε = base of 2.71828….

BK10 (Forward) Transformation for Oblique Mercator Projection

Input:

φ = geodetic latitude (positive north)
λ = geodetic longitude (positive east)

Compute:

$$L = \left(\lambda_0 - \lambda\right)B \tag{10.141}$$

$$Q = \left[\ln\frac{1 + \sin\phi}{1 - \sin\phi} - e\ln\frac{1 + e\sin\phi}{1 - e\sin\phi}\right] \tag{10.142}$$

$$J = \sinh\left(BQ + C\right) \tag{10.143}$$

$$K = \cosh\left(BQ + C\right) \tag{10.144}$$

Note: $\cosh x = \dfrac{\exp(x) + \exp(-x)}{2}$, where exp(x) = ε^x and ε = base of 2.71828…...

$$u = D\tan^{-1}\left[\frac{JG - F\sin L}{\cos L}\right] \tag{10.145}$$

$$v = \left(\frac{D}{2}\right)\ln\left[\frac{K - FJ - G\sin L}{K + FJ + G\sin L}\right] \tag{10.146}$$

$$E = E_0 + u\sin\alpha_c + v\cos\alpha_c \text{ easting of point } (\varphi, \lambda) \tag{10.147}$$

$$N = N_0 + u\cos\alpha_c - v\sin\alpha_c \text{ northing of point } (\varphi, \lambda) \tag{10.148}$$

$$\gamma = \tan^{-1}\left[\frac{F - JG\sin L}{KG\cos L}\right] - \alpha_c \text{ convergence at point } (\varphi, \lambda) \tag{10.149}$$

$$k = \frac{I\sqrt{1 - e^2\sin^2\phi}\ \ \cos\left(\dfrac{u}{D}\right)}{\cos\phi\cos L} \text{ grid scale factor at point } (\varphi, \lambda) \tag{10.150}$$

BK11 (Inverse) Transformation for Oblique Mercator Projection

Input:

E = easting of point within defined map projection
N = northing of point within defined map projection

Compute:

$$u = \left(E - E_0\right)\sin\alpha_c + \left(N - N_0\right)\cos\alpha_c \tag{10.151}$$

$$v = (E - E_0)\cos\alpha_c - (N - N_0)\sin\alpha_c \tag{10.152}$$

$$R = \sinh\left(\frac{v}{D}\right) \tag{10.153}$$

$$S = \cosh\left(\frac{v}{D}\right) \tag{10.154}$$

$$T = \sin\left(\frac{u}{D}\right) \tag{10.155}$$

$$Q = \frac{1}{B}\left[\frac{1}{2}\ln\frac{S - RF + GT}{S + RF - GT} - C\right] \tag{10.156}$$

$$\chi = 2\tan^{-1}\left(\frac{\varepsilon^Q - 1}{\varepsilon^Q + 1}\right) \quad \varepsilon = \text{base of natural logarithms} = 2.71828182 \tag{10.157}$$

$$\phi = \chi + \sin\chi\cos\chi\left(F_0 + \cos^2\chi\left(F_2 + \cos^2\chi\left(F_4 + \cos^2\chi\left(F_6 + F_8\cos^2\chi\right)\right)\right)\right)$$

geodetic latitude of point (E, N) (10.158)

$$\lambda = \lambda_0 + \left(\frac{1}{B}\right)\tan^{-1}\left[\frac{RG + TF}{\cos(u/D)}\right] \quad \text{east geodetic longitude of point } (E, N) \tag{10.159}$$

$$L = (\lambda - \lambda_0)B \tag{10.160}$$

$$J = \sinh(BQ + C) \tag{10.161}$$

$$K = \cosh\left(BQ + C\right) \tag{10.162}$$

$$\gamma = \tan^{-1}\left[\frac{F - JG\sin L}{KG\cos L}\right] - \alpha_c \text{ , convergence of meridian at point } (E, N) \tag{10.163}$$

$$k = \frac{I\sqrt{1 - e^2\sin^2\phi}\,\cos(u / D)}{\cos\phi\cos L} \text{ , grid scale factor at point } (E, N) \tag{10.164}$$

Alaska Zone 1 is the only state plane coordinate system zone in the United States that uses the oblique Mercator projection. Figure 10.10 is a computer printout showing example oblique Mercator BK10 and BK11 transformations at Station "JNU C" (PID AI4906) near Juneau, Alaska. The transformations were computed using the equations in this section, and numerical values match those shown on the NGS data sheet for the same station.

Low-Distortion Projections

The nuisance of working with grid distance and ground distance (and their differences) can be avoided when using the GSDM. However, until the GSDM becomes a part of mainstream professional practice, many see and have used a low-distortion projection (LDP) as a way of solving that problem. Several issues with using an LDP in the context of the 3-D GSDM are discussed in chapter 12. The discussion here looks at 2-D issues in the context of the standard map projection. Various methods have been used to compute what some call "surface coordinates" and some call "project datum coordinates." The corresponding lack of standardization is a drawback to using an LDP, and that issue is also addressed in chapter 12. The method described here is rigorous and simple, but implementation suffers for various reasons. Very simply, all the equations listed in this chapter are applicable to an LDP with one modification—the value of the ellipsoid semimajor axis is increased by a value (elevation) chosen by the user (Burkholder 1993a). Whether working with a Lambert conic conformal projection, a transverse Mercator projection, or an oblique Mercator projection, the substitution is the same—replace the semimajor axis value a with $(a + h_{ref})$, where $\boldsymbol{h_{ref}}$ **= value selected by the user**. Specific occurrences are as follows:

Lambert Conic Conformal Projection

Equation 10.12: in place of a, use $a_{ref} = a + h_{ref}$.
Equation 10.31: in place of a, use $a_{ref} = a + h_{ref}$.

```
                    PROGRAM: LOCALCOR
          COPYRIGHT 2001 BY GLOBAL COGO, INC.
              LAS CRUCES, NEW MEXICO 88003
                  WWW.GLOBALCOGO.COM


USER: Earl F. Burkholder
DATE: 12 June 2007

OBLIQUE MERCATOR COORDINATE TRANSFORMATIONS
PROJECTION ZONE: Alaska Zone 1 - 5001

REFERENCE ELLIPSOID: GEODETIC REFERENCE SYSTEM 1980

     A =  6378137.0000 METERS    1/F =  298.2572221008827

ZONE PARAMETERS:

     LATITUDE OF ZONE CENTER               57  0   .000000
     LONGIUTDE OF ZONE CENTER             133 40   .000000
     SCALE FACTOR ALONG AXIS                 .999900000000
     FALSE EASTING AT (U,V) ORIGIN            5000000.0000 METERS
     FALSE NORTHING AT (U,V) ORIGIN          -5000000.0000 METERS
     AZIMUTH OF AXIS AT ORIGIN            323  7 48.368475

ELLIPSOID AND ZONE CONSTANTS

     FIRST ECCENTRICITY SQUARED           .00669438002290341 6
     SECOND ECCENTRICITY SQUARED          .006739496775481622
     CONFORMAL LATITUDE CONSTANTS:    F(0) =    .006686920927
          F(2) =    .000052014583     F(4) =    .000000554458
          F(6) =    .000000006718     F(8) =    .000000000089

          A    =  6388718.8623050     F    = -.32701295544998
          B    = 1.00029646140436     G    =  .94501985532997
          C    =  .00442683392641      I    = 1.00155891766182
          D    =  6386186.7325316


TRANSFORMATIONS:

NAME OF STATION: JNU C - PID AI4906                 FORWARD (BK10)

     LATITUDE:     58 21 14.364490   NORTHING    726233.5912 METERS
     LONGITUDE:  134 34 26.891220    EASTING     765542.9612 METERS
     CONVERGENCE:       0-45 42.27   SCALE FACTOR:     .999928422781


NAME OF STATION: JNU C - PID AI4906                 INVERSE (BK11)

     LATITUDE:     58 21 14.364490   NORTHING    726233.5912 METERS
     LONGITUDE:  134 34 26.891220    EASTING     765542.9612 METERS
     CONVERGENCE:       0-45 42.27   SCALE FACTOR:     .999928422781
```

FIGURE 10.10 Example BK10 and BK11 Transformations for Oblique Mercator Projection

Transverse Mercator Projection

Equation 10.51: in place of a, use $a_{ref} = a + h_{ref}$.
Equation 10.55: in place of a, use $a_{ref} = a + h_{ref}$.
Equation 10.79: in place of a, use $a_{ref} = a + h_{ref}$.
Equation 10.99: in place of a, use $a_{ref} = a + h_{ref}$.

Oblique Mercator Projection

Equation 10.117: in place of a, use $a_{ref} = a + h_{ref}$.
Equation 10.133: in place of a, use $a_{ref} = a + h_{ref}$.
Equation 10.137: in place of a, use $a_{ref} = a + h_{ref}$.
Equation 10.139: in place of a, use $a_{ref} = a + h_{ref}$.

With these modifications and with reference to Figure 1.4, the map projection "forward" and "inverse" transformations for low-distortion projections are referred to as BK14 and BK15 computations. The intent for that naming distinction is to avoid computational confusion and to reinforce the fact that LDP and state plane computations (although very similar) may provide very different results.

Restatement: adoption and use of the GSDM will provide the same benefits as an LDP while avoiding the possible confusion caused by not knowing specifically what coordinate system is being used.

REFERENCES

Burkholder, Earl F. 1980a. A metric map projection for the State of Michigan. Master's thesis, School of Civil Engineering, Purdue University, West Lafayette, IN.

———. 1985. State Plane Coordinates on the NAD 1983. Paper presented at the ASCE Spring Convention, Denver, Colorado, 29 April–3 May. (Copy available from author.)

———. 1991. Computation of horizontal/level distances. *Journal of Surveying Engineering* 117 (3): 104–16.

———. 1993a. Design of a local coordinate system for surveying, engineering, and LIS/GIS. *Surveying & Land Information Systems* 53 (1): 29–40.

———. 1993b. Using GPS results in true 3-D coordinate system. *Journal of Surveying Engineering* 119 (1): 1–21.

———. 2004. Accuracy of elevation factor. *Journal of Surveying Engineering* 130 (3): 134–37.

Claire, Charles N. 1968. State plane coordinates by automatic data processing. Publication 62-4. Rockville, MD: Coast & Geodetic Survey, National Information Branch, National Geodetic Survey, National Oceanic and Atmospheric Administration.

Federal Register notice. 1959. *Federal Register* 24 (128): 5348.

Federal Register notice. 1998. *Federal Register*, May 16.

Mitchell, Hugh C., and Lansing Simmons. [1945] 1977. The state coordinate systems (a manual for surveyors). Special Publication no. 235. Washington, DC: Coast & Geodetic Survey, U.S. Department of Commerce.

Pearson, Fredrick II. 1990. *Map projections: Theory and applications.* Boca Raton, FL: CRC Press.

Richardus, Peter, and Ron K. Adler. 1972. *Map projections for geodesists, cartographers and geographers.* Amsterdam: North-Holland.

Snyder, John P. 1987. *Map projections: A working manual*. U.S. Geological Survey Professional Paper 1395. Washington, DC: Government Printing Office.

Stem, James E. 1989. State plane coordinate system of 1983. NOAA manual NOS NGS 5. Rockville, MD: National Geodetic Survey.

11 Using Spatial Data

INTRODUCTION

This book attempts to:

1. Provide spatial data users a concise source of information with respect to the mathematical and geometrical characteristics of geospatial data.
2. View the transition from the past to the future in terms of the digital revolution.
3. Identify a common spatial data model equally useful to those who build, operate, or use measuring systems and those who use spatial data.
4. Highlight the importance of spatial data quality while providing well-defined procedures for establishing, tracking, storing, and using spatial data accuracy.

Following comments related to the first three objectives, this chapter* focuses on objective number four, spatial data accuracy.

FORCES DRIVING CHANGE

The digital revolution is characterized by using the electronic computer and related devices to collect, store, analyze, and report information about the world and human activities—from the cradle to the grave. Using computers to keep track of spatial data is only a small part of the digital revolution—but it is an important part. Scientists, mathematicians, cartographers, and others have been observing, recording, and describing our world for generations. Although many others also deserve credit, two persons are recognized for specific contributions—Gerardus Mercator is widely acclaimed as the mapmaker who revolutionized cartography, and René Descartes (1596–1650), perhaps better known as a philosopher, was a mathematician who systematized analytical geometry and gave us the Cartesian coordinate system. The profound impact of their combined legacy has permeated mapping and the use of spatial data for the past 400 years. Without making light of their work, the digital revolution justifies a new look at fundamental assumptions, and the GSDM identifies additional innovative tools available for working with 3-D spatial data.

Although others can also be identified, forces driving the digital revolution include:

- The transistor
- Miniaturization of circuits and physical devices
- Development of information technology, science, and management

* This chapter makes extensive use of material from an article "Fundamentals of Spatial Data Accuracy and the Global Spatial Data Model" by the author and filed with the U.S. Copyright Office, Washington, DC, 2004 (Burkholder 2004). Used with permission.

- Electronic signal processing
- Satellite positioning
- Enhanced spatial literacy

Other factors may be identified as consequences of the digital revolution.

- Reduction in privacy
- Better knowledge of where things are
- More efficient movement of people, products, and resources
- Greater access to information
- Nanotechnology

Recognizing the impossibility of identifying mutually exclusive cause-and-effect properties of the various factors, the focus of this book is on spatial data. With that qualifier, the assumption is that electronic measurements (GNSS and others), computer databases, information management, and the exponentially growing demand for reliable spatial data are the forces driving this reevaluation of spatial data models.

TRANSITION

This book may not be an easy read for those with a low tolerance for detail. The attitude of some students is "Just tell me what I need to know and let's move on." An observation is that just-in-time learning is preferred to not learning. We buy a piece of new equipment or software and learn how to use it. With practice, we can become proficient, productive, and profitable. Given the rapid pace of technological developments, we all do it. However, whether learned in a formal classroom setting or in individual study, an understanding of fundamental concepts provides a foundation for such just-in-time learning. Building that foundation is the goal of this book. The first two chapters identify details and concepts of the GSDM. But, starting in chapter 3, fundamental mathematical concepts are summarized, and a careful logical building process is included in subsequent chapters. While not all things to all users, the goal is to focus on developing an understanding of the concepts. With an ever increasing number of persons using GPS and with the GSDM providing a context for better understanding of spatial data, our collective passion for spatial literacy will be enhanced and a multitude of readers will be able to apply innovative 3-D concepts to an exponentially expanding array of spatial data applications.

One could say society has been complacent in applying analytical geometry to mapping and geospatial data. The fundamental theorems of solid geometry and vector algebra have been long proven, and a map is a map is a map. Prior to the digital revolution and burgeoning use of 3-D data, there was little to get excited about. Oh yes, geodesists have developed complex mathematical expressions for describing the size and shape of the world, and cartographers have developed an endless array of map projections. Their work is truly impressive. But, for many, the most useful map projections are those that make it possible to perform "flat-Earth" computations using simple rules of plane 2-D Euclidean geometry. Because few of us really understand what it takes to make a good map, we take such beneficial features for granted, and, in part because we walk erect, we view the world in terms of horizontal and vertical—separately. Again, without complaining, we grow up learning those concepts, and the spatial data user community largely accepts the traditional 1-D/2-D view. It has been only recently, if at all, that high-level policy professionals have given serious consideration to using an integrated 3-D database. With implementation of the GSDM, spatial data disciplines all over the world will enjoy the luxury of working with a better set of rules with regard to the exchange and use of 3-D spatial data.

Of course, the digital revolution has already had an enormous impact on traditional mapping and the way spatial data are used. Existing processes were computerized and automated so that better work could be done more efficiently. Maps are now stored in a digital format in electronic files instead of flat files, and a new map can be plotted from the database at any time. Furthermore, the same database can support a wide range of map products having different themes and/or scales. This added flexibility is very beneficial and has had a significant impact on the use of maps. That is very important, but not the point.

The point is that digital spatial data have replaced the map as the primary storage medium. In August 2007, a web search of "digital earth" returned over 250,000 hits, and it seems that each organization has a preferred answer for the best format for digital spatial data. The process of formulating the GSDM was to back up and consider the underlying characteristics of digital 3-D spatial data. Starting with a single origin for 3-D data, rules of solid geometry and vector algebra were followed carefully in building a simple consistent logical 3-D model. Of course, the historic value of maps as a record of the development of civilization remains enormously significant and should be accommodated (Harvey 2000). But, so far, it appears that the transition to using digital maps is evolving in a fragmented manner and begs careful evaluation. A concise summary of three evolutionary stages is:

- In the past, spatial data were analog and separated into horizontal and vertical components. Maps or photographs were used as the primary storage medium.
- In the interim, spatial data are digital with separate horizontal and vertical components. Digitizing existing maps became an important professional and technical activity. Datums, projections, units, and coordinate systems all affect interoperability.
- In the future, the 3-D characteristics of spatial data measurements are preserved in the computational processes and spatial data are stored in an integrated 3-D database. Yes, interoperability details are still important, but meta data and/or covariances of all spatial data going into the database are the responsibility of the "owner" of the database. Then each user is able to retrieve spatial data from the database and is able to rely upon the proven quality of those data. From there, each user has the freedom of manipulating the spatial data bidirectionally according to specifications of the application at hand. The GSDM provides an efficient link between rigorous scientific uses and local "flat-Earth" applications.

A transition from the past to the future is both a challenge and an opportunity for spatial data users. Thomas Kuhn (1996) describes the processes involved in such a transition in "The Structure of Scientific Revolutions." Several quotes are:

Page 67: "[T]he awareness of anomaly had lasted so long and penetrated so deep that one can appropriately describe the fields affected by it as being in a state of growing crisis."

Page 84: "[A] crisis may end with the emergence of a new candidate for paradigm and with the ensuing battle over its acceptance."

Page 153: "Probably the single most prevalent claim advanced by proponents of a new paradigm is that they can solve problems that have led the old one to a crisis."

Page 158: "Because scientists are reasonable men [*sic*], one or another argument will ultimately persuade many of them. But there is no single argument that can or should persuade them all."

This author is encouraged by the adaptability of the younger generation and their capacity for visualizing spatial relationships. Yes, adapting to policies consistent with the efficient use of 3-D spatial data is more difficult for those having a 1-D/2-D mind-set, but, somehow, the younger generation seems to be saddled with fewer conceptual obstacles. In time, using the GSDM will be as comfortable for the spatial data user as is the automatic transmission for automobile drivers. Yes, there are still those who, for whatever reason, prefer to use the clutch and standard shift. Such 1-D/2-D derivative uses of spatial data remain fully supported by the 3-D GSDM. Understandably, while 1-D and 2-D applications are fully supported by a 3-D database, attempting to build a 3-D database from 1-D/2-D data is not recommended. The recommended procedure is to add competent 3-D observations to an existing 3-D database.

CONSEQUENCES

There will be many consequences of using the GSDM. Some have already occurred and are accepted as routine, some are a matter of recognizing the impact of using existing technologies and policies, and others will involve the realization of benefits due to careful planning and implementation. It is impossible to identify all the consequences, but several of the more obvious ones are as follows:

- All spatial data measurements going into the GSDM will need to be 3-D (or even time-stamped, making them 4-D).
- Spatial data are processed and stored under the assumption that there is a single origin for geospatial data. Elevation becomes a derived quantity. Local observed differential elevation differences remain valid subject to deflection-of-the-vertical considerations.
- A World Vertical Datum (xx) will be adopted in which ellipsoid height is used as the third dimension. Arguments in favor of such adoption are given by Burkholder (2002, 2006), Kumar (2005, 2007), and Soler (2007). The

need for geoid modeling will be enormously reduced, but the importance of geoid modeling remains. Geoid heights will continue to be used by those for whom the difference matters, e.g., those needing to relate current measurements to legacy data.

- Spatial data users all over the world will be able to work with local flat-Earth differences while enjoying specific, reliable connectivity to the world at large via the ECEF coordinates stored by the GSDM.
- Two groups of spatial data professionals are equally well served by information stored in a common 3-D database:
    1. Surveying, engineering, mapping, and navigation: The *relative* difference of one point with respect to another is critical and relied upon. The P.O.B. feature in Box 10 of Figure 1.4 means that flat-Earth plane surveying components can be used to obtain ground-level tangent plane direction and distance between points. Consequently, the grid-ground distance difference issue becomes moot, and the need for low-distortion projection coordinates goes away. Furthermore, the GSDM defines a mathematical process by which the local accuracy and network accuracy of such inversed quantities can be reliably determined.
    2. GIS, planning, inventory, and navigation: The *absolute* unique location of a point is of primary consideration and is preserved via the ECEF geocentric coordinates. Of course, a point defined by ECEF coordinates can be equivalently expressed by 2-D map projections such as state plane, UTM, or other map projection coordinates.

The alert reader will note that "navigation" appears in both categories.

SPATIAL DATA ACCURACY

Introduction

The stochastic model portion of the GSDM is described in chapter 1 and addresses issues of spatial data accuracy. Additional information on stochastic models is given by Mikhail (1976), Ghilani and Wolf (2006), and Burkholder (1999, 2004). "Spatial data accuracy" is an umbrella term that includes concepts such as uncertainty, standard deviation, positional tolerance, confidence intervals, and error ellipses. Of those, standard deviation is used as the underlying concept in the GSDM.

The stochastic model is used to answer the question "Accuracy with respect to what?" Two obvious possibilities are:

- What is the absolute accuracy of a point with respect to the datum?
- What is the relative accuracy of a point with respect to another identified point?

Although both answers are readily available, they are somewhat different. Absolute datum accuracy is given in terms of standard deviations at each point—one in each of three orthogonal directions—either in the ECEF reference frame or (easier for humans to visualize) in the local east/north/up reference frame. The relationship between the ECEF perspective and the local perspective is given by matrix equations 1.32 and 1.33.

Relative accuracy is the standard deviation of the computed distance and/or the direction from the standpoint to the forepoint. Relative accuracy between any two points can be computed using matrix equation 1.36. Depending upon user choices and available covariance information, matrix equation 1.36 can be used to obtain either network accuracy (the two points are statistically independent) or local accuracy (reflecting the correlation given by the covariance information) between the two points.

The stochastic model portion of the GSDM is optional. In some cases, the standard deviation information may be readily available, but is not needed or used. In other cases, no standard deviations are known or given. That means the default value of zero is used for the standard deviation and the associated spatial data element is used as an exact value. In either case, functional model computations can still be performed whether standard deviations are available or not. The GSDM can accommodate whatever stochastic information the user provides, but providing reliable stochastic data is ultimately the responsibility of each user. Like fire, competent use of the GSDM can provide enormous benefits but misuse can also be very harmful. Responsible use is essential and consists of inputting or using a reasonable estimate of the standard deviation of each observation.

Not restricted to any one discipline, the GSDM facilitates collection, storage, manipulation, exchange, and use of spatial data worldwide because the same 3-D model accommodates both those activities that generate spatial data and those activities that use spatial data, whether in high-level scientific research or in local "flat-Earth" applications. And, regardless of application, questions regarding spatial data accuracy can be handled with a common set of stochastic model equations. The spatial data accuracy discriminator is the magnitude of the standard deviation, component by component, and can be judged against FGDC (Federal Geographic Data Committee 1998, Table 2.1) standards for spatial data accuracy.

Definitions

While the intent is to use standard definitions and conventions, the following are used for the purposes of this chapter.

Spatial data uncertainty is given by its standard deviation in each of three dimensions. One standard deviation (1 sigma) provides a 68 percent confidence level. Many spatial data users routinely use a 95 percent (2 sigma) confidence level as the basis for making comparisons and/or inferences.

As stated in chapter 2, **spatial data** are defined as the distance between endpoints of a line in Euclidean space. Even though a line is the path of a moving point, a distance (not a point) is viewed as the spatial data primitive because the location of a point is meaningless unless or until described with coordinates (distances).

Physical geodesists use a definition of a geodetic datum that also includes the gravity field (National Imagery and Mapping Agency 1997). But, for purposes of describing location and spatial data accuracy, a **3-D GSDM datum** is taken to be an ECEF right-handed rectangular *X*/*Y*/*Z* coordinate system whose

1. origin is at the Earth's center of mass.
2. Z-axis coincides with the Earth's mean spin axis. That means *X*/*Y* coordinates are in the plane of the equator.
3. X-axis is coincident with zero degrees longitude (the Greenwich meridian). That means the Y-axis lies at 90° east longitude.
4. distance unit is meters.
5. ellipsoid is defined by two parameters that permit computation of equivalent *latitude*/*longitude*/*height* coordinates from geocentric *X*/*Y*/*Z* coordinates.

As discussed in chapter 7, the GSDM and error propagation concepts described herein work equally well with the NAD83, the WGS84, or the ITRF. It is not appropriate to mix coordinate values from different datums. See datum transformations in chapter 7.

Recognizing that the origin of any system is relative to some "larger" system (e.g., the center of mass of the Earth is relative to the center of mass of our solar system), **absolute quantities** are expressed by a numerical value in a defined system.

A **relative value** is the difference between two absolute quantities expressed in the same system. The value assigned to a well-defined origin may be an absolute quantity.

Comments on absolute and relative relationships:

- ECEF coordinate values are absolute, and the standard deviations of those absolute values are referred to as **datum accuracy**.
- Coordinate differences (in any given system) are relative.
- An angle, being the difference between two directions, is relative.
- It is possible for an absolute quantity to be treated as a relative quantity. This could happen if the origin has units of zero. If zero is subtracted from an absolute quantity, the result can be considered a relative value because it represents the difference of two absolute quantities.
- The accuracy of relative spatial data can be expressed in either of two ways. One expression, **network accuracy**, represents the uncertainty (standard deviation) of the difference between two statistically independent points in the same system. Another expression, **local accuracy**, uses statistical correlation between two points to represent the uncertainty of one point with respect to another.
- Elevations and time are similar in that each may look like an absolute value. But, in reality, both are used as relative values due to the ambiguity of their physical origins.
  1. Traditional vertical datums are referenced to an arbitrary zero elevation surface, which implies datum elevations are all relative.
  2. Time is counted from the "big bang" (Hawking 1988), from the birth of Christ (b.c. and a.d.), from the vernal equinox (the instant of the sun's zero declination), from the daily transit of the sun over a stated meridian (a.m. or p.m.), or from some arbitrary zero computed from the readings

of a group of atomic clocks. Whether in years, months, days, hours, or seconds, time is an interval between two specified events—a relative quantity.

- Mean sea level, the geoid, enjoys a simple physical definition as a "zero" equipotential surface. But, as yet, that origin has not been precisely located worldwide. Therefore, it can be said that precise absolute elevations do not exist.
- Time differences and elevation differences can each be measured quite precisely, and that information can be quite useful. But, accuracy statements regarding time and elevation should be limited to relative accuracy statements. In terms of absolute accuracy, there is nothing to be gained from adding a precise interval to an absolute quantity of dubious value.
- Ellipsoid height is a derived quantity with respect to the ellipsoid (and ultimately with respect to the Earth's center of mass). Because the origin is well defined and measurable, ellipsoid height can be considered an absolute quantity. Ellipsoid height differences are relative quantities.

Spatial Data Components and Their Accuracy

Spatial data components were listed in chapter 2 and have been subsequently used throughout the book. That list is repeated here with reference to the boxes identified in Figure 1.4.

1. **Absolute *X*/*Y*/*Z* geocentric coordinates** (Box 1) are perpendicular distances in meter units from the respective axes of the ECEF coordinate system.
2. **Absolute geodetic coordinates** (Box 2) of latitude/longitude/height are computed from ECEF coordinates with respect to some named datum or ellipsoid.
3. **Relative geocentric coordinate differences** (Box 3) are obtained by differencing compatible geocentric $X/Y/Z$ coordinate values, or they can be obtained by rotating relative local coordinate differences into the $X/Y/Z$ reference frame. Relative geocentric coordinate differences are also obtained directly as the $\Delta X/\Delta Y/\Delta Z$ components of a GPS vector.
4. **Relative geodetic coordinate differences**, $\Delta\varphi/\Delta\lambda/\Delta h$ (not shown in Figure 1.4), are obtained as the difference of compatible (common datum) geodetic coordinates.
5. **Relative local coordinate differences** (Box 9) are the local tangent plane components of conventional total station surveying measurements. If deflection-of-the-vertical is severe and if project requirements warrant same (e.g., establishing traditional geodetic control for a multi-discipline project covering a large site in a mountainous area), the vertical-based measurements of a total station instrument should be converted to normal-based measurements before calling them local geodetic horizon components. Relative local coordinate differences are also components of a geocentric $\Delta X/\Delta Y/\Delta Z$ vector rotated into the local geodetic horizon.

6. **Local coordinates**, *e*/*n*/*u*, are distances from some origin whose local definition may be sufficient in three dimensions, two dimensions, or one dimension. Burkholder (2001) calls these absolute coordinates, but, depending upon how they are viewed, they could also be considered relative values. Elevations are particularly difficult to categorize. The real underlying issue is how the local system is defined. Examples include the following:
   A. **Point-of-beginning (P.O.B.) datum coordinates** (Box 10) are defined by Burkholder (1997) as the local tangent plane components from any point (origin) selected by the user to any other point. These derived coordinates enjoy full mathematical definition in three dimensions and suffer no loss of geometrical integrity in the GSDM.
   B. **Map projection** (or state plane) **coordinates** (Box 11) are well defined in two dimensions with respect to some named origin and geodetic datum.
   C. **Ellipsoid heights** (Box 2) and **orthometric heights** (Boxes 8 and 11) are one-dimensional distances above or below some named surface. Ellipsoid heights can be considered absolute, but other elevations are considered relative.
7. **Arbitrary local coordinates** (not shown in Figure 1.4) may be 1-D, 2-D, or 3-D based upon some assumed origin. Although useful in some applications, arbitrary local coordinates are generally not compatible with other local coordinate systems and have limited value in the broader context of georeferencing. Many computer graphics and data visualization programs use arbitrary local coordinates.

With regard to all spatial data components, both absolute and relative, each one can have a standard deviation associated with it. If the standard deviation of any component is zero, either the quantity is known very precisely and/or the value (e.g., a control point) is being used as a "fixed" quantity. Standard deviations of subsequently computed spatial data components are based upon propagation of the measurement error, and standard deviations of the computed points are determined through the network adjustment process. Given a successful network adjustment and computation of coordinates, the implied accuracy statement is "with respect to the points held fixed by the user." Maybe the beginning point was a hub pounded in the ground. Maybe it was a section corner of the U.S. Public Land Survey System. Maybe it was a HARN point or a CORS point published by the NGS. Or maybe it was the orbit parameters of the GPS satellites. Understandably, the value of a completed project is greatly enhanced if explicit accuracy statements are made. But, making or not making an explicit statement is not the real issue.

The real issue is being able to make one of the following statements related to the FGDC standards (1998, Table 2.1) and supported by appropriate statistics.

1. "The absolute datum [choose one—NAD83, WGS84, or ITRF] accuracy of point *X* in three dimensions is σ_X = ________, σ_Y = ________, and σ_Z = ________." An equivalent statement, derived from the first, gives the standard deviations in the local reference frame as σ_e = ________, σ_n = ________, and σ_u = ________. The absolute accuracy statement involves only one point and is with respect to the datum selected or named by the user. If the project were a 2-D survey (i.e., state plane coordinates), only two components would be named.
2. "The relative network accuracy of the direction and distance from point 1 to point 2 is σ_{AZ} = ________ and σ_{DIST} = ________." Relative accuracy applies to the difference between two independent points having absolute accuracy values in the same datum.
3. From point 1 to point 2, "the relative network accuracy of the height difference (Δh) or perpendicular distance from the local tangent plane (Δu) is $\sigma_{\Delta h}$ = ________ or $\sigma_{\Delta u}$ = ________."
4. "The relative local accuracy of point 2 with respect to point 1 is σ_{AZ} = ________, σ_{DIST} = ________, $\sigma_{\Delta h}$ = ________, or $\sigma_{\Delta u}$ = ________." Relative local accuracy exploits and is largely governed by the statistical correlation that exists between two directly connected points in the same datum. The procedure for computing each of the listed accuracies is given in equation 1.36.

But Everything Moves

Most spatial data activities involve using a database such as a GIS. The importance of the basic geodetic control in a GIS is well documented by the National Research Council (NRC; 1983) and others. Ideally, the geodetic control information upon which the database is built should be of such quality that it could be held "fixed," that is, having a zero standard deviation. Here again, the question "With respect to what?" becomes relevant. A monumented point that is stable in one system (e.g., NAD83) may, in fact, be moving in another (WGS84 or ITRF). With the advent of GPS positioning, it is now possible to determine the location of control points much more accurately than before, and the scientific community now has conclusive evidence that points once thought to be permanent are, in fact, moving. With respect to what? An oversimplified answer is that "everything moves."

A better answer is required. More specifically, the administrators and users of a database (whether local, regional, national, or global) deserve explicit information as to the stability and accuracy for the various categories of points in the database. And, if they are moving, what is the velocity vector of the point? This chapter is primarily about 3-D uncertainties, but, given that points move, time must be added as the fourth dimension and the epoch must enjoy equal standing with the coordinates. Software for converting *X*/*Y*/*Z* coordinates from one epoch to another is called HTDP (horizontal time time-dependent positioning) and is available gratis from the U.S. National Geodetic Survey (NGS) at http://www.ngs.noaa.gov. HTDP can also be used to convert *X*/*Y*/*Z* coordinates from one 3-D datum to another (Snay 1999).

With respect to movement, a simple question must be asked: "Are we standing on the train watching the station go by, or are we standing at the station watching the train go by?" The Earth's center of mass is the location reference for the entire globe. Points on the Earth's surface or anywhere within the Earth may move with respect to the Earth's center of mass, but the reference is fixed by definition—it does not move. Admittedly, with respect to where we stand or with respect to monumented points, statements are made that the center of mass of the Earth moves. The implied perspective is considered subordinate to the explicit statement, "The Earth's center of mass does not move."

The ITRF is defined such that the net tectonic movement of all the Earth's continental plates is zero (Snay and Soler 1999). But, points on the Earth's surface still move with respect to the Earth's center of mass and with respect to each other. Therefore, the locations of the ITRF monuments are defined with both coordinates and velocities. Spatial data users in North America will be reassured to know that the NAD83 datum is the one to use because, except for areas of tectonic activity, points on the NAD83 remain "fixed" to the North American plate and move together. Such oversimplification is dangerous. The NAD83 monumented control points on the ground may be stable, but the GPS satellite orbits are expressed in (and the NGS CORS coordinates are published in) the ITRF reference frame. (NGS also publishes NAD83 coordinates for the CORS stations.) The issue of which to be aware is that the absolute coordinates (for points on the ground) may be in one reference frame and the relative coordinate differences (obtained from GPS) may be in a different reference frame. Since the NAD83 and ITRF relative coordinate differences are nearly identical, it is generally permissible to attach ITRF relative coordinate differences to absolute NAD83 datum coordinates, but mixing absolute datum coordinates in the same solution should be avoided.

The point here is that most spatial data users should be aware of three competing 3-D geodetic datums—NAD83, WGS84, and ITRF. Each has a reason for existing, and each has a role to fill. At a gross level of accuracy, it does not matter which of the three datums is used. But, as the tolerance for uncertainty gets smaller and smaller, it does matter which datum is used. The GSDM can be used with each datum individually and provides a systematic method for identifying and tracking the uncertainties in a given datum—whatever they are. Comparing uncertainties (standard deviations) between datums is beyond the scope of this book. Those larger issues are being addressed by others, such as Han, J. Y., et al (2008).

Observations, Measurements, and Error Propagation

In many ways, observations and measurements are very similar, and the terms are used interchangeably. But, a mathematical distinction is that observations are always independent quantities and measurements may be either independent or correlated. Stated differently, any observation may be called a measurement, but a measurement can be called an observation only if it is an independent quantity. As listed in chapter 2, there are only a limited number of quantities that can be directly measured. But, whether the measurement is length, time, voltage, temperature, or the like, spatial data components are determined indirectly from those measurements using appropriate models and computations. The standard deviation of each component is determined by propagating the measurement uncertainty through the variance/covariance equation given by the following matrix formulation:

$$\Sigma_{YY} = \boldsymbol{J}_{YX} \Sigma_{XX} J^{t}_{XY} \tag{11.1}$$

where

Σ_{YY} = covariance matrix of computed result,
$\boldsymbol{J}_{YX}$ = Jacobean matrix of partial derivatives of the result with respect to the variables (measurements), and
Σ_{XX} = covariance matrix of the variables (measurements) used in the computations.

To reiterate, the variables in the measurement covariance matrix are independent and considered to be observations if and only if there is no correlation in the measurement covariance matrix.

Finding the Uncertainty of Spatial Data Elements

In the process of establishing the spatial data uncertainty of each point, the user must first decide which datum will be used. Mixing datum values is permissible only if the datum differences are smaller than the resolution of data added to the database. For example, if 10 m data are being used and if the datum differences are at the 1 m level, it makes no difference which datum is used. On the other hand, if 10 mm data are being used and datum differences are at the 1 m level, the choice of datum does matter.

Second, each project should be based upon reliable control points having *X*/*Y*/*Z* geocentric coordinates in the appropriate datum. One control point may be sufficient to put a new project on the chosen datum, but making a connection to two or more points is standard practice. If the basic control points are assigned a zero standard deviation, then that means subsequent accuracy statements should be made "with respect to the control points selected and held fixed by the user." Better statements regarding datum accuracy can be made if realistic standard deviations are assigned to the points used to control the project. The covariance matrix for each new point and the correlation between points in the network are a standard by-product of a least squares adjustment. When the network adjustment is done in terms of geocentric coordinates and coordinate differences, the resulting covariance matrix is in terms of the geocentric reference frame. The geocentric environment is more efficient for storage and computer operations, but, because of the human perspective, the local covariance matrix is preferred as being more intuitive—giving sigma east, sigma north, and sigma up as the square root of the diagonal elements.

The GSDM includes both the geocentric and local covariance matrices for each point, but, since one can be derived from the other, a BURKORD™ database stores only the geocentric covariance matrix. The local covariance matrix is computed upon demand. Both covariance matrices contain the same datum accuracy of each point component by component, but, because of perspective, the numbers are different. Each of the two covariance matrices is a 3 × 3 symmetrical matrix containing the following elements:

Geocentric Covariance Matrix

$$\Sigma_{XYZ} = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & \sigma_Z^2 \end{bmatrix} \text{ and } \Sigma_{enu} = \begin{bmatrix} \sigma_e^2 & \sigma_{en} & \sigma_{eu} \\ \sigma_{en} & \sigma_n^2 & \sigma_{nu} \\ \sigma_{eu} & \sigma_{nu} & \sigma_u^2 \end{bmatrix}$$

Local Covariance Matrix (11.2 and 11.3)

where

σ_X^2, σ_Y^2, σ_Z^2 = variances for geocentric coordinates for the point,
σ_{XY}, σ_{XZ}, σ_{YZ} = covariance elements for geocentric coordinates,
σ_e^2, σ_n^2, σ_u^2 = local perspective variances for the point, and
σ_{en}, σ_{eu}, σ_{nu} = local perspective covariance elements for the point.

The two covariance matrices are related by the following rotation matrix evaluated at the latitude/longitude of the standpoint (local origin).

$$\boldsymbol{R} = \begin{bmatrix} -\sin\lambda & \cos\lambda & 0 \\ -\sin\phi \cos\lambda & -\sin\phi \sin\lambda & \cos\phi \\ \cos\phi \cos\lambda & \cos\phi \sin\lambda & \sin\phi \end{bmatrix} \quad (11.4)$$

The matrix expression for the relationship between the two covariance matrices is

$$\Sigma_{enu} = \boldsymbol{R} \Sigma_{XYZ} \boldsymbol{R}^t \quad (11.5)$$

$$\Sigma_{XYZ} = \boldsymbol{R}^t \Sigma_{enu} \boldsymbol{R} \quad (11.6)$$

Points that are part of a network adjustment enjoy an interrelationship described by correlation. The correlation is especially significant for adjacent points that have been connected by a direct measurement. Correlation exists between points not directly connected, but the influence drops rapidly as the number of courses between points increases (correlation is the reason why cross-ties serve to strengthen a network). If the significant correlations between points are stored along with the covariance matrix for each point, the local accuracy of one point with respect to the other is readily computed along with the inverse direction and distance. If correlations are not stored (or if they are assumed to be zero), an inverse computation will readily provide the direction and distance between points, and the two endpoint covariance matrices will provide the basis of the network accuracy associated with the relative differences.

Using Points Stored in the *X/Y/Z* Database

Each stored *X*/*Y*/*Z* location is unique within the birdcage of orbiting GPS satellites. Three application modes for using the stored *X*/*Y*/*Z* locations are as follows:

- Single point (e.g., a unique location for an inventory tag).
- Point-pair (used to create lines, surfaces, and objects).
- "Cloud" (mapping). Even though stored as *X*/*Y*/*Z*, the location of any point can also be readily expressed in 2-D latitude/longitude, UTM, or state plane coordinates along with 1-D ellipsoid height.

The uncertainty of a single point is given by the datum accuracy as computed from the geocentric covariance matrix. These uncertainties (standard deviations, variances, and other covariance elements) can be viewed in either the geocentric reference frame or the local reference frame. The geocentric reference frame is more efficient for data storage and computerized manipulation, but the local reference frame is more convenient for viewing because horizontal and vertical comprise the human perspective. A benchmark will have a small standard deviation on the vertical component. By contrast, a horizontal control point will have small standard deviations on the east and/or north components. A 3-D control point will have small standard deviations on all three components.

The point-pair application provides the relative location of one point with respect to another. A map is generated by extensive successive use of the point-pair mode, and an accuracy statement as applied to such a "cloud" of points is not addressed here.

Specifically, in the point-pair mode, point 1 is defined by $X_1/Y_1/Z_1$ and point 2 is defined by $X_2/Y_2/Z_2$. The matrix formulation of the 3-D geocentric inverse from point 1 to point 2 is

$$\begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \\ X_2 \\ Y_2 \\ Z_2 \end{bmatrix} \tag{11.7}$$

The matrix of coefficients to the variables is called the Jacobian matrix, and the general error propagation formulation in the form of equation 11.1 is

$$\Sigma_{\Delta} = \boldsymbol{J} \, \Sigma_{1\to 2} \, \boldsymbol{J} \tag{11.8}$$

Using the Jacobian matrix of 1's and 0's from equation 11.7, having the geocentric covariance matrix of both point 1 and point 2 available, and using the correlation between point 1 and point 2, the covariance matrix of the inverse is computed using equation 11.8 as

$$\Sigma_\Delta = \begin{bmatrix} -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1Y_1} & \sigma_{X_1Z_1} \\ \sigma_{X_1Y_1} & \sigma^2_{Y_1} & \sigma_{Y_1Z_1} \\ \sigma_{X_1Z_1} & \sigma_{Y_1Z_1} & \sigma^2_{Z_1} \end{bmatrix} & \begin{bmatrix} \sigma_{X_1X_2} & \sigma_{X_1Y_2} & \sigma_{X_1Z_2} \\ \sigma_{Y_1X_2} & \sigma_{Y_1Y_2} & \sigma_{Y_1Z_2} \\ \sigma_{Z_1X_2} & \sigma_{Z_1Y_2} & \sigma_{Z_1Z_2} \end{bmatrix} \\ \begin{bmatrix} \sigma_{X_1X_2} & \sigma_{Y_1X_2} & \sigma_{Z_1X_2} \\ \sigma_{X_1Y_2} & \sigma_{Y_1Y_2} & \sigma_{Z_1Y_2} \\ \sigma_{X_1Z_2} & \sigma_{Y_1Z_2} & \sigma_{Z_1Z_2} \end{bmatrix} & \begin{bmatrix} \sigma^2_{X_2} & \sigma_{X_2Y_2} & \sigma_{X_2Z_2} \\ \sigma_{Y_2X_2} & \sigma^2_{Y_2} & \sigma_{Y_2Z_2} \\ \sigma_{Z_2X_2} & \sigma_{Z_2Y_2} & \sigma^2_{Z_2} \end{bmatrix} \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{11.9}$$

The off-diagonal submatrices reflect the correlation between point 1 and point 2. The **datum accuracy** of point 1 and point 2 is included in equation 11.9 as their respective covariance submatrices.

The following concise mathematical statements comprise the basis for the definitions of local accuracy and network accuracy given earlier.

Local accuracy of the inverse between point 1 and point 2 is obtained by using the full covariance matrix in equation 11.9. Correlation between point 1 and point 2 is included.

Network accuracy of the inverse between point 1 and point 2 is obtained if the correlation between point 1 and point 2 is either nonexistent or taken to be zero.

Example

The following example is a summary of a fully documented least squares network solution posted at http://www.globalcogo.com/nmsunet1.pdf. The GPS network includes seven GPS vectors and is based upon two A-order HARN points: station "Reilly," located in the central horseshoe of the NMSU campus; and station "Crucesair," located at the Las Cruces, New Mexico, airport some 16 kilometers west of the campus. The network consists of seven independent baselines connecting four additional points to the existing HARN stations, as shown in Figure 11.1.

The GPS baselines shown and used were collected on four different dates over a period of five years. These are not the only baselines on campus, nor are they the only observations between the points in question. These baselines were selected because they show excellent consistency, are independent, and include often used points. The network is included here to illustrate use of the GSDM and computation of both network accuracy and local accuracy.

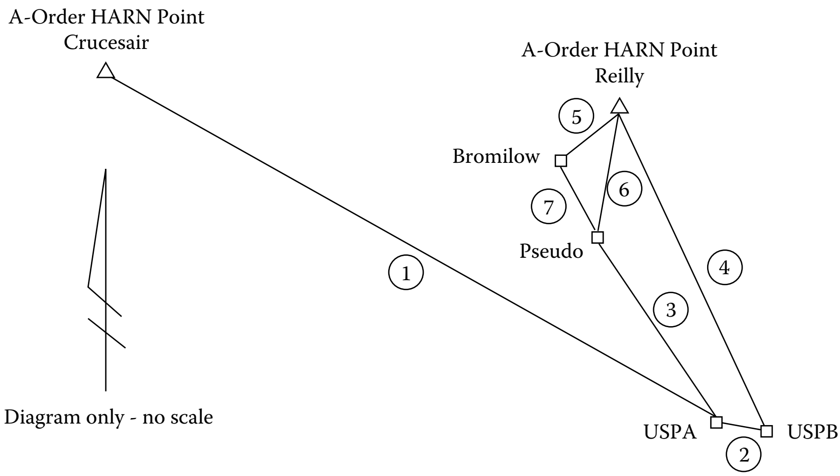


FIGURE 11.1 GPS Survey Network on NMSU Campus

Control Values and Observed Vectors

The NAD83 geocentric *X*/*Y*/*Z* coordinates for A-order HARN stations "Reilly" and "Crucesair" are as published by the National Geodetic Survey (NGS) and were held fixed in this exercise. They are as follows:

Station Reilly	Station Crucesair
X = –1,556,177.615 m	X = –1,571,430.672 m
Y = –5,169,235.319 m	Y = –5,164,782.312 m
Z = 3,387,551.709 m	Z = 3,387,603.188 m

Single-frequency Trimble GPS receivers were used to collect static data, 57 minutes being the shortest common observation time for any of the seven baselines. The baseline components and the covariance matrix for each observed baseline as determined by Trimble software using default processing parameters are as follows:

Baseline 1, Crucesair to USPA, Observed 28 March 2002 (use subscript $_{CA}$)

		Sxx	Syy	Szz
ΔX_{CA} = 15,752.080 m	Sxx	6.321492E-06		
ΔY_{CA} = –5,179.102 m	Syy	1.545948E-05	4.739877E-05	
ΔZ_{CA} = –903.089 m	Szz	–1.061303E-05	–3.184780E-05	2.388036E-05

Baseline 2, USPA to USPB, Observed 12 November 2003 (use subscript $_{AB}$)

		Sxx	Syy	Szz
ΔX_{AB} = 14.964 m	Sxx	1.412453E-06		
ΔY_{AB} = –15.365 m	Syy	1.285418E-06	4.653209E-06	
ΔZ_{AB} = –16.664 m	Szz	–5.669127E-07	–1.658118E-06	1.872469E-06

Baseline 3, USPA to Pseudo, Observed 28 March 2002 (use subscript $_{AP}$)

		Sxx	Syy	Szz
ΔX_{AP} = –528.036 m	Sxx	9.505016E-08		
ΔY_{AP} = 560.657 m	Syy	8.957064E-08	3.729339E-07	
ΔZ_{AP} = 585.897 m	Szz	–5.022282E-08	–2.221975E-07	3.363763E-07

Baseline 4, USPB to Reilly, Observed 28 March 2002 (use subscript $_{BR}$)

		Sxx	Syy	Szz
ΔX_{BR} = –514.003 m	Sxx	3.650165E-07		
ΔY_{BR} = 741.438 m	Syy	9.024127E-07	2.796189E-06	
ΔZ_{BR} = 868.293 m	Szz	–6.189027E-07	–1.881145E-06	1.410196E-06

Baseline 5, Bromilow to Reilly, Observed 10 December 1998 (use subscript $_{MR}$)

		Sxx	Syy	Szz
ΔX_{MR} = 32.134 m	Sxx	2.762550E-07		
ΔY_{MR} = 51.175 m	Syy	3.200312E-07	6.870545E-07	
ΔZ_{MR} = 94.198 m	Szz	–2.008940E-07	–4.006259E-07	4.661596E-07

Baseline 6, Pseudo to Reilly, Observed 23 January 2002 (use subscript $_{PR}$)

		Sxx	Syy	Szz
ΔX_{PR} = 29.000 m	Sxx	1.325760E-07		
ΔY_{PR} = 165.422 m	Syy	1.317165E-07	5.265054E-07	
ΔZ_{PR} = 265.719 m	Szz	–7.253348E-08	–3.020965E-07	5.006575E-07

Baseline 7, Bromilow to Pseudo, Observed 23 January 2002 (use subscript $_{MP}$)

		Sxx	Syy	Szz
ΔX_{MP} = 3.136 m	Sxx	3.367818E-07		
ΔY_{MP} = –114.242 m	Syy	3.937476E-07	8.766570E-07	
ΔZ_{MP} = –171.527 m	Szz	–5.186521E-07	–8.977932E-07	1.446501E-06

Blunder Checks

In order to verify the absence of blunders in the baselines, misclosures were computed for each component (*X*/*Y*/*Z*) as follows.

Traverse including baselines 1, 2, and 4 (from "Crucesair" to "Reilly"):

	X	*Y*	*Z*
Station Crucesair	–1,571,430.672 m	–5,164,782.312 m	3,387,603.188 m
Baseline 1	15,752.080 m	–5,179.102 m	–903.089 m
Baseline 2	14.964 m	–15.365 m	–16.664 m
Baseline 4	–514.003 m	741.438 m	868.293 m
Computed value	–1,556,177.631 m	–5,169,235.341 m	3,387,551.728 m
Station Reilly	–1,556,177.615 m	–5,169,235.319 m	3,387,551.709 m
Misclosures	–0.016 m	–0.022 m	0.019 m

The loop including baselines 2-3-7-5-4 (being careful to preserve sign convention):

Baseline 2	–14.964 m	15.365 m	16.664 m
Baseline 3	–528.036 m	560.657 m	585.897 m
Baseline 7	–3.136 m	114.242 m	171.527 m
Baseline 5	32.134 m	51.175 m	94.198 m
Baseline 4	514.003 m	–741.438 m	–868.293 m
Misclosures	0.001 m	0.001 m	–0.007 m

The loop including baselines 5-6-7 (being careful to preserve sign convention):

Baseline 5	32.134 m	51.175 m	94.198 m
Baseline 6	–29.000 m	–165.422 m	–265.719 m
Baseline 7	–3.136 m	114.242 m	171.527 m
Misclosures	–0.002 m	–0.005 m	0.006 m

All baselines have been included in the checks, and all misclosures are acceptable. Therefore, it is legitimate to perform a least squares adjustment of the seven baselines to determine the "best" adjusted position for points USPA, USPB, Pseudo, and Bromilow. Any adjustment should also provide information on the quality of the answers (i.e., "What is the standard deviation of the computed position?") in both the geocentric (*X*/*Y*/*Z*) reference frame and the local (*east*/*north*/*up*) reference frame. The full posted paper includes three different weighting schemes and shows a comparison of the various answers. The example here only shows the least squares results obtained using the full covariance matrix of each observed baseline in the "indirect observations" least squares model using one equation for each observation (3 observations per baseline × 7 baselines = 21 observations). The weight matrix was computed as the inverse of the covariance matrix of the observations with an a priori reference variance of 1.0.

$$\mathbf{v} + \mathbf{B}\boldsymbol{\Delta} = \mathbf{f} \tag{11.10}$$

$$\mathbf{W} = (1.0)\,\Sigma^{-1} \tag{11.11}$$

The posted paper shows a formulation of the matrices used in the solution for all three weighting possibilities. Those details are not included here, but the solution shown below was formulated as a linear problem and the matrix solution was obtained as

$$\boldsymbol{\Delta} = (\mathbf{B}^t\,\mathbf{W}\,\mathbf{B})^{-1}\,\mathbf{B}^t\,\mathbf{W}\,\mathbf{f}\text{, or, stated differently, } \boldsymbol{\Delta} = \mathbf{N}^{-1}\,\mathbf{B}^t\,\mathbf{W}\,\mathbf{f} \tag{11.12}$$

where

Δ = vector of parameters (answers),
$\boldsymbol{N}$ = matrix of normal equations ($\boldsymbol{B^tWB}$) ($\boldsymbol{N^{-1}}$ contains statistics for the answers),
$\boldsymbol{W}$ = weight matrix obtained from baseline covariance matrices,
$\boldsymbol{B}$ = matrix of coefficients for the unknown parameters,
$\boldsymbol{f}$ = vector of constants computed from known values and observations, and
$\boldsymbol{v}$ = vector of residuals.

$\boldsymbol{N^{-1}}$ is a normal part of a least squares adjustment and is shown in Figure 11.2.

The estimated (a posteriori) reference variance was computed as

$$\sigma_0^2 = \frac{\boldsymbol{v^t W v}}{r} = 12.8006 \text{ m}^2 \tag{11.13}$$

where

$\boldsymbol{v}$ = vector of residuals,
$\boldsymbol{W}$ = weight matrix,
$\boldsymbol{r}$ = the redundancy,

and the covariance matrix of the computed parameters is the product of the reference variance and the N^{-1} matrix, as shown in Figure 11.3.

Results

The geocentric *X*/*Y*/*Z* coordinates of the four unknown points as shown in Figure 11.4 were computed directly in the least squares adjustment, while the geodetic latitude, geodetic longitude, and ellipsoid height were computed from the *X*/*Y*/*Z* values using the BK2 transformation. The covariance matrix of each new point is the 3 × 3 submatrix shown in Figure 11.3, and the standard deviations of the geocentric *X*/*Y*/*Z* coordinates were computed as the square root of the variances as found in Figure 11.3. The standard deviations at each point in the local reference frame (*e*/*n*/*u*) were computed using equation 11.5 and the computed latitude/longitude at each station.

Network Accuracy and Local Accuracy

Datum accuracy, network accuracy, and local accuracy are defined mathematically in equation 11.9. Datum accuracy is a statement of how well the position of a single point is known with respect to the published datum. Network accuracy can be intuitively understood to be a statement of accuracy between points based upon how well the positions are known with respect to the control held by the user. It is presumed the points are independent—that is, there is no correlation of one with respect to the other as might be determined by a direct tie between them. Alternatively, local

	USPA			USPB			PSEUDO			BROMILOW		
	1.688E-7	1.834E-7	-1.169E-7	3.401E-8	8.217E-8	-5.782E-8	8.910E-8	1.004E-7	-6.974E-8	4.530E-8	5.233E-8	-2.315E-8
USPA	1.834E-7	6.620E-7	-3.920E-7	8.806E-8	2.747E-7	-1.918E-7	1.010E-7	3.312E-7	-2.012E-7	5.318E-8	1.559E-7	-7.339E-8
	-1.169E-7	-3.920E-7	5.321E-7	-8.636E-8	-2.657E-7	2.100E-7	-6.898E-8	-2.006E-7	2.815E-7	-1.119E-8	-5.992E-8	7.011E-8
	3.401E-8	8.806E-8	-8.636E-8	2.433E-7	5.443E-7	-3.649E-7	1.897E-8	4.519E-8	-4.598E-8	7.051E-9	1.778E-8	-1.296E-8
USPB	8.217E-8	2.747E-7	-2.657E-7	5.443E-7	1.705E-6	-1.109E-6	4.674E-8	1.392E-7	-1.395E-7	1.595E-8	5.384E-8	-3.967E-8
	-5.782E-8	-1.918E-7	2.100E-7	-3.649E-7	-1.109E-6	8.549E-7	-3.326E-8	-9.757E-8	1.107E-7	-9.419E-9	-3.494E-8	2.987E-8
	8.910E-8	1.010E-7	-6.898E-8	1.897E-8	4.674E-8	-3.326E-8	9.706E-8	1.049E-7	-7.197E-8	4.955E-8	5.538E-8	-2.402E-8
PSEUDO	1.004E-7	3.312E-7	-2.006E-7	4.519E-8	1.392E-7	-9.757E-8	1.049E-7	3.521E-7	-2.168E-7	5.477E-8	1.647E-7	-7.849E-8
	-6.974E-8	-2.012E-7	2.815E-7	-4.598E-8	-1.395E-7	1.107E-7	-7.197E-8	-2.168E-7	3.267E-7	-7.207E-9	-5.946E-8	7.943E-8
	4.530E-8	5.318E-8	-1.119E-8	7.051E-9	1.595E-8	-9.419E-9	4.955E-8	5.477E-8	-7.207E-9	1.626E-7	1.849E-7	-1.381E-7
BROMILOW	5.233E-8	1.559E-7	-5.992E-8	1.778E-8	5.384E-8	-3.494E-8	5.538E-8	1.647E-7	-5.946E-8	1.849E-7	4.339E-7	-2.820E-7
	-2.315E-8	-7.339E-8	7.011E-8	-1.296E-8	-3.967E-8	2.987E-8	-2.402E-8	-7.849E-8	7.943E-8	-1.381E-7	-2.820E-7	3.647E-7

FIGURE 11.2 **N Inverse** Matrix from Least Squares Adjustment

	USPA			USPB			PSEUDO			BROMILOW		
USPA	2.161E-6	2.347E-6	-1.496E-6	4.354E-7	1.052E-6	-7.402E-7	1.141E-6	1.285E-6	-8.927E-7	5.799E-7	6.698E-7	-2.964E-7
	2.347E-6	8.474E-6	-5.017E-6	1.127E-6	3.517E-6	-2.456E-6	1.293E-6	4.240E-6	-2.576E-6	6.807E-7	1.995E-6	-9.394E-7
	-1.496E-6	-5.017E-6	6.812E-6	-1.105E-6	-3.401E-6	2.689E-6	-8.829E-7	-2.568E-6	3.603E-6	-1.432E-7	-7.670E-7	8.975E-7
USPB	4.354E-7	1.127E-6	-1.105E-6	3.114E-6	6.968E-6	-4.671E-6	2.428E-7	5.785E-7	-5.886E-7	9.026E-8	2.276E-7	-1.659E-7
	1.052E-6	3.517E-6	-3.401E-6	6.968E-6	2.182E-5	-1.420E-5	5.983E-7	1.782E-6	-1.786E-6	2.042E-7	6.892E-7	-5.078E-7
	-7.402E-7	-2.456E-6	2.689E-6	-4.671E-6	-1.420E-5	1.094E-5	-4.258E-7	-1.249E-6	1.416E-6	-1.206E-7	-4.472E-7	3.824E-7
PSEUDO	1.141E-6	1.293E-6	-8.829E-7	2.428E-7	5.983E-7	-4.258E-7	1.242E-6	1.343E-6	-9.212E-7	6.343E-7	7.089E-7	-3.075E-7
	1.285E-6	4.240E-6	-2.568E-6	5.785E-7	1.782E-6	-1.249E-6	1.343E-6	4.506E-6	-2.775E-6	7.011E-7	2.109E-6	-1.005E-6
	-8.927E-7	-2.576E-6	3.603E-6	-5.886E-7	-1.786E-6	1.416E-6	-9.212E-7	-2.775E-6	4.182E-6	-9.225E-8	-7.611E-7	1.017E-6
BROMILOW	5.799E-7	6.807E-7	-1.432E-7	9.026E-8	2.042E-7	-1.206E-7	6.343E-7	7.011E-7	-9.225E-8	2.081E-6	2.367E-6	-1.768E-6
	6.698E-7	1.995E-6	-7.670E-7	2.276E-7	6.892E-7	-4.472E-7	7.089E-7	2.109E-6	-7.611E-7	2.367E-6	5.554E-6	-3.609E-6
	-2.964E-7	-9.394E-7	8.975E-7	-1.659E-7	-5.078E-7	3.824E-7	-3.075E-7	-1.005E-6	1.017E-6	-1.768E-6	-3.609E-6	4.668E-6

FIGURE 11.3 Covariance Matrix of Computed Points in Network

	Geocentric & ECEF sigma	Geodetic & local sigma
Station USPA:		
	X = -1,555,678.579 m +/- 0.0015 m	φ = 32° 16' 23."00019 N +/- 0.0027 m (N)
	Y = -5,169,961.396 m +/- 0.0029 m	λ = 106° 44' 48."90817 W +/- 0.0017 m (E)
	Z = 3,386,700.089 m +/- 0.0026 m	h = 1,178.015 m +/- 0.0028 m (U)
Station USPB:		
	X = -1,555,663.613 m +/- 0.0018 m	φ = 32° 16' 22."36244 N +/- 0.0037 m (N)
	Y = -5,169,976.761 m +/- 0.0047 m	λ = 106° 44' 48."19151 W +/- 0.0022 m (E)
	Z = 3,386,683.419 m +/- 0.0033 m	h = 1,177.908 m +/- 0.0042 m (U)
Station Pseudo:		
	X = -1,556,206.615 m +/- 0.0011 m	φ = 32° 16' 45."74650 N +/- 0.0020 m (N)
	Y = -5,169,400.740 m +/- 0.0021 m	λ = 106° 45' 14."39975 W +/- 0.0012 m (E)
	Z = 3,387,285.987 m +/- 0.0020 m	h = 1,165.641 m +/- 0.0020 m (U)
Station Bromilow:		
	X = -1,556,209.750 m +/- 0.0014 m	φ = 32° 16' 52."33407 N +/- 0.0022 m (N)
	Y = -5,169,286.496 m +/- 0.0024 m	λ = 106° 45' 15."77273 W +/- 0.0015 m (E)
	Z = 3,387,457.512 m +/- 0.0022 m	h = 1,165.523 m +/- 0.0023 m (U)

FIGURE 11.4 Geocentric and Local Reference Frames Positions and Standard Deviations

accuracy can be understood to be a statement of accuracy between points based upon a direct measurement between the points. The following paragraphs describe the results of computing both network accuracy and local accuracy from point USPA to point Pseudo. The Excel spreadsheet shown in appendix C (the file is called "3-D Inverse with statistics.xls") was used to generate the values in Table 11.1 and can be obtained gratis from the author at http://www.globalcogo.com.

When using the Excel spreadsheet, the user keys information into the spreadsheet and answers appear instantaneously. Input includes the names of the two stations, the geocentric *X*/*Y*/*Z* coordinates of the two points, and the covariance information. When computing the inverse, the direction and distance will remain the same but the standard deviations will be different depending upon the covariance information input by the user. Four choices (cases) for entering covariance information are:

1. All standard deviations are entered as zeros. That means there is no standard deviation available and the *X*/*Y*/*Z* coordinate data are used as being "fixed." The spreadsheet will still compute the local tangent plane direction and distance between points, but there will be no standard deviations associated with the inverse direction and distance.
2. The user can enter the standard deviations of the geocentric *X*/*Y*/*Z* coordinates as variances (standard deviations squared). These covariance data are entered on the diagonal of the geocentric covariance matrix for each point. The spreadsheet computes the local reference frame covariance matrix

TABLE 11.1
Comparison of Network and Local Accuracies

USPA (standpoint)	**Pseudo (forepoint)**
X = –1,555,678.579 m	X = –1,556,206.615 m
Y = –5,169,961.396 m	Y = –5,169,400.740 m
Z = 3,386,700.089 m	Z = 3,387,285.987 m

Coordinate differences from standpoint to forepoint are:

ΔX = –528.036 m	Δe = –667.190 m
ΔY = 560.656 m	Δn = 700.811 m
$Z\Delta$ = 585.898 m	Δu = –12.448 m

Note: $Dist = \sqrt{\Delta e^2 + \Delta n^2}$ and $\tan\alpha = \frac{\Delta e}{\Delta n}$ (same for each case following).

		Network Accuracy	Local Accuracy
1. No standard deviations	Distance = 967.615 m	+/– 0.0000 m	0.0000 m
	Direction = 316º 24’ 28.”2	+/– 0.00 sec.	0.00 sec.
2. Standard deviations of X/Y/Z values only	Distance = 967.615 m	+/– 0.0031 m	0.0031 m
	Direction = 316º 24’ 28.”2	+/– 0.53 sec	0.53 sec.
3. Full covariance matrix of each X/Y/Z point	Distance = 967.615 m	+/– 0.0018 m	0.0018 m
	Direction = 316º 24’ 28.”2	+/– 0.40 sec.	0.40 sec.
4. Full covariance matrix and correlation submatrix	Distance = 967.615 m	+/– 0.0018 m	0.0011 m
	Direction = 316º 24’ 28.”2	+/– 0.40 sec.	0.24 sec.

(showing the local component *e*/*n*/*u* standard deviations of each point), the inverse direction and distance standpoint to forepoint, and the standard deviation of the direction and the distance. Local and network accuracy will be identical because no correlation data were entered.

3. The user can enter the full covariance matrix for each point. This is the “best” inverse one can get without also providing correlation information. This answer is “network” accuracy and presumes that the coordinates of the two points are statistically independent of one another. Local accuracy will compute as being identical to network accuracy because no correlation data are provided.
4. Or, the user may enter the full covariance matrix at each point as well as the correlation matrices between points. The correlation of the forepoint with respect to the standpoint is the transpose of the correlation of the standpoint

with respect to the forepoint. It is redundant, but both correlation matrices need to be entered (the astute Excel user will quickly rekey the appropriate cells so that correlation data need to be entered only once).

REFERENCES

Burkholder, E. F. 1997. Definition and description of a global spatial data model (GSDM). Washington, DC: Filed with the U.S. Copyright Office. http://www.zianet.com/globalcogo/gsdmdefn.pdf.

———. 1999. Spatial data accuracy as defined by the GSDM. *Journal of Surveying and Land Information Systems* 59 (1). http://www.zianet.com/globalcogo/accuracy.pdf.

———. 2001. Spatial data, coordinate systems, and the science of measurement. *Journal of Surveying Engineering* 127 (4). http://www.nmsu.edu/~survey/asce_spatial_data_paper.pdf.

———. 2002. Elevations and the global spatial data model (GSDM). Paper presented at the 58th annual meeting of the Institute of Navigation, June 24–26, Albuquerque, NM. http://www.zianet.com/globalcogo/elevgsdm.pdf.

———. 2004. Fundamentals of spatial data accuracy and the global spatial data model (GSDM). Washington, DC: Filed with the U.S. Copyright Office. http://www.globalcogo.com/fsdagsdm.pdf.

———. 2006. The digital revolution—whither now? Global spatial data model. *GIM International* 20 (9): 25–27. http://www.globalcogo.com/Digital%20RevR.pdf.

Federal Geographic Data Committee (FGDC). 1998. *Geospatial positioning accuracy standards, part 2: Standards for geodetic networks*. Reston, VA: Federal Geographic Data Committee, U.S. Geological Survey. http://www.fgdc.gov/standards/standards_publications.

Ghilani, Charles, and Paul R. Wolf. 2006. *Adjustment computations*, 4th ed. New York: John Wiley.

Han, J. Y., B. M. van Gelder, T. Soler, R. Snay. (2008). On geometric combination of multiple terrestial network solutions. *Journal of Surveying Engineering*, in press.

Harvey, Miles. 2000. *The Island of Lost Maps: A true story of cartographic crime*. New York: Broadway.

Hawking, S. W. 1988. *A brief history of time*. New York: Bantam.

Kuhn, Thomas S. 1996. *The structure of scientific revolutions*, 3rd ed. Chicago: University of Chicago Press.

Kumar, Muneendra. 2005. When ellipsoid height will do the job, why look elsewhere! *Surveying and Land Information Science* 65 (2): 91–94.

———. 2007. Ellipsoidal heights and engineering applications: Research with real data proves they will work! *Journal of Surveying Engineering* 133 (3): 96–97.

Mikhail, E. 1976. *Observations and least squares*. New York: Harper & Row.

National Imagery and Mapping Agency. 1997. Department of Defense World Geodetic System 1984: Its definition and relationships with local geodetic systems. NIMA TR8350.2, 3rd ed., 4 July. Bethesda, MD: National Imagery and Mapping Agency.

National Research Council. 1983. *Procedures and standards for a multipurpose cadastre*. Panel on a Multipurpose Cadastre, National Research Council. Washington, DC: National Academies Press.

Snay, R. A. 1999. Using the HTDP software to transform spatial coordinates across time and between reference frames. *Surveying and Land Information Systems* 59 (1): 15–25.

Snay, R. A., and T. Soler. 1999. Modern terrestrial reference systems—part 3: WGS 84 and ITRS. *Professional Surveyor* 20 (3). http://www.ngs.noaa.gov/CORS/Articles/Reference-Systems-Part-3.pdf.

Soler, T. 2007. Editorial. *Journal of Surveying Engineering* 133 (3): 95.

12 Using the GSDM

INTRODUCTION

Using the GSDM is primarily a matter of choosing to do so. The technology is already in place, and all equations and procedures are in the public domain. Although the GSDM contains no new science, it does represent a different way of organizing and storing spatial data. The intent in defining the GSDM was to begin with the assumption of a single origin for 3-D data and, from that, to build a collection of procedures that can be used to handle 3-D spatial data more efficiently. Due to overlaps with existing practice, the decision to use the GSDM need not be "all or nothing." For many, the transition to using the GSDM will lie between two extremes.

- Start with a small project, and build a 3-D database while using the GSDM. Anyone can input autonomous latitude, longitude, ellipsoid height, and point names or descriptors as obtained from a simple handheld GPS receiver. Standard deviations are optional—the default value is 0.0. But provable values or reasonable estimates for standard deviations are appropriate. Each point record includes a point number, ECEF coordinates, an associated covariance matrix, and a descriptor.

  If using GPS vector data, the beginning control points are defined with ECEF coordinates as obtained from a reliable source and entered into the database. Subsequent field observations are processed to obtain ECEF coordinate differences (a baseline vector), and new points are defined by adding those differences to points already in the 3-D database. The *X*/*Y*/*Z* (ECEF) coordinates of any project (large or small) are compatible with all other ECEF coordinates on the same datum the world over. If included, the positional quality (standard deviation) of each point is described by its covariance matrix.
- Build a 3-D database before using it. Points in existing horizontal and vertical databases are systematically converted to ECEF coordinate values and stored in an integrated 3-D database. The challenge will be to make sure that each point has a legitimate ECEF 3-D definition. It may be possible to build a reliable 3-D database with few (or no new) field ties, but the mathematical conversion of existing data will need to be done with great care. For example, datum compatibility will be a huge issue, and the orthometric height of each vertical point will need to be converted to ellipsoid height using a proven (acceptable) geoid model. But the saving grace is that the GSDM accommodates the standard deviation of each point—whatever it is. So long as a standard deviation reliably reflects the quality of the newly defined point in the 3-D database, there is no need for datum conversions to be "perfect." Well-documented professional judgment and consistent application of adopted

policies will serve to protect the reputation of a named 3-D database. Each subsequent person who uses information from a given database must be able to rely on the statistical quality of data obtained therefrom.

Regardless of how the 3-D database is started or established, once information is stored as ECEF coordinates and covariances, then interoperability has been established and any or all users have the option of using that information in a multitude of applications. The user is responsible for subsequent manipulation of the data. Ideally, the rules of such manipulation will be easy to understand and will be bidirectional so that new data can be added to the 3-D database. (In this context, there needs to be a careful discussion of administrative responsibility for databases—master, global, national, regional, agency, discipline, project, task, local, and temporary. That discussion is beyond the scope of this book.) Of course, any data added to the database will be done under appropriate authority and will meet established quality control criteria. However, it should be understood that a *reliable* standard deviation of any quantity is more important than how big it is. The GSDM competently handles 3-D data with large standard deviations just as well as it handles data with small standard deviations. The difference lies in the quality of the answer obtained when using those data. Data with large standard deviations will provide answers with large standard deviations.

The actual transition to using the GSDM will vary among persons and organizations. Many will begin using the GSDM only when they see it as being in their self-interest to do so. Arriving at that conclusion or developing such a consensus will take time. At the beginning of a new project or on a given date, it could be decreed that GSDM policies and procedures will henceforth be used. That approach would be the most efficient way to implement the GSDM—given a compatible database has been developed. But, developing a 3-D database will also take time and resources. The presumption is that each user or organization will explore and discover what is best for the circumstance. Many of the individual procedures described in this book are already being used and will continue to be useful. However, as users become more comfortable with how the individual pieces relate to the whole, resistance to using the GSDM will be reduced. The goal is that spatial data practices will evolve in such a manner that the underlying features of the GSDM will be recognized as common ground for practice worldwide. It is conceivable and anticipated that the GSDM will eventually become the global standard for handling geospatial data.

Once we get beyond the "magic" of electronic signal processing and have access to the spatial data components from our GPS receivers (or other sensors), all we need are the rules of solid geometry to keep track of where we are or where we've been. No, it is not quite that easy, as we still need to deal with issues of datums, coordinate systems, units, and whether the data are relative or absolute. But, the GSDM provides a unique bridge between the builders and operators of measurement systems and spatial data users in many disciplines all over the world. Using the same GSDM and 3-D database, rocket scientists, engineers, photogrammetrists, surveyors, and others can continue working in a rigorous global environment while local users simultaneously enjoy the luxury of working with flat-Earth rectangular components. The GSDM will be useful to novice and expert alike.

FEATURES

The GSDM has two primary features—the functional model and the stochastic model. The functional model is a collection of equations and geometrical relationships that can be used to describe a unique position anywhere within the birdcage of orbiting GNSS (global navigation satellite systems) satellites. The optional stochastic model is a set of rules or procedures that can be used to keep track of standard deviations of the observations, the derived measurements, the computed or stored coordinates, and any quantity computed from them.

The Functional Model

Because geospatial data are connected to the Earth and because spatial data are 3-D, the "big-picture" geometrical relationships included in the GSDM involve a lot of geometrical geodesy. Given previous use of 2-D latitude and longitude on the ellipsoid surface, rules of classical geodesy are still very useful and need to be included. However, given that local users view a "flat Earth" from a given point, the underlying model needs to provide simple rectangular plane surveying answers with the same integrity as when providing answers on the ellipsoid. The GSDM does both. Functional model equations are listed in chapter 1 and derived in chapter 6—except for the rotation matrix, which is derived in appendix A.

The following is an oversimplified summary of models. In the first approximation, the world is considered flat. That assumption is appropriate for many local applications. A spherical Earth is a better model and is useful for many "big-picture" applications in geography and navigation. An ellipsoidal Earth model has been used for triangulation computations (and other applications requiring a high level of geometrical integrity) for over 200 years. The GSDM goes one step beyond those to model geospatial data in 3-D space. Given the digital revolution and enormous data storage capability, each user now has the option of viewing geospatial data from any origin or perspective. With simple solid geometry relationships at our disposal, each user can enjoy the luxury of unique ECEF absolute *X*/*Y*/*Z* coordinates while simultaneously working with local relative coordinate differences (in any previous model)—all without sacrificing the geometrical integrity of the observations, the measurements, or the data.

The Stochastic Model

The stochastic model can be used to assign standard deviations to the position of any point in any direction—that is, standard deviation in the north-south direction, in the east-west direction, and in the up-down direction or in the ECEF reference frame. With practice and appropriate software, determining those standard deviations can be fairly straightforward. Even so, spatial data users need to be specific when quoting standard deviations by stating the context (i.e., "with respect to what"). Possible options include the following:

- With respect to the NAD83
- With respect to the ITRF

- With respect to the WGS84
- With respect to the geoid
- With respect to the control held by the user
- With respect to an implied or unstated datum

Given the sheer volume of spatial data being generated and used, and given the possible consequences of bad answers, the issue of spatial data accuracy is becoming increasingly important, and statements of spatial data accuracy need to be unambiguous. In the past, the quality of geodetic surveys was often described with adjectives such as "first-order," "second-order," and so on. Such categories worked for those familiar with geodetic surveying, but the standards were largely distance dependent and were expressed as ratios—for example, one part in 100,000 is first-order. While that criterion was appropriate for long triangulation lines, it is very difficult to meet first-order specifications when measuring short lines. One part in 100,000 translates to 0.0001 meters in 10 meters.

In the 1990s, the Federal Geographic Data Committee (FGDC; 1998) developed "Geospatial Positioning Accuracy Standards, Part 2: Standards for Geodetic Networks." Several comments are as follows:

- The new standards are part of a larger effort to quantify spatial data accuracy in many categories, not just those used in geodetic surveying.
- The new standards have a better theoretical basis and use a positional tolerance criterion rather than a ratio of precision to describe spatial data quality.
- The first-order, second-order, and so on categories have been replaced with names that are more intuitive.
- The standards are more closely aligned with modern positioning (GNSS) technology and provide a more intuitive "meter stick" against which to make comparisons.
- The FGDC standards are quoted at the 95 percent confidence level and are applicable to 3-D data as well as 2-D data.
- The FGDC standards discuss local accuracy and network accuracy without providing a specific mathematical definition. While mostly compatible, the mathematical definition for network accuracy and local accuracy as provided by equation 1.36 in chapter 1 (and repeated in chapter 11) goes beyond that provided by the FGDC.

Table 12.1 is a portion of the accuracy standards taken from the FGDC (1998) web site.

Meta data are data about data. Concepts of meta data were developed in parallel with GIS, and meta data are a very important part of working with spatial data. One reason for the popularity of meta data is that meta data contain more than just metrical characteristics. For images and photogrammetric data, for example, meta data include information such as the name of the organization collecting the data, the equipment used to record the observations, the flying height of aircraft, the time of day, and other details. Given appropriate meta data, equipment calibration details, and knowledgeability of the data reduction overall process, reasonable professionals

TABLE 12.1
Summary of FGDC Accuracy Standards

Horizontal, Ellipsoid Height, and Orthometric Height Accuracy Classification	95% Confidence Less Than or Equal To
1 millimeter	0.001 meters
2 millimeters	0.002 meters
5 millimeters	0.005 meters
1 centimeter	0.010 meters
2 centimeters	0.020 meters
5 centimeters	0.050 meters
1 decimeter	0.100 meters
2 decimeters	0.200 meters
5 decimeters	0.500 meters
1 meter	1.000 meters
2 meters	2.000 meters
5 meters	5.000 meters
10 meters	10.00 meters

place justifiable reliance on the quality of the resulting information. Typically, meta data apply to a particular data set—often a very large data set. The accuracy of an individual point is stated as being representative of many points in the same data set sharing similar characteristics. Although not exclusive, meta data are often considered to be more appropriate when working with raster data as opposed to vector data.

Certainly meta data will continue to be important, but, when working with vector data, the GSDM provides some very powerful advantages. The GSDM contains algorithms for determining the standard deviation of each point in all three directions. Without knowing any of the meta data associated with points in the GSDM database, a numerical filter can be imposed to screen out any data not meeting a positional tolerance criterion selected by the user.

The following statements include some crystal ball speculation. As storage capacity becomes more affordable, it will be feasible to convert more raster data to vector data. At some point in the future, it will be possible (and sometimes warranted) for each pixel in a raster image file to have its geospatial location defined by ECEF coordinates and to have the associated point covariance matrix stored for each point. Since raster data are already stored, an alternative would be to develop, store, and use an algorithm for converting raster data to ECEF vector data (ECEF coordinates and covariances) on an "as needed" basis. That would avoid duplicate storage requirements while permitting raster data to be converted and used in a vector environment. With that capability in place, many silos of existing electronic imagery can be readily accessible to spatial data users for a multitude of applications.

DATABASE ISSUES

A statement of the obvious is "Don't mix datums in a 3-D database." One of the challenges of using GNSS data is knowing what datum to choose. In the United States, there are three obvious choices—the NAD83, the WGS84, and the ITRF. At the gross level, it does not matter which datum is used. As computed in chapter 3 using spherical trigonometry, the distance from New Orleans to Chicago is 1,359.4 kilometers. Later, in chapter 6, the GSDM was used to find the distance as 1,356.5 kilometers. That difference is primarily the effect of using an ellipsoidal Earth in place of a spherical Earth. Furthermore, the latitude-longitude positions in each city were listed only to the nearest 1 second of arc, giving an implied tolerance of about 30 meters at each end of the line. The GSDM inverse in chapter 6 assumed the latitude-longitude positions were "exact" and achieved a proven answer within a centimeter on the same line. Now, if the only difference is to recompute the inverse using the WGS84 as the ellipsoid (on the WGS84 datum) instead of the GRS80 ellipsoid (on the NAD83 datum), the GSDM answer is still within 1 cm (left as an exercise for the reader). So, why does the datum make a difference? The datum makes a difference because the origins of the two datums are not at the same place. The relative differences on the two datums are very nearly identical, but the absolute positions are different. In the New Orleans to Chicago example just cited, the latitude/longitude values were taken to be the same, but the same latitude/longitude values represent two different points on the ground in New Orleans and two different points on the ground in Chicago. The egregious mistake would be to inverse from New Orleans on one datum to Chicago on another datum. Of course, even a discrepancy of a meter or two in the distance from New Orleans to Chicago would hardly be significant if the comparison were being made to the nearest kilometer. The rule may be conservative but still stands: don't mix datums in a 3-D database.

Other database issues are related to records, fields, and format. A BURKORD™ database contains two kinds of records—a point record and a correlation record. There is no specified order for records in the database, but the first field in each record is reserved for flexibility and future use. Any refinements to the generic format listed here will be identified on the Global COGO, Inc., web site, http://www.globalcogo.com.

Additional uses may be defined in the future, but any such changes are intended to preserve compatibility with the following format. Any record beginning with a "p" is a point record, and any record beginning with a "c" is a correlation record. The specified format for each type is as follows:

A point record contains the following fields, space or comma delimited.

- Attribute field—string characters (no blanks):
  1. First character (required) is reserved.
     A. "p" is a point record.
     B. "c" is a correlation record—see below.
  2. Next three characters (optional) are project identifiers.
  3. Characters 5 to *n* (also optional) are the prerogative of the user.
- Point number must be an integer.
- *X*/*Y*/*Z* coordinate values—three double precision fields.

- Variances of *X*/*Y*/*Z*—three double precision fields.
- Covariances *XY*, *XZ*, and *YZ*—three double precision fields.
- Station name—string characters (blanks OK prior to end of record).

A point-to-point correlation record contains the following fields, space or comma delimited.

- Attribute field—string characters (no blanks):
  1. First character (required) is reserved.
     A. "p" is a point record—see above.
     B. "c" is a correlation record.
  2. Next three characters (optional) are project identifiers.
  3. Characters 5 to *n* (also optional) are the prerogative of the user.
- Two point numbers—two integer fields.
- X_1X_2, X_1Y_2, and X_1Z_2 covariances—three double precision fields.
- Y_1X_2, Y_1Y_2, and Y_1Z_2 covariances—three double precision fields.
- Z_1X_2, Z_1Y_2, and Z_1Z_2 covariances—three double precision fields.

An example of the BURKORD™ database file format is given in Table 12.2.

TABLE 12.2

Example of a BURKORD™ Database File

New Mexico State University (NMSU)—Las Cruces, New Mexico

GPS Network for Control Points on NMSU Campus—EFB 2005

p, 1001, –1571430.672, –5164782.312, 3387603.188, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, Crucesair

p, 1002, –1556177.615, –5169235.319, 3387551.709, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, Reilly

p, 1003, –1555678.579, –5169961.396, 3386700.089, 2.161E-6, 8.474E-6, 6.812E-6, 2.347E-6, –1.496E-6, –5.017E-6, USPA

p, 1004, –1555663.613, –5169976.761, 3386683.419, 3.114E-6, 2.182E-5, 1.094E-5, 6.968E-6, –4.671E-6, –1.420E-5, USPB

p, 1005, –1555206.615, –5169400.740, 3387285.987, 1.242E-6, 4.506E-6, 4.182E-6, 1.343E-6, –9.212E-7, –2.775E-6, Pseudo

p, 1006, –1556209.750, –5169286.496, 3387457.512, 2.081E-6, 5.554E-6, 4.668E-6, 2.367E-6, –1.768E-6, –3.609E-6, Bromilow

c, 1003, 1005, 1.141E-6, 1.285E-6, –8.927E-7, 1.293E-6, 4.240E-6, –2.576E-6, –8.829E-7, –2.568E-6, 3.603E-6

Note: The first two lines are headers—string data of user's choice. Only one correlation record is shown. Numbers for other point-pair combinations are in the matrix shown in Figure 11.3.

IMPLEMENTATION ISSUES

This section looks specifically at characteristics of the GSDM and discusses some of the implications associated with implementation. Issues include, but are not limited to, the following:

- The GSDM accommodates modern measurement procedures and digital data.
- The GSDM uses proven rules of solid geometry and vector algebra in a global rectangular environment.
- The 3-D database is simple and equally applicable the world over.
- While the underlying standard insures interoperability, geospatial data users in any discipline have complete freedom to be innovative in derivative applications. Traditional 2-D applications are fully supported.
- Orthometric heights are referenced to the elusive geoid, while ellipsoid heights are referenced to the Earth's center of mass. Full adoption of the GSDM presumes ellipsoid heights will be used to describe the third dimension. Geoid modeling will still be needed and used by those requiring precise hydraulic grade lines.
- The stochastic feature provides tools by which spatial data accuracy can be established, tracked, and used.
- The GSDM provides a concise mathematical definition of network accuracy and local accuracy.
- Absolute *X*/*Y*/*Z* coordinates for any point in the "birdcage" of satellites are globally unique. As appropriate, any user is free to convert those coordinates to other systems such as *latitude*/*longitude*/*height*, UTM, and state plane or other coordinate systems.
- The GSDM does *not* provide for transformations between datums. But, datum-to-datum relationships are best modeled in terms of ECEF coordinates using the standard seven-parameter transformation.
- Relative coordinate differences of $\Delta X/\Delta Y/\Delta Z$ are not very intuitive for humans. But, given the ease with which geocentric differences can be rotated to local tangent plane differences, the local flat-Earth user has immediate access to plane surveying rectangular components.
- Distances from standpoint to forepoint are in the local tangent plane and are identical to the HD(1) distance defined in Burkholder (1991). Other distances can be computed, if needed, without disturbing the coordinates in the database.
- *Important point!* The tangent plane from "here" to "there" is not the same as the tangent plane from "there" to "here." Within a very small tolerance, the 3-D azimuth (Burkholder 1997) is the true geodetic azimuth from standpoint to forepoint. To the geodesist, this is as it should be. However, for plane surveying applications, the P.O.B. datum feature should be used.
- When using the P.O.B. datum, all distances between points are in the same tangent plane through the P.O.B., and the azimuths are grid azimuths with respect to the true meridian through the P.O.B. The implication of this feature is that two surveys referenced to separate P.O.B.'s and sharing a

common line will have two azimuths for the same line. That difference of the two azimuths is the convergence of the meridian between the respective P.O.B.'s. That "problem" is resolved by identifying the P.O.B. for the survey on each plat. The underlying ECEF coordinates and their covariances remain unchanged in each case.

- With these features of the GSDM already defined, in place, and universally available, there is no need for a "low-distortion projection."
- Ellipsoid heights and their standard deviations are obtained directly from ECEF coordinates in the 3-D database. Geoid modeling procedures will be used to obtain orthometric heights if needed. The broader question is "Why are orthometric heights needed?" Unless grades are very critical for hydraulic grade lines (in which case dynamic heights and height differences should be used), an ellipsoid height difference readily approximates an orthometric height difference. Geoid modeling will still be used by those for whom the difference matters.
- Comments and rhetorical questions:
  1. Satellites orbit the physical center of mass of the Earth.
  2. The center of mass of the Earth is quite stable, whereas the geoid moves up and down during the Earth's daily rotation.
  3. CORS stations fixed to bedrock also go up and down during the day—Earth tides.
  4. Do CORS stations and the underlying geoid go up and down together?
  5. GPS data can be used to monitor the daily motion of a precisely surveyed CORS station.
  6. Is the mean ellipsoid height of a CORS station preferred to the instantaneous ellipsoid height? How precisely can an instantaneous ellipsoid height be determined?
  7. What difference, if any, does it make that ellipsoid heights are absolute (they move up and down daily relative to the center of mass) while orthometric heights are relative (benchmarks move up and down during the day along with the geoid)?

APPLICATIONS AND EXAMPLES

Examples of using the GSDM are posted on the Global COGO web site. Three of them are as follows:

- http://www.globalcogo.com/nmsunet1.pdf
- http://www.globalcogo.com/gpselev1.pdf
- http://www.globalcogo.com/3Dgps2Dplat.pdf

The first example is highlighted in chapter 11, but more details are included on the web site posting. This project illustrates the process of collecting GPS data, processing the baselines, checking the misclosures, performing a least squares adjustment (using three different weighting options), developing the statistics for the computed points, and using the GSDM to output results complete with a computation of network accuracy and local accuracy between two points.

The second example shows how GPS data were used in the context of the GSDM along with geoid modeling to determine the NAVD88 orthometric height of the A-order HARN station on the NMSU campus. The procedure used is not that sanctioned by the NGS, but all steps are documented and the final elevation relies on, among other criteria, the quality of GEOID03. Although more rigorous procedures (such as dual-frequency GPS, longer sessions, and more first-order benchmarks) could have been used, the observed or computed elevation has a standard deviation of 3 mm and the computed elevation is consistent with other methods of observation. Questions about the stability of the existing first-order benchmarks may be legitimate.

The third example noted above illustrates the direct connection between 3-D GPS-derived positions and a 2-D plat of the survey. The direct connection lies in the use of the rotation matrix and the P.O.B. datum feature of the GSDM. Very briefly, the project consists of two HARN stations and Section 31, T23S-R1E, New Mexico Principal Meridian—all in the Las Cruces area. Section 31 is BLM property, conveniently lies between the two HARN stations, and, although strictly vacant desert land, enjoys convenient vehicular access. Details are posted on the web site and summarized here.

GPS data were collected on three separate days, and trivial vectors were carefully avoided. The NAD83 geocentric *X*/*Y*/*Z* coordinates as published by the NGS were held fixed for the two HARN points, and a least squares adjustment of the interconnected baselines provided the geocentric coordinates and standard deviations as shown in Table 12.3. The latitude, longitude, and ellipsoid heights and standard deviations were computed using the GSDM.

Using the points in Table 12.3, the southwest corner of Section 31 was chosen as the P.O.B., and local tangent plane coordinates (eastings and northings) were computed for all points in the survey with respect to the SW corner. Those values are shown in Table 12.4 in meters.

Table 12.5 shows local tangent plane inverses around Section 31 referenced to the true meridian through the SW corner. Although meter units are the international

TABLE 12.3

3-D GPS Points for 2-D Survey

Crucesair (HARN pt)	
X = –1,571,430.6720 m fixed	Lat. 32º 16’ 54.”63123 N fixed
Y = –5,164,782.3120 m fixed	Long. 106º 55’ 22.”24784 W fixed
Z = 3,387,603.1880 m fixed	El Hgt h = 1,326.250 m fixed
Reilly (HARN pt)	
X = –1,556,177.6150 m fixed	Lat. 32º 16’ 55.92906 N fixed
Y = –5,169,235.3190 m fixed	Long. 106º 45’ 15.16070 W fixed
Z = 3,387,551.7090 m fixed	El Hgt h = 1,166.570 m fixed

NW Cor 31

ECEF Frame	Local Frame
X = –1,568,446.9652 m +/– 0.0033 m	Lat. 32º 16’ 16.51587 N +/– 0.0054 m
Y = –5,166,282.9266 m +/– 0.0077 m	Long. 106º 53’ 16.50858 W +/– 0.0039 m
Z = 3,386,573.0861 m +/– 0.0044 m	El Hgt h = 1,256.511 m +/– 0.0067 m

NE Cor 31

ECEF Frame	Local Frame
X = –1,566,906.8273 m +/– 0.0034 m	Lat. 32º 16’ 16.50308 N +/– 0.0057 m
Y = –5,166,748.4577 m +/– 0.0080 m	Long. 106º 52’ 15.04095 W +/– 0.0040 m
Z = 3,386,571.5363 m +/– 0.0046 m	El Hgt h = 1,254.233 m +/– 0.0070 m

SW Cor 31

ECEF Frame	Local Frame
X = –1,568,698.0864 m +/– 0.0035 m	Lat. 32º 15’ 24.28753 N +/– 0.0058 m
Y = –5,167,107.1198 m +/– 0.0083 m	Long. 106º 53’ 16.54155 W +/– 0.0041 m
Z = 3,385,214.0743 m +/– 0.0047 m	El Hgt h = 1,259.609 m +/– 0.0072 m

SE Cor 31

ECEF Frame	Local Frame
X = –1,567,157.4899 m +/– 0.0035 m	Lat. 32º 15’ 24.26696 N +/– 0.0058 m
Y = –5,167,571.1861 m +/– 0.0081 m	Long. 106º 52’ 15.08320 W +/– 0.0041 m
Z = 3,385,211.2732 m +/– 0.0047 m	El Hgt h = 1,255.365 m +/– 0.0071 m

N 1/4 Cor 31

ECEF Frame	Local Frame
X = –1,567,682.4363 m +/– 0.0036 m	Lat. 32º 16’ 16.72241 N +/– 0.0058 m
Y = –5,166,510.8119 m +/– 0.0080 m	Long. 106º 52’46.03144 W +/– 0.0042 m
Z = 3,386,578.0990 m +/– 0.0048 m	El Hgt h = 1,255.823 m +/– 0.0070 m

W 1/4 Cor 31

ECEF Frame	Local Frame
X = –1,568,572.7788 m +/– 0.0035 m	Lat. 32º 15’ 50.39908 N +/– 0.0058 m
Y = –5,166,694.9985 m +/– 0.0080 m	Long. 106º 53’ 16.53459 W +/– 0.0041 m
Z = 3,385,893.5160 m +/– 0.0048 m	El Hgt h = 1,258.017 m +/– 0.0070 m

S 1/4 Cor 31

ECEF Frame	Local Frame
X = –1,567,928.0513 m +/– 0.0038 m	Lat. 32º 15’ 24.28747 N +/– 0.0060 m
Y = –5,167,339.5732 m +/– 0.0083 m	Long. 106º 52’ 45.81734 W +/– 0.0044 m
Z = 3,385,213.3104 m +/– 0.0050 m	El Hgt h = 1,258.181 m +/– 0.0073 m

E 1/4 Cor 31

ECEF Frame	Local Frame
X = –1,567,032.6554 m +/– 0.0038 m	Lat. 32º 15’ 50.37719 N +/– 0.0059 m
Y = –5,167,160.8706 m +/– 0.0082 m	Long. 106º 52’ 15.06861 W +/– 0.0043 m
Z = 3,385,891.8435 m +/– 0.0048 m	El Hgt h = 1,255.952 m +/– 0.0072 m

TABLE 12.4
Local P.O.B. Coordinate Differences

Point Description	P.O.B. East (meters)	P.O.B. North (meters)	P.O.B. Up (meters)
1001 Crucesair	–3,290.100	2,783.992	65.183
1002 Reilly	12,598.767	2,831.217	–106.099
1013 NW Cor Sec 31	0.863	1,609.117	–3.302
1014 NE Cor Sec 31	1,609.819	1,608.850	–5.783
1015 SW Cor Sec 31	0.000	0.000	0.000
1016 SE Cor Sec 31	1,608.970	–0.506	–4.447
1017 N Quarter Cor Sec 31	798.622	1,615.512	–4.041
1018 W Quarter Cor Sec 31	0.182	804.478	–1.643
1019 S Quarter Cor Sec 31	804.355	0.030	–1.479
1020 E Quarter Cor Sec 31	1,609.224	803.931	–3.910

TABLE 12.5
Inverse Directions and Distances Based on P.O.B. Values

Azimuth From Point To	Point	DMS	Distance
SW Cor	W Quarter Cor	0 00 46.7	2,639.357 ft
W Quarter Cor	NW Cor	0 02 54.5	2,639.889 ft
NW Cor	N Quarter Cor	89 32 26.7	2,617.399 ft
N Quarter Cor	NE Cor	90 28 13.7	2,661.492 ft
NE Cor	E Quarter Cor	180 02 32.6	2,640.808 ft
E Quarter Cor	SE Cor	180 01 05.2	2,639.222 ft
SE Cor	S Quarter Cor	270 02 17.4	2,639.807 ft
S Quarter Cor	SW Cor	269 59 52.3	2,638.955 ft

standard, the distances were converted to feet to illustrate the flexibility of the GSDM with respect to units. When displaying the data, the user is free to use units of choice. A note here is that because coordinate differences were used and because none of those differences *on the plat* is more than 2 kilometers, it does not matter whether the U.S. Survey Foot definition or the International Foot is used. Chains or other units could also be used. Figure 12.1 is a plat of the survey.

WBK SOFTWARE

As stated in chapter 1, the term BURKORD™ has been trademarked to cover software developed and marketed by Global COGO, Inc. The original BURKORD

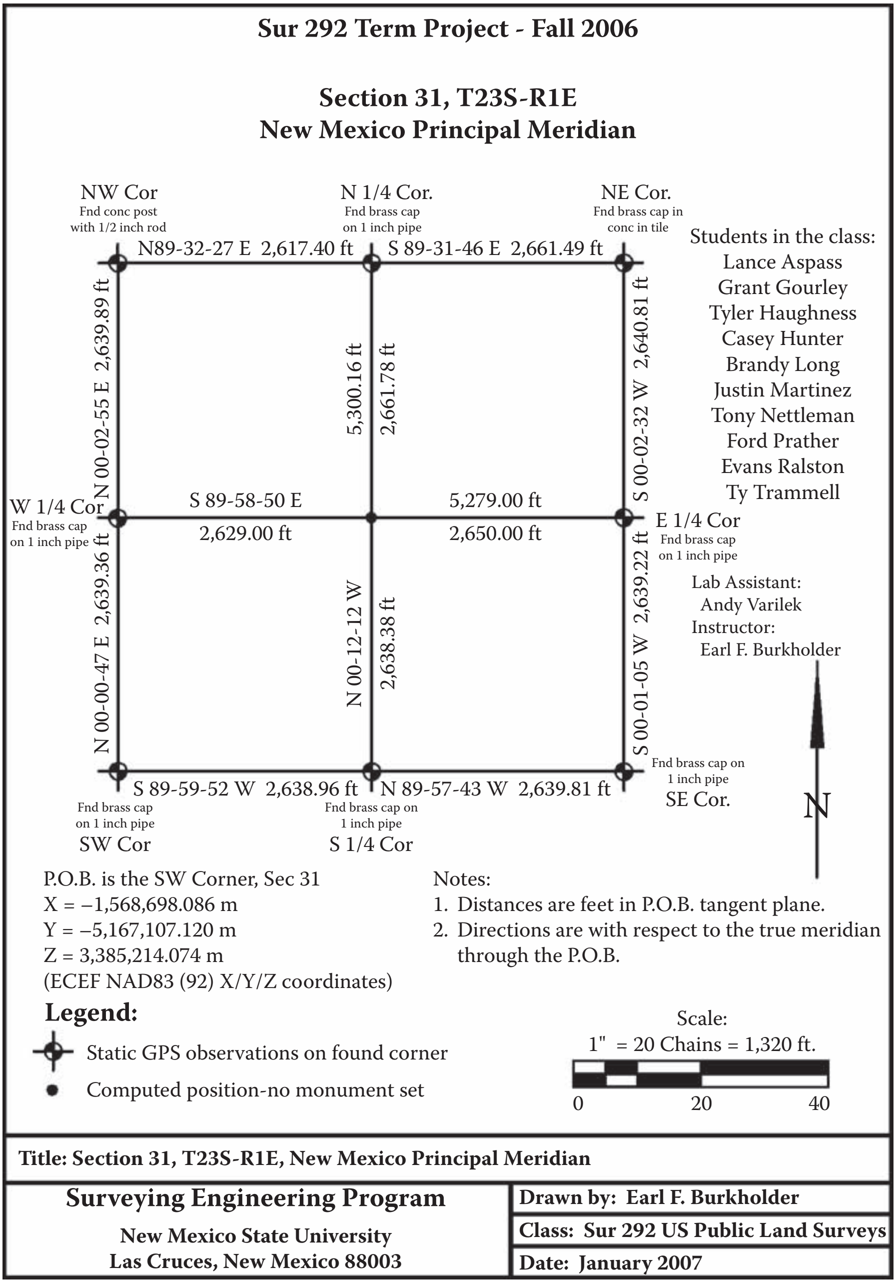


FIGURE 12.1 2-D Plat of 3-D Survey

program was DOS based, was menu driven, and served very well as a prototype for proof of concept. The Windows version of BURKORD is called WBK and will eventually be available as a commercial product. In the meantime, a fifty-point version is called WBK-Basic and is available free upon request. Follow the appropriate link on the Global COGO, Inc., web page, http://www.globalcogo.com.

REFERENCES

Burkholder, E. F. 1991. Computation of horizontal/level distances. *Journal of Surveying Engineering* 117 (3): 104–16.

———. 1997. Azimuth of a GPS vector. *Journal of Surveying Engineering* 123 (4): 139–46.

Federal Geographic Data Committee (FGDC). 1998. *Geospatial positioning accuracy standards, part 2: Standards for geodetic networks*. Reston, VA: Federal Geographic Data Committee, U.S. Geological Survey. http://www.fgdc.gov/standards/.

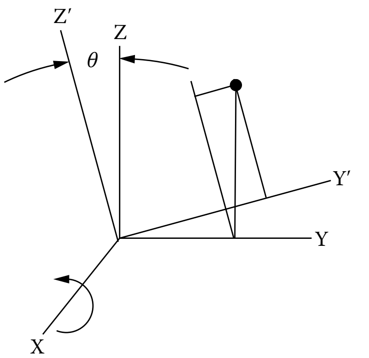
Appendix A: Rotation Matrix Derivation

A rotation matrix is a collection of equations expressed in matrix form and used to change the perspective associated with spatial data. In the case of the GSDM, the perspective changes from looking at points in the geocentric ECEF reference frame to viewing the same data from the perspective of one occupying a given point (called the point of beginning, or P.O.B.) and viewing all other points as if standing at the selected origin.

One way of applying the rotation matrix is to recompute the coordinates of each point with respect to the original origin. Another approach is to apply the rotation matrix to the vector from the P.O.B. to any other point in the database. The GSDM uses the second approach. That means the geocentric coordinate differences are found first. Then the rotation matrix is applied to the ECEF vector components, and the result is local "flat-Earth" components from the P.O.B. to the selected point. Although it could be argued that the requirement to compute values of the rotation matrix for each P.O.B. is a disadvantage, several advantages are as follows:

1. The underlying primary data (ECEF coordinates) are not modified and remain available for immediate recomputation—such as selecting a different P.O.B.
2. After using the rotation matrix, the local components of any vector appear to the user as rectangular flat-Earth plane-surveying components. These components can be used in a variety of operations as selected by the user.
3. The process is reversible in that local perspective components can also be rotated into the ECEF perspective, making such differences compatible with *X*/*Y*/*Z* points already stored in the database. The *X*/*Y*/*Z* coordinates of new points are established using simple addition and subtraction operations.
4. If the spatial data user needs more traditional values for a point, such as latitude-longitude or state plane coordinates, those values are also immediately available by rigorous computation (though not involving a rotation matrix) from the geocentric ECEF values stored in the database.

Admittedly, the derivation of the rotation matrix is tedious, but it really is not that difficult. The convention is that a positive rotation is counterclockwise as viewed looking at the origin from the positive direction along the axis being rotated (Leick 2004). This rule needs to be applied once for each of the three possible axes. That process yields three separate matrices—one for each rotation. This part is generic without regard to being attached to the Earth, and each rotation is illustrated in a separate diagram. Figure A.1 shows a positive rotation (R_1) about the X axis. Figure A.2 shows a positive rotation (R_2) about the Y axis. And, Figure A.3 shows a positive rotation (R_3) about the Z axis.

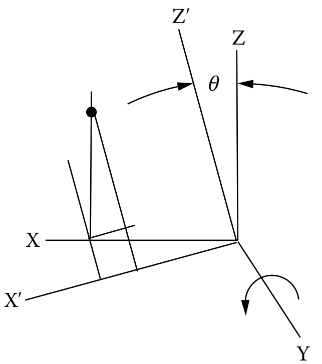


$$X' = X$$
$$Y' = Y\cos\theta + Z\sin\theta$$
$$Z' = -Y\sin\theta + Z\cos\theta$$

In matrix form,

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{A.1}$$

FIGURE A.1 R_1 Rotation about the X Axis

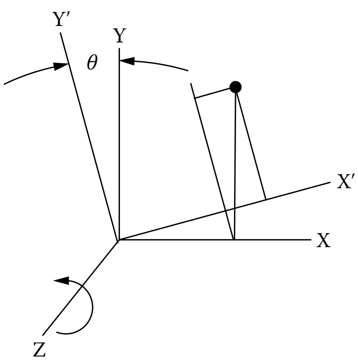


$$X' = X\cos\theta + Z\sin\theta$$
$$Y' = Y$$
$$Z' = X\sin\theta + Z\cos\theta$$

In matrix form,

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{A.2}$$

FIGURE A.2 R_2 Rotation about the Y Axis



$$X' = X\cos\theta + Y\sin\theta$$
$$Y' = -X\sin\theta + Y\cos\theta$$
$$Z' = Z$$

In matrix form,

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{A.3}$$

FIGURE A.3 R_3 Rotation about the Z Axis

With each rotation quantified, the process is applied to vectors (coordinate differences) attached to the Earth. Starting with the X/Y/Z axes of the ECEF reference frame, the first rotation is a positive rotation about the Z axis (R_3) to bring the Y axis into the vertical plane of the local meridian. The angular amount is determined by the longitude of the selected P.O.B. and is computed as east longitude + 90°. A full circle can be subtracted if the sum goes over 360°. The second rotation is also a positive rotation about the X axis (R_1) to bring the Y axis into the local geodetic horizon. The two rotations describe movement of the original Y axis in two rotations—first is a positive rotation about the Z axis, then a second positive rotation about the X axis. Other rotation sequences could be used to obtain the same result. The original $\Delta X/\Delta Y/\Delta Z$ vector is right-handed, and the rotated differences are also right-handed if used as $\Delta e/\Delta n/\Delta u$.

Multiplication of the two matrices is shown below. The rules of matrix multiplication require the sequence to be as shown in equation A.1. In equation A.2, the actual latitude and longitude of the selected P.O.B. are used, and equation A.3 is a simplification based upon making substitutions for trigonometric identities. Finally, equation A.4 is the form of the rotation matrix given in chapter 1, equation 1.21.

It is stated without proof here that equation 1.22 uses the transpose of the rotation matrix to convert the local geodetic horizon perspective to the ECEF perspective. Being able to use the transpose for the reverse computation is a consequence of the two right-handed systems both being orthogonal—see chapter 3 (Vanicek and Krakiwsky 1986).

Now, using the matrices in Figure A.1 and Figure A.3, the process is attached to the Earth in the following matrix statement.

$$\begin{bmatrix} \Delta e \\ \Delta n \\ \Delta u \end{bmatrix} = R_1(90^\circ - \phi) R_3(\lambda + 90^\circ) \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix} \tag{A.1}$$

$$\begin{bmatrix} \Delta e \\ \Delta n \\ \Delta u \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(90^\circ - \phi) & \sin(90^\circ - \phi) \\ 0 & -\sin(90^\circ - \phi) & \cos(90^\circ - \phi) \end{bmatrix}$$
$$\begin{bmatrix} \cos(\lambda + 90^\circ) & \sin(\lambda + 90^\circ) & 0 \\ -\sin(\lambda + 90^\circ) & \cos(\lambda + 90^\circ) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix} \tag{A.2}$$

$$\begin{bmatrix} \Delta e \\ \Delta n \\ \Delta u \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sin\phi & \cos\phi \\ 0 & -\cos\phi & \sin\phi \end{bmatrix} \begin{bmatrix} -\sin\lambda & \cos\lambda & 0 \\ -\cos\lambda & -\sin\lambda & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix} \tag{A.3}$$

$$\begin{bmatrix} \Delta e \\ \Delta n \\ \Delta u \end{bmatrix} = \begin{bmatrix} -\sin\lambda & \cos\lambda & 0 \\ -\sin\phi\cos\phi & -\sin\phi\sin\lambda & \cos\phi \\ \cos\phi\cos\lambda & \cos\phi\sin\lambda & \sin\phi \end{bmatrix} \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix} \tag{A.4}$$

REFERENCES

Leick, A. 2004. *GPS satellite surveying*, 3rd ed. New York: John Wiley.
Vanicek, P., and Edward Krakiwsky. 1986. *Geodesy: The concepts*. New York: North Holland.

Appendix B: 1983 State Plane Coordinate System Constants

Defining Constants for the 1983 State Plane Coordinate System
Source: NOAA Manual NOS NGS 5, State Plane Coordinate System of 1983.

Abbreviations for projection types: (T.M.) = transverse Mercator
(O.M.) = oblique Mercator
(L.) = Lambert conic conformal

Latitudes and longitudes are given as DD MM, and all longitudes are given as west longitudes.

State and Zone Name		Zone Code	Projection Type	Central Meridian and Scale Factor (T.M.) or Standard Parallels (L.)		Grid Origin Latitude Longtitude		Grid Origin Easting Northing	
Alabama:									
AL	East	0101	T.M.	85	50	85	50	200,000	m
				1:25,000		30	30	0.0	m
AL	West	0102	T.M.	87	30	87	30	600,000	m
				1:15,000		30	30	0.0	m
Alaska:									
AK	Zone 1	5001	O.M.	Axis azimuth = atan (–3/4)		133	40	5,000,000.0	m
				1:10,000		57	00	-5,000,000.0	m
AK	Zone 2	5002	T.M.	142	00	142	00	500,000.0	m
				1:10,000		54	00	0.0	m
AK	Zone 3	5003	T.M.	146	00	146	00	500,000.0	m
				1:10,000		54	00	0.0	m
AK	Zone 4	5004	T.M.	150	00	150	00	500,000.0	m
				1:10,000		54	00	0.0	m
AK	Zone 5	5005	T.M.	154	00	154	00	500,000.0	m
				1:10,000		54	00	0.0	m
AK	Zone 6	5006	T.M.	158	00	158	00	500,000.0	m
				1:10,000		54	00	0.0	m
AK	Zone 7	5007	T.M.	162	00	162	00	500,000.0	m
				1:10,000		54	00	0.0	m
AK	Zone 8	5008	T.M.	166	00	166	00	500,000.0	m
				1:10,000		54	00	0.0	m
AK	Zone 9	5009	T.M.	170	00	170	00	500,000.0	m
				1:10,000		54	00	0.0	m
AK	Zone 10	5010	L.	51	50	176	00	1,000,000.0	m
				53	50	51	00	0.0	m
Arizona:									
AZ	East	0201	T.M.	110	10	110	10	213,360.0	m
				1:10,000		31	00	0.0	m
AZ	Central	0202	T.M.	111	55	111	55	212,360.0	m
				1:10,000		31	00	0.0	m
AZ	West	0203	T.M.	113	45	113	45	213,360.0	m
				1:15,000		31	00	0.0	m

Note: State defines origin in International Feet. 213,360 m = 700,000 International Feet.

State and Zone Name		Zone Code	Projection Type	Central Meridian and Scale Factor (T.M.) or Standard Parallels (L.)		Grid Origin Latitude Longtitude		Grid Origin Easting Northing	
Arkansas:									
AR	North	0301	L.	34	56	92	00	400,000.0	m
				36	14	34	20	0.0	m
AR	South	0302	L.	33	18	92	00	400,000.0	m
				34	46	32	40	400,000.0	m

State and Zone Name		Zone Code	Projection Type	Central Meridian and Scale Factor (T.M.) or Standard Parallels (L.)		Grid Origin Latitude Longtitude		Easting Northing	
California:									
CA	Zone 1	0401	L.	40	00	122	00	2,000,000.0	m
				41	40	39	20	500,000.0	m
CA	Zone 2	0402	L.	38	20	122	00	2,000,000.0	m
				39	50	37	40	500,000.0	m
CA	Zone 3	0403	L.	37	04	120	30	2,000,000.0	m
				38	26	36	30	500,000.0	m
CA	Zone 4	0404	L.	36	00	119	00	2,000,000.0	m
				37	15	35	20	500,000.0	m
CA	Zone 5	0405	L.	34	02	118	00	2,000,000.0	m
				35	28	33	30	500,000.0	m
CA	Zone 6	0406	L.	32	47	116	15	2,000,000.0	m
				33	53	32	10	500,000.0	m
Colorado:									
CO	North	0501	L.	39	43	105	30	914,401.8289	m
				40	47	39	20	304,800.6096	m
CO	Central	0502	L.	38	27	105	30	914,401.8289	m
				39	45	37	50	304,800.6096	m
CO	South	0503	L.	37	14	105	30	914,401.8289	m
				38	26	36	40	304,800.6096	m
Connecticut:									
CT		0600	L.	41	12	72	45	304,800.6096	m
				41	52	40	50	152,400.3048	m
Delware:									
DE		0700	T.M.	75	25	75	25	200,000.0	m
				1:200,000		38	00	0.0	m
Florida:									
FL	East	0901	T.M.	81	00	81	00	200,000.0	m
				1:17,000		24	20	0.0	m
FL	West	0902	T.M.	82	00	82	00	200,000.0	m
				1:17,000		24	20	0.0	m
FL	North	0903	L.	29	35	84	30	600,000.0	m
				30	45	29	00	0.0	m
Georgia:									
GA	East	1001	T.M.	82	10	82	10	200,000.0	m
				1:10,000		30	00	0.0	m
GA	West	1002	T.M.	84	10	84	10	700,000.0	m
				1:10,000		30	00	0.0	m
Hawaii:									
HI	Zone 1	5101	T.M.	155	30	155	30	500,000.0	m
				1:30,000		18	50	0.0	m
HI	Zone 2	5102	T.M.	156	40	156	40	500,000.0	m
				1:30,000		20	20	0.0	m
HI	Zone 3	5103	T.M.	158	00	158	00	500,000.0	m
				1:100,000		21	10	0.0	m
HI	Zone 4	5104	T.M.	159	30	159	30	500,000.0	m
				1:100,000		21	50	0.0	m
HI	Zone 5	5105	T.M.	160	10	160	10	500,000.0	m
				1:infinity		21	40	0.0	m

State and Zone Name		Zone Code	Projection Type	Central Meridian and Scale Factor (T.M.) or Standard Parallels (L.)		Grid Origin Latitude Longtitude		Grid Origin Easting Northing	
Idaho:									
ID	East	1101	T.M.	112	10	112	10	200,000.0	m
				1:19,000		41	40	0.0	m
ID	Central	1102	T.M.	114	00	114	00	500,000.0	m
				1:19,000		41	40	0.0	m
ID	West	1103	T.M.	115	45	115	45	800,000.0	m
				1:15,000		41	40	0.0	m
Illinois:									
IL	East	1201	T.M.	88	20	88	20	300,000.0	m
				1:40,000		36	40	0.0	m
IL	West	1202	T.M.	90	10	90	10	700,000.0	m
				1:17,000		36	40	0.0	m
Indiana:									
IN	East	1301	T.M.	85	40	85	40	100,000.0	m
				1:30,000		37	30	250,000.0	m
IN	West	1302	T.M.	87	05	87	05	900,000.0	m
				1:30,000		37	30	250,000.0	m
Iowa:									
IA	North	1401	L.	42	04	93	30	1,500,000.0	m
				43	16	41	30	1,000,000.0	m
IA	South	1402	L.	40	37	93	30	500,000.0	m
				41	47	40	00	0.0	m
Kansas:									
KS	North	1501	L.	38	43	98	00	400,000.0	m
				39	47	38	20	0.0	m
KS	South	1502	L.	37	16	98	30	400,000.0	m
				38	34	36	40	400,000.0	m
Kentucky:									
KY	North	1601	L.	37	58	84	15	500,000.0	m
				38	58	37	30	0.0	m
KY	South	1602	L.	36	44	85	45	500,000.0	m
				37	56	36	20	500,000.0	m
KY		1600	L.	37	05	85	45	1,000,000.0	m
Note: Zone 1600 is a single new zone.				38	40	36	20	1,500,000.0	m
Louisiana:									
LA	North	1701	L.	31	10	92	30	1,000,000.0	m
				32	40	30	30	0.0	m
LA	South	1702	L.	29	18	91	20	1,000,000.0	m
				30	42	28	30	0.0	m
LA	Offshore	1703	L.	26	10	91	20	1,000,000.0	m
				27	50	25	30	0.0	m
Maine:									
ME	East	1801	T.M.	68	30	68	30	300,000.0	m
				1:10,000		43	40	0.0	m
ME	West	1802	T.M.	70	10	70	10	900,000.0	m
				1:30,000		42	50	0.0	m

State and Zone Name		Zone Code	Projection Type	Central Meridian and Scale Factor (T.M.) or Standard Parallels (L.)		Grid Origin Latitude Longtitude		Easting Northing	
Maryland:									
MD		1900	L.	38	18	77	00	400,000.0	m
				39	27	37	40	0.0	m
Massachusetts:									
MA	Mainland	2001	L.	41	43	71	30	200,000.00	m
				42	41	41	00	750,000.00	m
MA	Island	2002	L.	41	17	70	30	500,000.00	m
				41	29	41	00	0.0	m
Michigan:									
MI	North	2111	L.	45	29	87	00	8,000,000.0	m
				47	05	44	47	0.0	m
MI	Central	2112	L.	44	11	84	22	6,000,000.0	m
				45	42	43	19	0.0	m
MI	South	2113	L.	42	06	84	22	4,000,000.0	m
				43	40	41	30	0.0	m
Minnesota:									
MN	North	2201	L.	47	02	93	06	800,000.0	m
				48	38	46	30	100,000.0	m
MN	Central	2202	L.	45	37	94	15	800,000.0	m
				47	03	45	00	100,000.0	m
MN	South	2203	L.	43	47	94	00	800,000.0	m
				45	13	43	00	100,000.0	m
Mississippi:									
MS	East	2301	T.M.	88	50	88	50	300,000.0	m
				1:20,000		29	30	0.0	m
MS	West	2302	T. M.	90	20	90	20	700,000.0	m
				1:20,000		29	30	0.0	m
Missouri:									
MO	East	2401	T.M.	90	30	90	30	250,000.0	m
				1:15,000		35	50	0.0	m
MO	Central	2402	T.M.	92	30	92	30	500,000.0	m
				1:15,000		35	50	0.0	m
MO	West	2403	T.M.	94	30	94	30	850,000.0	m
				1:17,000		36	10	0.0	m
Montana:									
MT		2500	L.	45	00	109	30	600,000.0	m
				49	00	44	15	0.0	m
Nebraska:									
NE		2600	L.	40	00	100	00	500,000.0	m
				43	00	39	50	0.0	m
Nevada:									
NV	East	2701	T.M.	115	35	115	35	200,000.0	m
				1:10,000		34	45	8,000,000.0	m
NV	Central	2702	T.M.	116	40	116	40	500,000.0	m
				1:10,000		34	45	6,000,000.0	m
NV	West	2703	T.M.	118	35	118	35	800,000.0	m
				1:10,000		34	45	4,000,000.0	m

State and Zone Name		Zone Code	Projection Type	Central Meridian and Scale Factor (T.M.) or Standard Parallels (L.)		Grid Origin Latitude Longtitude		Easting Northing	
New Hampshire:									
NH		2800	T.M.	71	40	71	40	300,000.0	m
				1:30,000		42	30	0.0	m
New Jersey:									
NJ	(NY East)	2900	T.M.	74	30	74	30	150,000.0	m
				1:10,000		38	50	0.0	m
New Mexico:									
NM	East	3001	T.M.	104	20	104	20	165,000.0	m
				1:11,000		31	00	0.0	m
NM	Central	3002	T.M.	106	15	106	15	500,000.0	m
				1:10,000		31	00	0.0	m
NM	West	3003	T.M.	107	50	107	50	830,000.0	m
				1:12,000		31	00	0.0	m
New York:									
NY	East	3101	T.M.	74	30	74	30	150,000.0	m
	(New Jersey)			1:10,000		38	50	0.0	m
NY	Central	3102	T.M.	76	35	76	35	250,000.0	m
				1:16,000		40	00	0.0	m
NY	West	3103	T.M.	78	35	78	35	350,000.0	m
				1:16,000		40	00	0.0	m
NY	Long Island	3104	L.	40	40	74	00	300,000.0	m
				41	02	40	10	0.0	m
North Carolina:									
NC		3200	L.	34	20	79	00	609,601.22	m
				36	10	33	45	0.00	m
North Dakota:									
ND	North	3301	L.	47	26	100	30	600,000.0	m
				48	44	47	00	0.0	m
ND	South	3302	L.	46	11	100	30	600,000.0	m
				47	29	45	40	0.0	m
Ohio:									
OH	North	3401	L.	40	26	82	30	600,000.0	m
				41	42	39	40	0.0	m
OH	South	3402	L.	38	44	82	30	600,000.0	m
				40	02	38	00	0.0	m
Oklahoma:									
OK	North	3501	L.	35	34	98	00	600,000.0	m
				36	46	35	00	0.0	m
OK	South	3502	L.	33	56	98	00	600,000.0	m
				35	14	33	20	0.0	m
Oregon:									
OR	North	3601	L.	44	20	120	30	2,500,000.0	m
				46	00	43	40	0.0	m
OR	South	3602	L.	42	20	120	30	1,500,000.0	m
				44	00	41	40	0.0	m
Pennsylvania:									
PA	North	3701	L.	40	53	77	45	600,000.0	m
				41	57	40	10	0.0	m
PA	South	3702	L.	39	56	77	45	600,000.0	m
				40	58	39	20	0.0	m

State and Zone Name		Zone Code	Projection Type	Central Meridian and Scale Factor (T.M.) or Standard Parallels (L.)		Grid Origin Latitude Longtitude		Easting Northing	
Rhode Island:									
RI		3800	T.M.	71	30	71	30	100,000.0	m
				1:160,000		41	05	0.0	m
South Carolina:									
SC		3900	L.	32	30	81	00	609,600.0	m
				34	50	31	50	0.0	m
South Dakota:									
SD	North	4001	L.	44	25	100	00	600,000.0	m
				45	41	43	50	0.0	m
SD	South	4002	L.	42	50	100	20	600,000.0	m
				44	24	42	20	0.0	m
Tennessee:									
TN		4100	L.	35	15	86	00	600,000.0	m
				36	25	34	20	0.0	m
Texas:									
TX	North	4201	L.	34	39	101	30	200,000.0	m
				36	11	34	00	1,000,000.0	m
TX	North Centra	4202	L.	32	08	98	30	600,000.0	m
				33	58	31	40	2,000,000.0	m
TX	Central	4203	L.	30	07	100	20	700,000.0	m
				31	53	29	40	3,000,000.0	m
TX	South Centra	4204	L.	28	23	99	00	600,000.0	m
				30	17	27	50	4,000,000.0	m
TX	South	4205	L.	26	10	98	30	300,000.0	m
				27	50	25	40	5,000,000.0	m
Utah:									
UT	North	4301	L.	40	43	111	30	500,000.0	m
				41	47	40	20	1,000,000.0	m
UT	Central	4302	L.	39	01	111	30	500,000.0	m
				40	39	38	20	2,000,000.0	m
UT	South	4303	L.	37	13	111	30	500,000.0	m
				38	21	36	40	3,000,000.0	m
Vermont:									
VT		4400	T.M.	72	30	72	30	500,000.0	m
				1:28,000		42	30	0.0	m
Virginia:									
VA	North	4501	L.	38	02	78	30	3,500,000.0	m
				39	12	37	40	2,000,000.0	m
VA	South	4502	L.	36	46	78	30	3,500,000.0	m
				37	58	36	20	1,000,000.0	m
Washington:									
WA	North	4601	L.	47	30	120	50	500,000.0	m
				48	44	47	00	0.0	m
WA	South	4602	L.	45	50	120	30	500,000.0	m
				47	20	45	20	0.0	m

State and Zone Name		Zone Code	Projection Type	Central Meridian and Scale Factor (T.M.) or Standard Parallels (L.)		Grid Origin Latitude Longtitude		Easting Northing	
West Virginia:									
WV	North	4701	L.	39	00	79	30	600,000.0	m
				40	15	38	30	0.0	m
WV	South	4702	L.	37	29	81	00	600,000.0	m
				38	53	37	00	0.0	m
Wisconsin:									
WI	North	4801	L.	45	34	90	00	600,000.0	m
				46	46	45	10	0.0	m
WI	Central	4802	L.	44	15	90	00	600,000.0	m
				45	30	43	50	0.0	m
WI	South	4803	L.	42	44	90	00	600,000.0	m
				44	04	42	00	0.0	m
Wyoming:									
WY	East	4901	T.M.	105	10	105	10	200,000.0	m
				1:16,000		40	30	0.0	m
WY	East Central	4902	T.M.	107	20	107	20	400,000.0	m
				1:16,000		40	30	100,000.0	m
WY	West Central	4903	T.M.	108	45	108	45	600,000.0	m
				1:16,000		40	30	0.0	m
WY	West	4904	T.M.	110	05	110	05	800,000.0	m
				1:16,000		40	30	100,000.0	m
Puerto Rico and Virgin Islands:									
PR		5200	L.	18	02	66	26	200,000.0	m
				18	26	17	50	200,000.0	m

Appendix C: Example Computation—Network Accuracy and Local Accuracy

Appendix C: 3-D Inverse with Statistics

Used to compute values in Table 11.1.

80 ellipsoid:	a =	6,378,137.000 m	ccentricity squared =	0.006694380022903
	1/f =	298.257222100883	econds per radian =	206,264.806247096

ndpoint:	Pseudo				ECEF Covariance Matrix for Standpoint		
X =	-1,556,206.6150	m +/-	0.0011	m	0.00000124	0.00000134	-0.00000092
Y=	-5,169,400.7400	m +/-	0.0021	m	0.00000134	0.00000451	-0.00000278
Z=	3,387,285.9870	m + -	0.0020	m	-0.00000092	-0.00000278	0.00000418
	(user inputs these values)				(Default values = zero implies point is held fixed.)		

repoint:	USPA				ECEF Covariance Matrix for Forepoint		
X =	-1,555,678.5790	m +/-	0.0015	m	0.00000216	0.00000235	-0.00000150
Y=	-5,169,961.3960	m +/-	0.0029	m	0.00000235	0.00000847	-0.00000502
Z=	3,386,700.0890	m +/-	0.0026	m	-0.00000150	-0.00000502	0.00000681
	(user inputs these values)				(Default values = zero implies point is held fixed.)		

s zero and implies the point positions are independent.
re present. Correlation of the standpoint with respect to
tion of the forepoint with respect to the standpoint.

Correlation of Forepoint wrt Standpoint			Correlation of Standpoint wrt Forepoint		
1.1405E-06	1.2934E-06	-8.8294E-07	1.1405E-06	1.2848E-06	-8.9269E-07
1.2848E-06	4.2401E-06	-2.5676E-06	1.2934E-06	4.2401E-06	-2.5759E-06
-8.9269E-07	-2.5759E-06	3.6030E-06	-8.8294E-07	-2.5676E-06	3.6030E-06

Standpoint:			Pseudo			From local co-variance matrix	
	D	M	Sec.				
φ =	32	16	45.746506	N	+/-	0.0013	m
λ =	253	14	45.600250	E	+/-	0.0009	m
	106	45	14.399750	W	+/-	0.0009	m
h =			1,165.6411	m	+/-	0.0027	m
Forepoint:			**USPA**				
	D	M	Sec.				
φ =	32	16	23.000192	N	+/-	0.0017	m
λ =	253	15	11.091830	E	+/-	0.0014	m
	106	44	48.908170	W	+/-	0.0014	m
h =			1,178.0154	m	+/-	0.0036	m

$$\lambda = \tan^{-1}\left(\frac{Y}{X}\right); \quad 0° \text{ to } 360° \text{ East}$$

$$\tan\phi = \frac{Z}{P}\left(1 + \frac{e^2 N \sin\phi}{Z}\right); \quad P = \sqrt{X^2 + Y^2}$$

$$h = \frac{P}{\cos\phi} - N$$

Must iterate to find latitude and height.

$$\phi_0 = \tan^{-1}\left(\frac{Z}{P(1-e^2)}\right) \quad N_0 = \frac{a}{\sqrt{1-e^2\sin^2\phi_0}}$$

$$R = \begin{bmatrix} -\sin\lambda & \cos\lambda & 0 \\ -\sin\phi\cos\lambda & -\sin\phi\sin\lambda & \cos\phi \\ \cos\phi\cos\lambda & \cos\phi\sin\lambda & \sin\phi \end{bmatrix}$$

$$\phi_i = \tan^{-1}\left[\frac{Z}{P}\left(1 + \frac{e^2 N_{i-1}\sin\phi_{i-1}}{Z}\right)\right] \text{ and } N_i = \frac{a}{\sqrt{1-e^2\sin^2\phi_i}}$$

Local covariance = $\Sigma_{e/n/u} = R\,\Sigma_{X/Y/Z}\,R^t$

R =	Standpoint		R =	Forepoint	
0.957551239	-0.288263117	0.000000000	-0.53395479	-0.288144775	0.000000000
0.153946349	0.511378351	0.845454141	0.153856283	0.51130809	0.845513029
-0.243713246	-0.80956566	0.534048028	-0.243630161	-0.809652164	0.533954790
R(t) =			**R(rt) =**		
0.957551239	0.153946349	-0.243713246	-0.53395479	0.153856283	-0.243630161
-0.288263117	0.511378351	-0.80956566	-0.288144775	0.511308090	-0.809652164
0.000000000	0.845454141	0.534048028	0.000000000	0.845513029	0.533954790

Local Reference Frame Covariance:

Standpoint			Forepoint		
0.0000007718	0.0000000474	-0.0000002289	0.000002042	-0.000000273	0.000004636
0.0000000474	0.0000017695	0.0000008970	-0.000000273	0.000002778	0.000001151
-0.0000002289	0.0000008970	0.0000073896	0.000004636	0.000001151	0.000013279

Local Tangent Plane Inverse from Standpoint to Forepoint: **Pseudo** to **USPA**

$$\begin{bmatrix}\Delta e\\ \Delta n\\ \Delta u\end{bmatrix} = R_{stdpt}\begin{bmatrix}\Delta X\\ \Delta Y\\ \Delta Z\end{bmatrix}$$

R =	0.957551239	-0.288263117	0.000000000	ΔX =	528.036
	0.153946349	0.511378351	0.845454141	ΔY =	-560.656
	-0.243713246	-0.80956566	0.534048028	ΔZ =	-585.898

$$Dist = \sqrt{\Delta e^2 + \Delta n^2} \qquad Azi = \tan^{-1}\left(\frac{\Delta e}{\Delta n}\right)$$

Δe = 667.238 m
Δn = -700.768 m
Δu = 12.301 m

Distance =	**967.6168**			**m**
Azimuth =	**136**	**24**	**14.57**	

Use ATAN2 function: Azi = 2.380699743 radians

Equation 1.36 from chapter 1 and repeated as equation 11.9 in chapter 11 is used as the basis for computing both network accuracy and local accuracy standard deviations. If the submatrices on the upper right and lower left are zero, the computed answer will be network accuracy. Local accuracy is obtained if the full covariance matrix is used.

$$\Sigma_\Delta = \begin{bmatrix}-1&0&0&1&0&0\\0&-1&0&0&1&0\\0&0&-1&0&0&1\end{bmatrix}\begin{bmatrix}\begin{bmatrix}\sigma^2_{X_1}&\sigma_{X_1Y_1}&\sigma_{X_1Z_1}\\ \sigma_{X_1Y_1}&\sigma^2_{Y_1}&\sigma_{Y_1Z_1}\\ \sigma_{X_1Z_1}&\sigma_{Y_1Z_1}&\sigma^2_{Z_1}\end{bmatrix} & \begin{bmatrix}\sigma_{X_1X_2}&\sigma_{X_1Y_2}&\sigma_{X_1Z_2}\\ \sigma_{Y_1X_2}&\sigma_{Y_1Y_2}&\sigma_{Y_1Z_2}\\ \sigma_{Z_1X_2}&\sigma_{Z_1Y_2}&\sigma_{Z_1Z_2}\end{bmatrix}\\ \begin{bmatrix}\sigma_{X_1X_2}&\sigma_{Y_1X_2}&\sigma_{Z_1X_2}\\ \sigma_{X_1Y_2}&\sigma_{Y_1Y_2}&\sigma_{Z_1Y_2}\\ \sigma_{X_1Z_2}&\sigma_{Y_1Z_2}&\sigma_{Z_1Z_2}\end{bmatrix} & \begin{bmatrix}\sigma^2_{X_2}&\sigma_{X_2Y_2}&\sigma_{X_2Z_2}\\ \sigma_{Y_2X_2}&\sigma^2_{Y_2}&\sigma_{Y_2Z_2}\\ \sigma_{Z_2X_2}&\sigma_{Z_2Y_2}&\sigma^2_{Z_2}\end{bmatrix}\end{bmatrix}\begin{bmatrix}-1&0&0\\0&-1&0\\0&0&-1\\1&0&0\\0&1&0\\0&0&1\end{bmatrix} = \begin{bmatrix}\sigma^2_{\Delta X}&\sigma_{\Delta X\Delta Y}&\sigma_{\Delta X\Delta Z}\\ \sigma_{\Delta X\Delta Y}&\sigma^2_{\Delta Y}&\sigma_{\Delta Y\Delta Z}\\ \sigma_{\Delta X\Delta Z}&\sigma_{\Delta Y\Delta Z}&\sigma^2_{\Delta Z}\end{bmatrix}$$

$$\Sigma_{3D-INV} = \begin{bmatrix}\sigma^2_{\Delta e}&\sigma_{\Delta e\Delta n}&\sigma_{\Delta e\Delta u}\\ \sigma_{\Delta e\Delta n}&\sigma^2_{\Delta n}&\sigma_{\Delta n\Delta u}\\ \sigma_{\Delta e\Delta u}&\sigma_{\Delta n\Delta u}&\sigma^2_{\Delta u}\end{bmatrix} = R\begin{bmatrix}\sigma^2_{\Delta X}&\sigma_{\Delta X\Delta Y}&\sigma_{\Delta X\Delta Z}\\ \sigma_{\Delta X\Delta Y}&\sigma^2_{\Delta Y}&\sigma_{\Delta Y\Delta Z}\\ \sigma_{\Delta X\Delta Z}&\sigma_{\Delta Y\Delta Z}&\sigma^2_{\Delta Z}\end{bmatrix}R^t \quad \text{where} \quad R = \begin{bmatrix}-\sin\lambda & \cos\lambda & 0\\ -\sin\phi\cos\lambda & -\sin\phi\sin\lambda & \cos\phi\\ \cos\phi\cos\lambda & \cos\phi\sin\lambda & \sin\phi\end{bmatrix}$$

$$S = \sqrt{\Delta e^2 + \Delta n^2} \qquad \alpha = \tan^{-1}\left(\frac{\Delta e}{\Delta n}\right)$$

$$J_3 = \begin{bmatrix}\frac{\partial S}{\partial \Delta e}&\frac{\partial S}{\partial \Delta n}&\frac{\partial S}{\partial \Delta u}\\ \frac{\partial \alpha}{\partial \Delta e}&\frac{\partial \alpha}{\partial \Delta n}&\frac{\partial \alpha}{\partial \Delta u}\end{bmatrix} = \begin{bmatrix}\frac{\Delta e}{S}&\frac{\Delta n}{S}&0\\ \frac{\Delta n}{S^2}&\frac{-\Delta e}{S^2}&0\end{bmatrix}$$

$$\begin{bmatrix}\sigma^2_S&\sigma_{S\alpha}\\ \sigma_{S\alpha}&\sigma^2_\alpha\end{bmatrix} = J_3\,\Sigma_{3D-INV}\,J^t$$

Network accuracy standpoint to forepoint is computed assuming no correlation between points.
Local accuracy is computed using the full covariance matrix (including correlation input by user).

Pseudo to **USPA**					**Network Accuracy**			**Local Accuracy**	
Tangent Plane Distance =	**967.6168**		**m**	+/-	**0.0018**	m	+/-	**0.0011**	m
3-D Azimuth =	**136**	**24**	**14.57**	+/-	**0.40**	sec	+/-	**0.24**	sec

Index

B

C

D

E

F

G

H

I

O

P

Q

R

S

T

U

V

W

X

Z

Environmental Engineering

THE 3-D GLOBAL SPATIAL DATA MODEL

Foundation of the Spatial Data Infrastructure

Traditional methods for handling spatial data are encumbered by the assumption of separate origins for horizontal and vertical measurements. Modern measurement systems operate in a 3-D spatial environment. **The 3-D Global Spatial Data Model: Foundation of the Spatial Data Infrastructure** offers a new model for handling digital spatial data, the global spatial data model or GSDM.

The GSDM preserves the integrity of three-dimensional spatial data while also providing additional benefits such as simpler equations, worldwide standardization, and the ability to track spatial data accuracy with greater specificity and convenience. This groundbreaking spatial model incorporates both a functional model and a stochastic model to connect the physical world to the ECEF rectangular system.

Combining horizontal and vertical data into a single, three-dimensional database, this authoritative monograph provides a logical development of theoretical concepts and practical tools that can be used to handle spatial data more efficiently. The book clearly describes procedures that can be used to handle both ECEF and flat-Earth rectangular components in the context of a rigorous global environment.

Features

- Uses standard measurements that can be applied to disciplines all around the world
- Supports both analog map plots and computer visualization of digital data
- Combines horizontal and vertical data into a single three-dimensional database, facilitating rigorous error propagation and standard deviation computations
- Utilizes the BURKFORD™ software, readily available for download on a companion website, to perform 3-D coordinate geometry and error propagation computations

63014

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

ISBN: 978-1-4200-6301-1
90000
9 781420 063011